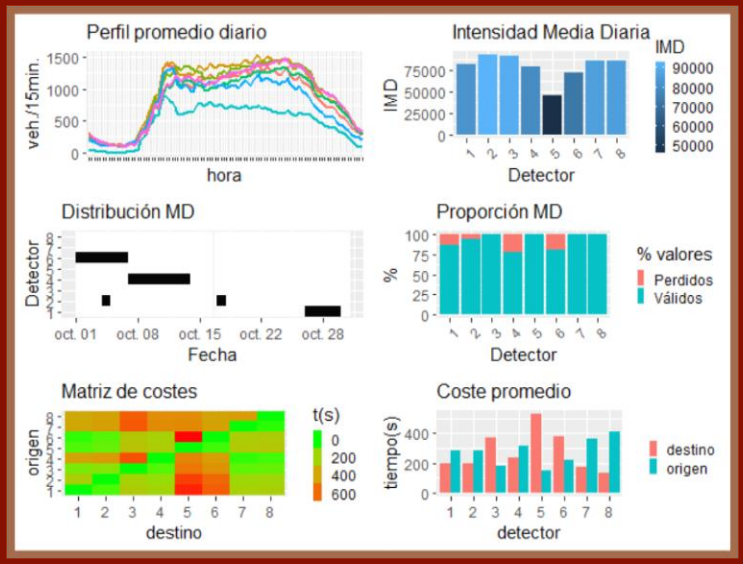
Tesis Doctoral

Marco para la predicción del flujo de tráfico con datos incompletos basado en técnicas Machine Learning



Autor: Cayetano Ruiz de Alarcón Quintero

Directores: Noelia Cáceres Sánchez

Luis Miguel Romero Pérez

Dpto. Ingeniería y Ciencia de los Materiales y del Transporte

Escuela Técnica Superior de Ingeniería

Universidad de Sevilla

2022

Tesis Doctoral

Marco para la predicción del flujo de tráfico con datos incompletos basado en técnicas Machine Learning

Autor:

Cayetano Ruiz de Alarcón Quintero

Dirigida por:

Noelia Cáceres Sáncehz

Dr. Ingeniero de Telecomunicación

Luis Miguel Romero Pérez

Dr. Ingeniero Industrial

Dpto. de Ingeniería y Ciencia de los Materiales y del Transporte

Escuela Técnica Superior de Ingeniería

Universidad de Sevilla

Sevilla, 2022

A Cayetano y Mariló

Agradecimientos

Me gustaría mostrar mi agradecimiento, en primer lugar, a Francisco García Benítez, Luis Miguel Romero Pérez y Noelia Cáceres Sánchez por orientarme con este trabajo.

También querría hacer mención a mis compañeros de departamento, durante los años que pasé por la Universidad, a Fran, Antonio, Antonio, Jesús y José Ignacio con los que compartí dependencias y cafés durante aquellos años y que hicieron más llevadero el trabajo diario.

Vorrei anche ricordare i miei colleghi della IUAV, Olga, Federico e Silvio.

Me gustaría agradecer al Ministerio de Ciencia Innovación y Universidades por convocar la ayuda FPI que me permitió inciar la escritura de esta tesis.

Desearía hacer un agradecimiento especial a mi familia, por la paciencia que han tenido durante los años que ha durado la maduración de esta tesis, que ha conllevado robarles un tiempo precioso. Por eso me gustaría reconocer el apoyo que mi madre, mi padre, que ya no está con nosotros y mis hermanos me han brindado durante estos años.

A mis amigos que me han acompañado y animado a seguir, hasta acabar este trabajo.

Me gustaría dar un especial agradecimiento a Mariló y a Cayetano, a los que quiero muchísimo y son los que más han sufrido todos los inconvenientes del tiempo dedicado a este trabajo. Aún así, siempre me han aapoyado y han sido mi impulso para poder culminarlo.

Resumen

Este trabajo ha tratado de obtener una visión actual y global de los procesos de modelado de predicción de tráfico a corto plazo. Para ello, mediante la consulta de la literatura científica y casos prácticos, se han identificado los factores comunes que intervienen en el proceso de predicción.

Se trata de un campo dinámico, en continua evolución que se ha visto empujado por el desarrollo de las tecnologías, provocando varios efectos como el aumento de la presencia dispositivos de monitorización de las carreteras, el incremento de la capacidad de cálculo de los equipos informáticos, el desarrollo de nuevos conceptos relacionados con la conectividad y la abundancia de de información.

Todo ello ha devenido en una explosión de métodos guiados por datos que han mejorado la eficiencia de los modelos existentes, aunque han puesto a la luz diversos problemas derivados del tratamiento de volúmenes extensos de información

En este contexto se ha realizado una puesta en común en torno a las metodologías más existosas y a los retos que se plantean, con el objetivo de identificar las principales fases que se deben integrar el proceso de predicción y determinar la dirección de este campo en un futuro próximo.

Para el desarrollo de esta idea se ha creado un marco de predicción orientado a las téncicas guiadas por datos, en el que se definen las fases funcionales que componen el proceso. Este marco permite la incorporación de distintas metodologías pera el desarrollo de modelos predictivos. Adicionalmente, se ha enriquecido con un modelo de datos que permite la fácil incorporación de casos prácticos que sirvan para testar estos modelos. De esta forma se dispone de un banco de ensayo que facilita la evaluación real de metodologías y la generalización de las conclusiones

de un modo más eficaz favoreciendo el desarrollo de esta materia.

Índice

ÍNDICE DE FIGURAS

1 INTRODUCCIÓN

El estado del tráfico ejerce una influencia significativa sobre la dinámica de las áreas urbanas, determinando el grado de movilidad de personas y mercancías. Los agentes responsables de la gestión de la movilidad aseguran el buen funcionamiento del tráfico apoyándose en los Sistemas Inteligentes de Transporte (ITS) para la monitorización y gestión del transporte de forma eficiente, tanto en el interior de las ciudades, como en sus áreas metropolitanas.

Para desarrollar sus funciones, los ITS necesitan herramientas que proporcionen información sobre el estado del tráfico. Los intrumentos de predicción realizan esta función, apoyándose en la modelización de una serie de parámetros, permitiendo anticiparse a las incidencias que influyen en el grado de movilidad de la zona gestionada por el ITS en el que se integran. Esta instrumentos se conocen como modelos de predicción de parámetros de tráfico a corto plazo.

La literatura científica sobre este tema constata, por una parte, la enorme variedad de soluciones existentes, fruto de la rápida proliferación de avances técnicos y, por otra parte, la falta de homogeneización de dichas soluciones y técnicas, así como de las fuentes de información utilizadas. Estos dos aspectos combinados dificultan la comparación de los avances y, por tanto, la extensión de los resultados y conclusiones derivados de ellos. Se obstaculiza así la obtención de conclusiones generales que orienten, de una manera más concreta, la dirección en la que se deben invertir los esfuerzos para un desarrollo más eficiente de esta disciplina.

A pesar de la evolución que han experimentado las herramientas de predicción, aún queda terreno por explorar, en cuanto al potencial de las tecnologías de la información y comunicación en el campo de la modelización predictiva del tráfico. Las políticas de datos abiertos, la conectividad a través de internet y la publicación de datos en servidores accesibles han promovido el uso e intercambio de datos y servicios en todo el mundo por parte de distintos agentes, potenciando la confluencia de información relativa a diferentes temáticas y zonas geográficas. Como consecuencia, se está generando un volumen y una variedad de información sin precedentes, que se incrementa exponencialmente a medida que pasa el tiempo, favoreciendo el uso de información contextual para desarrollar herramientas de monitorización y predicción más precisas.

Esta coyuntura ha abierto una serie de oportunidades que están siendo explotadas en los últimos años con las técnicas más avanzadas, promoviendo la aparición de conceptos como *Big Data*, *Internet of Things* (IoT) y el uso generalizado de herramientas basadas en técnicas guiadas por datos, o *Data Driven*. Todos estos conceptos deben atender a problemas comunes como la adquisición y gestión de grandes volúmenes de datos y su utilización eficiente en base al objetivo que persiguen.

Este trabajo pretende abarcar de forma generalizada el problema de la predicción de parámetros de tráfico a corto plazo, determinando sus principales características y dificultades, su situación actual y aportando una propuesta en forma de marco de predicción en la que integrar todas las fases que intervienen en el proceso de predicción.

1.1 Motivación

La principal motivación para la realización de esta tesis es el establecimiento de unas bases comunes sobre las características y dificultades que atañen a la predicción de parámetros de tráfico a corto plazo. Se trata de una cuestión que ha sido abordada desde multiples aproximaciones, existiendo una vasta y variada producción científica acerca de este tema, un hecho beneficioso, aunque plantea una serie de dificultades.

En los trabajos que versan sobre este campo, generalmente, se proponen multiples y variadas soluciones, extrayéndose conclusiones sobre su aplicación a casos de estudio concretos que, por una serie de dificultades, son difícilmente

generalizables.

El propósito fundamental de este trabajo consiste, en primer lugar, en la identificación de las características y dificultades de las fases que intervienen en el proceso de predicción de parámetros de tráfico y en segundo lugar, en la realización de una propuesta que plantee las bases para la generalización de técnicas validadas sobre condiciones concretas.

1.2 Ámbito

Desde una perspectiva técnica, este trabajo se enmarca en el área temática de la Ingeniería del Transporte, más concretamente en las técnicas de predicción del tráfico a corto plazo. Se trata de un campo que comprende una amplia diversidad de técnicas que, se aplican a un problema crucial en la vida real de las personas, la gestión del tráfico.

La gestión de grandes volúmenes de datos se ha convertido en un elemento fundamental para las técnicas guiadas por datos, hacia las que se ha ido desplazando el foco en el campo de la predicción del tráfico. El nivel de calidad de los conjuntos de datos utilizados y su correcta gestión son determinantes en los modelos desarrollados a partir de las técnicas guiadas por datos. Para su manejo eficiente se requieren técnicas de adquisición y análisis de datos, que exigen un alto grado de conocimiento de las características de la información y de las fuentes que los suministran.

En relación a la gestión de la información, es fundamental dominar los estándares existentes y las técnicas de armonización de datos, debido a que permiten una correcta comprensión e importación de la información y servicios, a partir de distintas fuentes, favoreciendo las virtudes de los modelos guiados por datos y facilitando la obtención de resultados generalizables.

Desde el punto de vista de la escala geográfica, este problema se trata desde la perspectiva del contexto urbano, en sus diferentes variantes, desde vías residenciales a autovías de gran capacidad. Es en este entorno en el que se producen las principales problemáticas relativas al tráfico y en el que los modelos de predicción resultan de mayor utilidad.

1.3 Objetivos

Mediante el desarrollo de este trabajo se pretenden determinar las principales técnicas de predicción del tráfico a corto plazo, así como las etapas en las que se subdivide este proceso y los aspectos que influyen, directa o indirectamente, en su elaboración. Para ello se realiza una profunda revisión de la bibliografía científica, identificando aquellos elementos que muestran una incidencia significativa sobre el proceso de predicción.

Un objetivo secundario es la determinación del estado actual, en cuanto a disponibilidad, calidad y grado de implementación de los conjuntos de datos sobre información de tráfico. La información relativa al tráfico ha sufrido una rápida evolución, que se ha visto impulsada por los avances tecnológicos, las políticas de datos abiertos y el desarrollo de estándares de datos y servicios auspiciadas tanto desde el sector público, como del privado. Esto ha derivado en la creación y publicación de conjuntos de datos relativos al tráfico, accesibles, estructurados y de gran volumen. Para satisfacer este objetivo se consultan distintas fuentes de datos reales, entre éstas, aquellas consideradas como paradigmáticas, por su uso generalizado en estudios de tráfico.

El tratamiento de la información se revela como un aspecto fundamental en este campo, al combinarse el desarrollo de las técnicas guiadas por datos y la evolución de los conjuntos de información relativos al tráfico. Por este motivo, se plantea como objetivo la identificación de procesos que han ofrecido buenas prestaciones de manera general en el tratamiento de datos en el campo de la predicción del tráfico.

El elemento central de este trabajo es el modelado predictivo de parámetros de tráfico. Este aspecto se aborda mediante la identificación de las fortalezas y debilidades de las técnicas más extendidas, su adecuación a las características de cada caso de estudio y la identificación de las fases que intervienen en ellas.

Para conseguir este objetivo, se plantea la determinación de un esquema integral del proceso de predicción de tráfico a corto plazo, mediante la identificación de las pautas comunes para las que se han observado mejores resultados. Dentro de este proceso se incluyen las etapas de importación y tratamiento de la información relativa a un escenario, la generación de los modelos de predicción y el análisis de los resultados obtenidos una vez se han aplicado, tomando cuerpo en una propueste de marco general de predicción, que funcione como base de

comparación universal entre técnicas y casos de estudio, favoreciendo la generalización de las conclusiones.

1.4 Estructura del documento

Este documento está estructurado en seis capítulos.

En el primero presentan los aspectos que motivan el desarrollo de la tesis (sección 1.1), el ámbito en el que se enmarca (sección 1.2), los objetivos que se desean obtener mediante su elaboración (sección 1.3) y la estructura del documento (sección 1.4).

En el capítulo 2 se realiza una descripción de las tecnologías utilizadas en la gestión del tráfico, centrada en los aspectos que guardan relación con la predicción y prestando una especial atención a su evolución dentro del desarrollo de los ITS. Se realiza una revisión de las metodologías de monitorización de parámetros de tráfico y de los conjuntos de datos más significativos, utilizados como casos de estudio en publicaciones y casos prácticos reales. Se pone el foco en características como el formato, el volumen de datos, su grado de accesibilidad, la escala geográfica que abarca la información contenida, o el nivel de agregación física y temporal de la información.

El capítulo 3 se centra en las técnicas de predicción del tráfico a corto plazo teniendo en cuenta la presencia de datos incompletos. En primer lugar, se realiza una revisión de técnicas que han sido utilizadas en la predicción (sección 3.1), analizando sus características metodológicas y las conclusiones derivadas de su utilización. En segundo lugar, se profundiza en las problemáticas relativas los datos corruptos y erróneos incluidos en los conjuntos de datos que se utilizan en el proceso de predicción (sección 3.2), estudiando las características y tipologías de los registros erróneos y revisándose las técnicas de imputación que minimizan el efecto de los valores perdidos. Se finaliza el capítulo realizando una síntesis de las características comunes observadas en las técnicas presentadas (sección 3.3).

En el capítulo 4 se desarrolla un marco metodológico que recopila los procesos y subprocesos más frecuentes y eficientes, de entre aquellos que se han revisado en el capítulo 3. En dicho marco se integran cinco etapas referentes a: i) la adquisición de la información proveniente de casos reales (secciones 4.1.1), ii) el análisis de la información adquirida (sección 4.1.2), iii) el proceso de imputación de valores perdidos (sección 4.1.3), iv) el proceso de predicción (sección 4.1.4), y v) la

evaluación de todo el proceso (sección 4.1.5).

En el capítulo 5 se muestra el funcionamiento práctico del esquema integral de predicción, por medio de su aplicación a casos de estudio. En primer lugar, se describe el esquema de aplicación de manera abstracta (sección 5.1) y consecutivamente se detallan las configuraciones utilizadas de cada una de las fases: i) las técnicas de preprocesamiento en la sección 5.2, que engloban la adquisición y el análisis del conjunto de datos; ii) las aproximaciones de modelado en la sección 5.3, describiéndose tanto las técnicas de imputación como las de predicción, iii) y la metodología de evaluación del rendimiento en la sección 5.4. Posteriormente, se describen los casos de estudio en los que se aplica la predicción, exponiéndose las características de cada caso, y la especificación de la información de tráfico disponible en cada una de ellas. En cada uno de los casos de estudio se detalla el proceso de modelado mediante el análisis de parámetros descriptivos y los resultados obtenidos en cada fase; finalmente se analizan los parámetros relativos al rendimiento de los modelos y se evalúa el proceso en función de las características del caso de estudio.

Por último, en el capítulo 6 se extraen las principales conclusiones derivadas del desarrollo del marco metodológico, de la aplicación de los métodos con los que se concreta, analizándose las ventajas encontradas en su utilización, así como las dificultades halladas en el proceso. Por último, se muestra una relación de los posibles avances y propuestas de líneas futuras de investigación.

En la parte final de este documento se incluyen varios anexos para facilitar la comprensión del mismo, y aportar, de manera pormenorizada, descripciones y parámetros que dificultarían la lectura y entendimiento si se ubicaran en el desarrollo de la metodología o el desarrollo de las fases de modelado.

2. ITS Y ADQUISICIÓN DE PARÁMETROS DE TRÁFICO

Los Sistemas Inteligentes de Transporte (ITS) son aplicaciones avanzadas cuyo objetivo es proporcionar servicios innovadores para los diferentes modos de transporte y gestión del tráfico, además de facilitar información a los usuarios para hacer un uso más seguro, coordinado e inteligente de las redes de transporte.

El núcleo de los ITS radica en la transferencia de información entre los sistemas de captación de datos de la red y los centros de gestión, en los que la información es tratada previamente a su transmisión al resto de agentes que actúan sobre la red (Dölger and Geißler, 2012).

Las herramientas de predicción son elementos fundamentales para el funcionamiento de los ITS, adelantando información el estado del tráfico y permitiendo la anticipación de la toma de decisiones. Estas herramientas dependen, en gran medida, de la información sobre parámetros de tráfico que se adquiere a través de dispositivos instalados en la red.

En este capítulo se presenta la importancia de la adquisición de información en el correcto funcionamiento de los ITS. En primer lugar, se realiza una aproximación a los sistemas de adquisición de información, presentándose sistemas reales, que se encuentran integrados en ITS. Posteriormente se describen varios estándares y esquemas armonizados de datos que se han implementado en la Unión Europea y los beneficios obtenidos con ellos.

2.1. Adquisición de información de tráfico

La información disponible determina, en gran medida, el buen funcionamiento de las herramientas utilizadas por los ITS en la gestión del tráfico. La calidad de los conjuntos de datos se puede medir según diferentes características como el nivel de agregación de los registros, el grado de compleción de la información, el nivel de accesibilidad y la presencia de metadatos. El contexto en el que se adquiere, se almacena, publica y comparte la información ha cambiado notablemente en los últimos años, influyendo positivamente en su disponibilidad y calidad (Bearn et al., 2018; Boukerche and Wang, 2020; Jin et al., 2021; Negrete and Subirana, 2012).

Para la adquisición de datos se requieren sensores y sistemas que los gestionen. En el caso de los ITS, los sensores se clasifican en dos tipos (Figura 2-1), según su función (Sastre García et al., 2011):

i) Detección y monitorización de vehículos:

   a. Autónomos, en los que el sensor está situado en la infraestructura.

   b. Dependientes, requieren la presencia de un dispositivo embarcado en el vehículo.

ii) Medición de condiciones cronológicas y ambientales.

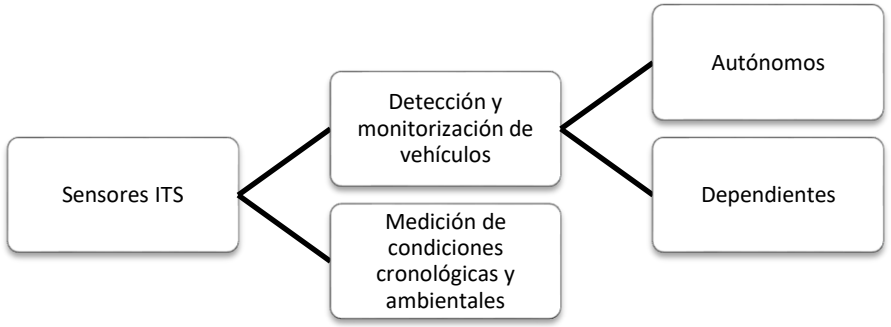


Figura 2-1: Tipología de sensores utilizados en los ITS.

El presente estudio se centra en los detectores de tipo autónomo, ya que son la fuente fundamental de información de los modelos de imputación y predicción descritos en el capítulo 3. Entre los sensores autónomos, aquellos que presentan un

mayor grado de implantación son las espiras magnéticas y las cámaras de detección de vehículos.

Las espiras magnéticas tienen capacidad para proveer información sobre diversos parámetros en la sección de vía en la que se encuentran instaladas:

i) Volumen de vehículos.

ii) Velocidad individual y velocidad media de los vehículos.

iii) Ocupación de la vía por parte de los vehículos.

Estos elementos de detección presentan como mayor inconveniente el procedimiento invasivo de su instalación y su propensión a fallos y errores de funcionamiento (Muspratt, 2011).

Las cámaras de vídeo emplean técnicas de visión artificial para extraer, de manera automática, distintos parámetros relacionados con el tráfico (Muspratt, 2011). Se usan en los ITS para detección y aforo de vehículos mediante su identificación por extracción automática de matrícula y tipo de vehículo. Su instalación es menos invasiva que en el caso de las espiras magnéticas, aunque son más sensibles a las condiciones atmosféricas y necesitan un mayor grado de mantenimiento.

Existen numerosas herramientas de adquisición de datos de tráfico sobre las cuales no se profundizará en este estudio ya que, debido a su heterogeneidad, complejidad y estado de desarrollo, requieren de un análisis pormenorizado, no siendo el objetivo de este trabajo. Se proponen como fuentes de datos para una posible extensión a la metodología aquí presentada. Entre estos elementos de adquisición, los más usuales son:

i) Los vehículos flotantes, que suministran información de trayectos de vehículos individuales. Su utilización, de forma combinada con detectores fijos, ha sido tratada en numerosos estudios (Herrmann et al., 2012; Vázquez et al., 2020; Xiao et al., 2015).

ii) Los vehículos autónomos y conectados, que presentan la capacidad de actuar como vehículos flotantes, además de poder configurarse con tipos de conducción predeterminadas para observar su influencia sobre el resto del tráfico (Calvert et al., 2020; Gora et al., 2020; Miglani and Kumar, 2019), proporcionando una rica variedad y volumen de datos respecto a los dispositivos utilizados previamente a su aparición.

iii) Drones para la detección de datos de tráfico, que constituyen una buena opción de toma de datos debido a su versatilidad para posicionarse, no requiriendo instalación. Presenta el inconveniente de disponer de poca autonomía. (Barmpounakis and Geroliminis, 2020; Fan et al., 2021; Kamnik et al., 2020).

Es importante señalar que el proceso de adquisición de datos conlleva la presencia de cierto nivel de ruido y errores, independientemente del dispositivo utilizado. Entre los orígenes más habituales de los errores se encuentran:

i) El mal funcionamiento de los dispositivos. Estos errores son difíciles de detectar, ya que se registran los valores en el conjunto de datos, aunque éstos no coinciden con la situación real de la vía.

ii) Errores de comunicación en el sistema de adquisición. Estos errores producen valores perdidos, es decir, no se registran en el conjunto de datos. Este fenómeno se desarrolla en el apartado 3.2.1 y es uno de los puntos fundamentales de este estudio, ya que repercute en el rendimiento de las herramientas de predicción.

Desde los años 50 se han desarrollado numerosos sistemas de monitorización del tráfico debido al incremento gradual de la presencia de vehículos motorizados privados y a la proliferación de vías destinadas a su desplazamiento (Lockwood et al., 2016). En las últimas décadas se han desplegado sistemas ITS en carreteras de todo el mundo, contribuyendo a una evolución constante, aumentándose el nivel de detalle de los datos capturados y la velocidad de comunicación con los sistemas de procesamiento.

Dos ejemplos representativos de estos sistemas son:

i) Performance Measure System o PeMS (Berner and Hart, 2013; California Department of Transportation, 2018), creado y administrado por Caltrans, el organismo encargado de la gestión del tráfico en el estado de California.

ii) Motorway Incident Detection and Automatic Signalling o MIDAS (The Highways Agency et al., 1994; Tucker et al., 2006), una red distribuida de sensores instalados en las carreteras del Reino Unido.

PeMS proporciona información captada en 39.000 detectores desplegados en las autovías de California (California Department of Transportation, 2018). Estos

detectores contribuyen a la elaboración de un conjunto de datos compuesto por registros agregados cada 30 s, con valores sobre volumen, velocidad y ocupación (Berner and Hart, 2013). La información se suministra en tiempo real, pudiendo consultarse igualmente mediante una serie histórica, publicada en formato abierto y accesible. Debido al volumen de datos publicados, su nivel de desagregación y el área geográfica abarcada, es un ejemplo magnífico sobre gestión e intercambio de información de tráfico. Por estos motivos, se considera un sistema óptimo para la validación de metodologías de predicción a corto plazo, y por ello ha sido profusamente utilizado como banco de ensayo para herramientas de este tipo (Castro-Neto et al., 2009; Lippi et al., 2013; Lv et al., 2015; Tian and Pan, 2015; Wu and Tan, 2016; Zhang, 2011).

El sistema MIDAS, por otra parte, se encuentra desplegado en la red de carreteras del Reino Unido. Se trata de una red distribuida de sensores de tráfico, principalmente espiras inductivas, que están diseñadas para adquirir datos sobre flujos de tráfico y velocidades, que se proporcionan a los centros de control regionales. Varios estudios sobre predicción del tráfico utilizan este sistema debido a la amplitud de su despliegue, cubriendo más de 910 km de la red inglesa de autovías (Muspratt, 2011; Williams and Hoel, 2003; Yangzhou Chen, Jiang Luo, Wei Li, 2014).

Los sistemas descritos se restringen al ámbito de la región o el estado en el que están instalados. En la actualidad, se desarrollan herramientas que pretenden alcanzar un ámbito transfronterizo, permitiendo la gestión del tráfico de manera coordinada entre regiones o países.

Habiéndose superado el reto tecnológico de la eficiencia y desarrollo de las herramientas de adquisición de datos, se requiere un esfuerzo en la armonización de la información y la coordinación entre las herramientas de gestión, con el objetivo de intercambiar información y servicios entre distintos países y niveles de administración. En el apartado 2.2 se presentan las herramientas y sistemas que permiten este intercambio.

2.2. Armonización de datos de tráfico

Los sistemas ITS muestran la importancia de disponer de información homogénea de un área geográfica amplia. Ésta homogeneización permite la comparación y

evaluación de herramientas y medidas implementadas en distintas zonas. Este aspecto convierte a los procesos de armonización en estratégicos para los organismos que gestionan las carreteras (Steenberghen et al., 2013).

Tanto desde el punto de vista de la Administración pública, como desde el sector privado, se observa la necesidad de unificar formatos y utilizar estándares en el intercambio y los procesos de adquisición, almacenaje y explotación de información relativa al tráfico (Avazpour et al., 2019). El uso de un marco común favorece la compatibilidad e intercambio de información entre las agencias públicas, privadas y la ciudadanía (World Customs Organization, 2007). De este mode, además, se favorece la visualización y explotación de los datos, facilitando la toma de decisiones por parte de los organismos gestores (Avazpour et al., 2019).

La armonización de datos facilita la incorporación de nuevos dispositivos en sistemas ya existentes, como ya está ocurriendo con la información que registran los vehículos autónomos, que se presentan como una fuente de datos muy prometedora (Ghiasi et al., 2019).

Con el objetivo de incrementar el intercambio de información y conseguir un marco europeo común de datos se han emprendido varias iniciativas. Un ejemplo representativo es el acuerdo al que llegaron los Estados Miembros de la Unión Europea en 2014 (Shilton et al., 2015), cuya meta es establecer un método estandarizado para el cálculo del ruido del tráfico en la carretera. Este método requiere el establecimiento de una categorización común de vehículos, además de la elaboración de metodologías para la equiparación de los datos obtenidos a partir de elementos de adquisición de datos heterogéneos, con la finalidad de hacerlos compatibles en toda Europa.

Para dotar a las autoridades de tráfico europeas de un formato homogéneo de datos se desarrolló DATEXII, el estándar de datos específico para ITS en la UE, publicado en octubre de 2012 (Dölger and Geißler, 2012). Su objetivo es proporcionar una manera estandarizada de comunicación e intercambio de información entre centros de control, proveedores de servicios, operadores de tráfico y medios de comunicación. Este intercambio se produce de forma transfronteriza, con lo que se mejora la administración de las redes de transporte europeas hasta ese momento. La armonización se asume como una meta fundamental en la UE, tanto para la sociedad de la información en sí, como para los ITS. DATEXII se ha propuesto como estándar para los operadores de tráfico de la

denominada Red Trans-Europea de transporte, o TEN-T (European Commission, 2014) y ya ha sido implementada por numerosos países.

Los procesos de armonización se vuelven aún más importantes teniendo en cuenta que en los últimos años están emergiendo numerosos portales de datos abiertos en los distintos niveles de la Administración pública (Ayuntamiento de Madrid, 2016; Ayuntamiento de Málaga, 2016; Ayuntamiento de Sevilla, 2016; Ministerio de Política Territorial y Función Pública and Ministerio de Economía y Empresa, 2018). Estos portales persiguen dotar de accesibilidad a la información pública a través de conjuntos de datos y servicios, para que sea utilizada y compartida por parte de los ciudadanos y otros actores públicos y privados.

Con el fin de facilitar la reutilización de la información, ésta debe ofrecerse en un formato legible automáticamente y en un estándar de metadatos aceptado, de este modo se favorece la interoperabilidad y descubribilidad[1]. Estos portales se han creado de manera descoordinada (European Commission, 2014) y los potenciales usuarios (ciudadanos o entidades públicas y privadas) están teniendo dificultades para encontrar y reutilizar información del sector público (Berends et al., 2017).

A pesar de que se están haciendo considerables esfuerzos por parte de los estados miembros, el estado de armonización e interoperabilidad de datos y servicios todavía es incompleto (Veeckman et al., 2017). Desde 2007 se está desarrollando una directiva en la UE para corregir la falta de disponibilidad, calidad, organización y accesibilidad de datos espaciales en la Unión Europea, a través de la implementación de una Infraestructura de Datos Espaciales, o IDE, común, denominada INSPIRE (European Parliament and Council of the European Union, 2007). En este IDE se incluye una especificación de datos concreta para las redes de transporte que comprende una red integrada de los distintos modos, y los elementos relacionados, con continuidad entre límites nacionales. Los elementos topográficos de la red de transportes son relativos a la carretera, el tren y el transporte por agua y por aire (European Commission, 2007). La red facilita la referencia de los flujos de transporte en sus arcos para habilitar servicios de navegación lo que permite enlazar los datos de los detectores a la red.

Para complementar la implementación de INSPIRE, desde de la Unión Europea se están desarrollando varias herramientas para la armonización de fuentes de datos,

[1] La descubribilidad es la cualidad de un conjunto de datos para ser fácilmente encontrado mediante un motor de búsqueda, una aplicación o un sitio web (Deirdre Lee, 2016).

requiriéndose coordinación de la gestión e intercambio de esta información entre agencias gubernamentales y otros tipos de entidades (empresas, fundaciones, entidades de investigación, etc.).

El proceso de armonización se compone de varios pasos, que parten de mapear los esquemas de datos de origen y, mediante procesos de transformación automatizados, obtener el formato homogéneo final (Hintz, 2010).

El desarrollo de estas herramientas y su aplicación a casos prácticos, mediante proyectos impulsados por la Comisión Europea dentro del programa *EU Horizon* 2020, como *PoliVisu*, permiten observar la potencialidad de esta gestión coordinada como apoyo a la implementación de políticas relativas al tráfico. Mediante este tipo de proyectos se hace posible la incorporación de las técnicas *Big Data* a la toma de decisiones políticas sobre la gestión del tráfico (McAleer et al., 2018).

El desarrollo de las IDE en España ha estado impulsado, por las agencias públicas generadoras de información geográfica. Las campañas de seguimiento de INSPIRE reflejan un alto número de Conjuntos de Datos Espaciales (CDE) y servicios puestos a disposición pública a través de internet desde esos geo-portales. Ya se encuentra totalmente actualizada, armonizada y publicada la información relativa a las redes de transporte (Consejo Nacional de Información Geográfica de España, 2017).

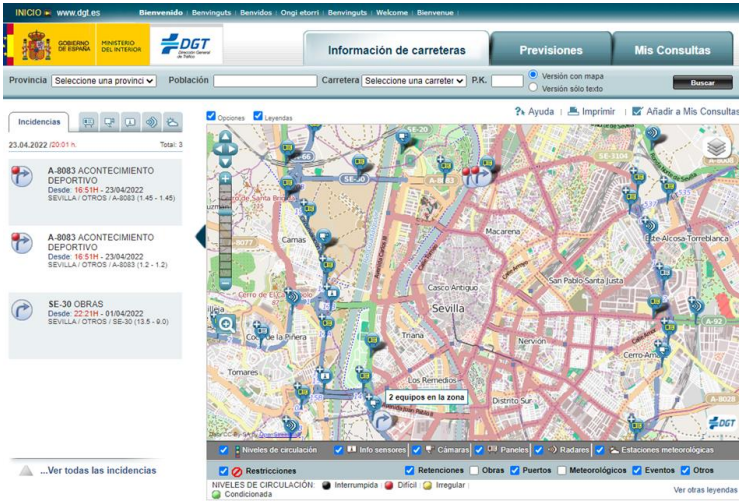


Figura 2-2: Mapa de tráfico interactivo publicado en el portal web de la Dirección General de Tráfico (Dirección General de Tráfico, 2022).

En paralelo, también se está produciendo la implantación progresiva del estándar DATEXII por parte de la DGT en España (Subdirección General de la Gestión de la Movilidad y Tecnología, 2020). Los datos sobre parámetros de tráfico son consultables a través del portal web de la DGT, mostrándose en formato cartográfico y a título informativo en el Mapa de Tráfico ubicado en el mismo portal (Figura 2-2).

La adquisición y armonización de datos se identifica como una fase crucial dentro del proceso de predicción de parámetros de tráfico a corto plazo. Por este motivo, se contempla como la fase inicial en el marco de predicción que se propone en este trabajo, detallándose en la sección 4.1.1 , en la que se presenta un modelo de datos sobre información de tráfico basado en los estándares de datos de tráfico descritos previamente. Este modelo se describe pormenorizadamente en un artículo previo a la publicación de esta tesis (Ruiz-Alarcon-Quintero, 2016).

Este modelo se ofrece como un modo de armonizar e integrar la red de transportes propuesta en la directiva INSPIRE y el estándar de datos de tráfico DATEXII, además de facilitar el intercambio de información con otras fuentes externas a la UE.

3. PREDICCIÓN DEL FLUJO DE TRÁFICO CON DATOS INCOMPLETOS

En este capítulo se presenta una revisión de los principales métodos de predicción del flujo de tráfico a corto plazo, en la que se incluyen las herramientas auxiliares que estos métodos utilizan para optimizar su rendimiento.

La evolución de las técnicas de predicción a corto plazo más relevantes se detalla en el sección 3.1, enmarcando las características de esta disciplina y los avances que se han ido produciendo desde sus inicios. En la sección 3.2 se aborda la problemática relativa a los valores perdidos, o corruptos, presentes en los conjuntos de datos de tráfico, revisándose los principales métodos de imputación y sus efectos sobre los modelos de predicción.

Por último, en la sección 3.3 se realiza una puesta en común sobre las principales tipologías y características observadas en las principales familias de métodos, tanto de predicción como de imputación, estableciéndose como punto de partida para el desarrollo del marco predicción con datos incompletos, que constituye el elemento central de este trabajo, expuesto en el capítulo 4.

3.1. Técnicas de predicción de parámetros de tráfico a corto plazo

Como ya se ha comentado en el capítulo 2, los ITS han vivido una intensa evolución desde su origen. Los avances en los dispositivos de monitorización y en las

herramientas de gestión de información han supuesto un gran impulso en el desarrollo de estos sistemas.

Atendiendo a la definición de la PIARC (*Permanent International Association of Road Congresses, antigua World Road Association*) (Chowdhury and Sadek, 2021), las funciones principales de los ITS son:

i) La mejora de la movilidad de personas y bienes

ii) El incremento de la seguridad, reduciendo la congestión del tráfico y administrando eficientemente las incidencias,

iii) El establecimiento de los objetivos y metas de las políticas de transporte, como la administración de la demanda de las medidas prioritarias del transporte público.

Para todas estas funciones se necesitan realizar previsiones del estado del tráfico a corto plazo, con el objetivo de anticiparse a posibles incidencias y optimizar el servicio prestado.

Se entienden como predicciones a corto plazo, aquellas realizadas con un horizonte temporal limitado. De un modo general se considera que dicho horizonte debe variar entre pocos segundos y un día. El horizonte de predicción depende de varios factores, principalmente de la escala de la zona estudiada y del propósito principal de la predicción (Vlahogianni et al., 2004).

El punto de partida de esta disciplina se produce con la publicación del artículo *Analysis of Freeway Traffic Time-Series Data by using Box-Jenkins Techniques* (Ahmed and Cook, 1979), en el que por primera vez se diseña un modelo que proporciona predicciones del volumen de tráfico y de la ocupación de las vías estudiadas.

Los elementos sobre los que se necesitan predicciones en los ITS son los parámetros fundamentales del tráfico: la intensidad, densidad y velocidad media de los vehículos que transcurren por una sección, o carril de una vía. Las relaciones entre los tres parámetros (Figura 3-1) se definen mediante el diagrama fundamental del tráfico (Greenshields et al., 1934). Las relaciones representadas en el diagrama sumadas a las características de la red de transportes constituyen la base de la teoría del tráfico (Imeers and Logghe, 2002).

Los avances tecnológicos en algorítmica y potencia de computación, añadidas a la disponibilidad y accesibilidad de grandes volúmenes de datos han propiciado la

proliferación de técnicas guiadas por datos, o *data-driven modelling*, en el campo del modelado de los parámetros de tráfico. Este desarrollo se ha visto impulsado por la expansión de diversas políticas orientadas a potenciar la hiperconectividad y accesibilidad a la información, descritas en la sección 2.2.

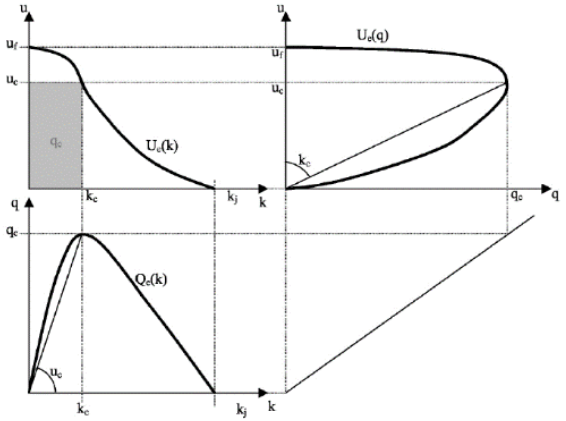


Figura 3-1: Diagramas fundamentales del tráfico, en el que se representa la intensidad q, la densidad k y la velocidad media u (Imeers and Logghe, 2002).

Este campo de estudio ha experimentado un notable avance gracias al desarrollo de la tecnología y la accesibilidad a las colecciones de datos, aunque, por otra parte, estos dos factores han derivado en una serie de problemas metodológicos, debido a que el volumen de datos disponible excede la capacidad de las aproximaciones clásicas de modelado. Las principales consecuencias de la gestión de grandes volúmenes de datos en e l modelado de parámetros de tráfico son:

i) El aumento de los costes de computación.

ii) El incremento de requerimientos de espacio de almacenamiento.

iii) El impacto de los datos corruptos sobre el rendimiento de los modelos implementados.

Ante estas dificultades, se han desarrollado una serie de técnicas, conocidas como *data mining* o minería de datos, que permiten el tratamiento de grandes volúmenes de datos, mediante la optimización de la gestión de información utilizada. Se encuentran enmarcadas en las técnicas *Big Data*, que son aquellas utilizadas para lograr costes admisiblesn en la gestión de grandes cantidades de información. La

predicción a corto plazo se ha apoyado en este tipo de técnicas de una manera recurrente, permitiendo optimizar el proceso de selección y gestión de datos para mejorar la precisión de los valores modelados y reducir los costes de computación.

Entre la vasta cantidad de publicaciones sobre esta área, en las que se presentan continuamente nuevas técnicas y metodologías, se llegan a numerosas conclusiones parciales, no alcanzándose consensos claros sobre la predominancia de una familia de técnicas, o de un método concreto, respecto a otras. Las conclusiones que se extraen de los resultados de cada publicación no son generalizables debido a que las condiciones de cada experimento son difícilmente reproducibles por las características determinadas de cada caso de estudio (Vlahogianni et al., 2014). Es común que las técnicas presentadas como nuevos métodos de predicción sean probadas en casos de estudio ad-hoc, que favorecen la comprensión del método y la obtención de resultados positivos, aunque no proporcionan una clara muestra de la validez del método en casos reales y en condiciones variadas.

Se observa la necesidad de establecer una base comparativa, que favorezca la aplicación de múltiples técnicas a un mismo escenario, y de una misma técnica a múltiples escenarios reales.

Por estos motivos, se estima la necesidad de establecer un marco de predicción común, en el que se establezcan unas bases comunes que permita la comparación entre diversas técnicas y su aplicación a diferentes escenarios. En este trabajo se realiza una propuesta a este respecto que se desarrolla en el capítulo 4. En este sentido ya existen propuestas que integran distintas fases que participan en la predicción (Boquet et al., 2020), danto una especial importancia al tratamiento previo de la información, antes de ser provista a la técnica de predicción.

El rendimiento de una técnica, depende en gran medida de las condiciones del caso de estudio sobre el que se aplica, por ello, los casos de estudio se deben definir con precisión. Caracterizar cada caso es un asunto complejo debido, a los numerosos aspectos que intervienen en la predicción y en el comportamiento del tráfico. Se deben seguir una serie de pasos que guíen en su elección (Vlahogianni et al., 2004), atendiendo a tres factores:

i) El objetivo del modelo de predicción.

ii) La información de salida.

iii) El proceso de desarrollo del modelo.

3.1.1. Objetivo del modelo de predicción

El objetivo que se persigue con la predicción es determinante para el desarrollo del modelo; este factor se define en torno a dos aspectos (Vlahogianni et al., 2004):

i) El tipo de sistema en el que se integra el modelo.

ii) Área de implementación del algoritmo de predicción.

De un modo general, existen dos tipos de sistemas ITS en los que se integran los modelos de predicción (Lockwood et al., 2016):

i) Sistemas de gestión de tráfico. El objetivo de la predicción es la determinación del estado del tráfico en una red viaria determinada, ya sea a través de los parámetros fundamentales, u otras variables derivadas

ii) Sistemas de asistencia y guiado de viajeros. Las principales finalidades de este tipo de sistemas son el cálculo del tiempo de viaje entre dos puntos dados, por una parte, y, por otra, la provisión de recomendaciones durante la ruta, para lo que se requiere información en tiempo real sobre las características y el estado de las vías.

Ambas metas demandan información en tiempo real para ofrecer predicciones actualizadas.

Los objetivos del modelo de predicción deben ser compatibles con el área de implementación en el que se integra el ITS. Esta área está caracterizada por aspectos como la disposición geográfica de la población, la estructura de la red de transportes, la cantidad y tipo de detectores existentes y la tipología de las vías. Las aproximaciones se adecúan de una manera diferente a cada ámbito de aplicación. El conjunto de datos disponibles en el área es un factor determinante en las metodologías a desarrollar, ya que establece los niveles mínimos de agregación y precisión y establece la información disponible en el escenario.

Considerando la escala de mayor a menor detalle, en el primer nivel se tienen las vías e intersecciones urbanas. Las herramientas de predicción a esta escala requieren información muy detallada desde el punto de vista de la agregación temporal y del número de parámetros que intervienen, al tratarse de un problema con situaciones más especializadas y complejas (Bowman and Miller, 2016; Ghosh et al., 2009; Gilmore et al., 1993; Okutani and Stephanedes, 1984; Yin et al., 2002;

Zhu et al., 2014). Se deben tener en cuenta factores como las fases semafóricas y la distribución del volumen de vehículos en cada intersección. A este nivel, se utilizan grados de desagregación temporal en la toma de datos que van desde pocos segundos a 5 minutos, siendo ideal la toma de los datos individualizados a nivel de carril.

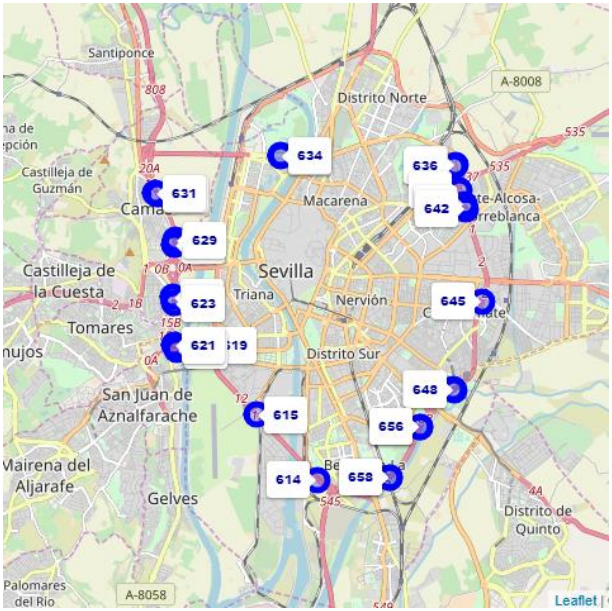


Figura 3-2: Ejemplo de caso de estudio a escala de circunvalación, presentado en el caso de estudio 5.5.4.

El siguiente grado se refiere a las vías urbanas arteriales y carreteras de circunvalación (Figura 3-2), redes compuestas de vías de una mayor capacidad que la escala descrita anteriormente. Presentan valores superiores de velocidad máxima en las vías y, debido a su proximidad a zonas pobladas, contienen intersecciones separadas por pocos km y un límite de velocidad más limitado que el de una autovía por motivos prácticos y de seguridad. Requieren un grado menor del nivel de detalle de los datos para desarrollar herramientas precisas de predicción (Hofleitner et al., 2012; Kamarianakis and Prastacos, 2005; Qiu et al., 2016; Smith and Demetsky, 1994; Stathopoulos and Karlaftis, 2003; Vlahogianni et al., 2005; Wang et al., 2014; Williams et al., 1998). A esta escala, se considera que una agregación temporal en intervalos de 5 y 15 minutos de los parámetros de tráfico es

suficiente, ya que el comportamiento del tráfico se ve menos influenciado por las fases semafóricas que en la escala urbana; la adquisición de datos a nivel de carril sigue siendo la ideal, aunque en las fuentes consultadas es más común disponer de una agregación de los datos a nivel de sentido de circulación.

Figura 3-3: Ejemplo de caso de estudio a escala de autovía, correspondiente al caso de estudio 5.5.1.

El área de implementación con un menor grado de detalle se refiere a las vías de alta capacidad, autovías y autopistas, sobre las que se han desarrollado numerosos estudios de metodologías de predicción, debido a la fácil accesibilidad a las fuentes de datos a esta escala y a la importancia de su monitorización (Ahmed and Cook, 1979; Bing et al., 2015; Castro-Neto et al., 2009; Davis and Nihan, 1991; Dougherty and Cobbett, 1997; Kwon et al., 2000; Lv et al., 2015; Roncoli et al., 2016; Sun et al., 2006).

Estas vías sirven para la comunicación a larga distancia, presentando unas dinámicas características que se ven menos influenciadas por las intersecciones. Necesitan un nivel de agregación menor que en los otros dos casos para realizar predicciones con una precisión admisible, manejando horizontes temporales más amplios, debido a que el comportamiento del tráfico a esta escala es bastante más estable (Figura 3-3).

Tabla 3-1: Cuadro resumen de los objetivos de los modelos de predicción del tráfico.

Objetivo	Características	Función del modelo de predicción
Gestión del tráfico	Gestión agregada del tráfico para optimizarlo a nivel de red.	Predicción centrada a nivel de red, sinergia entre teoría del tráfico y modelos de predicción. Se necesitan predecir los parámetros fundamentales y medidas derivadas como tiempos de viaje en tramos.
Asistencia en viaje/ enrutamiento de vehículo	Guía de un único vehículo para optimizar el viaje individual.	Proporcionar información sobre el estado de las vías, incidentes, y tiempos de trayectos, principalmente, del tiempo de viaje total entre origen y destino. Para ello se necesitan los tiempos de viaje a corto plazo de los arcos por los que trascurre el trayecto.
Relación con otros modos de transporte	Influencia entre distintos modos de transportes.	Influencia de unos modos de transporte en otros, medida del impacto de diversas situaciones modeladas y de futuras actuaciones. Predicción de situaciones de congestión en vías compartidas por modos público y privado.
Gestión de parkings	Monitorización y previsión del nivel de ocupación de parkings.	Predicción del nivel de servicio según las condiciones del tráfico y la distribución de vehículos.
Cálculo de emisiones	Cálculo del nivel de emisiones producidas por el tráfico, con el objetivo de disminuirlo.	Modelización y predicción de las emisiones derivadas del tráfico de los vehículos de una vía. Predecir el volumen de vehículos por clasificación de emisiones.
Objetivo	Características	Función del modelo de predicción
Gestión del tráfico	Gestión agregada del tráfico para optimizarlo a nivel de red.	Predicción centrada a nivel de red, sinergia entre teoría del tráfico y modelos de predicción. Se necesitan predecir los parámetros fundamentales y medidas derivadas como tiempos de viaje en tramos.

En cuanto a los objetivos de los modelos de predicción a corto plazo, éstos han dejado de limitarse simplemente a la asistencia en viaje o a la gestión del tráfico, orientándose a otros propósitos tales como el modelado de emisiones derivadas del tráfico, la gestión de parkings y la relación con otros modos de transporte (Tabla

3-1), con el objetivo final de favorecer una movilidad más sostenible mediante programas impulsados por la Administración pública (European Commission, 2013).

El presente estudio se centra en el primer objetivo, denominado *Gestión del tráfico* en la Tabla 3-1, aunque los modelos y técnicas desarrolladas pueden servir como apoyo para herramientas cuyo objetivo es cualquiera de los presentados en dicha tabla.

3.1.2. Información de salida

La elección de la metodología adecuada para afrontar el modelado en un caso de estudio depende, en gran medida, de la información que se espera obtener como resultado de su aplicación. Normalmente, la información de salida se define en función de los parámetros que se necesitan predecir y del horizonte de predicción establecido.

El primer elemento a considerar para definir la información de salida, es el tipo de parámetro que se establece como objetivo de la predicción, en función de la finalidad del modelado. En sistemas de guiado de viajeros el parámetro de mayor interés es el tiempo de viaje a través de la red; por otra parte, para la gestión del tráfico se predicen los parámetros fundamentales en toda la red considerada para describir las condiciones del tráfico. La mayoría de las publicaciones y modelos aplicados en la práctica se centran en la estimación del volumen de tráfico. Se ha demostrado que las predicciones sobre este parámetro son más robustas que las realizadas sobre los otros parámetros fundamentales, mostrándose más estables que las obtenidas de modelos implementados para la ocupación y más precisos que aquellos cuyo parámetro de salida es la velocidad media (Habtemichael and Cetin, 2016; Kumar and Vanajakshi, 2015; Qiu et al., 2016).

El nivel de agregación de la información utilizada en los modelos está supeditado al conjunto de datos disponible en el caso de estudio. Dependiendo de su accesibilidad y nivel de agregación de la información, se definen los factores básicos del problema:

i) El horizonte de predicción.

ii) El paso, tiempo que transcurre entre un registro del parámetro y el

siguiente.

Se debe tener en cuenta que la amplitud del horizonte y el nivel de precisión de los modelos son inversamente proporcionales; además, la predicción es más costosa, desde el punto de vista computacional, cuanto más corto es el paso, al producirse un nivel excesivo de información con poco valor explicativo (Ishak and Al-Deek, 2003).

La definición de la resolución del dato es un aspecto fundamental para la determinación de la metodología, especialmente en los algoritmos guiados por datos, porque afecta a la calidad de la información sobre las condiciones del tráfico.

Debido a la fuerte variabilidad del dato en cortos intervalos se recomienda agregar los resgistros sobre velocidad y flujo en intervalos más amplios; por ello en numerosos estudios se usan agregaciones de 5 minutos, aunque se disponga de desagregaciones para cada minuto o incluso más detalladas (Mantecchini, 2011; Tarko and Perez-Cartagena, 2005).

Por último, se debe definir la cantidad de parámetros de salida del modelo, dividiéndose en dos tipos básicos:

i) Salida univariable, cuando solo se modela solo un parámetro.

ii) Salida multivariable, cuando se modelan varios parámetros simultáneamente.

3.1.3. Proceso de desarrollo del modelo

El desarrollo del modelo se refiere a la elección de la aproximación metodológica utilizada para la predicción, basándose en el objetivo del modelo y en la información de salida, descritos en las secciones anteriores.

La estructura básica de modelado (Figura 3-4) está compuesta por:

i) La adquisición y tratamiento de la información de entrada.

ii) un módulo de cálculo que realiza predicciones.

iii) La información de salida, que se compone de los valores modelados, además del índice de rendimiento con el que se evalúa el modelo.

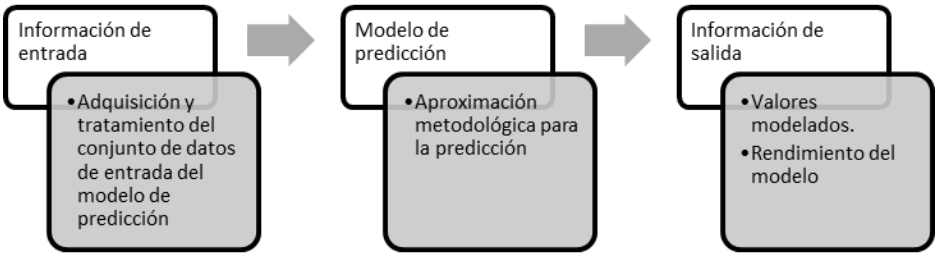


Figura 3-4: Esquema básico de modelado de predicción.

Siguiendo esta estructura como planteamiento base, se han implementado evoluciones consistentes en la modificación de la segunda etapa (modelo de predicción), con el propósito de mejorar el rendimiento u optimizar la gestión de la información disponible. Estas modificaciones se representan mediante dos tipos de modelos:

i) Modelos híbridos (Figura 3-5). En este esquema la etapa intermedia se subdivide en dos. El objetivo de la primera es la clasificación y/o selección de información del conjunto de datos de entrada. En esta fase se agrupan los datos con características similares y/o se seleccionan los parámetros de entrada con mayor valor explicativo respecto al parámetro de salida. La información clasificada/seleccionada se suministra como conjunto de datos de entrada al modelo de predicción. Existen múltiples versiones de modelos híbridos, según la forma en la que se combinan las técnicas en sus fases intermedias. En la descripción de modelos de predicción, que se realiza en las siguientes secciones, se especificarán los casos en los que se sigue este tipo de estructura.

ii) Modelos de combinación (Figura 3-6). Consisten en el desarrollo en paralelo de varios modelos de predicción basados en aproximaciones metodológicas diferentes cuyas salidas se combinan en una fase posterior. Con este proceso se busca dotar de mayor robustez a la predicción, ya que es ratificada por varios métodos distintos.

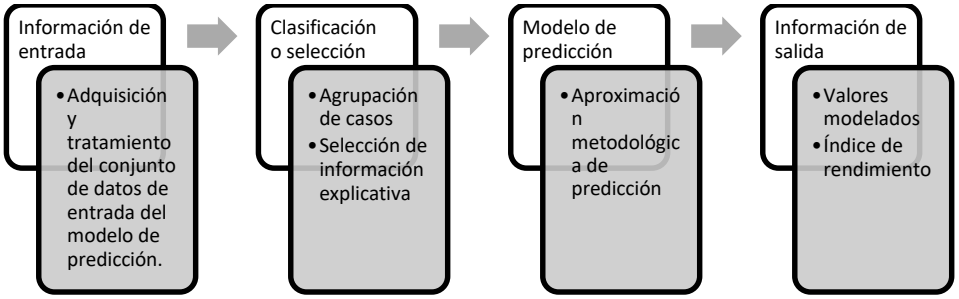


Figura 3-5: Esquema de modelado híbrido de predicción.

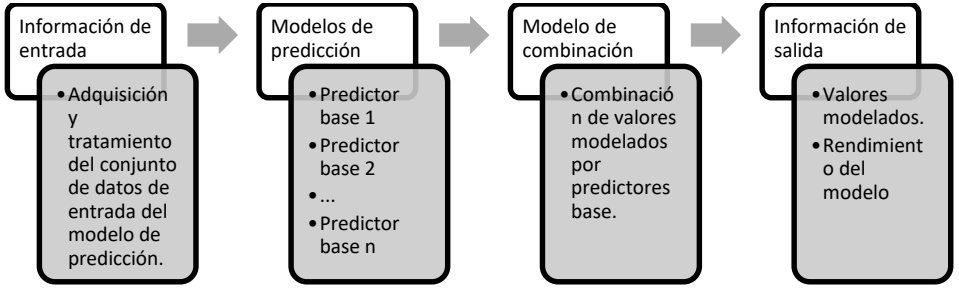


Figura 3-6: Esquema de modelado combinado de predicción.

En lo relativo al método aplicado en la fase de predicción, existe un cierto consenso en dividirlos de dos grandes categorías de modelos (Smith et al., 2002; Vlahogianni et al., 2014):

i) Técnicas paramétricas (sección 3.1.3.1), Se basan en la definición de una forma funcional que, mediante una serie de parámetros, describen el comportamiento de uno o varios parámetros de salida.

ii) Técnicas no paramétricas (sección 3.1.3.2), No asumen una forma funcional entre los parámetros de entrada y los de salida.

A continuación, se revisan los métodos utilizados en la predicción del tráfico siguiendo esta clasificación; realizando una descripción de los principales métodos y presentándose ejemplos de su utilización en problemas de predicción de parámetros de tráfico.

3.1.3.1. Técnicas paramétricas

Las técnicas paramétricas de predicción buscan la definición de una formulación matemática que describa el comportamiento de un parámetro. Entre las aproximaciones más simples de esta categoría se encuentran los caminos aleatorios o *Random Walk* (RW), la media histórica o *Historical Average* (HA), algoritmos de suavizado y, con un mayor grado de complejidad, los algoritmos basados en la correlación y regresión lineal (Smith et al., 2002). Consisten en aproximaciones sencillas, con un bajo coste computacional y un rendimiento aceptable en condiciones normales, debido al comportamiento recurrente del tráfico.

Una manera habitual de representar la información de tráfico son las series temporales, en las que se basan numerosas aproximaciones. Una de las familias de métodos más usados desde los inicios del desarrollo de esta disciplina (Ahmed and Cook, 1979; Levin and Tsao, 1980), es la basada en el método ARIMA (*Auto Regressive Integrated Moving Average*), desarrollado originariamente para la predicción de parámetros económicos. Su aplicación está especialmente indicada para modelar series de datos con un comportamiento estacional y recurrente. Ambos factores están presentes en las series temporales de tráfico en condiciones normales.

En diversas aproximaciones basadas en métodos de esta familia se constata una mejora de rendimiento, respecto a métodos más sencillos, en condiciones normales de tráfico a distintas escalas (Kirby et al., 1997; Laña et al., 2018b; Smith et al., 2002; Tselentis et al., 2015; Vlahogianni et al., 2014, 2004), a la vez que presentan un bajo coste computacional. Estos factores han prolongado la utilización de esta aproximación hasta hoy en día, siendo una buena alternativa en los casos en los se dispone de una serie histórica de datos pequeña y se requieren predicciones en tiempo real. Se utiliza habitualmente como elemento con el que comparar el rendimiento de otras aproximaciones más complejas.

El mayor defecto de ARIMA es su tendencia a concentrarse en la media, y no ser tan preciso en los extremos. Este aspecto provoca resultados erráticos en las predicciones realizadas en los periodos de pico de tráfico (Vlahogianni et al., 2004). Para atenuar este efecto se han desarrollado diversas variantes, entre las que se encuentra el método híbrido KARIMA (*Kohonen Maps* + ARIMA) que introduce una primera fase de clasificación de información mediante una red neuronal de tipo Mapa de Kohonen. En el artículo en el que se desarrolla por primera vez este método con el objetivo de predecir el flujo de tráfico a corto plazo (Van Der Voort et

al., 1996), se aplica a cuatro ubicaciones de autovías francesas. La fase de clasificación busca la definición de patrones en el comportamiento del flujo en los 30 minutos previos al intervalo del valor modelado. Se desarrollan un modelo ARIMA simple y un KARIMA para cada una de las localizaciones, en todos los casos, el método KARIMA muestra un rendimiento superior.

Siguiendo la misma tendencia, se han desarrollado otros métodos híbridos basados en ARIMA. En uno de ellos se utiliza el método de clasificación KNN (K Nearest Neighbours) en la primera fase, que se integra en el sistema de predicción ATHENA (Kirby et al., 1997). El método KNN-ARIMA supera a la aproximación ARIMA simple en predicciones con horizontes de predicción de 30 y 60 minutos. La misma configuración (KNN + ARIMA), también ha sido compara en otro trabajo con una red neuronal BPNN (*Back Propagation Neural Network*) y con el método HA, con un horizonte de predicción de 15 minutos (Williams et al., 1998), superando a ambos en las condiciones específicas de los escenarios expuestos.

(Williams and Hoel, 2003) aplicaron el método SARIMA (ARIMA con una componete estacional) a la predicción del flujo de tráfico. Para tratar de mejorar el rendimiento del modelo, se incluye la recurrencia del comportamiento del tráfico en la formulación del modelo. Se observa que usando la recurrencia semanal se consigue mejorar el rendimiento respecto al método ARIMA simple. En el mismo sentido (Min et al., 2007) proponen el método MSTAR (*Multivariate State-Space Auto Regressive*), en el que se introduce una formulación espacio-estado para representar la dependencia entre las observaciones.

(Chen et al., 2017) proponen dos implementaciones de modelos híbridos, en los que la fase de clasificación se realiza mediante técnicas *Big Data*, y la fase de predicción con métodos ARIMA y SARIMA respectivamente. En ambos casos se realiza una fase de reconocimiento de patrones en la información contenida en una gran base de datos sobre tiempos de viaje en Inglaterra. Se compara cada método con su versión sin fase de clasificación (ARIMA y SARIMA), observándose una mejora del rendimiento en los casos en los que sí la incluyen, atribuida a una selección más precisa de la información de entrada de los modelos.

Otra subcategoría de los métodos paramétricos son los métodos espacio-estado, que utilizan variables de estado para describir el comportamiento de un sistema mediante un conjunto de ecuaciones diferenciales de primer orden (Hamilton, 1994). Las variables de estado se caracterizan por poder ser reconstruidas desde la

información medida en la entrada/salida del sistema. El objetivo de este tipo de modelado es inferir la información sobre los estados, dadas las observaciones reales, cuando se tiene nueva información en la entrada. Se consideran apropiadas para el modelado de parámetros de tráfico por su naturaleza multivariable y su habilidad para modelar series temporales univariables más simples (Dong et al., 2014; Kawasaki et al., 2017; Tampere and Immers, 2007).

En (Dong et al., 2014) los métodos espacio-estado multivariable consiguen rendimientos superiores a aproximaciones ARIMA simples en la predicción del flujo de tráfico y la velocidad media sobre el conjunto de datos en el que se prueban.

Dentro de la subcategoría de los métodos espacio-estado se incluye el Filtro de Kalman (KF), que se ha mostrado superior al ARIMA simple cuando se modelan datos de tráfico clasificados en diferentes periodos del día (Stathopoulos and Karlaftis, 2003). Se ha utilizado en el proceso de predicción, realizando la función de predictor principal, así como realizando la función de clasificador de información de los modelos híbridos, por su habilidad para eliminar el ruido de una muestra (Habtemichael and Cetin, 2016; Jabari and Liu, 2013; Lippi et al., 2013; Okutani and Stephanedes, 1984; Tampere and Immers, 2007). La técnica KF se sigue intentando mejorar en aplicaciones para la predicción de parámetros de tráfico en estudios recientes, para la generación de predictores inmunes al ruido en los datos (Cai et al., 2019).

3.1.3.2. Técnicas no paramétricas

La otra gran categoría de modelos de predicción está formada por los métodos no paramétricos, que han ganado peso en su utilización para la predicción del flujo de tráfico en los últimos años respecto a los paramétricos.

Se caracterizan por la no asunción de una forma funcional que relacione a las variables dependientes con las independientes, por lo que se adaptan mejor que los métodos paramétricos al comportamiento caótico de las condiciones del tráfico en los cambios de régimen de congestión (Hoogendoorn and Bovy, 2001; Vlahogianni et al., 2015).

El impulso a los métodos no paramétricos, de los que la mayoría se enmarcan en las

técnicas de modelado guiado por datos[2], ha venido facilitado por diversos factores, que ya se han comentado en la sección 2.1 :

i) La mejora en el rendimiento computacional de los equipos informáticos y de la computación en paralelo.

ii) Los avances en técnicas de minería de datos (Zhang et al., 2011), con la aparición del análisis *Big Data* en el campo del transporte (OECD/ITF, 2015) y, más específicamente, en el de la predicción del tráfico (Abirami and Sridevi, 2017; Lv et al., 2015; Schimbinschi et al., 2015).

iii) El asentamiento del concepto *Internet of Things* (Chen et al., 2017; European Commission, 2017), que permite a numerosos dispositivos volcar sus datos, en tiempo real, en sistemas ITS y en portales accesibles a través de internet.

Estos factores han contribuido al desarrollo de técnicas que en décadas anteriores se abandonaron debido a la imposibilidad de implementaciones reales por su elevado coste computacional y a la ausencia de herramientas para la gestión eficiente del volumen de datos que requieren.

La mayoría de las aproximaciones no paramétricas aplicadas a la predicción del tráfico se basan en los principios del aprendizaje automático, que se fundamentan en el análisis de las características principales de conjuntos de datos reales, obteniendo una forma general que se aproxima gradualmente a una forma más precisa, usando un conjunto de datos creciente para el aprendizaje. Los métodos basados en el aprendizaje automático se separan, a su vez, en dos tipos de aproximaciones:

i) Aprendizaje no supervisado.

ii) Aprendizaje supervisado.

El aprendizaje no supervisado está basado en los principios del reconocimiento de patrones y sistemas caóticos. Se diferencia del aprendizaje supervisado por no

[2] El modelado guiado por datos o *data-driven modeling* se fundamenta en el análisis de los datos de un sistema, en particular, en la búsqueda de conexiones entre las variables de estado del sistema (de entrada, internas y de salida) sin un conocimiento explícito de su comportamiento físico. Estos métodos han constituido una gran contribución en distintas áreas de estudio como la Inteligencia Artificial, la Computación Inteligente, el Aprendizaje Automático, la Minería de Datos o el *Big Data* (Solomatine et al., 2008).

asumir un conocimiento a priori del sistema estudiado y en que el modelado está compuesto sólo de variables independientes, identificando las dinámicas del conjunto de entrada. Dentro de esta categoría se incluyen las técnicas de agrupamiento o *clustering* que, basándose en las relaciones estadísticas de las variables estudiadas, determinan patrones en los datos examinados (Tan et al., 2005). Los métodos de agrupamiento se consideran idóneos para detectar las pautas en el comportamiento recurrente y estacional del tráfico, así como para revelar las relaciones entre parámetros registrados en distintos arcos de la red. Dentro de este tipo de técnicas se integran el agrupamiento por k vecinos más próximos (KNN), k medias (*K-Means*), c medias o agrupamiento difuso (*fuzzy C-Means*) y el análisis de componentes principales (*PCA*), como técnicas que han sido usadas de forma recurrente en métodos de predicción de parámetros de tráfico.

Las técnicas de agrupamiento cumplen dos funciones en el campo del modelado del tráfico:

i) Método de predicción, identificando pautas de comportamiento de una variable y asumiendo que éstas se repiten.

ii) Método auxiliar de clasificación y selección de información en métodos híbridos, junto a otro modelo principal de predicción.

El método KNN, consiste en la clasificación de cada nuevo elemento según el grupo mayoritario al que pertenecen los k elementos más cercanos, que ya han sido agrupados previamente. En el campo de la predicción de parámetros de tráfico, se propuso como alternativa a los métodos paramétricos existentes a principios de los 90s (Davis and Nihan, 1991), mediante una clasificación de perfiles promedios diarios de los parámetros del tráfico. En este caso no se observa una significativa respecto a los métodos paramétricos, aunque se considera que se abre una vía interesante y con posibilidades de mejora en el momento de su publicación.

Ante la proliferación de métodos paramétricos en este campo, varios trabajos comenzaron a profundizar en la comparación del rendimiento de este tipo de modelos con el de los no paramétricos. (Smith et al., 2002) comparan una implementación de KNN con un SARIMA, mostrando un comportamiento más eficiente el método paramétrico. En otro estudio (Habtemichael and Cetin, 2016), la capacidad de los métodos paramétricos para la identificación de patrones es usada por un método al que se denomina KNN mejorado para la predicción de flujo de tráfico, valiéndose de la distancia euclídea ponderada para dar mayor valor a las

mediciones más recientes. KNN mejorado presenta un comportamiento más robusto que implementaciones de modelos ARIMA y KF. Otra versión enriquecida de KNN (Cai et al., 2016), que se apoya en la correlación espacio-temporal de los parámetros del tráfico de distintos arcos, también consigue mejores rendimientos que implementaciones de HA, SVM (*Support Vector Machine*), ANN (red neuronal artificial) y KNN simple. Las fortalezas de este modelo son su facilidad de implementación, su bajo coste computacional y su robustez; que lo hacen idóneo para ser integrado en ITS que necesitan predicciones en tiempo. Otro aspecto reseñable de las conclusiones de este trabajo es que el KNN mejorado se considera eficiente en la estimación de valores en métodos con salida univariable y mejor que ARIMA en condiciones de picos de tráfico.

El método K-means consiste en la agrupación de elementos de una muestra alrededor de k centros, formando k grupos con características similares (Figura 3-7). Ha sido utilizado como método de clasificación de información, en métodos híbridos, para proporcionar una entrada más eficiente al método principal de predicción (Elhenawy et al., 2014). Se ha usado para identificar las distintas fases del día de los flujos de tráfico en una intersección (Zhu et al., 2016). En este mismo sentido, una evolución de este método, denominada *fuzzy C-Means*, basada en la lógica difusa que permite a un mismo elemento pertenecer a varios grupos a la vez, se utiliza como algoritmo de agrupación de perfiles promedios diarios (Caceres et al., 2012) para ser proporcionados a un algoritmo de estimación basado en una Red Neuronal de Base Radial (RBFNN).

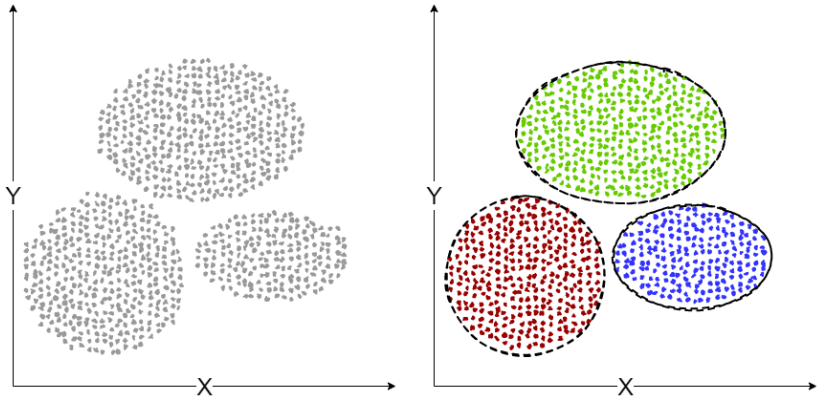


Figura 3-7: imagen representativa de aplicación de método k-means.

En la familia de modelos basados en el aprendizaje no supervisado también se enmarcan las técnicas de descomposición. El análisis de componentes principales, o PCA, es la más ampliamente usada para reducir el número de variables, mientras se conserva la mayor parte de la información del conjunto original de datos (Jolliffe, 2011). Se ha utilizado como clasificaro y como predictor en diversas aplicaciones a aproximaciones de modelado de parámetros de tráfico. Como ejemplo de este segundo uso, se ha utilizado en la fase de clasificación de un método híbrido en el que el método principal es una Red Neuronal con Retardo Temporal, oTDNN, (Dia, 2001), mejorando el rendimiento respecto a una versión sin fase de clasificación. En cuanto a su uso como método de predicción principal, PCA es capaz de identificar los eventos especiales que se producen en la red (Tsekeris and Stathopoulos, 2006) mediante la medición de la variabilidad del flujo de tráfico urbano en toda la red de transporte. En el mismo sentido, mediante PCA se pueden predecir los picos de tráfico en una autovía de dos carriles (Mantecchini, 2011), estableciéndose resultados interesantes en cuanto a la modelización de la variación del tráfico a largo y medio plazo.

La otra gran familia de métodos basados en el aprendizaje automático es la conocida como aprendizaje supervisado. En este caso se infiere una función a través de datos que componen el conjunto de ejemplos de entrenamiento (Russell and Norvig, 2010). A diferencia del aprendizaje no supervisado, se le proporciona el parámetro desalida, utilizándose en el proceso de entrenamiento para establecer la relación entre las variables dependientes e independientes.

La tendencia en el campo de la predicción del flujo de tráfico se está orientando hacia este tipo de aproximaciones, aprovechando las condiciones actuales de potencia computacional, desarrollo de nuevos algoritmos, accesibilidad y volumen de información. Dentro de esta familia de modelado se consideran, principalmente, las Máquinas de Vectores de Soporte (SVM), las Redes Neuronales (NN), los Bosques Aleatorios (RF) y los Algoritmos Genéticos (GA).

El uso de las Máquinas de Vectores de Soporte (SVM) ha sido esporádico en la predicción del flujo de tráfico, aunque ha ofrecido un rendimiento robusto en condiciones concretas.

Las SVM basan su funcionamiento en la determinación de hiperplanos óptimos mediante el conjunto de entrenamiento (Figura 3-8), con el objetivo de clasificar los nuevos ejemplos a partir de dichos hiperplanos (Hastie et al., 2009). Esta técnica se

ha adaptado para trabajar con datos de tráfico en tiempo real en el método denominado *On Line - Support Vector Regression*, u OL-SVR (Castro-Neto et al., 2009), probándose en dos escenarios reales con datos provenientes de PeMS. Se comprueba que, en predicciones con un horizonte de 5 minutos y condiciones típicas de tráfico, es superado por otras aproximaciones paramétricas más sencillas; mientras que, en condiciones atípicas, se muestra superior al resto de los algoritmos con los que se compara (tanto paramétricos con no paramétricos). SVM ha dado evidencias de ser un modelo solvente cuando tiene en cuenta la estacionalidad del tráfico, presentando un buen compromiso entre la precisión de la predicción y la eficiencia computacional con entrada univariable (Lippi et al., 2013).

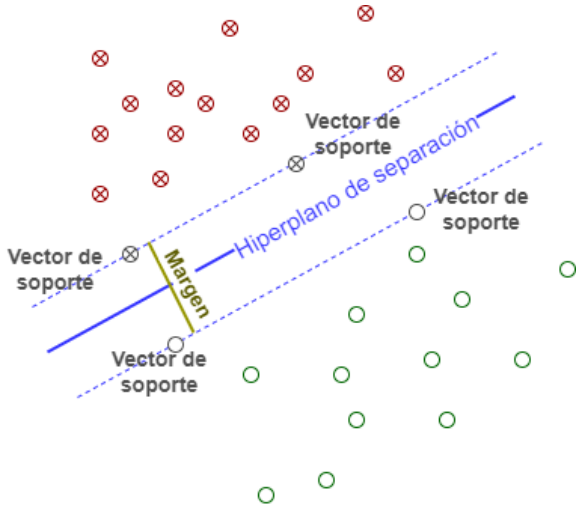


Figura 3-8: Hiperplano óptimo obtenido por mediante el método SVM.

Un ejemplo paradigmático de métodos basados en aprendizaje automático supervisado, aplicado a la predicción del tráfico, es el de los modelos basados en redes neuronales artificiales. Existen numerosos casos de aplicación a este campo desde hace décadas y sus evoluciones siguen siendo una de las alternativas más utilizadas en la actualidad.

Una red neuronal es un algoritmo basado en una estructura de grafo que se compone de una capa de entrada, una capa de salida y varias capas oculta, cada una de ellas compuesta por un número determinado de unidades (Rosenblatt, 1957). Las unidades de una capa se unen mediante enlaces ponderados a todas las

unidades de las capas anterior y posterior, mientras que no existen enlaces entre las unidades de la misma capa. Su representación más simple es el Perceptrón Multicapa, o MLP, que actualiza los pesos de las capas sólo hacia adelante, mientras que la siguiente evolución y la que se ha usado más extensamente es la Red Neuronal con Retro-Propagación, o BPNN, (Benvenuto and Piazza, 1996), en la que los pesos de los enlaces se actualizan hacia adelante con cada registro de entrada, a continuación se calcula el valor esperado, que se compara con el valor real, efectuándose otra actualización de los pesos de los enlaces hacia atrás, con el objetivo de minimizar el error respecto a la salida real (Figura 3-9).

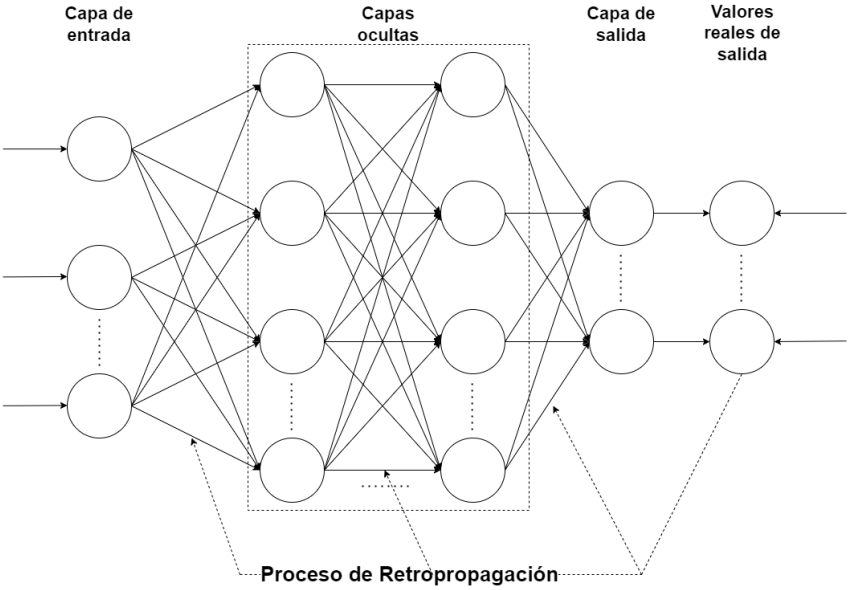


Figura 3-9: Estructura de red BPNN.

Existen numerosas evoluciones de este método, cuyos acercamientos a la predicción de parámetros de tráfico se explican en los siguientes párrafos.

Las primeras aproximaciones que integran predictores basados en redes neuronales en sistemas ITS se producen en 1993 (Gilmore et al., 1993; Gilmore and Elibiary, 1993). En el primer, se utiliza una BPNN que efectúa predicciones de la ocupación de una vía con un horizonte de 30 minutos, baasándose en agregaciones de 5 minutos de los valores de los parámetros de entrada. Se observa que en las predicciones realizadas con un horizonte menor se obtiene un mayor grado de

precisión, atribuido a la fuerte correlación entre los valores de congestión de un intervalo y el siguiente. En otro ejemplo de utilización de un método basado en BPNN, este tipo de red neuronal se ha mostrado útil para para reproducir comportamientos no lineales en la predicción de las variaciones de flujo de la red (Smith and Demetsky, 1994). Ofreciendo un buen rendimiento para horizontes de 15 minutos en condiciones dinámicas del tráfico, superando a las opciones paramétricas de modelado con las que se compara.

Tras los primeros años de su aplicación a la predicción de parámetros de tráfico con resultados diversos, (Kirby et al., 1997) plantean una cuestión fundamental: ¿qué tipo de modelos ofrecen mejores rendimientos en cuanto a la predicción del tráfico? ¿las redes neuronales o modelos estadísticos/paramétricos? En dicha publicación se compara una implementación de BPNN con el sistema ATHENA, un modelo híbrido basado en la combinación de K-Means + ARIMA para la predicción del volumen de tráfico. ATHENA rinde mejor en la mayoría de los casos presentados, especialmente con horizontes de predicción superiores a los 30 minutos. Los métodos son probados en 48 localizaciones de distintas autovías francesas conectadas entre sí. Si bien se esperaba un mejor rendimiento del modelo basado en redes neuronales, se llega a la conclusión de que el número de casos del conjunto de datos de entrenamiento es insuficiente para que la red neuronal generalice el comportamiento del tráfico en los arcos estudiados.

En una aplicación de una BPNN a la predicción del tráfico en varias autovías interurbanas de Holanda (Dougherty and Cobbett, 1997), se utilizan todos los parámetros disponibles en la zona (intensidad, velocidad y ocupación en cada uno de los 16 puntos estudiados), como conjunto de entrada. Los resultados obtenidos son satisfactorios desde el punto de vista de la precisión del modelo, aunque el tamaño de la red neuronal la hace poco práctica para su aplicación real, debido al excesivo esfuerzo computacional que requiere su estructura. Se propone una fase previa de selección de información del área de estudio, probando una técnica de reducción del tamaño de la red mediante un testeo de elasticidad; comprobándose que la reducción no repercute negativamente en el rendimiento del modelo, a la vez que simplifica su estructura.

Las Red Neuronal con Retardo Temporal (TDNN), tratan de aprovechar las dependencias temporales entre los parámetros del conjunto de entrada (Waibel et al., 1990) . Se diferencian de la configuración del MLP en que todas las unidades, en cada capa, obtienen entradas de una ventana contextual de salidas de la capa

anterior. Es decir, cada unidad recibe una serie de entradas que son las salidas de varios periodos pasados de la unidad anterior. De esta manera se introduce la variación de los valores de entrada a lo largo del tiempo en la estructura de la red neuronal. La invariabilidad del sistema se consigue eliminando explícitamente la dependencia de la posición en el proceso de retropropagación del entrenamiento, mediante la construcción de copias temporales de la red a lo largo de la dimensión del tiempo.

Las TDNN se han utilizado para predecir el flujo y la ocupación basándose en su perfil temporal más reciente y la información aportada por otras localizaciones cercanas (Abdulhai et al., 1999), encontrándose que:

i) El rendimiento del modelo es inversamente proporcional a la extensión de la contribución espacial.

ii) La inclusión de tres detectores en ambas en ambos sentidos respecto del detector objetivo es suficiente para obtener un rendimiento que cumpla con los objetivos prácticos.

iii) Cuanto más se amplía la ventana temporal mayor es la similitud entre la predicción y la media, perdiéndose precisión en casos específicos, por lo que este valor no debe ser excesivamente amplio.

iv) Para predicciones con un horizonte más extenso resulta conveniente tratar los datos con un mayor grado de agregación, concluyendo que el nivel de agregación de los datos debe ser del mismo orden que el del horizonte de predicción considerado, para mejorar la precisión.

En el mismo sentido, en (Dia, 2001) se proponen dos versiones diferentes de métodos basados en TDNN para comprobar si se produce una mejora mediante la introducción de una fase previa de clasificación, en concreto se comparan un método de predicción simple consistente en TDNN, un método híbrido con una primera fase de clasificación consistente en PCA+TDNN y un método de predicción basado en MLP. Todos emplean como información de entrada los valores de flujo y velocidad del propio detector y de los que se encuentran en su entorno más próximo. Presenta como parámetros de salida los valores de velocidad media y de tiempo de viaje en el tramo con horizontes de 5 y 15 minutos. Los resultados obtenidos muestran que el método híbrido obtiene resultados ligeramente peores que el TDNN simple, pero ambos mejoran los resultados obtenidos por MLP. En el caso de la velocidad se asegura un alto grado de

precisión para un horizonte de 5 minutos, mientras que para el tiempo de viaje el horizonte en el que la precisión se considera alta se amplía hasta los 15 minutos.

El aprendizaje basado en Redes Neuronales Convolucionales o Convolutional Neural Networks (CNN) también ha realizado su contribución en la predicción del flujo de tráfico por su capacidad para explotar las propiedades espaciales en las matrices que representan imágenes y vídeos. El método denominado *Convolutional Long-Term neural network for traffic Flow Predicion* CLTFP (Wu and Tan, 2016) utiliza las capacidades de este tipo de redes en la predicción del flujo de tráfico. La aproximación propuesta en la publicación en la que se presenta dicho método (Gers, 2001) se basa en una combinación de tres métodos que se centran en diferentes características del flujo de tráfico:

i) Una CNN de una dimensión que captura las características espaciales.

ii) Una red de tipo *Long Short-Term Memory Neural Network*, o LSTMNN, para capturar la variabilidad a corto plazo.

iii) Una LSTMNN para capturar la periodicidad que se produce en el flujo de tráfico.

Como conclusiones se determina que las características capturadas por LSTMNN presentan unos resultados modestos en cuanto a la precisión de la predicción, que el flujo de tráfico se ve afectado por muchos otros factores como el tiempo o los eventos sociales, que se deben integrar como en los modelos de predicción y el modelo se debe probar en una red de más amplia escala para comprobar su validez.

Este tipo de modelos ha seguido evolucionando para integrar, de manera más eficiente las relaciones, espacio-temporales entre los datos de entrada del modelo. Esta evolución ha derivado en un método denominado *Dynamic-GRCNN*, que supera a otros modelos basados en técnicas *Deep-Learning* en un caso de estudio sobre *Beijing* (Peng et al., 2020).

El Modelo Neuronal Difuso, o FNM, (Yin et al., 2002), constituye una aproximación, basada en un essquema híbrido, para la predicción del flujo en las salidas de una intersección urbana. El método consiste en una primera fase de agrupamiento difuso mediante la que se clasifican localizaciones con perfiles promedios diarios similares para el flujo de tráfico y una fase de predicción basada en una BPNN. A cada uno de los grupos definidos en la primera fase le corresponde un nodo en la

capa de entrada de una red BPNN utilizada para la predicción. Los resultados muestran una mejora considerable respecto a la red neuronal convencional en el mismo escenario.

En otra aproximación, se propone un método híbrido, que combina un KF con una red neuronal cuya estructura se define mediante una búsqueda dirigida basada en un sistema de reglas difusas (Stathopoulos et al., 2008). En este mismo sentido, (Kolidakis et al., 2019) presenta una técnica basada en el Análisis Singular Espectral o SSA y una ANN, orientada a la predicción del tráfico en autovías de peaje. Se trata de un metodo preparado para manejar grandes volúmenes de datos. Se concluye que SSA supera a la ANN simple en el caso de estudio sobre el que se prueba. En esta publicación se incorpora la optimización de los hiperparámetros del modelo usado como una fase más integrada en el proceso de modelado.

(Vlahogianni et al., 2005) destacan que el diseño de la estructura de los algoritmos de aprendizaje supervisado constituye un campo tratado con poca profundidad en el modelado de predicción de tráfico hasta ese momento. A partir de este estudio, y ante el creciente número de propuestas de distintas versiones de algoritmos basados en AI, se comienza a prestar una mayor atención a este aspecto. El ajuste de hiperparámetros[3] es un asunto complejo, que se afronta en en este mediante una búsqueda dirigida a través de un algoritmo genético que encuentra de forma eficiente una estructura de red neuronal, adaptada al problema real presentado, con una metodología extensible a otros casos de estudio.

La evolución del aprendizaje automático en los últimos años ha cristalizado en una nueva familia de técnicas denominada Aprendizaje Profundo o *Deep Learning* (DL). Este tipo de aproximación intenta modelar abstracciones de alto nivel en conjuntos de datos, usando arquitecturas compuestas por transformaciones múltiples no lineales (Bengio et al., 2012). En el campo de la predicción del tráfico, este tipo de aprendizaje se ha materializado en forma de BPNNs con numerosas capas ocultas, o también llamadas redes profundas y en Autocodificadores Apilados o *Stacked Auto Encoders* (SAE).

Los SAE se fundamentan en una estructura de red profunda cuyas capas ocultas

[3] En Aprendizaje Automático, un hiperparámetro es un parámetro cuyo valor es definido antes de que comience el proceso de aprendizaje (en contraste con los parámetros que se actualizan con el entrenamiento). El proceso de ajuste de hiperparámetros es la búsqueda de los valores que optimizan el rendimiento del método en el caso concreto en el que se utiliza, normalmente definiendo su estructura.

están compuestas por autocodificadores. Un autocodificador consiste en una capa que es entrenada de forma individual mediante el conjunto de entrenamiento, o una porción de éste, con el objetivo de inicializar los pesos de los enlaces, previamente a su inclusión en la estructura profunda. Mediante los autocodificadores se aumenta la velocidad de aprendizaje, en comparación con el uso de capas ocultas no preinicializadas, reduciendo el coste computacional del algoritmo al favorecer una convergencia más rápida (Le, 2015).

Dentro del campo de la predicción de parámetros de tráfico, este tipo de redes ha sido utilizada para la detección y predicción de posibles estados de congestión en redes de transporte de gran escala (Ma et al., 2015), mostrándose superior a otros métodos de como BPNN y SVM.

En otro estudio, (Lv et al., 2015) implementan una aproximación basada en SAE, que aprende cada característica genéricas del flujo del tráfico a través de un autocodificador individualizado. En el artículo se justifica esta propuesta al considerar que las ANNs poco profundas no han obtenido resultados completamente satisfactorios en sus aplicaciones en escenarios reales. Las alternativas SAE presentan mejores resultados que métodos basados en BPNN, SVM y RBFNN para horizontes de predicción de 15, 30, 45 y 60 minutos, probándose sobre tres detectores, localizados en autovías gestionadas por el PeMS.

El método *Variational Auto-Encoder*, o VAE (Boquet et al., 2020), aporta un esquema previo de preprocesamiento que reduce el ruido en el conjunto de datos de entrada, completa los valores perdidos mediante un modelo de imputación y comprime los datos para reducir su volumen. Presenta buenos resultados en los tres casos de estudio incluidos en el artículo, del que se extraen como conclusiones el preprocesamiento supone una mejora, además de ayudar a identificar características prácticas para el modelado y características descriptivas del problema, útiles para los modeladores del tráfico.

Los métodos basados en el aprendizaje profundo se han mostrado capaces de capturar las transiciones no lineales entre los distintos regímenes de tráfico (Polson and Sokolov, 2017). Se presenta un algoritmo basado en el aprendizaje profundo para realizar predicciones de flujo a corto plazo. Este método que muestra buen rendimiento en condiciones especiales del tráfico, como eventos deportivos. En esta

aproximación se ha optimizando la arquitectura mediante una regularización l1[4], que utiliza la función de activación de tipo tanh. En el método desarrollado también se utilizan las relaciones espacio-temporales de los parámetros que intervienen en el problema, introduciendo, como parámetros de entrada, los valores de los detectores vecinos con una ventana de tiempo amplia. Se determina que los datos de las observaciones incluidos en una ventana formada por los 40 minutos más recientes explican mejor el parámetro de salida que ventanas más amplias o, más restringidas.

En estudios más actuales, se han integrado los valores provenientes de vehículos autónomos en el conjunto de entrada de modelos basados en SAE (Vázquez et al., 2020). Abriéndose una vía para la utilización efectiva de ambos tipos de datos como entrada de modelos basados en el aprendizaje automático

Los árboles de decisión (Breiman et al., 1984)constituyen una opción de método basado en el aprendizaje automático con una estructura y lógica diferente a las redes neuronales. Consisten en una estructura de diagrama de construcciones lógicas que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva para la resolución de un problema. Su evolución inmediata es la técnica de Bosque Aleatorio o *Random Forest* (RF) cuyo funcionamiento se basa en una combinación de árboles predictores, tales que los casos del conjunto de entrada de cada árbol dependen de los valores de un vector aleatorio, probado independientemente y con la misma distribución para cada una de éstos (Breiman, 2001).

Este tipo de modelos ha sido usado en la predicción de parámetros de tráfico ofreciendo resultados equivalentes, o superiores, a los mostrados por modelos basados en otros métodos, cuando la ventana temporal de información anterior no es muy amplia (Schimbinschi et al., 2015). En esta publicación, el método RF presenta peores valores de rendimiento a medida que se amplía la ventana de información anterior, frente a dos modelos basados en Regresión Logística y BPNN.

La robustez y eficiencia de RF se utiliza como base del algoritmo *Selective Random*

[4] La regularización consiste en añadir una penalización a la función de coste (Hua Yang et al., 1998). El objetivo de esta penalización es la genración de modelos más simples, que generalizan mejor. Las regularizaciones más usadas en técnicas *Machine Learning* son: *Lasso* (también conocida como l_1), *Ridge* (conocida también como l_2) y *ElasticNet* que combina tanto l_1 como l_2.

Subspace Predictor (Sun and Zhang, 2007) para la predicción del flujo de tráfico usando la información completa del área de implementación como conjunto de entrada. Se realiza un preprocesamiento basado en la correlación espacio-temporal entre los datos.

3.1.3.3. Modelos de combinación

La combinación de modelos se propopne ante la discusión no resuelta sobre la superioridad clara de una aproximación metodológica frente a las demás (Dietterich, 2000). Se trata de un acercamiento que evita la toma de dicha decisión, al mismo tiempo que permite aprovechar las fortalezas de distintas técnicas en un mismo modelo.

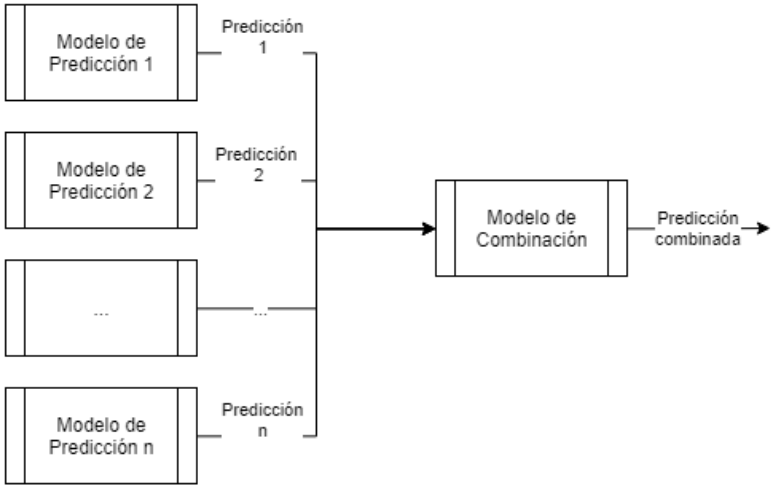


Figura 3-10: Esquema de combinación de modelos de predicción.

Se basa en la combinación de los valores de salida de los predictores para la obtención de un único valor modelado (Figura 3-10). Éste se calcula mediante un algoritmo de combinación, con el que se consigue un modelo más eficiente y robusto y que oculta las debilidades de cada predictor simple.

Como se ha observado previamente en este mismo capítulo, las distintas aproximaciones de modelado ofrecen un rendimiento variable que es dependiente de las características del caso de estudio que modelan. Estas características se

refieren a la información del caso de estudio, que depende de la escala, el tipo de red, el horizonte de predicción, el parámetro específico que se modela, las condiciones del tráfico observadas, el nivel de desagregación de los datos, la cantidad de casos previos observados, el nivel de valores perdidos, etc...

Este esquema de modelado compone de varios predictores simples y un método de combinación. El método de combinación es fundamental, ya que es la herramienta optimiza el valor modelado en función de los valores obtenidos por cada predictor simple. Este método se debe elegir siguiente un compromiso entre el grado de mejora obtenido y el nivel de complejidad introducido en el proceso.

A continuación, se detallan algunos métodos de combinación que han sido utilizados en el campo de la predicción de parámetros de tráfico a corto plazo.

Las Redes Bayesianas son uno de los modelos más extendidos como método de combinación. Modelan un fenómeno mediante un conjunto de variables y las relaciones de dependencia entre ellas, estimando la probabilidad posterior de las variables no conocidas, en base a las variables conocidas.

Se comienzan a utilizar en la predicción del tráfico mediante un método denominado BCP o Predictor de Combinación Bayesiano (Petridis et al., 2001) que utiliza predicciones de diversos modelos simples y produce una predicción final que consiste en la combinación ponderada de sus valores modelados. BCP obtiene mejores resultados que los modelos simples en los que se basa sobre dos casos reales.

En otra versión de la combinación bayesiana, se propone la denominada Red Neuronal de Combinación Bayesiana, o BCNN, cuya particularidad radica en el uso de distintas variantes de redes neuronales como métodos simples, concretamente una BPNN y una RBFNN (Zheng et al., 2006). Se utiliza para la predicción de flujo de tráfico, con un horizonte de 15 minutos, en una autovía. Observándose que el modelo combinado presenta un rendimiento superior al de los predictores simples en los que se basa.

Debido a que el método BCP produce una redundancia en el cálculo del error de los distintos predictores simples, en (Wang et al., 2014) se propone una implementación mejorada, más sensible a la perturbación del rendimiento de los predictores simples y que actualiza sus factores más rápidamente. La implementación concreta se realiza usando como predictores simples ARIMA, KF y BPNN, demostrándose que el nuevo método mejora considerablemente los

resultados del BCP original en precisión y estabilidad.

Además de las Redes Bayesianas se han propuesto aproximaciones para la combinación de modelos desde otro tipo de lógicas, como el Sistema Basado en Reglas Difusas o *Fuzzy-Rule Based System* (FRBS) (Stathopoulos et al., 2008). La combinación se basa en los Sistemas Difusos (Bellman and Zadeh, 1970), técnicas fundamentadas en la inteligencia artificial con la habilidad de incorporar conocimiento experto; un concepto que, teóricamente, se parece a la estructura del proceso humano en la toma de decisiones. El método es probado en una vía arterial urbana de 3 carriles con semáforos, con datos agregados cada 3 minutos. En todos los resultados obtenidos, FRBS muestra un rendimiento superior a los modelos individuales en los tres regímenes considerados (baja congestión, alta congestión y periodo total de predicción). El uso de una predicción combinada puede representar más adecuadamente las rápidas fluctuaciones asociadas a la no estacionalidad y no linealidad del flujo de tráfico en condiciones de congestión severa. También presenta una respuesta más rápida y sincronizada en las grandes fluctuaciones de los datos observados, particularmente durante el régimen de aumento de congestión en comparación con otros modelos; particularmente el KF presenta el defecto de tener un retardo temporal para adaptarse a este tipo de situaciones

La fusión de métodos es una metodología de combinación que combina la predicción de tráfico con métodos de predicción de variables explicativas, asociadas a otras temáticas, como métodos correctores de la salida. En este sentido, se han utilizado los valores de precipitación para corregir las predicciones de flujo de tráfico (Qiu et al., 2016) utilizando parámetros relativos a la pluviometría en la zona de estudio. El elemento diferenciador de este método es que la corrección mediante la variable explicativa extra se realiza previamente a la combinación de las predicciones de los modelos base.

Se extrae como conclusión que los modelos de combinación han ido estableciéndose con firmeza entre las aproximaciones de predicción a corto plazo. Aunque cabe plantearse si la robustez que se obtiene mediante la combinación compensa el aumento de complejidad introducido mediante el mecanismo de combinación. Esta cuestión se ha tratado de dilucidar mediante la aplicación de su utilización en diversos casos prácticos (Tselentis et al., 2015), concluyéndose que merece la pena desarrollar y aplicar modelos de combinación, ya que el esfuerzo de la implementación y la complejidad asumida es menor que el riesgo de elegir una

mala opción de modelado que necesite un excesivo aporte de información espacio-temporal para suplir sus defectos.

3.2. Imputación de conjuntos con datos incompletos

Las técnicas de predicción de tráfico a corto plazo se ven afectadas por los errores en los conjuntos de datos que utilizan como información de entrada. En este capítulo se analizan las causas que provocan estos errores y las técnicas utilizadas para atenuar sus efectos sobre la predicción de parámetros de tráfico.

Se toman en consideración dos tipologías de fuentes de errores; por una parte, los datos perdidos, que son aquellos valores que no se registran en el conjunto de datos y, por otro lado, los datos corruptos, que se llegan a registrar, aunque con un valor erróneo. Ambos tipos de errores son detallados en la sección 3.2.1. En la sección 3.2.2 se revisan los métodos utilizados en el campo de la predicción del tráfico para atenuar los efectos de los valores perdidos y para completarlos en el caso en el que se estime conveniente.

3.2.1. Datos perdidos y corruptos

Los conjuntos de datos que utilizan, como información de entrada, los modelos de predicción de parámetros de tráfico presentan registros perdidos y corruptos (van Lint et al., 2005) de forma generalizada.

En un caso de estudio de 2005 en el sistema MONICA, que monitoriza las autovías holandesas, se detectaba un promedio de un 15% de registros perdidos por los detectores desplegados en dicho sistema (van Lint et al., 2005). En esta publicación se observa que todos los sensores, sin excepción, presentan intervalos con valores perdidos. En otro ejemplo, en el sistema de control de datos de la ciudad de Gliwice, en Polonia, se detectó un valor del 10% entre valores perdidos y corruptos sobre el total de registros (Pamuła, 2018). En el sistema de control de tráfico de la ciudad de Madrid se advierte que varios detectores presentan una tasa de valores perdidos o corruptos mayor al 50% del total de los registros (Laña et al., 2018a), aunque la cantidad de detectores que presentan esta característica es variable, dependiendo del área observada. En zonas concretas se observa que el 9% de los detectores presentan una tasa de valores perdidos inferior al 3%. Por lo que este

efecto se produce de manera dispar en el conjunto de detectores de un sistema de control.

Estos ejemplos sirven para dar una idea de la presencia generalizada de valores perdidos y corruptos en los conjuntos de datos de los sistemas de gestión de tráfico, dificultando el cumplimiento de sus funciones. Desde el punto de vista del modelado de predicción, los valores perdidos y corruptos influyen negativamente en su rendimiento, por lo que se hacen necesaris la detección, clasificación y tratamiento de este tipo de fallos.

3.2.1.1. Datos perdidos

Los datos perdidos son aquellos que no llegan a ser reflejados en el conjunto de datos. Se producen por diversos motivos, los más generalizados son debidos a fallos de funcionamiento de los sensores, o a errores de transmisión de información en el sistema de adquisición de datos (Chen et al., 2001).

Tabla 3-2: Cuadro resumen de los tipos de valores perdidos según la aleatoriedad con la que se presentan.

Tipo	Grado de aleatoriedad	Características	Causa
MCR	Se producen de forma totalmente aleatoria.	Se identifican al aparecer periodos puntuales de valores perdidos aleatoriamente dispersos.	Fallos puntuales del dispositivo, o fallos de comunicación momentáneos.
MR	Se producen de manera parcialmente aleatoria.	suelen aparecer como una serie de periodos puntuales de valores perdidos no tan dispersos o de manera secuencial.	Fallo del dispositivo o del sistema de comunicación durante un intervalo de tiempo.
MNR	No se producen de forma aleatoria.	Siguen un patrón identificable. Suelen mostrarse como un largo periodo seguido de valores perdidos. O como varios intervalos continuados de valores perdidos cuya aparición se produce en algún patrón temporal observable.	Fallos continuados en el tiempo, debidos a situaciones de saturación del dispositivo, o a fallos de comunicación concretos, etc…

En la Tabla 3-2 se realiza una síntesis de los tipos de datos perdidos, así como de la

causa principal que los provoca. Los datos perdidos se dividen en tres tipos (Qu et al., 2009) según la aleatoriedad de su aparición:

i) Datos perdidos de forma completamente aleatoria o *Missing Completely at Random* (MCR): cada ocurrencia de dato perdido se produce de forma totalmente independiente de cualquier otra.

ii) Datos perdidos de forma parcialmente aleatoria o *Missing at Random* (MR): los datos se pierden de forma parcialmente aleatoria, guardando un cierto grado de relación con otras ocurrencias.

iii) Datos perdidos de forma no aleatorios o *Missing Not at Radom* (MNR): los datos perdidos siguen un patrón observable y claramente identificable.

Las tres tipologías de fallos pueden producirse de manera simultánea en casos reales (van Lint et al., 2005).

3.2.1.2. Datos corruptos

Además de los datos perdidos, en los conjuntos datos de parámetros de tráfico se producen datos corruptos. En este caso, el valor sí queda registrado en el conjunto de datos, aunque no refleja la situación real de la vía, debido a problemas de transmisión, o a errores en su manipulación desde que es emitido por el dispositivo hasta que se registra.

Los datos corruptos pueden ser detectados y tratados. Es posible identificarlos atendiendo a la relación entre los valores de los parámetros fundamentales (Chen et al., 2003).

Se encuentran cuatro tipos de situaciones:

i) La ocupación y el flujo presentan, mayoritariamente, el valor '0'.

ii) Ocupación mayor que '0' y flujo con valor '0'.

iii) Muy alta ocupación durante un largo periodo de tiempo.

iv) Ocupación y flujos constantes.

En la Tabla 3-3 se representan de manera sintética los 4 tipos de fallo y las causas que los provocan. En dicha tabla, el parámetro q representa la intensidad, el parámetro k la ocupación y el parámetro t el tiempo.

Tabla 3-3: Cuadro resumen de las tipologías de datos corruptos (Chen et al., 2003).

Tipo de dato corrupto	Descripción	Causa
$q == 0$	Mayoría de valores de intensidad y ocupación a '0'.	Sensor bloqueado (el sensor deja de tomar datos, aunque sigue registrándolos en el conjunto de datos).
$k == 0$	Valores de intensidad a 0, y ocupación mayor que 0.	Sensor saturado (el sensor se queda sin capacidad para almacenar valores de alguna variable).
$q == 0$	Ocupación mucho mayor que 0 durante un largo periodo	Sensor bloqueado, el sensor se ha bloqueado y ha dejado de actualizar el valor aunque sigue registrando.
$k > 0$	Intensidad y ocupación constantes durante varios intervalos.	Sensor saturado o bloqueado.

3.2.2. Métodos de imputación

Cada tipología de valor perdido produce un impacto diferente sobre los métodos de predicción. Desde los ITS se han utilizado diversas metodologías para tratar los valores erróneos y corruptos, normalmente sustituyéndolos mediante agregaciones o con valores estimados.

Debido a la generalización de la presencia de datos perdidos y corruptos en conjuntos de datos reales de tráfico, se considera que el esquema básico de predicción debe contemplar una etapa de tratamiento de valores perdidos, siendo la imputación la de uso más generalizado. El esquema de modelado de predicción, que se presenta en la Figura 3-4, se debe añadir una etapa de imputación, quedando configurada tal y como se presenta en la Figura 3-11.

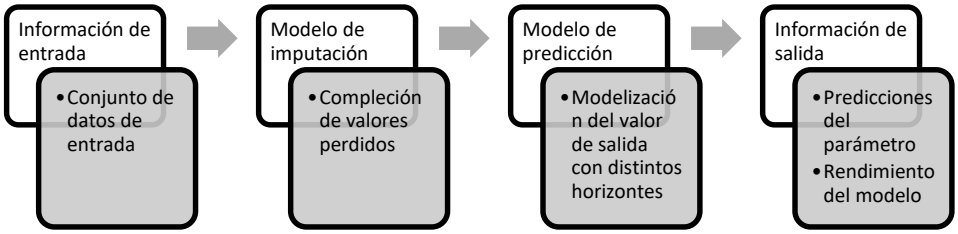


Figura 3-11: Esquema básico de modelado de predicción incluyendo la etapa de imputación de valores perdidos

Los primeros métodos de imputación implementados por los ITS seguían lógicas muy básicas de sustitución, siendo habitual el uso de la sustitución por el valor '0' y la sustitución por el valor promedio del parámetro, aunque se observó que generaban un cierto grado de distorsión sobre los principales indicadores de datos de tráfico a largo plazo como el Promedio Anual de Tráfico Diario o *Average Annual Daily traffic* (AADT), o el volumen en la hora de diseño o *Design Hour Volume* (DHV) (Zhong et al., 2004). Surge la necesidad de introducir métodos de imputación que produzcan la menor alteración posible sobre la tendencia general de la serie de datos afectada.

Antes de realizar la sustitución de los valores perdidos, se necesita conocer su tipología y su influencia sobre el conjunto de datos. En varios estudios se apunta a esa necesidad, proponiéndose la inclusión de una herramienta de diagnosis de datos corruptos y erróneos en los conjuntos de parámetros de tráfico (Bie et al., 2016; Chen et al., 2003; Li et al., 2014), con el objetivo de analizar su estructura y ser tenida en cuenta en el proceso de modelado de predicción.

Se han realizado diversas clasificaciones de las aproximaciones para la imputación de parámetros de tráfico, debido a la variedad de métodos existentes. En primer término, se propone una clasificación entre métodos simples/factores, métodos de regresión y métodos espacio-estado (Kamarianakis and Prastacos, 2005).

En una actualización posterior, (Zhang, 2012) incluye una nueva categoría para considerar los métodos basados en aprendizaje automático. Posteriormente, en (Asif et al., 2016) se vuelve a revisar esta clasificación, dividiendo a los métodos de imputación en dos categorías:

i) Modelos basados en Función de Estimación, para los que se asume que los datos perdidos se producen aisladamente, desde un punto de

vista espacial y temporal, y que pueden generalizar el comportamiento de los valores perdidos los datos históricos (en el que se incluirían todas las aproximaciones citadas hasta ahora en este párrafo).

ii) Métodos de Compleción mediante Cálculo Tensorial, que no requieren datos entrenados para calcular la imputación, se basan en las relaciones multidimensionales, subyacentes en los datos, explotando varias medidas temporales (semana, tipo de día, hora), y la dimensión espacial.

En este trabajo se asume esta última clasificación, aunque se desagrega la primera categoría en los subgrupos más representativos, proponiéndose la siguiente clasificación de métodos basados en:

i) Sustitución Simple.

ii) Métodos Numéricos y Estadísticos.

iii) Series Temporales.

iv) Aprendizaje Automático.

v) Tensores.

A continuación, se describe cada categoría, incluyéndose diversos ejemplos de aplicación a casos de estudio relacionados con parámetros de tráfico y las características observadas en cada uno de ellos

3.2.2.1. Imputación basada en sustitución simple

Hasta finales de los 90s, la práctica habitual de los agentes de gestión de tráfico había sido completar los datos con métodos simples, utilizando valores agregados o constantes (Chen et al., 2001).

Los tres métodos más usados en este sentido son:

i) *Hot-deck:* sustitución del valor perdido con un valor aleatorio de la misma serie de datos.

ii) *Cold-deck:* el valor de sustitución se toma de la serie histórica del parámetro que se sustituye.

iii) Sustitución por una constante: se reemplazan los datos perdidos con un valor constante, normalmente '0'.

iv) Sustitución por el promedio: Se utiliza la media de todos los datos históricos del propio parámetro como valor de sustitución.

Dentro de esta última forma de sustitución existen diversas variantes, como la sustitución por el promedio del valor total de la serie de datos, o por intervalos en los que se observa cierta recurrencia (Chen et al., 2001). De entre estas técnicas, la sustitución por la media es la que lleva a los mejores resultados de imputación y la que se muestra menos sensible al aumento de la tasa de datos perdidos.

3.2.2.2. Imputación basada en métodos numéricos y estadísticos

En este grupo se incluyen diversas técnicas, de complejidad dispar, basadas en métodos con base numérica, estadística y probabilística. Se suelen diferenciar entre:

i) Métodos con entrada univariable, en los que el valor imputado se estima solo con los valores anteriores del propio parámetro.

ii) Métodos con entrada multivariable, se utilizan valores anteriores del propio parámetro y de otros relacionados espacio-temporalmente con el parámetro de salida.

Una de las primeras aproximaciones interesantes en este sentido consiste en la estimación de los valores perdidos en una intersección (Whitlock and Queen, 2000), que se representa mediante un sistema multivariable, en el que se modela el comportamiento estocástico del tráfico a través de una Cadena de Markov en combinación con el Método de Monte Carlo. Las relaciones de probabilidad se modelan a través de una Red Bayesiana, presentando un buen rendimiento en los casos en los que se dispone de suficiente información previa.

Las relaciones lineales espacio-temporales presentes en los conjuntos de datos se comienzan a explotar con este tipo de modelos de imputación, comprobándose que los métodos que utilizan correlaciones basadas en datos históricos (Qu et al., 2009) son más precisos que aquellos que utilizan interpolación basada en la utilización de sensores vecinos, los que utilizan valores referidos al mismo intervalo y métodos de sustitución simple (Chen et al., 2003).

La inclusion de cierta recurrencia considerando, como parámetros de entrada, el día de la semana y la hora en la que se produce el registro del valor, aumenta la precisión de los modelos de imputación, reflejándose en una reducción de la influencia sobre AADT y DHV (Zhong et al., 2004). En este mismo sentido, la

influencia entre los valores de los distintos parámetros, registrados en localizaciones cercanas de una red, se modela mediante el concepto de vecindad de sensores. Estas relaciones se establecen por medio de las correlaciones espacio-temporales entre sus parámetros, ya sea con los datos anteriores del propio parámetro (Chen et al., 2003), o de otros parámetros del mismo detector y detectores vecinos (Bie et al., 2016; Henrickson et al., 2015).

Para optimizar la búsqueda de estas relaciones, se ha propuesto un método más sofisticado basado en la descomposición del conjunto de datos para identificar la información con mayor valor explicativo, ayudando a los modelos a realizar una gestión eficiente de la información existente. Se han usado diversas técnicas para dicha descomposición como la descomposición por media simples y PCA (Li et al., 2014) y la descomposición basada en onda (Qu et al., 2009).

La descomposición mediante PCA se ha mostrado muy útil como clasificador de información en un primer paso de métodos híbridos de imputación. En esta tipología de métodos se enmarca el Análisis Probabilístico de Componentes Principales o PPCA (Qu et al., 2009), para extraer las características espacio-temporales del flujo del tráfico. Este método presenta la ventaja de tener un bajo coste computacional y buen rendimiento en condiciones no típicas del tráfico. También se ha usado el método PCA en la detección de tendencias a corto plazo de los parámetros del tráfico (Li et al., 2014), definidas por los perfiles promedios de la tendencia inmediatamente anteriores a la estimación.

Se han desarrollado diversas evoluciones de PPCA; en una de ellas se presentan cuatro tipos (Li et al., 2013):

i) Una versión univariable.

ii) Una versión que introduce la variable temporal (teniendo en cuenta una ventana temporal consistente en valores de varios intervalos anteriores al que se pretende imputar).

iii) Una versión que introduce la variable espacial (que tiene en consideración los valores de los sensores vecinos para el mismo intervalo de la imputación).

iv) Una versión que combina todas las anteriores (se considera la ventana temporal del propio detector y de sus vecinos).

Se obtiene como conclusión que la inclusión de las variables espacio-temporales, en

este tipo de métodos, mejora considerablemente los resultados, a cambio de un notable aumento del coste computacional.

Las técnicas de imputación basadas en métodos numéricos y estadísticos presentan, como característica común, la representación de una manera fehaciente de las relaciones entre las series de datos y una correspondencia aceptable entre coste computacional y precisión de la imputación, evidenciándolas como una buena opción para su inclusión en sistemas que requieren imputaciones en tiempo real.

3.2.2.3. Imputación basada en series temporales

Los modelos de la familia ARIMA han sido una opción usada de manera recurrente en la imputación de valores perdidos. Presentan como mayor ventaja su bajo coste computacional y un rendimiento aceptable con un histórico de datos no excesivamente grande. Esta técnica muestra resultados deficientes, o la imposibilidad de imputar, en caso en que el parámetro objetivo presente valores perdidos durante una serie de intervalos continuados (Chen et al., 2003; Haworth and Cheng, 2012).

Los índices AADT y DHV, muestran menos sensibilidad cuando el conjunto de datos ARIMA, que cuando se hace con modelos basados en sustitución simple (Zhong et al., 2004). Los métodos ARIMA se han utilizado en diversas aproximaciones mediante modelos híbridos. (Chen et al., 2001) presentan un método que combina un Mapa Auto Organizado, o SOM y un modelo ARIMA, al que denomina SOM/ARIMA. SOM es una técnica de aprendizaje automático no supervisado que se utiliza para reducir las dimensiones del conjunto de datos sobre el que se aplica, en este caso se utiliza como método de clasificación. El conjunto clasificado se proporciona a un modelo de predicción basado en ARIMA. Se observa que SOM/ARIMA es superior a los ARIMA simples con los que se compara, aunque sigue presentando una sensibilidad alta al aumento de la tasa de valores perdidos.

A pesar de sus limitaciones, los métodos ARIMA son una buena herramienta cuando el parámetro objetivo muestra tasas bajas de valores perdidos, un histórico de datos reducido y los periodos con datos perdidos se producen de forma puntual (Zhang and Zhang, 2016). En casos en los que la cantidad de datos disponibles es limitada, presentan un mejor rendimiento que los métodos basados en el aprendizaje automático, que necesitan una muestra suficiente para generalizar el comportamiento del parámetro de salida.

Dentro de este grupo también se incluyen los métodos espacio-estado. Éstos introducen la toma en consideración los retardos temporales, con medidas agregadas de información histórica, información espacial, etc... Si el volumen de información espacio-temporal es suficientemente completo, pueden modelar el tráfico muy fehacientemente en vías arteriales urbanas (Zhang and Liu, 2009) aunque su uso no ha sido muy extendido en el modelado de imputación.

3.2.2.4. Imputación basada en aprendizaje automático

El uso de métodos de imputación basados en aprendizaje automático no ha sido tan pródigo como en los métodos de predicción. Aun así, se han generado estudios interesantes, que tratan de explotar las relaciones existentes en los conjuntos de datos, basadas en SVM y redes neuronales.

Los métodos de imputación presentan ciertas deficiencias en casos en los que se produce una alta tasa de valores perdidos, como la dificultad de ser aplicados en tiempo real y la imposibilidad de calcular imputaciones cuando se produce un largo periodo sin datos. Para dar respuesta a dichas carencias (Haworth and Cheng, 2012) proponen un método basado en SVM para la búsqueda de patrones en el conjunto de datos, con el objetivo de producir estimaciones ponderadas por una función *Kernel*. Se desarrolla el modelo combinado compuesto por KNN y SVM denominado KNN-Kernel, que muestra resultados prometedores en los casos en los que la tasa de datos perdidos es alta, aunque presenta como inconveniente su alto coste computacional.

Los métodos de imputación basados en aprendizaje automático mejoran su rendimiento mediante la incorporación de las correlaciones espacio-temporales existentes en el tráfico de una red (Zhang and Liu, 2009). El método denominado Máquina de Vectores de Soporte de Mínimos Cuadrados, o LS-SVM, que considera dichas correlaciones, imputando de una manera eficiente los valores perdidos en el caso de una intersección urbana. Una de las mayores ventajas de este método, es que evita los mínimos locales ayudando a una convergencia rápida hacia el mínimo global, además de mostrar un comportamiento robusto con distintas tasas de valores perdidos.

Las aproximaciones basadas en redes neuronales también han presentado resultados favorables. En una primera aproximación (Chen et al., 2001) presentan dos tipos, un Perceptrón Multicapa (MLP) y una Red Neuronal de Base Radial (RBFNN) cuyas imputaciones muestran un menor impacto en AADT y DHV que

los métodos ARIMA con los que se comparan en un caso particular.

(Zhang and Zhang, 2016) introducen las relaciones espaciales del flujo de tráfico como mediante una Red Neuronal de Regresión General GRNN, que presenta mejores resultados que implementaciones de HA y ARIMA en condiciones de tasas medias y altas de datos perdidos. Para tasas bajas, ARIMA se muestra superior.

Las redes neuronales tienen la capacidad de explotar el contexto espacial del parámetro de salida, utilizando el concepto de sensores vecinos y permiten inferir en las relaciones entre los parámetros de tráfico. En (Laña et al., 2018a), se proponen dos métodos basados en redes neuronales que muestran un rendimiento superior a técnicas basadas en sustitución simple cuando la tasa de datos perdidos es relativamente elevada. Observándose que para tasas bajas (menores al 10%), es recomendable utilizar técnicas más sencillas, sobre todo cuando se debe completar el conjunto de datos en tiempo real. Se prueba que con suficiente información contextual es posible imputar valores perdidos, incluso cuando no hay datos disponibles para un largo periodo del parámetro objetivo.

Los métodos basados en el aprendizaje profundo han sido utilizados recientemente como modelos de imputación. (Duan et al., 2016) presentan un SAE capaz de representar las relaciones espacio-temporales del área de studio, mejorando el rendimiento mostrado por una BPNN y un modelo ARIMA. SAE también se ha mostrado útil para representar las relaciones espacio-temporales presentes en el tráfico, (Pamuła, 2018) muestra su efectividad en la imputación de valores de flujo en autovías de gran capacidad.

3.2.2.5. Imputación basada en tensores

El cálculo tensorial ha ganado un enorme protagonismo recientemente en la imputación de datos de tráfico. Su principal característica es la capacidad para detectar relaciones multidimensionales subyacentes en los conjuntos de datos, además de poder operar con conjuntos de datos incompletos.

Un tensor es un vector multidimensional, se entiende como la generalización de más alto orden dentro del cálculo vectorial. Los vectores son los elementos de primer orden en este tipo de cálculo, en un orden mayor se encuentran las matrices y los tensores constituyen el elemento de mayor rango con una aritmética propia, adaptada al cálculo multidimensional.

Este tipo de cálculo es una opción adaptable a la estimación de parámetros de

tráfico debido a la capacidad para tratar las múltiples dimensiones, ya que puede representar las relaciones heterogéneas que se producen entre los parámetros (Figura 3-12). Básicamente, se han producido dos tipos de acercamientos en este tipo de métodos:

i) Compleción de Tensores, utilizan modelos de pocas dimensiones que representan las fuertes correlaciones que existen entre el estado del tráfico de arcos vecinos (Tan et al., 2014). (Asif et al., 2016) utilizan estos patrones para estimar los valores perdidos mediante la obtención de una aproximación de bajo rango que se adapte de manera adecuada al tensor incompleto. (Ran et al., 2016) proponen una metodología se basa en la norma de la traza del tensor y la relajación convexa del rango de un tensor; el problema de compleción del tensor se transforma en un problema de optimización convexa.

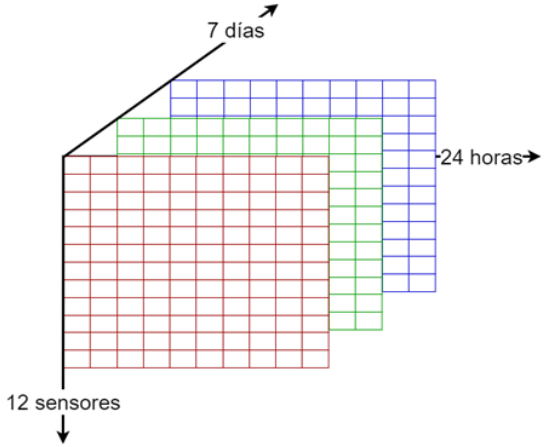


Figura 3-12: Dimensiones del tráfico representadas mediante un tensor. Se representan los valores de 12 sensores, durante las 24 horas de un día, separados en los 7 días de la semana.

ii) Descomposición de Tensores, desde menor rango se pueden aproximar los datos perdidos en un rango mayor (Ran et al., 2016). Existen diversas variaciones de esta metodología, como como el Método de Imputación por Descomposición de Tensores, o TDI (Tan et al., 2013), y la Compleción mediante Tensores de bajo rango, o LRTC

(Liu et al., 2013); ambos presentan buenos resultados al utilizarse como métodos de imputación (Asif et al., 2016; Ran et al., 2016), destacando LRTC.

En combinación con otras técnicas de descomposición, como PCA, *Variatonial Bayesian* PCA (VBPCA), el método de Mínimos Cuadrados (LS) y *Canonical Polyadic Weighted OPTimization* (CP-WOPT), se potencia la capacidad de encontrar estructuras y relaciones entre los datos (Asif et al., 2016). En esta publicación se identifican otras características que influyen en la precisión de la imputación, como es la categoría de la vía en la que está ubicado el sensor o el tipo de red del entorno del arco.

Los métodos basados en tensores se han revelado como una solución solvente en intersecciones (Qu et al., 2009), corredores (Ran et al., 2016; Tan et al., 2014) y redes a gran escala en las que se tiene un volumen de datos mucho más grande y con mayor diversidad diversidad (Asif et al., 2016). Su mayor ventaja es la poca sensibilidad mostrada ante el increment de la tasa de valores perdidos en el conjunto de datos.

3.3. Conclusiones

En el campo de la predicción a corto plazo, de un modo general, los esfuerzos en el desarrollo de modelos se han dirigido principalmente a:

i) mejorar la precisión de las predicciones (i.e., reducir el error de los valores modelados respecto al valor real).

ii) completar conjuntos de datos mediante imputación, asegurando la integridad de los datos que se completan.

En primer lugar se realiza una puesta en común sobre las principales ideas extraídas de la revisión de modelos de predicción que se realiza en la sección 3.1.3.

En la Tabla 3-4 se recogen las categorías y características de los métodos de predicción analizados, basándose en los consensos alcanzados tras la revisión de las conclusiones de numerosos estudios presentados en este capítulo.

Tabla 3-4: Cuadro resumen de tipologías de métodos de predicción de tráfico a corto plazo.

Categoría	Subcategoría	Características
Paramétrica	Modelos Simples	Bajo coste computacional
		Rendimientos aceptables en condiciones normales
		Tendencia a modelar con la media
	Series temporales	Bajo coste computacional
		Mejora rendimientos de modelos simples en condiciones normales
		Rendimientos poco aceptables en situaciones atípicas
		Mejora con la inclusión de la estacionalidad
		Rendimientos competitivos con pocos datos
No Paramétrica	Aprendizaje no supervisado	Bajo coste computacional
		Mejora rendimientos de modelos simples en condiciones normales
		Rendimientos poco aceptables en situaciones atípicas
		Mejora con la inclusión de la estacionalidad
		Rendimientos competitivos con pocos datos
	Aprendizaje supervisado	Alto coste computacional
		Requiere un amplio conjunto de datos para generalizar la solución
		Con suficientes datos mejora al resto de las alternativas
		Capacidad para adaptarse a distintos regímenes de tráfico

Analizando la categoría del modelado paramétrico se observa que:

i) Los modelos simples presentan el menor coste computacional y la mayor facilidad de implementación, aunque son muy deficientes en condiciones particulares del estado del tráfico.

ii) Los modelos basados en series temporales mejoran el rendimiento de

los simples en condiciones normales y con una alta tasa de valores perdidos, aunque siguen mostrando un rendimiento insuficiente en situaciones atípicas.

Por otra parte, poniendo el foco en las técnicas no paramétricas:

i) Los modelos basados en el aprendizaje no supervisado presentan ciertas mejoras respecto a los paramétricos en condiciones atípicas, aunque son usados más comúnmente como herramientas auxiliares para la identificación de patrones y simplificación de la información.

ii) los modelos basados en el aprendizaje supervisado presentan un mejor rendimiento en todas las condiciones del tráfico siempre que la muestra sea suficiente. Su coste computacional es elevado perjudicándolos en su utilización en herramientas que deben predecir valores en tiempo real.

Una vez revisadas las distintas categorías de métodos de imputación, sus principales características se sintetizan en la Tabla 3-5, resumiéndose sus fortalezas y debilidades

Se observa, como característica común, una mejora de todos los tipos de métodos mediante la introducción de información contextual en el proceso, ya sea para la selección y clasificación de información, o como parámetros de entrada de los modelos.

En cuanto a los modelos basados en la sustitución simple y series temporales, se observa, de manera general que:

i) Pueden ser aplicables en condiciones muy específicas con rendimientos competitivos, en situaciones en las que se dispone de un histórico de datos pequeño y una baja tasa de valores perdidos.

ii) Su sensibilidad al crecimiento de la tasa de valores perdidos los hace menos competitivos que otros métodos más complejos, en situaciones con una tasa superior al 10%.

Los métodos basados en aprendizaje automático y tensores presentan un coste computacional alto, haciéndolos poco adecuados para su integración en sistemas que necesitan conjuntos de datos completos en tiempo real, si la potencia computacional de la que se dispone no es alta.

Tabla 3-5: Cuadro resumen de las categorías de modelos de imputación.

Categoría	Características
Sustitución simple	Muy bajo coste computacional. Implementación sencilla. Introducen excesivo ruido, distorsión en la muestra. Aplicable en cualquier contexto de ausencia de datos. Muy sensibles a la tasa de valores perdidos.
Numéricos y Estadísticos	Bajo coste computacional. Representación fehaciente de las relaciones espacio-temporales entre parámetros. Buen compromiso entre coste computacional y precisión.
Series Temporales	Bajo coste computacional. No aplicable en una situación continuada de valores perdidos. Rendimiento muy sensible al aumento de la tasa de valores perdidos. Buen compromiso de coste/precisión con bajas tasas de valores perdidos y con poca información contextual.
Aprendizaje automático	Coste computacional relativamente alto. Capacidad de representación de relaciones no lineales entre los parámetros. Poco sensibles al aumento de la tasa de valores perdidos. Mejor precisión cuanto mayor sean el volumen de datos y la cantidad información contextual disponibles.
Tensores	Coste computacional relativamente alto. Capacidad para detectar relaciones multidimensionales subyacentes a los conjuntos de datos. Posibilidad de operar con conjuntos incompletos de datos. Representación de las relaciones multidimensionales entre los parámetros que participan en la imputación. Poco sensibles al aumento de tasa de valores perdidos.

La elección del método de imputación depende de diversos factores referentes al caso concreto al que se aplica, como:

i) Los conjuntos de datos disponibles.

ii) El tipo de sistema en el que se integra.

iii) La tasa de datos incompletos en el conjunto de datos.

iv) Los tipos de valores perdidos predominantes.

v) La escala y tipología de la red en la que se implementa.

Una gran parte de la bibliografía más reciente sobre modelización de parámetros de tráfico, tanto en la imputación, como en la predicción, está dedicada a la comparación entre diferentes metodologías En las secciones 3.1 y 3.2 se exponen numerosas publicaciones en las que se presentan técnicas novedosas que se comparan con otras aproximaciones ya conocidas, mediante su aplicación a casos de estudio específicos.

Se constatan las ideas expuestas en (Vlahogianni et al., 2014) sobre la necesidad de la definición de unas bases estrictas para la comparación entre aplicaciones de métodos, debido a las siguientes razones:

i) Los resultados empíricos se aplican a áreas específicas de implementación que usualmente incluyen ciertas restricciones topológicas y conceptuales.

ii) El esfuerzo y tiempo requerido para un solo artículo no permite la implementación de la gran variedad de algoritmos de predicción existentes, o al menos, una variedad representativa de los métodos principales, con el objetivo de comparar su rendimiento.

iii) No todos los investigadores usan las mismas bases de comparación, no usan los mismos índices estadísticos para la comparación de algoritmos.

Sin embargo, cada estudio comparativo aporta una serie de indicios que, observados en conjunto, apuntan a una dirección y una tendencia implícita, que puede ser usado como punto de partida para extraer las reglas de modelado generales.

Entre todos los métodos expuestos en 3.1 y 3.2 se advierten una serie de puntos comunes sobre sus características:

i) Los métodos paramétricos simples tienen un bajo coste computacional, son competitivos cuando se dispone un histórico limitado de datos y se requieren imputaciones en tiempo real.

ii) Los métodos más complejos, basados en aprendizaje automático o tensores, presentan mayor precisión en situaciones en las que se dispone de un número de casos suficiente.

iii) Las relaciones espacio-temporales entre los parámetros de tráfico son cruciales para mejorar el rendimiento de los modelos y se debe hacer un esfuerzo en detectarlas e incorporarlas en el proceso de modelado.

iv) Para evitar estructuras y costes inasumibles, se debe seleccionar cuidadosamente la información de entrada de los modelos, reduciendo el volumen de información y evitando una merma de la capacidad explicativa de ésta.

v) La combinación de modelos aporta robustez al resultado y permite al modelo adaptarse a las distintas condiciones del tráfico.

En las aproximaciones que integran las relaciones espacio-temporales en la predicción se deben definir dos parámetros clave:

i) La ventana temporal, intervalos anteriores al del parámetro de salida del modelo que se incluyen en el conjunto de entrada.

ii) La vecindad de un detector, extensión espacial dentro de la que se consideran los detectores como vecinos del detector que registra los valores del parámetro de salida.

El proceso de modelado de predicción del tráfico a corto plazo debe tener en cuenta todas las características del problema, debiéndose invertir el esfuerzo en la correcta definición y diseño de cada etapa que compone el proceso.

El proceso de predicción debe incluir las siguientes fases:

i) Adquisición y armonización de la información proveniente de las fuentes de datos.

ii) Análisis de las características del conjunto de datos que define al caso de estudio.

iii) Detección y tratamiento de los valores perdidos.

iv) Elección del método de predicción en base a las características del caso de estudio y de la información disponible.

v) Proceso de evaluación, mediante índices de rendimiento específicos, de todo el proceso y especialmente, del modelado de predicción.

En el capítulo 4 se propone un marco de predicción que da respuesta a los requerimientos comentados, integrando cada una de las fases descritas previamente.

4. MARCO PARA LA PREDICCIÓN DE FLUJO DE TRÁFICO CON DATOS INCOMPLETOS

En este capítulo se propone un marco metodológico para afrontar de manera integral el proceso de predicción del flujo de tráfico a corto plazo.

El marco propuesto se compone de las principales fases en las que se divide el proceso de predicción, identificadas en el capítulo 3, otorgando una especial atención a la gestión de la información.

Para el desarrollo del marco se tienen en cuenta las conclusiones extraidas en los dos capítulos anteriores, en concreto:

i) La evolución de los ITS y la gestión de los conjuntos de datos dentro de éstos (capítulo 2).

ii) Las herramientas de predicción del tráfico a corto plazo (sección 3.1).

iii) Las técnicas de imputación (sección 3.2).

En el diseño del marco se ha priorizado la flexibilidad, mediante fases totalmente modulares con funciones bien determinadas. Su diseño está orientado a la utilización de técnicas basadas en el Aprendizaje Automático.

El marco consiste en una descripción del procedimiento de predicción, en el que se exponen las funciones, los conjuntos de datos y acciones que participan en cada fase, así como el orden en que deben ejecutarse.

El marco se compone de cinco fases (Figura 4-1):

i) Adquisición de información (sección 4.1.1).

ii) Análisis de información (sección 4.1.2).

iii) Imputación de valores perdidos (sección 4.1.3).

iv) Predicción del flujo de tráfico a corto plazo (sección 4.1.4).

v) Evaluación del proceso (sección 4.1.5).

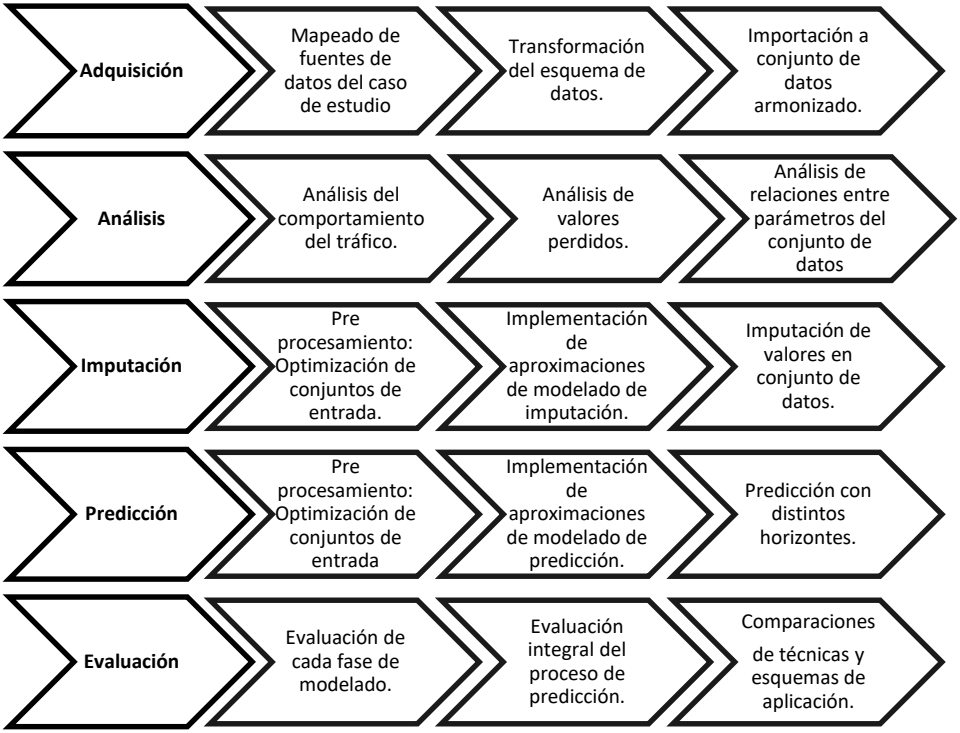


Figura 4-1: Esquema básico del marco de predicción, indicando funciones de acada fase.

A continuación, en la sección 4.1, se describen las fases del marco, sus funciones y los elementos que participan en cada una de ellas.

4.1. Descripción del marco de predicción

Desde un punto de vista general, el objetivo del marco es la implementación de modelos de predicción de flujo de tráfico robustos y precisos, basados en técnicas de Aprendizaje Automático, adaptados a la información del escenario para el que se desarrollan.

Adicionalmente, el marco presenta la capacidad de evaluar la eficiencia del proceso general y de cada una de las fases, apoyándose en la generación de medidas de rendimiento y en otros criterios cualitativos.

Se trata de una herramienta versátil, su esquema de datos permite la importación de información de distintas fuentes de datos de una manera ágil y la flexibilidad en la definición de las fases permite incluir en ellas diferentes técnicas que ejecuten la función indicada, sirviendo como banco de ensayo para la comparación de técnicas en diferentes casos de estudio.

4.1.1. Adquisición y armonización de información

La primera fase consiste en la importación de la información real de un escenario, que se integra en el esquema de datos del marco, con un formato adecuado para servir como información de entrada en el resto de fases del marco de predicción.

Este proceso se compone de una serie de etapas, como la consulta y el mapeado de datos de la fuente de información y las transformaciones necesarias para su importación al esquema de datos del marco de predicción. Este proceso se detalla en una publicación anterior (Ruiz-Alarcon-Quintero, 2016), en la que se presenta el modelo de datos utilizado por el marco de predicción.

Una vez se ha realizado la importación, se genera el conjunto de datos de flujos, que se proporciona al resto de fases (Figura 4-2). La información que se necesita conocer sobre un escenario es la siguiente:

a. Los valores de flujo registrados por los detectores de tráfico, incluyendo las características de la toma de datos (frecuencia de adquisición, frecuencia de registro y la unidad en la que se almacena la información).

b. Ubicación y características de los detectores.

c. Red viaria en la que se ubican los detectores, esta red debe ser conexa y disponer de información sobre el coste de desplazamiento entre puntos de la red.

d. Marco temporal de referencia del caso de estudio.

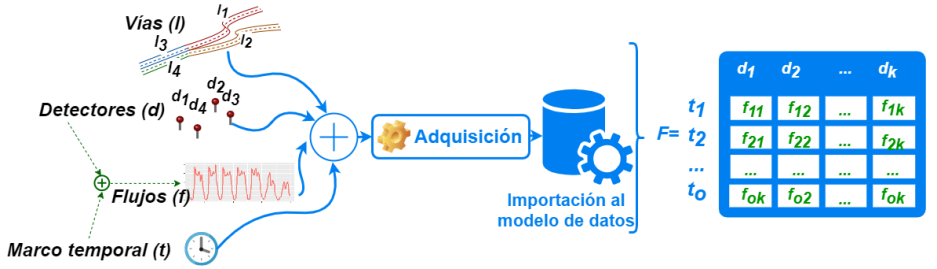


Figura 4-2: Esquema de la fase de adquisición de información..

El conjunto de datos ***F*** es resultado principal de este proceso. Está compuesto por los valores de flujo captados por los detectores en cada intervalo que compone el marco temporal.

4.1.2. Análisis de información del escenario

Tras la fase de adquisición, se realiza un análisis de la información que contiene el escenario, orientado a aportar datos de interés al resto de fases del marco (Figura 4-3).

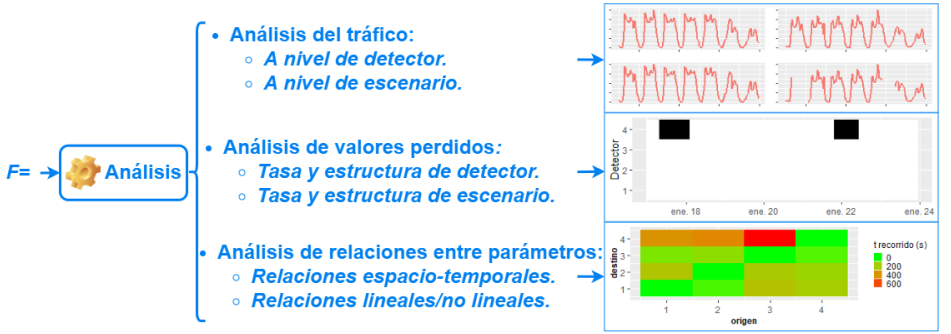


Figura 4-3: Esquema de la fase de análisis

Como entrada del proceso se toma el conjunto de flujos ***F***, además del resto de elementos de información contenidos en el esquema de datos, expuestos en la sección 4.1.1. A continuación, se enumeran los tipos de análisis que se realizan en esta fase:

4.1.2.1. Análisis del comportamiento del tráfico

Se realiza un análisis del comportamiento del tráfico en el escenario a dos niveles:

a. A nivel de detector: Se analiza el comportamiento del tráfico registrado en cada detector, con el objetivo de detectar pautas individuales y el volumen del tráfico registrado.

b. A nivel de escenario: Se analiza el comportamiento del tráfico de manera global, con el objetivo de detectar pautas generales, o que afectan a una porción del escenario.

4.1.2.2. Análisis de valores perdidos

Se examinan los valores perdidos presentes en el conjunto de datos ***F***, con el objetivo de establecer, cuantitativa y cualitativamente, una medida de su presencia y considerarla en el resto de fases del marco. El análisis se abarca desde dos perspectivas.

a. Análisis individual: Se extraen varios factores relativos a la estructura de la presencia de valores perdidos para un detector. Se calcula la tasa de valores perdidos respecto al total, la duración media, en intervalos, de cada periodo continuado de valores perdidos y el número de intervalos que transcurren entre un periodo de valores perdidos y el siguiente. Estas medidas permiten establecer un diagnóstico de cada detector respecto a este fenómeno.

b. Análisis general: Relaciona la aparición de valores perdidos a nivel de escenario, estableciéndose la tasa general del escenario respecto al total de valores que es posible registrar y las tasas de valores perdidos coincidentes entre pares de detectores. Estas medidas permiten identificar pautas respecto a su aparición.

4.1.2.3. Análisis de relaciones entre parámetros del escenario

Se analizan otros aspectos mediante los que se pueden establecer relaciones entre

parámetros que intervienen en el escenario. Básicamente, se plantean los análisis sobre dos aspectos:

a. Relaciones de coste entre las localizaciones de los detectores, como el tiempo de recorrido o distancia física entre éstos, a través de las vías del escenario.

b. Relaciones lineales/no lineales entre parámetros, que permitan identificar la intensidad de la relación o determinar patrones, no perceptibles de manera directa, entre ellos.

4.1.3. Imputación

La fase de imputación consiste en la sustitución de los valores perdidos en el conjunto de datos ***F*** por valores estimaciones generadas mediante un modelo basado en técnicas de Aprendizaje Automático.

En esta fase se elaboran modelos de imputación individuales para cada detector del escenario. En adelante, con el objetivo de ilustrar los ejemplos de las distintas fases, se denomina d_a a un detector genérico del escenario, sobre el que se ejecutan las diferentes acciones de las fases del marco de predicción.

Tras el proceso de imputación, al ser ejecutada sobre todos los detectores del escenario, se obtiene como resultado el conjunto $\boldsymbol{F^I}$, en el que se reduce la cantidad de valores perdidos, o se elimina totalmente, en el mejor de los casos.

La generación del modelo de imputación de un detector se compone de dos etapas, preprocesamiento e implementación del modelo, que se describen a continuación.

4.1.3.1. Preprocesamiento

Se realizan una serie de procesos sobre el conjunto ***F***, basados en la información obtenida en la fase de análisis. El objetivo del preprocesamiento es optimizar la información utilizada en los conjuntos de entrada de los modelos de imputación, se compone de los siguientes pasos:

a. Incorporación de la ventana temporal: Se añaden nuevaos parámetros, o columnas, al conjunto de datos ***F*** (Figura 4-4), que representan a los valores de flujo registrados por los detectores del escenario los intervalos anteriores a los correspondientes en cada registro. El

número de periodos anteriores añadidos se determina por la longitud de la ventana temporal *v*. Se toman los valores desde el intervalo *t-1*, que representa un periodo anterior al actual, hasta *t-v*, que representa el valor registrado *v* intervalos anteriores al actual. Al conjunto de datos resultante se le ***F'***.

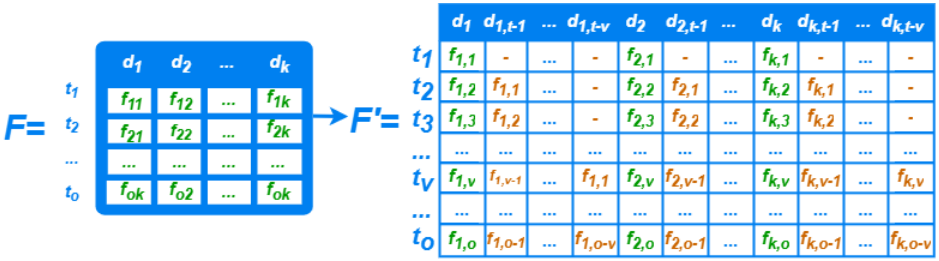


Figura 4-4: Incorporación de la ventana temporal al conjunto de datos.

b. Clasificación: consiste en la división de los casos (filas) del conjunto ***F'*** en subconjuntos con características similares entre sí. Se realiza en función de las particularidades del flujo de d_a. Como resultado de la clasificación, el conjunto de datos ***F'*** se divide en *n* subconjuntos (Figura 4-5), que se denominan $F'_a{}^1$, $F'_a{}^2$, …, $F'_a{}^n$.

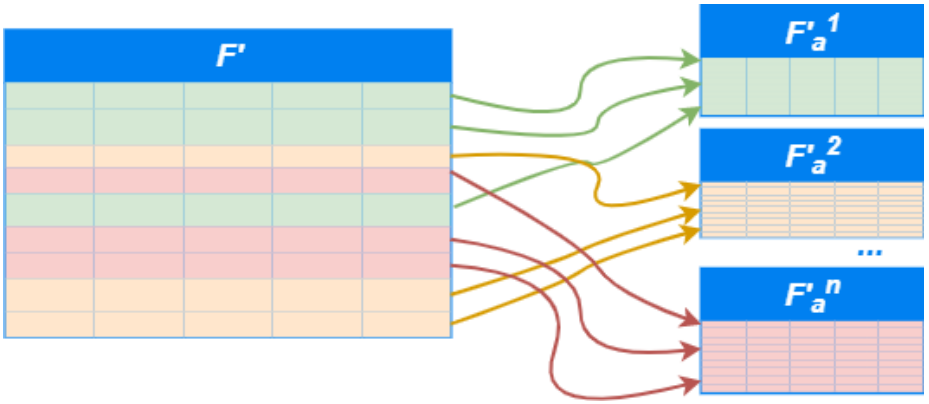


Figura 4-5: Clasificación de información en el modelado de imputación.

c. Selección de parámetros: este paso se realiza de forma paralela para cada subconjunto obtenido en el paso anterior $F'_a{}^1$, $F'_a{}^2$, …, $F'_a{}^n$.

Consiste en la selección de aquellos parámetros (columnas) que muestran un mayor grado de relación con el parámetro d_a. El grado de relación se basa en las características observadas y estudiadas durante la fase de análisis del conjunto de datos (4.1.2). La selección de parámetros parte de los subconjuntos $F'_a{}^1$, $F'_a{}^2$, …, $F'_a{}^n$ y obtiene como resultado los subconjuntos $F''_a{}^1$, $F''_a{}^2$, …, $F''_a{}^n$ con un número de parámetros más reducido (Figura 4-6).

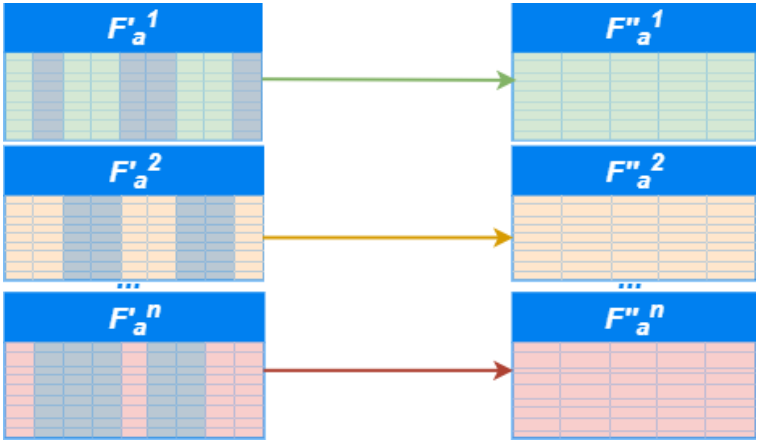


Figura 4-6: Selección de parámetros en el modelado de imputación.

4.1.3.2. Implementación del modelo

Este proceso se realiza de manera paralela para cada uno los subconjuntos derivados del preprocesamiento, $F''_a{}^1$, $F''_a{}^2$, …, $F''_a{}^n$. Para ilustrar los siguientes se utiliza un subconjunto genérico que se representa como $F''_a{}^*$, sobre el que se realizan las acciones descritas, siendo extensibles al resto de subconjuntos. La implementación del modelo consiste en los siguientes pasos:

a. Selección del método: Se debe elegir un método basado en técnicas de Aprendizaje Automático que sea idóneo para el conjunto de datos que recibe como entrada.

b. Ajuste de hiperparámetros: Se debe realizar un ajuste de hiperparámetros del método seleccionado, con el objetivo de optimizar su configuración en base al conjunto de datos del escenario (Figura 4-7).

Figura 4-7: Ajuste de hiperparámetros de la técnica de modelado de imputación.

c. Entrenamiento del modelo: Se toma el método ajustado en el paso anterior y se sigue el esquema típico de entrenamiento de modelo basado en el aprendizaje automático supervisado (Figura 4-8). Se separa el conjunto de entrada F''_{a^*} en dos subconjuntos: el subconjunto de entrenamiento F''_{a^*e}, que suele incluir un 80% de los casos y el subconjunto de test F''_{a^*t}, formado por un 20% de los casos y se proporciona. Al subconjunto de entrenamiento se le aplica un proceso de validación cruzada en función al conunto F''_{a^*e}, obteniéndose un modelo ajustado a este conjunto de datos. Posteriormente se obtiene una medida de rendimiento sobre casos no conocidos mediante la aplicación del modelo al conjunto de test.

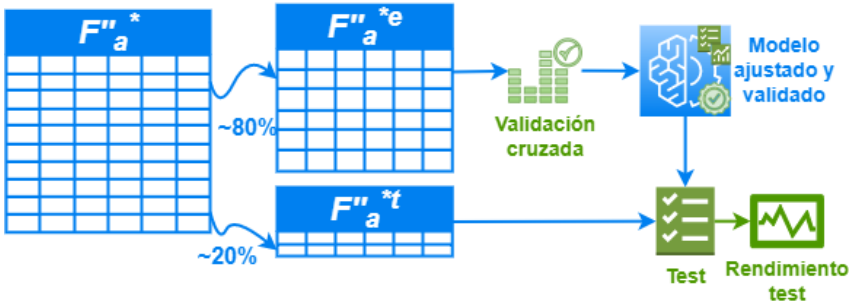


Figura 4-8: Implementación del modelo de predicción.

La validación cruzada consiste en un proceso iterativo de entrenamiento y validación, en base al conjunto de entrenamiento. Siendo *k* el número de iteraciones o *folds*, en cada una de ellas se separa el conjunto de entrenamiento en *k* subconjuntos, de los que *k-1* se utilizan como conjuntos de entrenamiento y un subconjunto como conjunto de validación, en cada iteración (Figura 4-9). Se obtiene como resultado un modelo entrenado y

medidas de rendimiento asociadas a los procesos de entrenamiento y validación, que ofrecen información sobre la eficiencia y ajuste que el modelo ha mostrado durante la validación cruzada

Figura 4-9: Ejemplo de división del conjunto de entrenamiento para un proceso de validación cruzada de 4 iteraciones.

d. Imputación de valores perdidos de d_a: Como último paso de esta fase, se debe realizar la imputación de los valores perdidos de d_a en el conjunto $\boldsymbol{F}$. En el paso anterior se ha generado un modelo ajustado a cada uno de los subconjunto $\boldsymbol{F''_{a^1}}$, $\boldsymbol{F''_{a^2}}$, …, $\boldsymbol{F''_{a^n}}$. Estos modelos se utilizan para simular los valores perdidos de cada subconjunto, que son añadidos al conjunto de datos $\boldsymbol{F}$. El proceso es extensible a cada uno de los detectores del escenario d_1, d_2, ..., d_k, generándose los modelos y valores simulados correspondientes. Al final del proceso se obtiene el conjunto de datos $\boldsymbol{F^I}$, que representa al conjunto de datos imputado (Figura 4-10).

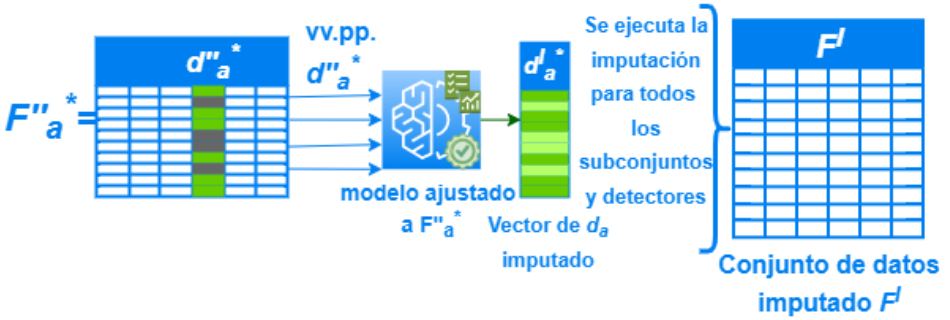


Figura 4-10: Imputación de valores en el conjunto de datos del escenario.

4.1.4. Predicción

La fase de predicción consiste en la elaboración de un modelo que permite pronosticar los valores de flujo de un detector del escenario, *da*, para un horizonte de predicción *h*, partiendo del conjunto de datos derivado de la fase de imputación $\boldsymbol{F^I}$.

La elaboración del modelo de predicción sigue el mismo proceso que el de imputación, componiéndose de una etapa de preprocesamiento y otra de implementación, que se describen a continuación.

4.1.4.1. Preprocesamiento

Se realizan una serie de procesos sobre el conjunto imputado de datos de flujos $\boldsymbol{F^I}$, basados en la información obtenida en la fase de análisis. El objetivo del preprocesamiento es optimizar la información utilizada como entrada en los modelos de predicción de d_a, teniendo en cuenta el horizonte de predicción *h*. Consta de los siguientes pasos:

a. Incorporación de la ventana temporal: Se añaden nuevas columnas al conjunto de datos $\boldsymbol{F^I}$, correspondientes a los valores de los detectores registrados en intervalos anteriores. Se diferencia de la etapa homóloga de imputación en que el primer periodo que se añade es *t-h*, con el fin de simular el efecto del horizonte de predicción. Por tanto, se añaden los valores registrados entre *t-h* y *t-v* al conjunto de datos $\boldsymbol{F^I}$. De esta etapa se obtiene como resultado el conjunto de datos $\boldsymbol{F^{I\prime}}$, que se proporciona a la etapa de clasificación (Figura 4-11).

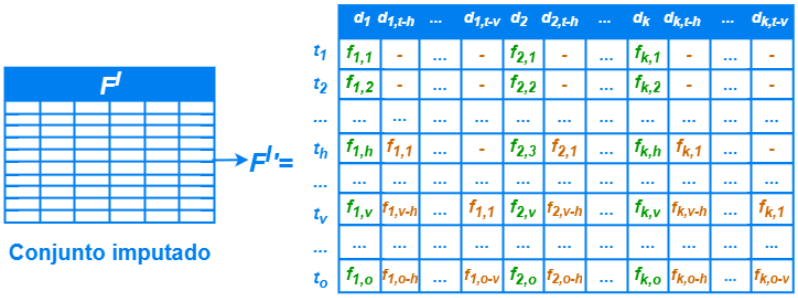


Figura 4-11: Incorporación de la ventana temporal a $\boldsymbol{F^I}$ en la fase de predicción.

b. Clasificación: consiste en la división de los casos (filas) del conjunto F' en subconjuntos con características similares entre sí. Se realiza en función de las particularidades del flujo de d_a. Como resultado de la clasificación, el conjunto de datos F' se divide en n subconjuntos (Figura 4-12), que se denominan F'_{a^1}, F'_{a^2}, …, F'_{a^n}.

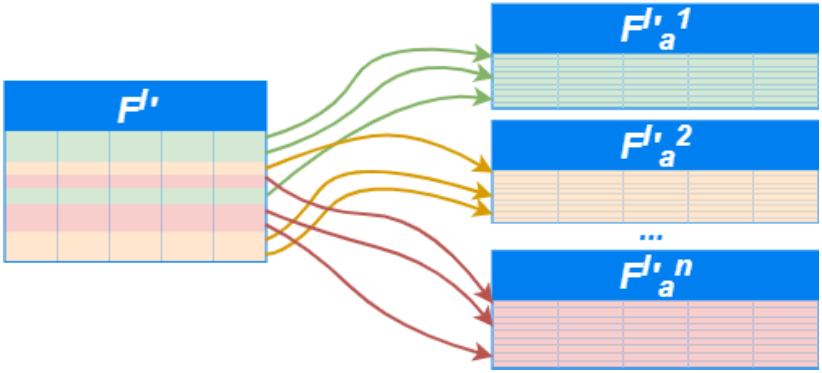


Figura 4-12: Clasificación de información en el modelado de predicción.

c. Selección de parámetros: este paso se realiza de forma paralela para cada subconjunto obtenido en el paso anterior F'_{a^1}, F'_{a^2}, …, F'_{a^n}. Consiste en la selección de aquellos parámetros (columnas) que muestran un mayor grado de relación con el parámetro d_a.

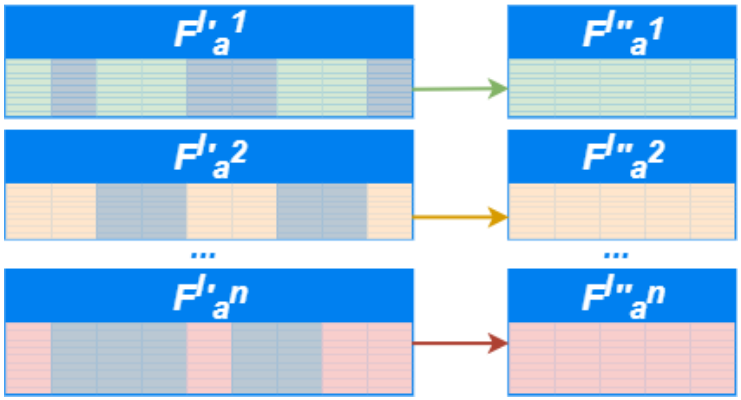


Figura 4-13: Selección de parámetros en el modelado de predicción.

El grado de relación se basa en las características observadas y estudiadas durante la fase de análisis del conjunto de datos (4.1.2). La selección de parámetros parte de los subconjuntos F'_{a^1}, F'_{a^2}, …, F'_{a^n} y obtiene como resultado los subconjuntos F''_{a^1}, F''_{a^2}, …, F''_{a^n} con un número de parámetros más reducido (Figura 4-13).

4.1.4.2. Implementación del modelo

En esta etapa se desarrolla un modelo de predicción para un detector genérico d_a y un horizonte de predicción h. Parte de los subconjuntos F''_{a^1}, F''_{a^2}, …, F''_{a^n} obtenidos en el preprocesamiento de d_a. Las etapas descritas se ejecutan en paralelo para cada subconjunto, generándose un modelo ajustado a cada uno de ellos al final del proceso. Se divide en las siguientes etapas:

a. Selección del método: Se debe elegir un método de predicción basado en técnicas de Aprendizaje Automático, que sea idóneo para las características observadas en el conjunto de datos que toma como entrada.

b. Ajuste de los hiperparámetros: Se debe realizar un ajuste de hiperparámetros del método elegido, con el objetivo de escoger una configuración que se ajuste de forma precisa al conjunto de datos del escenario (Figura 4-14).

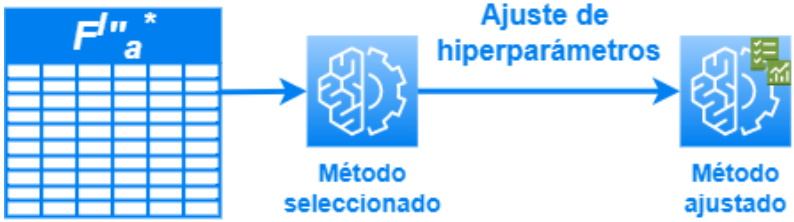


Figura 4-14: Ajuste de hiperparámetros de la técnica de modelado de predicción.

c. Entrenamiento del modelo: Se toma el modelo ajustado en el paso anterior y se sigue el esquema típico de entrenamiento de modelo basado en el aprendizaje automático supervisado. Se separa el conjunto de entrada F''_{a^*} en dos subconjuntos: uno de entrenamiento $F''_{a^{*e}}$, que suele incluir un 80% de los casos y otro de test $F''_{a^{*t}}$ que suele estar formado por un 20% de los casos (Figura 4-15). Sobre $F''_{a^{*e}}$ se realiza un proceso de validación cruzada, con el objetivo de entrenar el modelo al mismo tiempo que se

evita el sobreajuste. Una vez se ha entrenado y validado el modelo, éste toma como entrada el conjunto de test, que contiene casos no conocidos para el modelo. De los procesos de validación cruzada y de test se obtienen varias medidas de rendimiento relativas al proceso de entrnamiento, validación y test, que aportan información sobre la precisión del modelo y el grado de ajuste al conjunto de datos.

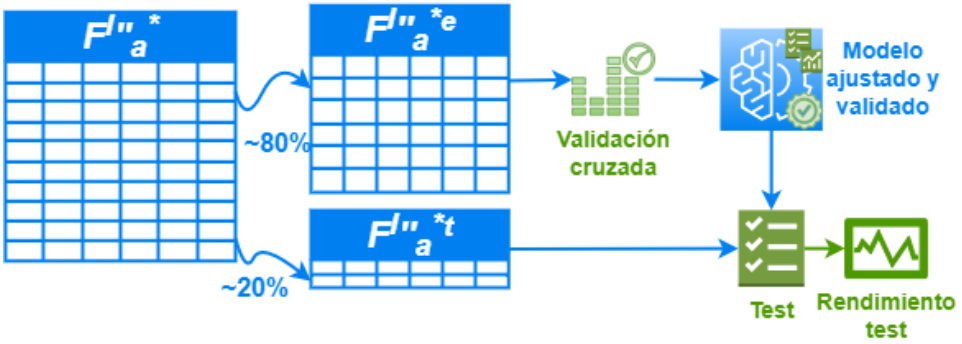


Figura 4-15: Implementación del modelo de predicción.

d. Predicción de valores de da con horizonte *h*: El resultado final del proceso de predicción es un modelo que permite simular valores del detector d_a con un horizonte de predicción *h* (h intervalos hacia adelante), para cada subconjunto F''_{a^1}, F''_{a^2}, …, F''_{a^n} obtenido en el preprocesamiento (Figura 4-16).

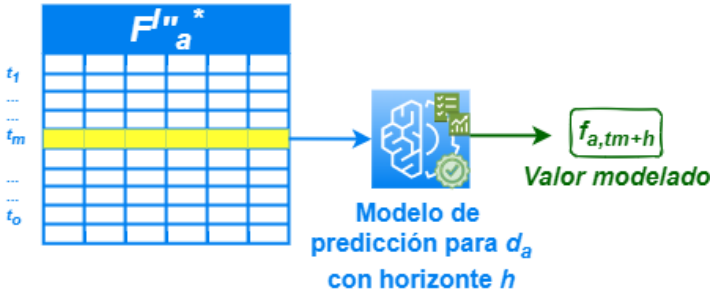


Figura 4-16: Predicción realizada mediante el modelo implementado.

4.1.5. Evaluación

La evaluación se realiza con dos tipos de perspectiva:

i) Una evaluación individualizada de cada fase del (desrita en la sección 4.1.5.1), en la que se analizan los índices generados en dicha fase y se realiza una valoración cualitativa de sus características.

ii) Una evaluación general del proceso de aplicación del marco de predicción (sección 4.1.5.2), en la que se contempla cada etapa en función de su influencia sobre la eficiencia de los modelos de predicción generados al final del proceso.

4.1.5.1. Evaluación específica de cada fase

Para cada fase se analizan una serie de aspectos, cualitativos y cuantitativos, que sirven para valorar la ejecución de cada fase sobre el escenario (Tabla 7)

Tabla 4-1: Características de la evaluación individualizada de cada fase.

Etapa	Características evaluadas
Adquisicón	Grado de accesibilidad de la fuente del dato.
	Grado de compatibilTransformaciones necesarias para importar la información al esquema de datos (compatibilidad entre formatos y esquemas de datos).
	Adecuación de la información al proceso de predicción.
	Nivel de detalle de los metadatos.
Análisis	Número de casos en el conjunto de datos.
	Tasa de valores perdidos y distribución, individuales en cada detector y general del escenario.
	Grado de relación entre parámetros.
Imputación	Parámetros relativos a los conjuntos de entrada obtenidos tras el preprocesamiento.
	Tasa valores imputados vs valores perdidos en el conjunto original.
	Índices de rendimiento de modelos de imputación (MAE, MSE, RMSE, R2) en las distintas etapas de implementación del modelo. Comparativa entre valores obtenidos para los distintos detectores.
	Comparación entre el rendimiento de distintas técnicas utilizadas en el método de imputación en un mismo escenario.
Predicción	Parámetros relativos a los conjuntos de entrada obtenidos tras el preprocesamiento.
	Índices de rendimiento de modelos de predicción (MAE, MSE, RMSE, R2) en las distintas etapas de implementación del modelo.
	Comparación entre el rendimiento de distintas técnicas utilizadas en el preprocesamiento para un mismo método de predicción.
	Comparación entre el rendimiento de distintas técnicas utilizadas en el método de predicción en un mismo escenario.

4.1.5.2. Evaluación general del marco

Se realiza un análisis general del marco, considerando la influencia de cada fase en el rendimiento obtenido por los modelos de predicción generados. Se tienen en cuenta distintas alternativas de ejecución del marco, alternando técnicas en las distintas fases y observando su efecto sobre el resto (Tabla 8).

Tabla 4-2: Evaluación global de la del marco de predicción.

Etapa	Características evaluadas
Marco completo	Observación de influencia de fases entre sí, ejecutando alternativas del marco de modelado en la que se elude la ejecución de fases determinadas y su efecto sobre el resto de fases.
	Comparación de rendimiento en fases mediante ejecuciones del marco con distintas alternativas de modelado en sus fases.
	Comparación de rendimiento en fases mediante ejecuciones idénticas del marco sobre distintos escenarios.

4.2. Ejemplo de aplicación del marco de predicción

Se ha diseñado un ejemplo sencillo de aplicación del marco sobre un pequeño caso de estudio real compuesto por 4 detectores en una sección de autovía (Figura 4-17), con el objetivo de ilustrar su aplicación práctica. El ejemplo mostrado permite exponer, paso a paso, el proceso de ejecución de cada fase, presentando los conjuntos de datos que intervienen y especificando técnicas concretas en cada una de ellas.

Figura 4-17: Escenario de ejemplo para ilustrar la aplicación práctica del marco de predicción.

4.2.1. Adquisición y armonización de información

El conjunto de datos utilizado en este ha sido proporcionado por el Centro de Gestión de Tráfico del Suroeste y se refiere a detectores gestionados por la Dirección General de Tráfico. Se ha suministrado información sobre intensidades de tráfico, registradas por detectores de tipo espira que se localizan en la SE-30, la circunvalación del área metropolitana de Sevilla.

En la Tabla Tabla 4-3 se muestra un cuadro resumen de las características del

escenario. Mediante la fase de adquisición es incorporada al esquema de datos del marco de predicción.

Tabla 4-3: Descripción del caso de estudio de ejemplo.

Característica	Valor	Descripción
Localización	SE-30, entre los Pks 6.7 y 8.05	Tramo de la circunvalación de Sevilla entre los pk 6.7 y 8.05
Detectores	SE-30 pk 6.7 C. SE-30 pk 6.7 D. SE-30 pk 8.05 C. SE-30 pk 8.05 D.	4 detectores situados en 2 localizaciones (pk 6.7 y pk 8.05), registrando información de flujo en cada sentido de circulación, cada sentido se compone de 3 carriles con una velocidad máxima de circulación de 80 km/h.
Agregación temporal	1 hora	Cada detector registra un valor de flujo de tráfico cada 15 minutos. Se han agregado con frecuencia horaria para simplificar la información tratada en el ejemplo.
Marco temporal	1 semana	Entre el 17/01/2022 y el 23/01/2022.

La información se aporta en ficheros de texto con extensión *.csv*, proporcionándose un fichero individual con los registros de cada detector. Cada uno de ellos contiene información sobre intensidades en una tabla con la estructura mostrada en la Tabla 4-4. En ésta se observa un registro cada 15 minutos, durante el marco temporal establecido.

Tabla 4-4: Ejemplo de formato en el que la DGT aporta información sobre aforos.

FECHA	INTENSIDAD	LIGEROS	PESADOS	VMED	TOCUP
…	…	…	…	…	…
23/01/2022 19:45	36	36	0	54.33	5.06
23/01/2022 20:00	43	40	1	54.28	6.06
23/01/2022 20:15	116	114	0	54.01	6.80
23/01/2022 20:45	80	79	1	53.94	6.00
…	…	…	…	…	…

Cada registro contiene información sobre la fecha de adquisición del dato, la intensidad total de vehículos, la intensidad desagregada en vehículos ligeros y

pesados, la velocidad media y el tiempo de ocupación del detector.

En cuanto a la información que requiere el modelo de datos, no se han aportado las coordenadas de cada detector, un hecho que dificulta localización, aunque se ha solventado utilizando la información sobre la vía, el pk y el sentido.

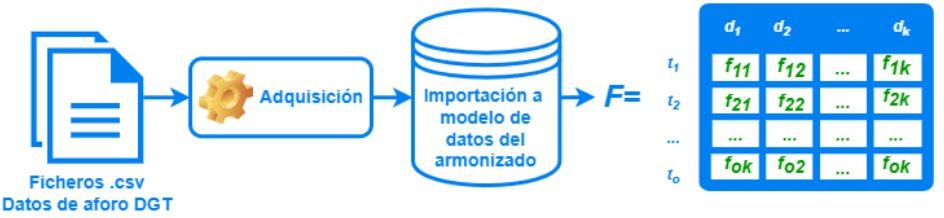


Figura 4-18: Esquema del proceso de adquisición e importación de información en el escenario de ejemplo.

En el proceso de adquisicón (Figura 4-18) se importa la información relativa a la "INTENSIDAD" y almacenándos con valores agregados de una hora, tal y como se refleja en la Tabla 4-5, que representa al conjunto de datos ***F*** para este escenario.

Tabla 4-5: Conjunto de datos ***F*** del escenario de ejemplo.

Fecha	SE-30 pk 6.7 C	SE-30 pk 6.7 D	SE-30 pk 8.05 C	SE-30 pk 8.05 D
17/01/2022 0:00	400	460	408	428
17/01/2022 1:00	188	232	188	224
17/01/2022 2:00	136	192	132	180
17/01/2022 3:00	160	192	160	188
…	…	…	…	…
23/01/2022 20:00	2204	2356	2216	2144
23/01/2022 21:00	1588	1432	1608	1320
23/01/2022 22:00	1004	988	1008	912
23/01/2022 23:00	636	568	572	480

4.2.2. Análisis de información del escenario

En primer lugar, se realiza una normalización de cada columna de ***F***, mediante el método de normalización Min-Max, con el objetivo de reducir la complejidad de los

cálculos en todas las fases del marco. De esta forma, todos los valores de ***F*** se encuentran entre 0 y 1.

A continuación, se muestra la aplicación de la fase de análisis sobre este escenario.

4.2.2.1. Análisis del comportamiento del tráfico

Para observar el comportamiento del flujo, se representan los valores registrados por cada detector durante el marco temporal (Figura 4-19). Se perciben perfiles muy parecidos entre los detectores que se encuentran en el mismo sentido de circulación. La intensidad registrada por todos los detectores muestra un comportamiento muy distinto entre los días laborables y no laborables.

El detector *SE-30 pk 8.05 D*, presenta una peculiaridad respecto al resto, al observarse dos periodos de valores perdidos que se prolongan durante varias horas.

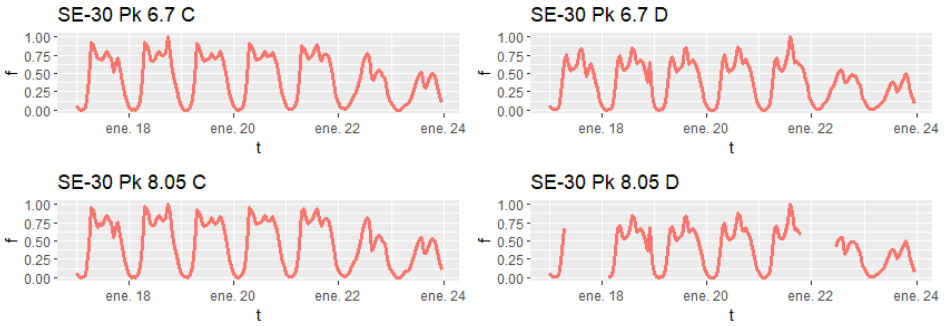


Figura 4-19: Representación de la intensidad registrada por los detectores del escenario.

4.2.2.2. Análisis de valores perdidos

Se realiza un análisis de la tasa y estructura de los valores perdidos que se producen en cada detector. En la Figura 4-20 se muestran en color blanco los periodos compuestos por intervalos con valores válidos y en negro aquellos con intervalos con valores perdidos. En el eje *y* e presentan los detectores, mientras que en el eje *x* se presentan las fechas correspondientes al marco temporal.

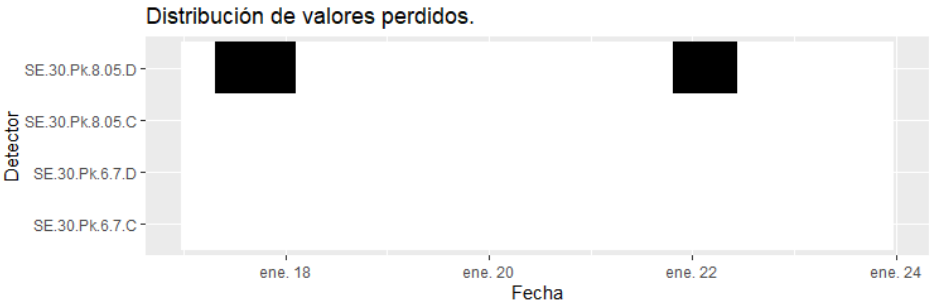


Figura 4-20: Representación de periodos de valores perdidos (en negro) en cada detector durante el marco temporal.

Se observa que el detector *SE-30 pk 8.05 D* es el único que muestra dos periodos continuados con valores perdidos, mientras que el resto de detectores no presenta ninguno.

La estructura de los periodos con valores perdidos en cada detector se muestra en la Tabla 4-6. Proporciona información sobre el total de registros con valores perdidos de cada detector (*vv.pp.*), la tasa de valores perdidos respecto al total *(% vv.pp.*), el número de periodos con valores perdidos (*Periodos vv.pp.*), la longitud media en intervalos de cada periodo (*Longitud periodos*) y la distancia media entre periodos con valores perdidos (*Distancia periodos*).

Tabla 4-6: Parámetros relativos a la estructura de valores perdidos registrados en cada detector.

Det.	vv.pp.	% vv.pp.	Periodos vv.pp.	Longitud periodos	Distancia periodos
SE-30 pk 6.7 C	0	0%	0	0	-
SE-30 pk 6.7 D	0	0%	0	0	-
SE-30 pk 8.05 C	0	0%	0	0	-
SE-30 pk 8.05 D	34	20.24%	2	17	108

4.2.2.3. Análisis de relaciones entre parámetros

En esta sección del análisis se identifican dinámicas de relación entre los parámetros que componen el escenario. Se ha decidido establecer una medida

sencilla de relación para ilustrar el ejemplo, mediante el cálculo coeficiente de correlación de Pearson (Tabla 4-7) entre parámetros. Se observan valores muy altos en todos los casos, debido a las características del escenario, debido a la localización cercana de los detectores en la misma vía. Se aprecia que es más alta entre detectores que se encuentran en el mismo sentido.

Tabla 4-7: Matriz de correlación de los parámetros del escenario.

Correlación	SE-30 pk 6.7 C	SE-30 pk 6.7 D	SE-30 pk 8.05 C	SE-30 pk 8.05 D
SE-30 pk 6.7 C	1.000	0.964	0.999	0.962
SE-30 pk 6.7 D	0.964	1.000	0.964	0.999
SE-30 pk 8.05 C	0.999	0.964	1.000	0.962
SE-30 pk 8.05 D	0.962	0.999	0.962	1.000

Para reflejar la relación espacio-temporal entre los detectores, se ha determinado utilizad el tiempo de recorrido entre las ubicaciones a través de la red de carreteras de la zona. La medida de coste utilizada es el tiempo de recorrido entre pares de detectores en régimen de flujo libre. Esta información se representa en la Figura 4-21, en la que se observan relaciones de tipo variado, que dependen de la estructura de la red, de la localización del detector y del sentido de circulación de la vía.

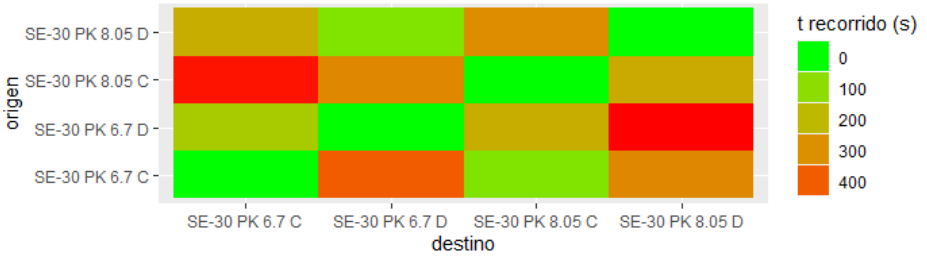


Figura 4-21: Matriz de costes, tiempos de recorrido a través de la red de carreteras entre cada par de detectores.

La información generada se utiliza en las etapas de preprocesamiento de las fases de imputación (sección 4.2.3.1) y predicción (sección 4.2.4.1).

4.2.3. Imputación

El proceso de imputación parte del conjunto de datos ***F*** mostrado en la Tabla 4-5. El objetivo principal de esta fase es la generación de modelos con la capacidad de simular con precisión los valores perdidos en ***F*** y completar este conjunto de datos de cara a proporcionar información más completa a la fase de predicción.

Esta fase se fundamenta en dos etapas, el preprocesamiento y la implementación del modelo (descritas en la sección 4.2.3). A continuación, se presenta una concreción de ambas, definiéndose las técnicas específicas para llevar a cabo todas las funciones que cumplen dentro del marco.

La aplicación de esta fase se ilustra mediante la generación del modelo de imputación para el detector *SE-30 pk 6.7 C*, que en las figuras se representa como d_a, siendo extensible la fase de imputación mostrada a todos los detectores del scenario.

4.2.3.1. Preprocesamiento

Esta etapa se compone de tres pasos: incorporación de la ventana temporal, clasificación de la información y selección de parámetros.

a. **Incorporación de la ventana temporal**: se ha decidido definir una ventana temporal de 2 horas. Incorporándose los registros correspondientes a dos intervalos anteriores para cada detector. En la Tabla 4-8 se representa el conjunto de datos ***F'*** que se obtiene como resultado de la incorporación de la ventana temporal. Los nuevos parámetros se representan con el nombre del detector, al que se le ha añadido y el texto *"(lag x)"*, siendo x el número de horas anteriores correspondientes al valor que representa el registro.

   El conjunto de datos ***F'*** queda configurado con doce parámetros, tres por detector, refiriéndose al valor registrado por el detector para t, para t-1 y para t-2 respectivamente.

Tabla 4-8: Incorporación de la ventana temporal al conjunto de datos de imputación. Conjunto ***F'***.

Fecha	SE-30 pk 6.7 C	SE-30 pk 6.7 C (LAG 1)	SE-30 pk 6.7 C (LAG 2)	SE-30 pk 6.7 D	SE-30 pk 6.7 D (LAG 1)	SE-30 pk 6.7 D (LAG 2)	SE-30 pk 8.05 C	SE-30 pk 8.05 C (LAG 1)	SE-30 pk 8.05 C (LAG 2)	SE-30 pk 8.05 D	SE-30 pk 8.05 D (LAG 1)	SE-30 pk 8.05 D (LAG 2)
17/01/22 0:00	400	-	-	460	-	-	408	-	-	428	-	-
17/01/22 1:00	188	400	-	232	460	-	188	408	-	224	428	-
17/01/22 2:00	136	188	400	192	232	460	132	188	408	180	224	428
17/01/22 3:00	160	136	188	192	192	232	160	132	188	188	180	224
...	-	-	-	-	-	-	-	-	-	-	-	-
23/01/22 20:00	2204	2424	2264	2356	2764	2392	2216	2428	2272	2144	2600	2280
23/01/22 21:00	1588	2204	2424	1432	2356	2764	1608	2216	2428	1320	2144	2600
23/01/22 22:00	1004	1588	2204	988	1432	2356	1008	1608	2216	912	1320	2144
23/01/22 23:00	636	1004	1588	568	988	1432	572	1008	1608	480	912	1320

b. **Clasificación:** Se realiza un proceso de clasificación para agrupar los casos con características comunes en función del parámetro *SE-30 pk 6.7 C*.

Se ha decidido clasificar el conjunto en dos grupos, para ilustrar este paso de la manera más simple posible y para que contengan un número de casos suficientes para el entrenamiento de los modelos, que recibirán esta información como entrada en etapas posteriores.

Se ha decidido utilizar un criterio de clasificación basado en el perfil diario del detector. La clasificación de perfiles se realiza mediante el algoritmo

fuzzy C-Means, imponiendo un número máximo de dos grupos.

Los perfiles diarios se agrupan en torno a dos perfiles característicos, representados a la izquierda de la Figura 4-22, mientras que a la derecha se muestra la clasidicación en torno a estos dos perfiles de los valores de *SE-30 pk 6.7 C*. El grupo 1 se identifica con los días laborables, mientras que el grupo 2 con los no laborables.

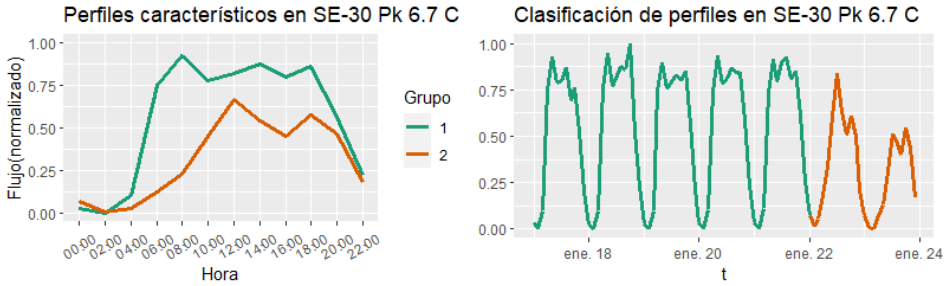


Figura 4-22: Clasificación de la información según perfiles característicos para el detector localizado en *SE-30 pk 6.7 C*.

Tras la clasificación, de los 168 registros que contiene F' originalmente, 120 se clasifican en el grupo 1, es decir en F'_{a^1} y 48 registros en el grupo 2, o lo que es lo mismo, en F'_{a^2} (Figura 4-23).

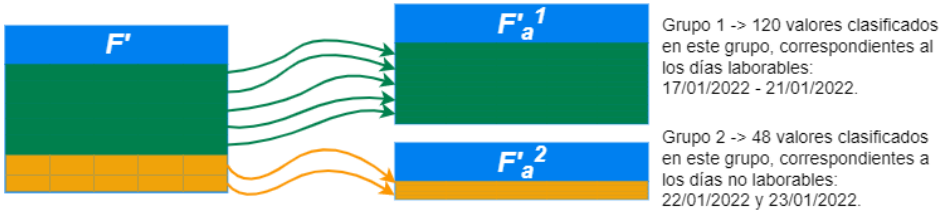


Figura 4-23: Clasificación del conjunto F' en subconjuntos basados en los perfiles característicos del detector *SE-30 pk 6.7 C*.

Esta etapa se aplica de forma idéntica en imputación y predicción, por lo que no se volverá a detallar en la fase de preprocesamiento de predicción.

c. **Selección de parámetros:** Este paso se aplica en paralelo a F'_{a^1} y F'_{a^2}. Consiste en realiza una selección basada en tres criterios, que se aplican de manera consecutiva el conjunto de datos: selección basada en coste,

selección basada en tasa coincidente de valores perdidos y selección basada en correlación.

Se deben definir varias restricciones en este paso, con el objetivo de establecer un compromiso entre la cantidad y la calidad de información resultante del preprocesamiento. Las restricciones impuestas se muestran en la Tabla 4-9 junto a una breve justificación de su elección.

Tabla 4-9: Restricciones impuestas en el paso de selección de parámetros

Restricción	Descripción	Valor
Parámetros en el conjunto de entrada	Se impone un número máximo de 4, para ilustrar de forma clara la etapa de selección de parámetros, reduciendo considerablemente el conjunto inicial..	4
Umbral de coste	Se selecciona los parámetros relativos a detectores cuyo tiempo de recorrido, como origen o como destino, es menor a un umbral dado..	Promedio de los tiempos de recorridos relacionados con d_a.
Tasa máxima de valores perdidos coincidentes	Tasa de valores perdidos coincidentes entre los parámetros del escenario y el total de casos de valores perdidos del parámetro de salida del modelo.	30%

A continuación, se muestra la aplicación de los criterios de selección descritos sobre el escenario de ejemplo.

Selección basada en coste: Este paso es común para ambos grupos, ya que el tiempo de recorrido no varía en función de cada uno de ellos. Para que un parámetro sea seleccionado, el coste de recorrido hasta/hacia el detector *SE-30 pk 6.7 C* debe ser menor a un umbral definido. El umbral se define como el tiempo de recorrido promedio de todos los trayectos relacionados en el detector relativo al parámetro de salida, que se establece en 283.1 s, en la Figura 4-24 se muestra mediante una línea azul discontinua.

En este paso, todos los detectores cumplen el criterio de relación de coste, sea como origen, o como destino de los trayectos de *SE-30 pk 6.7 C* (Figura 4-24) y, por tanto, no se discrimina ninguno de los parámetros relacionados con ellos ni en $\boldsymbol{F'}_a^1$ ni en $\boldsymbol{F'}_a^2$.

Esta etapa se aplica de forma idéntica en para los distintos subconjuntos de un mismo detector, del mismo modo no se producen cambios respecto a su aplicación en el preprocesamiento de imputación y predicción, por lo que

no se volverá a detallar en los casos posteriores.

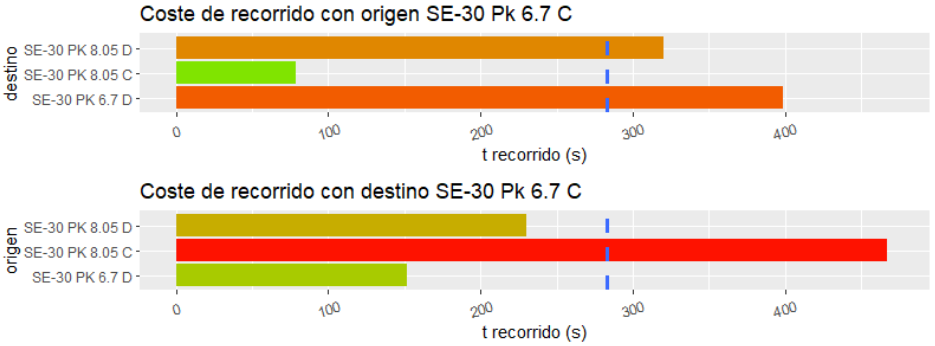


Figura 4-24: Coste de los recorridos que tienen a *SE-30 pk 6.7 C* como origen en el gráfico superior y como destino en el gráfico inferior.

Selección basada en valores perdidos y en correlación. Grupo 1: La selección basada en la tasa de coincidencia de valores perdidos no se aplica para este detector ya que no presenta ningún valor perdido en los periodos comprendidos en el grupo 1.

La selección basada en correlación es el último paso del preprocesamiento, por lo que consiste en seleccionar los 4 (tal y como se impone en las restricciones descritas en la Tabla 4-9) parámetros cuyo índice de correlación sea más alto respecto a *SE-30 pk 6.7 C*.

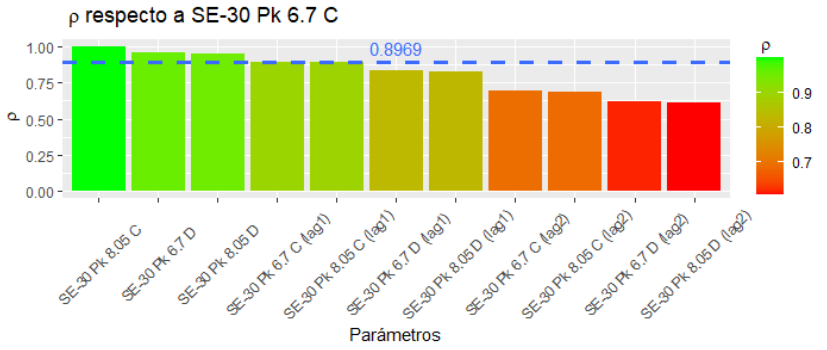


Figura 4-25: Coeficiente de correlación de Pearson de los parámetros del conjunto de datos del grupo 1 respecto al parámetro de salida SE-30 pk 6.7 C

El valor mínimo de correlación entre los parámetros que componen el conjunto de entrada, sirve como indicador de la intensidad de la relación entre los parámetros que lo componen respecto y el parámetro de salida. En el caso del grupo 1 se establece que ese valor es 0.8969 (Figura 4-25). En la Tabla 4-10, se muestran un resumen de las características de los parámetros seleccionados para el conjunto F''_{a^1}, que se proporciona a la etapa de implementación del modelo de imputación.

Tabla 4-10: Características de los parámetros incluidos en el conjunto F''_{a^1}.

Características	Descripción	Valor
Parámetros	Parámetros que componen el conjunto F''_{a^1}.	SE-30 pk 6.7 D (lag 0) SE-30 pk 8.05 C (lag 0) SE-30 pk 8.05 D (lag 0) SE-30 pk 6.7 C (lag 1)
Detectores	Detectores del escenario que aportan información al conjunto F''_{a^1}.	SE-30 pk 6.7 C SE-30 pk 6.7 D SE-30 pk 8.05 C SE-30 pk 8.05 D
Intervalos anteriores	Intervalos anteriores utilizados en el conjunto F''_{a^1}.	lag 0 lag 1
ϱ mínimo	Mínimo valor de coeficiente de correlación respecto a d_a de los parámetros que integran el conjunto F''_{a^1}.	0.08969
Coste umbral	Umbral de coste del conjunto F''_{a^1}.	283.1 s

En la Tabla 4-11 se presenta una muestra del conjunto F''_{a^1} que se obtiene como resultado del preprocesamiento de imputación, integrado por los cuatro parámetros selecciones, además del parámetro *SE-30 pk 6.7 C*, que se necesita aportar para el aprendizaje supervisado y la etapa de test en la implementación del modelo.

Tabla 4-11: Muestra del conjunto F''_{a^1} obtenido en el preprocesamiento de imputación.

Fecha	SE-30 pk 6.7 C	SE-30 pk 6.7 C (lag1)	SE-30 pk 6.7 D	SE-30 pk 8.05 C	SE-30 pk 8.05 D
17/01/2022 0:00	0.0589	-	0.0610	0.0646	0.0605
17/01/2022 1:00	0.0119	0.0589	0.0182	0.0126	0.0187
17/01/2022 2:00	0.0000	0.0119	0.0106	0.0000	0.0099
…	…	…	…	…	…
21/01/2022 21:00	0.4748	0.6236	0.4148	0.4991	-
21/01/2022 22:00	0.2779	0.4748	0.3330	0.3035	-
21/01/2022 23:00	0.1845	0.2779	0.1802	0.1962	-

Selección basada en valores perdidos y en correlación. Grupo 2: Al no producirse valores perdidos en este grupo, el paso de selección basado en tasa coincidente de valores perdidos no se aplica este paso de selección.

La selección basada en correlación se aplica del mismo modo que en el grupo 1, seleccionándose los 4 parámetros cuyo índice de correlación de Pearson respecto a *SE-30 pk 6.7 C* es más alto. En la Figura 4-26, se muestran los parámetros seleccionados, obteniéndose un índice mínimo de correlación de 0.9156.

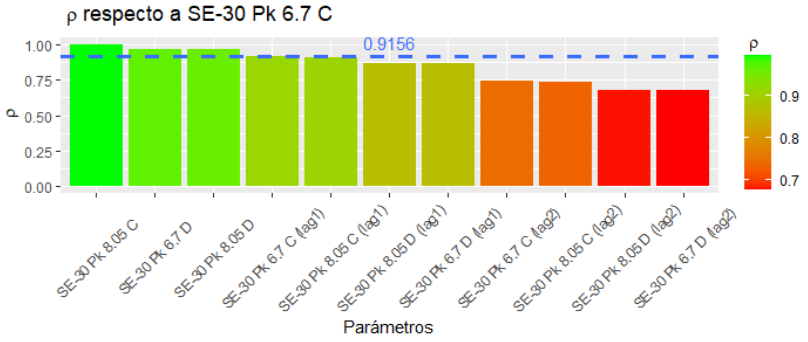


Figura 4-26: Coeficiente de correlación de Pearson de los parámetros del conjunto de datos del grupo 2 respecto al parámetro de salida SE-30 pk 6.7 C

En la Tabla 4-12 se exponen las características del conjunto de datos F''_{a^2}, obtenido tras la etapa de preprocesamiento del grupo 2.

Tabla 4-12: Características de los parámetros incluidos en el conjunto de entrada del modelo de imputación para el grupo 2.

Características	Descripción	Valor
Parámetros	Parámetros que componen el conjunto F''_{a^2}.	SE-30 pk 6.7 D (lag 0) SE-30 pk 8.05 C (lag 0) SE-30 pk 8.05 D (lag 0) SE-30 pk 6.7 C (lag 1)
Detectores	Detectores del escenario que aportan información al conjunto F''_{a^2}.	SE-30 pk 6.7 C SE-30 pk 6.7 D SE-30 pk 8.05 C SE-30 pk 8.05 D
Intervalos anteriores	Intervalos anteriores utilizados en el conjunto F''_{a^2}.	lag 0 lag 1
ϱ mínimo	Mínimo valor de coeficiente de correlación respecto a d_a de los parámetros que integran el conjunto F''_{a^2}.	0.9156
Coste umbral	Umbral de coste del conjunto F''_{a^2}.	283.1 s

En la Tabla 4-13, se presenta una muestra del conjunto F''_{a^2}, compuesto por los valores de los parámetros seleccionados en el preprocesamiento del grupo 2 y los del parámetro de salida.

Tabla 4-13: Muestra del conjunto $F''_a{}^2$, obtenido tras el preprocesamiento de imputación del grupo 2.

Fecha	SE-30 pk 6.7 C	SE-30 pk 6.7 C (lag1)	SE-30 pk 6.7 D	SE-30 pk 8.05 C	SE-30 pk 8.05 D
17/01/2022 0:00	0.0589	-	0.0610	0.0646	0.0605
17/01/2022 1:00	0.0119	0.0589	0.0182	0.0126	0.0187
17/01/2022 2:00	0.0000	0.0119	0.0106	0.0000	0.0099
…	…	…	…	…	…
21/01/2022 21:00	0.4748	0.6236	0.4148	0.4991	-
21/01/2022 22:00	0.2779	0.4748	0.3330	0.3035	-
21/01/2022 23:00	0.1845	0.2779	0.1802	0.1962	-

4.2.3.2. Implementación del modelo de imputación

La etapa de implementación del modelo de imputación parte de los subconjuntos $F''_a{}^1$ y $F''_a{}^2$ con el objetivo de generar un modelo de imputación por subconjunto, tras una serie de pasos que se detallan a continuación.

El proceso de implementación del modelo se basa en tres pasos, tal y como se detalla en 4.1.3.2, la selección del método, el ajuste de hiperparámetros y el entrenamiento del modelo. A continuación se describe su aplicación genérica para este caso de estudio y, posteriormente, se exponen los resultados de su aplicación a $F''_a{}^1$ y $F''_a{}^2$.

a. **Selección del método de imputación**: Para implementar el modelo de imputación se ha decidido utilizar una BPNN. Se trata de un método basado en el aprendizaje automático supervisado de uso muy generalizado que, concretamente en la modelación de parámetros de tráfico, ha sido utilizado de forma recurrente.

b. **Ajuste de hiperparámetros**: se ha realizado un ajuste de hiperparámetros de la red neuronal basado en el método *Grid Search* (o búsqueda exhaustiva). Se han probado cinco configuraciones distintas de BPNN (Tabla 4-14), en las que se varían el número de capas ocultas y el número de unidades que las integran. Para evitar una mayor complejidad en este paso, se ha decidido no modificar el resto de hiperparámetros, que permanecen fijos en las cinco configuraciones.

Tabla 4-14: Configuraciones probadas en la búsqueda exhaustiva del modelo de imputación.

Configuración	Distribución	Función de pérdida	Nº de pesos/sesgos	Capas (unidades)
1	Gaussiana	Cuadrática	601	3 (4-100-1)
2	Gaussiana	Cuadrática	10.701	4 (4-100-100-1)
3	Gaussiana	Cuadrática	20.801	5 (4-100-100-100-1)
4	Gaussiana	Cuadrática	281	5 (4-10-10-10-1)
5	Gaussiana	Cuadrática	301	3 (4-50-1)

Se realiza un proceso de validación cruzada de 5 iteraciones para cada una de ellas sobre el mismo conjunto de entrenamiento y se elige la configuración cuyo rendimiento en la validación cruzada sea más alto para establecerla en el modelo de imputación.

c. **Proceso de aprendizaje**: Se realiza un proceso de aprendizaje automático en el que se separa el conjunto de entrada en dos subconjuntos: entrenamiento ($F''_a{}^{*e}$) con un 80% de los casos y el de test ($F''_a{}^{*t}$) con un 20% de los casos.

Se ejecuta un proceso de validación cruzada sobre el conjunto de entrenamiento, mediante el que se realiza el proceso de aprendizaje, al mismo tiempo que se controla el sobreajuste. Como resultado de este paso se obtiene un modelo de imputación para el parámetro *SE-30 pk 6.7*, además de una serie de índices de rendimiento sobre su entrenamiento, validación y test.

A continuación, se especifica el resultado de aplicar los tres pasos de implementación del modelo sobre los conjuntos F''_{a^1} y F''_{a^2} obtenidos en el preprocesamiento

Implementación del modelo. Grupo 1: La estructura de la red neuronal para el grupo 1 (Figura 4-27) consiste en:

i) Una capa de entrada compuesta por cuatro unidades que reciben los parámetros de entrada de F''_{a^1}.

ii) Varias capas ocultas, cuya estructura se define mediante la etapa de ajuste de hiperparámetros.

iii) Una capa de salida, con una unidad, que simula el valor del parámetro de salida y efectúa el paso de retroporpagación comparándola con el valor real.

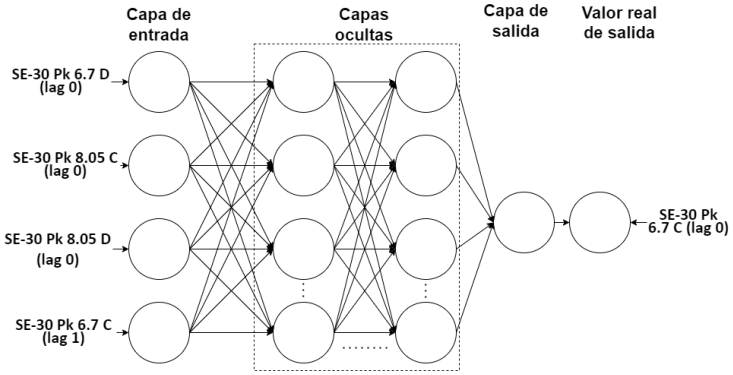


Figura 4-27: Estructura de la red neuronal para la imputación del grupo 1.

Se ha decidido utilizar el índice de rendimiento RMSE (*Root Mean Square Error*), como medida para la selección del modelo. Para la validación cruzada se utiliza el conjunto de entrenamiento $\boldsymbol{F''_{a}}^{1e}$, compuesto por un 80% de los casos de $\boldsymbol{F''_{a}}^{1}$, seleccionados aleatoriamente.

Tabla 4-15: Clasificación de modelos según la validación cruzada para el grupo 1.

Configuración	Capas (unidades)	RMSE cv
1	4 (4-100-100-1)	0.0571
2	3 (4-100-1)	0.0591
3	5 (4-100-100-100-1)	0.0605
4	5 (4-10-10-10-1)	0.0642
5	3 (4-50-1)	0.0653

Tras aplicarse la validación cruzada al conjunto $\boldsymbol{F''_{a}}^{1e}$ con las

configuraciones presentadas en la Tabla 4-14, se obtienen los resultados mostrados en la Tabla 4-15. Se determina que la configuración que mejor se ajusta es 4-100-100-1, es decir, 4 unidades en la capa de entrada, 2 capas ocultas de cien unidades cada una y una capa de salida.

Rendimiento en etapa de test. Grupo 1: Tras el proceso de validación cruzada, el modelo se prueba con el conjunto de test $F''_{a}{}^{1t}$, compuesto por el 20% de los casos de $F''_{a}{}^{1}$ que no han sido incluidos en $F''_{a}{}^{1e}$. En la Tabla 4-16, se presenta una comparativa entre el índice RMSE obtenido en el proceso de validación cruzada y el de test, resultando muy similares, por lo que se estima que se ha realizado un buen ajuste.

Tabla 4-16: Valores de RMSE para la validación cruzada y en la etapa de test.

Indice	Valor
RMSE cv	0.0571
RMSE Test	0.0563

Implementación del modelo. Grupo 2: La estructura de la red neuronal para el grupo 2 (Figura 4-28) consiste en:

i) Una capa de entrada compuesta por cuatro unidades que reciben los parámetros de entrada de $F''_{a}{}^{2}$.

ii) Varias capas ocultas, cuya estructura se define mediante la etapa de ajuste de hiperparámetros.

iii) Una capa de salida, con una unidad, que simula el valor del parámetro de salida y efectúa el paso de retroporpagación comparándola con el valor real.

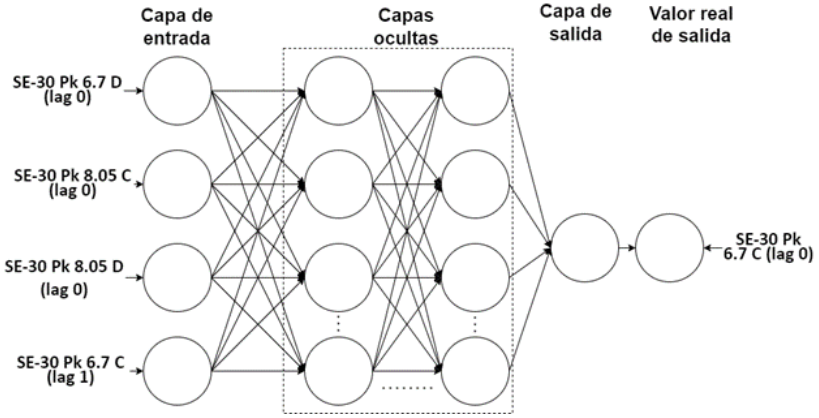


Figura 4-28: Estructura de la red neuronal para la imputación del grupo 2.

Al aplicar la validación cruzada sobre el conjunto de entrenamiento $F''_{a^{2e}}$, a las configuraciones de red neuronal descritas en la Tabla 4-14, se obtienen los resultados mostrados en la Tabla 4-17. El modelo que presenta mejores prestaciones está compuesto de cuatro capas, con la configuración 4-100-100-1.

Tabla 4-17: Clasificación de modelos según la validación cruzada para el grupo 2.

Configuración	Capas (unidades)	RMSE cv
1	4 (4-100-100-1)	0.03525635
2	3 (4-50-1)	0.03569150
3	5 (4-100-100-100-1)	0.04100628
4	3 (4-100-1)	0.04738980
5	5 (4-10-10-10-1)	0.14884007

Rendimiento en etapa de test. Grupo 2: Tras el proceso de validación cruzada, el modelo se prueba con el conjunto de test $F''_{a^{2t}}$, compuesto por el 20% de los casos de F''_{a^2} que no han sido incluidos en $F''_{a^{2e}}$. En la Tabla

4-18 se presenta una comparativa entre el índice RMSE obtenido en el proceso de validación cruzada y en el de test. Presentan valores muy parecidos, considerándose que se ha llegado a un ajuste correcto.

Tabla 4-18: Valores de RMSE para la validación cruzada y en la etapa de test.

Indice	Valor
RMSE cv	0.0352
RMSE Test	0.0396

4.2.3.3. Imputación de valores perdidos

Una vez se dispone de un modelo de imputación ajustado, entrenado y validado para cada uno de los subconjuntos, se completan los valores perdidos del parámetro de salida. Cada modelo se utiliza para imputar, los valores del grupo correspondiente.

La etapa de imputación que se ha descrito para el detector *SE-30 pk 6.7 C*, es extensible al resto de detectores del escenario. En la Figura 4-29, se muestra el resultado del proceso de imputación sobre el esenario, en el que se completan todos los valores perdidos que se daban en el detector *SE-30 Pk 8.05 D*. El conjunto completado, al que se denomina $\boldsymbol{F^I}$, se proporciona a la fase de predicción.

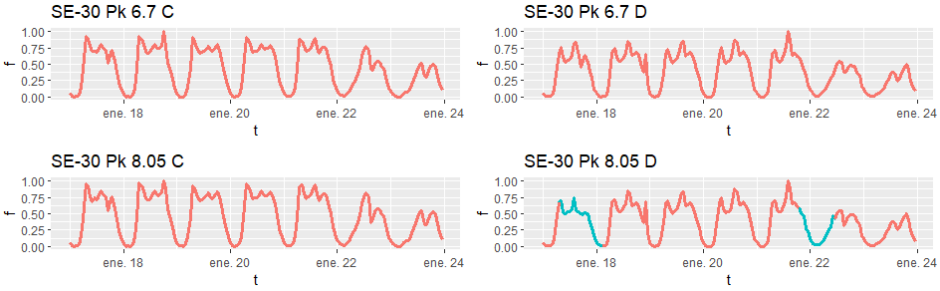


Figura 4-29: Valores de flujo de los detectores del escenario tras completar los valores perdidos mediante los modelos de imputación.

4.2.4. Predicción

El proceso de predicción parte del conjunto de datos imputado en la fase de imputación $\boldsymbol{F^I}$. El objetivo principal de esta fase es la generación de modelos precisos de predicción que, basándose en las características del escenario, tengan la capacidad de realizar predicciones precisas con un horizonte *h*.

Dentro del marco de predicción, el desarrollo del modelo de predicción se fundamenta en dos etapas, muy similares a las descritas para la imputación: preprocesamiento y implementación del modelo de predicción.

A continuación, se describe cada una de las etapas en su adaptación concreta a este escenario de ejemplo, ilustrándose con la generación de los modelos de predicción para el detector *SE-30 pk 6.7 C* con un horizonte de predicción de una hora. El proceso descrito es extensible a la generación de los modelos del resto de detectores del escenario.

4.2.4.1. Preprocesamiento

Esta etapa se compone de tres pasos: incorporación de la ventana temporal, clasificación de casos y selección de parámetros:

a. **Incorporación de la ventana temporal:** se ha decidido definir una ventana temporal de 3 horas para la predicción. Incorporándose los registros correspondientes a tres intervalos anteriores para cada detector, entre *t-h* y *t-v*, esto es, entre *t-1* y *t-3*. En la Tabla 4-19 se representa el conjunto de datos $\boldsymbol{F^{I\prime}_a}$ que se obtiene como resultado de la incorporación de la ventana temporal, teniendo en cuenta que se deben exluir los parámetros más próximos a *t-h*, en ese caso se exluyen los parámetros originales, es decir, los correspondientes a *t-0*, de los detectores diferentes a *SE-30 pk 6.7 C*. Los nuevos parámetros se representan con el nombre del detector, al que se le ha añadido el texto *"(lag x)"*, siendo x el número de horas anteriores correspondientes al valor que representa el registro.

   El conjunto de datos $\boldsymbol{F^{I\prime}_a}$ queda configurado con trece parámetros, tres por detector, refiriéndose al valor registrado por los detectores para *t-1, t-2 y t-3* respectivamente y el parámetro de salida *SE-30 pk 6.7 C*.

Tabla 4-19: Conjunto F'_a en el que se incorporan los parámetros correspondientes a la ventana temporal de predicción.

Fecha	SE-30 pk 6.7 C	SE-30 pk 6.7 C (LAG 1)	SE-30 pk 6.7 C (LAG 2)	SE-30 pk 6.7 C (LAG 3)	SE-30 pk 6.7 D (LAG 1)	SE-30 pk 6.7 D (LAG 2)	SE-30 pk 6.7 D (LAG 3)	SE-30 pk 8.05 C (LAG 1)	SE-30 pk 8.05 C (LAG 2)	SE-30 pk 8.05 C (LAG 3)	SE-30 pk 8.05 D (LAG 1)	SE-30 pk 8.05 D (LAG 2)	SE-30 pk 8.05 D (LAG 3)
17/01/22 0:00	400	-	-	-	-	-	-	-	-	-	-	-	-
17/01/22 1:00	188	400	-	-	460	-	-	408	-	-	428	-	-
17/01/22 2:00	136	188	400	-	232	460	-	188	408	-	224	428	-
17/01/22 3:00	160	136	188	400	192	232	460	132	188	408	180	224	428
...	...	...	...	...	...	...	...	...	...	...	...	...	...
23/01/22 20:00	2204	2424	2264	1964	2764	2392	2056	2428	2272	1952	2600	2280	2000
23/01/22 21:00	1588	2204	2424	2264	2356	2764	2392	2216	2428	2272	2144	2600	2280
23/01/22 22:00	1004	1588	2204	2424	1432	2356	2764	1608	2216	2428	1320	2144	2600
23/01/22 23:00	636	1004	1588	2204	988	1432	2356	1008	1608	2216	912	1320	2144

a. **Clasificación:** El proceso de clasificación es idéntico al realizado para la imputación (sección 4.2.3.1), ya que las condiciones de la predicción no influyen en la clasificación de los perfiles diarios de d_a. Es decir los conjuntos F'_{a^1} y F'_{a^2}, clasifican a F''_a del mismo modo que el mostrado en la imputación, en días laborables y no laborables.

b. **Selección de parámetros:** Se realiza una selección basada en tres criterios, aplicados en paralelo sobre los conjuntos F'_{a^1} y F'_{a^2}, ejecutándose consecutivamente la selección basada en el coste, la selección basada en valores perdidos y la selección basada en correlación. En el proceso de selección de parámetros de predicción se imponen las mismas restricciones

que en el de imputación, presentadas en la Tabla 4-9.

Selección basada en coste: Este paso se ejecuta de manera idéntica a la expuesta en el preprocesamiento de imputación (sección 4.2.3.1), seleccionándose todos los parámetros de $F'_{a}{}^{1}$ y $F'_{a}{}^{2}$ para el siguiente paso, al cumplor todos los detectores el criterio impuesto en cuanto al coste umbral.

Selección basada en valores perdidos y en correlación. Grupo 1: La selección basa en la tasa de coincidencia de valores perdidos no se aplica para este detector ya que no presenta ningún valor perdido en $F'_{a}{}^{1}$.

El criterio de selección en base al índice de correlación se ejecuta seleccionando los 4 parámetros de $F'_{a}{}^{1}$ que presentan un índice más alto respecto a *SE-30 pk 6.7 C*. En la Figura 4-30, se muestran los parámetros y su índice de correlación en orden descendente. Se seleccionan todos los parámetros correspondientes a *t-1*, estableciéndose el valor de correlación mínimo entre los parámetros del conjunto de entrada en 0.8322.

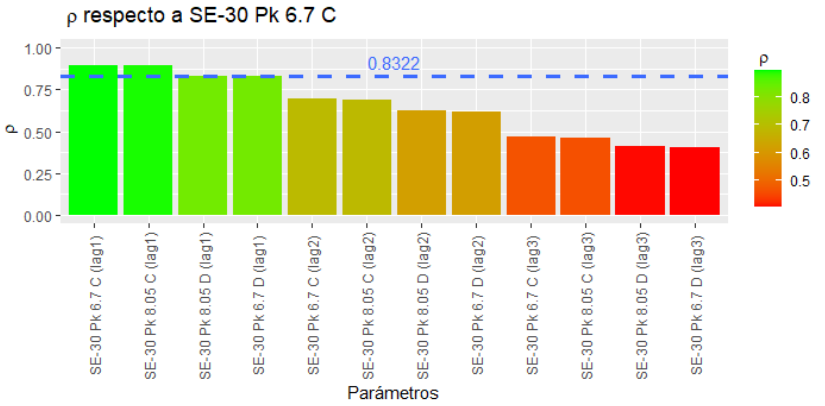


Figura 4-30: Coeficiente de correlación de Pearson de los parámetros del conjunto $F'_{a}{}^{1}$ respecto al parámetro de salida *SE-30 pk 6.7 C*.

En la Tabla 4-20 se exponen las características del conjunto de datos $F''_{a}{}^{1}$ obtenido tras la etapa de preprocesamiento para el grupo 1. Se observa que todos los detectores del escenario aportan valores al conjunto $F''_{a}{}^{1}$, que todos ellos corresponden a *t-1* y que el índice de correlación mínimo es inferior al que se observaba para la imputación del grupo 1.

Tabla 4-20: Características de los parámetros incluidos en $F''_{a}{}^{1}$.

Características	Descripción	Valor
Parámetros	Parámetros que componen el conjunto $F''_{a}{}^{1}$.	SE-30 pk 6.7 C (lag 1) SE-30 pk 6.7 D (lag 1) SE-30 pk 8.05 C (lag 1) SE-30 pk 8.05 D (lag 1)
Detectores	Detectores del escenario que aportan información al conjunto $F''_{a}{}^{1}$.	SE-30 pk 6.7 C SE-30 pk 6.7 D SE-30 pk 8.05 C SE-30 pk 8.05 D
Intervalos anteriores	Intervalos anteriores utilizados en el conjunto $F''_{a}{}^{1}$.	lag 1
ϱ mínimo	Mínimo valor de coeficiente de correlación respecto a d_a de los parámetros que integran el conjunto $F''_{a}{}^{1}$.	0.8322
Coste umbral	Umbral de coste del conjunto $F''_{a}{}^{1}$.	283.1 s

Selección basada en coste y en correlación. Grupo 2: La selección basa en la tasa de coincidencia de valores perdidos no se aplica para este detector ya que no presenta ningún valor perdido en $F'_{a}{}^{2}$.

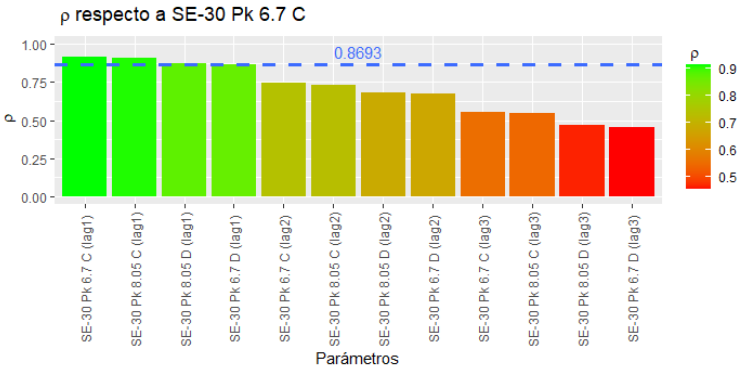


Figura 4-31: Coeficiente de correlación de Pearson de los parámetros del conjunto $F'_{a}{}^{2}$ respecto al parámetro de salida SE-30 pk 6.7 C.

El criterio de selección en base al índice de correlación se ejecuta seleccionando los 4 parámetros de $F^{I'a^2}$ con un valor más alto respecto a *SE-30 pk 6.7 C*. En la Figura 4-31, se muestra el índice de correlación de cada parámetro en orden descendente. Se seleccionan todos los parámetros correspondientes a *t-1*, estableciéndose el valor de correlación mínimo entre los parámetros del conjunto de entrada en 0.8693.

Tabla 4-21: Características de los parámetros incluidos en el conjunto de entrada del modelo de imputación para el grupo 2.

Características	Descripción	Valor
Parámetros	Especificación de los parámetros que componen el conjunto de entrada del modelo M.L. de imputación.	SE-30 pk 6.7 C (lag 1)
		SE-30 pk 6.7 D (lag 1)
		SE-30 pk 8.05 C (lag 1)
		SE-30 pk 8.05 D (lag 1)
Detectores	Detectores del escenario que aportan información al conjunto de datos de entrada de los modelos M.L.	SE-30 pk 6.7 C
		SE-30 pk 6.7 D
		SE-30 pk 8.05 C
		SE-30 pk 8.05 D
Intervalos anteriores	Periodos utilizados en el conjunto de entrada.	lag 1
ϱ mínimo	Mínimo valor de coeficiente de correlación de los parámetros que integran el conjunto de entrada respecto al parámetro de salida.	0.8693
Coste umbral	Valor del coste umbral utilizado.	283.1 s

En la Tabla 4-21 se exponen las características del conjunto de datos $F^{I''a^2}$ obtenido tras la etapa de preprocesamiento para el grupo 2. Se observa que todos los detectores del escenario aportan valores al conjunto $F^{I''a^2}$, que todos ellos corresponden a *t-1* y que el índice de correlación mínimo es inferior al que se observaba para la imputación del grupo 2.

4.2.4.2. Implementación del modelo de predicción

A continuación, se expone el proceso de implementación del modelo de predicción, para los dos subconjuntos obtenidos en la etapa de preprocesamiento $F^{I''}{}_{a}{}^{1}$ y $F^{I''}{}_{a}{}^{2}$.

a. **Selección del método de imputación**: Para implementar el modelo de imputación se ha decidido utilizar una BPNN o red neuronal con retropropagación, se trata de un método basado en el aprendizaje automático de uso muy generalizado que, concretamente en la modelación de parámetros de tráfico, ha sido utilizado de forma recurrente.

b. **Ajuste de hiperparámetros**: se ha realizado un ajuste de hiperparámetros de la red neuronal basado en el método Grid Search (o búsqueda exhaustiva). Se prueban las mismas cinco configuraciones ya expuestas para la implementación del modelo en la fase de imputación (Tabla 4-14).

   Se realiza un proceso de validación cruzada de 5 iteraciones, para cada una de las configuraciones y se elige la que presente un mejor rendimiento.

c. **Proceso de aprendizaje**: Se realiza un proceso de aprendizaje automático en el que se separa el conjunto de entrada en dos subconjuntos: entrenamiento $F^{I''}{}_{a}{}^{1e}$, con un 80% de los casos y el de test $F^{I''}{}_{a}{}^{1t}$ con un 20% de los casos.

   Implementación del modelo de predicción. Grupo 1: La estructura de la red neuronal para el grupo 1 (Figura 4-32) consiste en:

   i) Una capa de entrada compuesta por cuatro unidades que reciben los parámetros de entrada de $F^{I''}{}_{a}{}^{1}$.

   ii) Varias capas ocultas, cuya estructura se define mediante la etapa de ajuste de hiperparámetros.

   iii) Una capa de salida, con una unidad, que simula el valor del parámetro de salida y efectúa el paso de retroporpagación comparándola con el valor real.

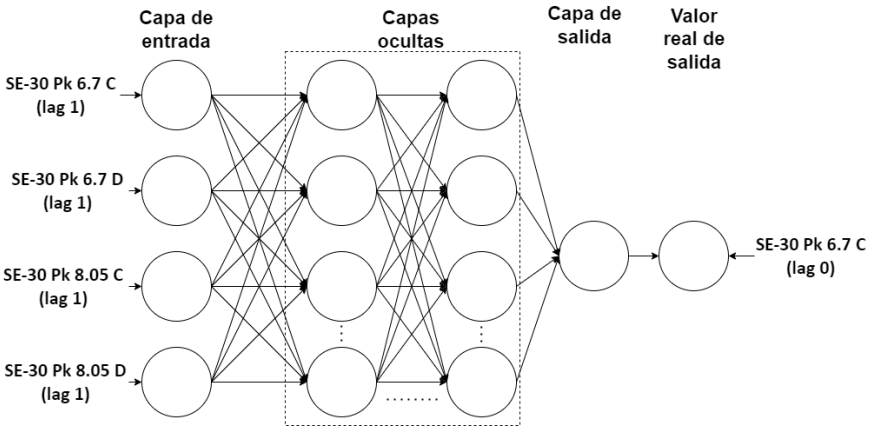


Figura 4-32: Estructura de la red neuronal para la predicción del grupo 1.

En la Tabla 4-22, presentan los resultados obtenidos tras la búsqueda exhaustiva, en la que la configuración de 4 capas 4-100-100-1 muestra el mejor rendimiento.

Tabla 4-22: Selección de modelo según la validación cruzada para el grupo 1.

Configuración	Capas (unidades)	RMSE cv
1	4 (4-100-100-1)	0.1042
2	5 (4-100-100-100-1)	0.1062
3	3 (4-100-1)	0.1394
4	5 (4-10-10-10-1)	0.1472
5	3 (4-50-1)	0.1547

Rendimiento en etapa de test: Tras el proceso de validación cruzada, el modelo se prueba con el conjunto de test $F^{I''}{}_{a}{}^{1t}$, compuesto por el 20% de los casos de $F^{I''}{}_{a}{}^{1}$ que no han sido incluidos en $F^{I''}{}_{a}{}^{1e}$. En la Tabla 4-23, se presenta una comparativa entre el índice RMSE obtenido en el proceso de validación cruzada y en el de test, resultando muy similares, por lo que se estima que se ha realizado un buen ajuste. Se observa que el error

acumulado ha crecido considerablemente si se compara con el modelo de imputación del grupo 1.

Tabla 4-23: Valores de RMSE para la validación cruzada y en la etapa de test.

Indice	Valor
RMSE cv	0.1042
RMSE Test	0.0968

Implementación del modelo de predicción. Grupo 2: La estructura de la red neuronal para el grupo 2 (Figura 4-33) consiste en:

i) Una capa de entrada compuesta por cuatro unidades que reciben los parámetros de entrada de $F^{I''}{}_a{}^2$.

ii) Varias capas ocultas, cuya estructura se define mediante la etapa de ajuste de hiperparámetros.

iii) Una capa de salida, con una unidad, que simula el valor del parámetro de salida y efectúa el paso de retroporpagación comparándola con el valor real.

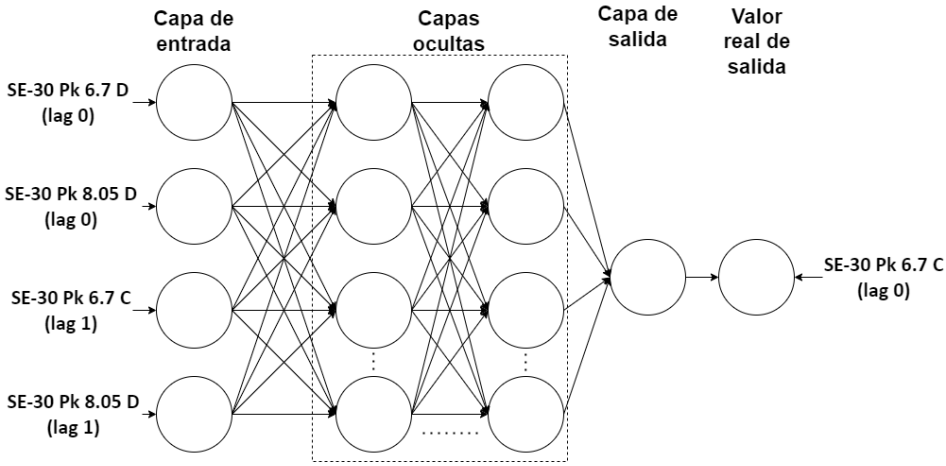


Figura 4-33: Estructura de la red neuronal para la predicción del grupo 2.

El resultado del paso de ajuste de hiperparámetros de predicción para el grupo 2 se muestra en la Tabla 4-24. La configuración que presenta el mejor resultado está compuesta por cuatro capas con la configuración 4-100-100-1.

Tabla 4-24: Selección de modelo de predicción para grupo 2.

Configuración	Capas (unidades)	RMSE cv
1	4 (4-100-100-1)	0.0923
2	5 (4-100-100-100-1)	0.0986
3	3 (4-50-1)	0.1046
4	5 (4-10-10-10-1)	0.1058
5	3 (4-100-1)	0.1059

Rendimiento en etapa de test. Grupo 2: Tras el proceso de validación cruzada, el modelo se prueba con el conjunto de test $F^{l''}{}_{a^2t}$. En laTabla 4-24, se presenta una comparativa entre el índice RMSE obtenido en el proceso de validación cruzada y en el de test, resultando muy similares, por lo que se estima que se ha realizado un buen ajuste.

Se observa que el error acumulado ha crecido considerablemente si se compara con el modelo de imputación del grupo 2.

Tabla 4-25: Valores de RMSE para la validación cruzada y en la etapa de test del grupo 2.

Indice	Valor
RMSE cv	0.0923
RMSE Test	0.0964

4.2.5. Evaluación

En esta fase se realiza una evaluación de todo el proceso, analizando las características de cada una de las fases anteriores y su influencia sobre el proceso general de predicción

4.2.5.1. Evaluación individualizada de cada fase

En la Tabla 40, se expone un cuadro resumen de esta evaluación, mostrando los aspectos reseñables observados en cada una de las fases.

1. **Adquisición.** A continuación se exponen los principales aspectos relativos a la evaluación del escenario, respecto a la fase de adquisicón. Los criterios a evaluar se refieren a la accesibilidad que presenta el dato original y a la compatibilidad el formato orginal con el esquema de datos del marco de predicción (Tabla 4-26).

Tabla 4-26: Criterios evaluados en la fase de adquisición para el caso de estudio de ejemplo.

Criterio de evaluación	Evaluación
Grado de accesibilidad de la fuente del dato.	Información cedida por el Centro de Gestión de Tráfico del Suroeste, no disponible públicamente.
Coste de importación al esquema de datos.	Formato sencillo facilita el mapeado de datos y la importación automatizada.
Adecuación de la información al proceso de predicción.	La frecuencia de registro (15 minutos) es adecuada para objetivo del horizonte de predicción de una hora. La fuente contiene toda la información requerida.
Nivel de detalle de los metadatos.	Se ofrece una guía con el significado de cada campo de información y con la ubicación de los detectores. El nivel de detalle de los metadatos es adecuado.

2. **Análisis.** En esta fase se evalúan las características del conjunto de datos importado y su idoneidad para llevar a cabo el proceso de predicción, los criterios evaluados se exponent en la (Tabla 4-27).

Tabla 4-27: Criterios evaluados en la fase de análisis para el caso de estudio de ejemplo

Criterio de evaluación	Evaluación
Características del conjunto de datos	Se dispone de 672 registros relativos a 4 detectores en el conjunto de datos del esecenario. Se ha creado un ejemplo reducido para ilustrar el funcionamiento del marco de predicción. Es una cantidad escasa para el entrenamiento de modelos de aprendizaje automático. Se necesitaría un número de casos mayor y más variado para generalizar correctamente el comportamiento de los parámetros de salida.
Tasas de valores perdidos	Tasa de valores perdidos es baja. Solo se producen valores perdidos (vv.pp. en uno de los detectores del escenario. Tasa vv.pp. escenario: 5,05% Tasa vv.pp. SE-30 6.7 C: 0% Tasa vv.pp. SE-30 6.7 D: 0% Tasa vv.pp. SE-30 8.05 C: 0% Tasa vv.pp. SE-30 8.05 D C: 20.24%
Grado de relación entre parámetros.	Los detectores se encuentran en posiciones muy cercanas, en la misma vía, 2 por sentido de circulación. Las relaciones de coste y correlación son intensas y favorables para el modelado del flujo de un detector en base a la información aportada por los demás.

3. **Imputación.** Para esta fase se evalúan las etapas internas de la fase de imputación, extrayendo y analizando las medidas de rendimiento y características observadas en su ejecución (Tabla 4-28).

Tabla 4-28: Criterios evaluados en la fase de análisis para el caso de estudio de ejemplo.

Criterio de evaluación	Evaluación
Parámetros relativos a los conjuntos de entrada obtenidos tras el preprocesamiento.	**Grupo1** Casos: 120 Parámetros entrada: 4. SE-30 pk 6.7 D, SE-30 pk 8.05 C, SE-30 pk 8.05 D y SE-30 pk 6.7 C (lag 1). ρ mínimo: 0.08969. Tiempo de recorrido máximo: 283.1s **Grupo2** Casos: 48 Parámetros entrada: 4. SE-30 pk 6.7 D, SE-30 pk 8.05 C, SE-30 pk 8.05 D y SE-30 pk 6.7 C (lag 1). ρ mínimo: 0.9156. Tiempo de recorrido máximo: 283.1s
Tasa valores imputados	Tasa de valores imputados = 100%
Índices de rendimiento de modelos de imputación.	**Grupo1** Configuración del modelo *BPNN*: 4-100-100-1 RMSE cv: 0.0571 RMSE test: 0.0563 **Grupo2** Configuración del modelo *BPNN*: 4-100-100-1 RMSE cv: 0.0352 RMSE test: 0.0396

4. **Predicción:** Para esta fase se evalúan las etapas internas de la fase de predicción, extrayendo y analizando las medidas de rendimiento y características observadas en su ejecución (Tabla 4-29).

Tabla 4-29: Criterios evaluados en la fase de análisis para el caso de estudio de ejemplo.

Criterio de evaluación	Evaluación
Parámetros relativos a los conjuntos de entrada	**Grupo1** Casos: 120. Parámetros entrada: 4. SE-30 pk 6.7 C (lag 1), SE-30 pk 6.7 D (lag 1), SE-30 pk 8.05 C (lag 1) y SE-30 pk 8.05 D (lag 1). ρ mínimo: 0.8322. Tiempo de recorrido máximo: 283.1s **Grupo2** Casos: 48. Parámetros entrada: 4. SE-30 pk 6.7 C (lag 1), SE-30 pk 6.7 D (lag 1), SE-30 pk 8.05 C (lag 1) y SE-30 pk 8.05 D (lag 1). ρ mínimo: 0.8693. Tiempo de recorrido máximo: 283.1s
Rendimiento de modelos de predicción.	**Grupo1** Configuración del modelo *BPNN*: 4-100-100-1 RMSE cv: 0.1041956 RMSE test: 0.09684678 **Grupo2** Configuración del modelo *BPNN*: 4-100-100-1 RMSE cv: 0.1042 RMSE test: 0.0968

5. APLICACIÓN A CASOS PRÁCTICOS

En este capítulo, el marco de predicción descrito en el capítulo 4 se aplica a varios casos prácticos (sección 5.5), con el objetivo de mostrar su utilidad y potencial.

Uno de los principales objetivos que se persigue con el desarrollo del marco de predicción es el establecimiento de unas bases homogéneas para la comparación de diferentes aproximaciones de modelado. Esta comparación se realiza a dos niveles:

i) Comparación entre diferentes técnicas de predicción aplicadas al mismo caso práctico.

ii) Comparación de la misma técnica aplicada a diferentes escenarios o situaciones.

A lo largo de este capítulo se demuestra es que el marco de predicción presentado en este trabajo cumple esta función solventemente.

La primera parte del capítulo se centra en describir los distintos esquemas de aplicación del marco de predicción sobre los casos prácticos incluidos en este capítulo (sección 5.1). Un esquema de aplicación consiste en una descripción de los diferentes flujos de ejecución de las fases del marco, con el objetivo de comparar diversas técnicas o estrategias.

En la secciones 5.2 se describen las técnicas utilizadas en las etapas de preprocesamiento de las fases de imputación y predicción, mientras que en la sección 5.3 se definen los métodos usados para la implementación de los modelos

de imputación y predicción.

La sección 5.4 se centra en la descripción del procedimiento de evaluación de la aplicación del marco de predicción en los casos prácticos.

En la sección 5.5 se presentan los escenarios sobre los que se ha realizado un aplicación del marco de predicción siguiendo las especificaciones descritas en las secciones 5.1-5.4.

5.1. Esquemas de aplicación

El marco de predicción descrito en el capítulo 4 se caracteriza por su modularidad y flexibilidad, permitiendo la ejecución de diversas técnicas en cada una de las fases funcionales que lo componen.

En las siguientes secciones se exponen los esquemas de aplicación específicos que se han implementado. Se definen diferentes técnicas, o formas de ejecución para una una misma fase. Estas variaciones se aplican en paralelo sobre los casos de estudio, con el objetivo de compararlas.

Se establecen diversos flujos de ejecución en las fases de adquisición y análisis (sección 5.1.1), en la fase de imputación (sección 5.1.2) y en la fase de predicción (5.1.3).

5.1.1. Adquisición y análisis

Se proponen dos variantes de ejecución a partir de la fase de adquisición y análisis de información, en función del horizonte h. Se establece que el nivel de agregación del conjunto de datos $\boldsymbol{F}$ en cada ejecución del marco debe presentar una agregación temporal de los registros de flujo equivalente al del valor de h.

Se disponen dos horizontes de predicción para las aplicaciones del marco:

i) h=15 minutos. Los valores de flujo en $\boldsymbol{F}$ se agregan 15 en 15 minutos.

ii) h=1 hora. Se configura $\boldsymbol{F}$ con un nivel de agregación temporal de 1 hora de los valores de flujo.

Mediante estas dos variantes, se persigue evaluar el rendimiento del marco sobre

un mismo escenario, con dos horizontes diferentes de predicción y con dos agregaciones diferentes de los valores de flujo en el conjunto de datos.

5.1.2. Esquemas de aplicación de la fase de imputación

Se proponen dos alternativas de ejecución de esta fase.

i) NI: *No* Imputar, en este caso se obvia la fase de imputación dentro de la ejecución del marco. Se proporciona el conjunto de datos ***F*** a la fase de predicción.

ii) DLI*: Deep Learning Imputation*, los modelos que se implementan en la fase de imputación se basan en redes neuronales profundas, de tipo BPNN. Las especificaciones de este modelo se detallan en la sección 5.3.1.

5.1.3. Esquemas de aplicación de la fase de predicción

En la fase de predicción se aplican dos métodos de forma:

i) DLP: *Deep Learning Prediction*, los modelos que se implementan en la fase de predicción se basan en redes neuronales profundas de tipo BPNN, las especificaciones de este modelo se detallan en la sección 5.3.2.

ii) RFP: *Random Forest Prediction*, los modelos que se implementan en la fase de predicción se basan en técnica de Bosque Aleatorio, las especificaciones de este modelo se detallan en la sección 5.3.2.

Se establecen dos horizontes de predicción distintos:

i) h=15 minutos, en la fase de predicción se generan modelos de predicción cuyo horizonte es 15 minutos.

ii) h=1 hora, en la fase de predicción se generan modelos de predicción cuyo horizonte es 60 minutos.

En la Tabla 5-1, se muestran los 8 esquemas de aplicación definidos, en base al horizonte, a la técnica utilizada en la fase de imputación y a la técnica utilizada en la

fase de predicción. Esto significa que para cada detector de un escenario se generan 8 versiones distintas de modelo de predicción, una por cada esquema que se haya aplicado.

Los modelos de predicción que se obtienen como resultado de cada uno de estos esquemas de aplicación reciben una denominación mostrada en la primera fila de la tabla Tabla 5-1.

Tabla 5-1: Esquemas de apliación del marco de predicción a los casos prácticos.

Denominación del modelo para d_a	h	Imputación	Predicción
NI-DLP-15$_a$	15′	NI	DLP
NI-RFP-15$_a$	15′	NI	RFP
DLI-DLP-15$_a$	15′	DLI	DLP
DLI-RFP-15$_a$	15′	DLI	RFP
NI-DLP-1h$_a$	1h	NI	DLP
NI-RFP-1h$_a$	1h	NI	RFP
DLI-DLP-1h$_a$	1h	DLI	DLP
DLI-DLP-1h$_a$	1h	DLI	RFP

La evaluación del rendimiento obtenido para los modelos de predicción en cada una de las ocho ejecuciones permite comparar la influencia sobre el proceso completo de los aspectos que son diferentes, que se enumeran a continuación:

i) Influencia del nivel de agregación de la información sobre el proceso de predicción.

ii) Influencia del horizonte de predicción sobre el rendimiento del modelo de predicción.

iii) Impacto de la fase de imputación sobre el proceso general de

predicción.

iv) Comparación de rendimiento de una misma técnica de modelado en las distintas situaciones planteadas previamente.

v) Comparación del rendimiento de dos técnicas distintas de predicción sobre la misma situación.

En los casos de estudio mostrados en la sección 5.5, se ejecutan estos esquemas de aplicación sobre una selección reducida de detectores, con el objetivo de ilustrar el proceso con claridad y concisión. Se han seleccionado detectores con características diversas, como su localización dentro del escenario, el comportamiento del tráfico que registra y la tasa y estructura de los valores perdidos que presentan, permitiendo evaluar la aplicación del marco a detectores en situaciones diversas.

5.2. Técnicas de preprocesamiento

En esta sección se detallan las técnicas concretas que realizan las función de preprocesamiento en las fases de imputación y predicción de la aplicación práctica del marco de predicción.

5.2.1. Clasificación

El proceso de clasificación de información, integrado en la etapa de preprocesamiento, se realiza mediante una técnica de agrupamiento o *clustering*.

Este tipo de técnicas ha contribuido a la mejora del rendimiento de los modelos con los que se combina, tal y como se observa en los múltiples ejemplos mostrados en las secciones 3.1 y 3.2.2.

El elemento sobre el que se realiza la calsificación es el perfil diario del flujo del detector que representa el parámetro de salida para el que se desarolla el modelo d_a. Se trata de un elemento sencillo de calcular y es de utilidad en la predicción del tráfico, ya que ayuda a identificar el comportamiento recurrente de éste (Figura 5-1).

Las técnicas de *clustering*, se caracterizan por separar los elementos en función de su cercanía a un elemento central, o centroide. Entre las múltiples opciones que

presentan rendimiento y coste de computación parecido, se ha decidido utilizar la técnica *fuzzy C-Means* porque, además de realizar de forma eficiente la clasificación, proporciona información sobre el grado de pertenencia de cada elemento a cada grupo.

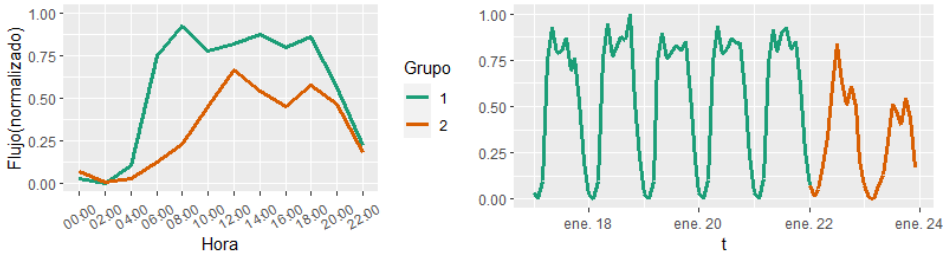


Figura 5-1: Ejemplo de perfiles característicos derivados de la clasificación de perfiles diarios de un detector.

Se ha decidido establecer el número máximo de grupos en los que se separa la muestra en dos, en todas las aplicaciones del marco. De esta forma se asegura que los grupos presenten un volumen suficiente de información en la entrada de los modelos de imputación y predicción de cada uno de los grupos.

Se debe tener en cuenta que el método *fuzzy C-Means*, no tiene la capacidad de clasificar elementos que presentan valores perdidos en su entrada, que, como se ha observado en la sección 3.2.1, es una circunstancia usual en los conjuntos de datos de información de tráfico.

Para subsanar este aspecto se ha considerado una agregación temporal amplia de los valores en los perfiles de entrada, de 2 en 2 horas, calculando la mediana de los valores presentes en cada intervalo. Se observa que este nivel de agregación no afecta notablemente a la clasificación de los perfiles ya que sigue reflejando los elementos característicos del perfil como los picos de tráfico y los cambios bruscos de condiciones. Los perfiles que siguen mostrándose como incompletos a pesar de este nivel de agregación, se tratan posteriormente, integrándose en el grupo cuyo centroide se encuentre a una distancia euclídea menor (comparando los valores de los intervalos presentes en ambos perfiles), considerándose que aquellos perfiles en los que no se dispone de más del 40% de los valores.

Este valor se ha elegido de una manera arbitraria, permitiendo la introducción de un número considerable de elementos que quedarían fuera del análisis y, por tanto,

de la clasificación. Es necesario un estudio más profundo sobre la idoneidad del valor elegido, para establecer una medida concreta óptima, al tratarse de un elemento secundario de una herramienta auxiliar se ha decidido no profundizar en este aspecto.

En la implementación concreta se ha utilizado el software estadístico R, concretamente la librería *"clúster"* (Rousseeuw et al., 2019) y su función *cmeans()*, que clasifica el conjunto de datos proporcionado en su entrada, en un número de grupos indicado mediante el método *fuzzy C-Means*.

5.2.2. Selección de parámetros

La selección de parámetros consiste en la ejecución secuencial de tres pasos:

i) Selección basada en coste.

ii) Selección basada en la tasa de coincidencia de valores perdidos.

iii) Selección basada en relaciones de tipo lineal/no lineal.

En esta sección se concretan las técnicas que se utilizan en cada una de ellas, debiendo presentar la capacidad de adaptarse a las características de los conjuntos de datos los distintos escenarios.

La selección de parámetros se debe basar en un compromiso entre el volumen y la calidad de la información que se aporta a los modelos de imputación y predicción. Se debe asegurar la presencia de un número de casos y parámetros suficiente para permitir que los modelos generalicen el comportamiento del parámetro de salida, al mismo tiempo que se debe evitar un volumen excesivo de información que introduzca una complejidad excesiva, suponiendo un aumento de coste computacional de las operaciones que se realizan sobre él, además de favorecer la introducción de ruido e información redundante y superflua que podrían perjudicar el proceso de entrenamiento de los modelos.

A continuación, se exponen las técnicas utilizadas en cada uno de los pasos que componen la etapa de selección de parámetros:

5.2.2.1. Selección basada en coste

La selección basada en coste, se fundamenta en el cálculo del tiempo de recorrido

entre las localizaciones de cada par de detectores, en ambos sentidos, utilizando los segundos como unidad de medida.

Se obtiene como resultado una matriz de costes entre las localizaciones de los pares de detectores del caso de estudio. En la Figura 5-2 se muestra una de matriz de costes, correspondiente al ejemplo expuesto en la sección 4.2.

Se considera que un detector presenta una relación de mayor intensidad con otro, cuanto menor sea el coste del desplazamiento entre ambos

Figura 5-2: Ejemplo de matriz de costes de un escenario.

El tiempo de recorrido entre dos pares de detectores se calcula estableciendo la ruta válida más corta entre las localizaciones en la que se ubican los detectores. Este cálculo debe tener en cuenta las restricciones existentes en la red de carreteras del escenario. El cálculo se realiza asumiendo un régimen de flujo libre.

Se deben seleccionar los parámetros relativos a detectores cuyos trayectos hacia, o desde, d_a presente los valores más bajos de coste. El objetivo de esta selección es reducir el volumen de datos de $\boldsymbol{F'}$, asegurando que los parámetros seleccionados guardan una relación de alta intensidad respecto a d_a en el plano espacio-temporal, reduciéndose significativamente el volumen de datos tratado. El número de parámetros seleccionados depende de las características del caso de estudio, atendiendo a:

i) Su escala.

ii) Al número de detectores que lo componen.

iii) Al nivel de accesibilidad, a través de la red, de cada detector.

La selección mediante este criterio se establece definiendo un umbal máximo de coste. Se seleccionan los parámetros relativos a los detectores que presenten un

valor de coste hacia, o desde, d_a inferior a dicho umbral.

El umbral para d_a se establece mediante el cálculo del promedio del coste de todos los tiempos de recorrido que tienen como origen/destino a d_a. De esta forma el umbral se define dinámicamente para cada detector, adaptándose a las características de éste respecto a su ubicación en el escenario y asegurando la inclusión los parámetros relativos a aquellos detectores más cercanos.

Para realizar el cálculo del coste entre detectores se ha utilizado la API *OSRM* (OSRM, 2019) que es instanciada a través del software de cálculo estadístico y matemático R. OSRM es una herramienta de cálculo de rutas entre puntos de la red de carreteras de *OpenStreetMap*, de acceso abierto y que está debidamente caracterizada para este tipo de cálculos.

5.2.2.2. Selección basada en tasa de valores perdidos coincidencentes

La aparición simultánea de valores perdidos en distintos detectores es un suceso común, más frecuente entre detectores que comparten una misma localización, ya sea en diferentes carriles de una misma sección de la vía, o bien, en posiciones cercanas de un mismo tramo.

En la Figura 5-3 se muestra un ejemplo en el que se producen varios periodos de valores perdidos (representados en negro), en múltiples detectores de forma simultánea.

La tasa de valores perdidos coincidentes de un detector d_a, respecto a otro detector d_b, se establece mediante el cociente entre el total de intervalos con valores perdidos de d_a y el número de ellos en los que coincide en el mismo intervalo con valores perdidos de d_b.

Los parámetros que presentan una alta tasa de valores perdidos coincidentes respecto a d_a, son candidatos deficientes para formar parte de los conjuntos de entrada de los modelos generados para ese detector, ya que reducirían el número de casos completos, además de existir una alta probabilidad que presenten valores perdidos al mismo tiempo en el que se necesita simular el valor de d_a.

Por estos motivos, la tasa coincidente de valores perdidos es un factor determinante a la hora de considerar un parámetro como candidato al conjunto de entrada de los modelos de otro parámetro.

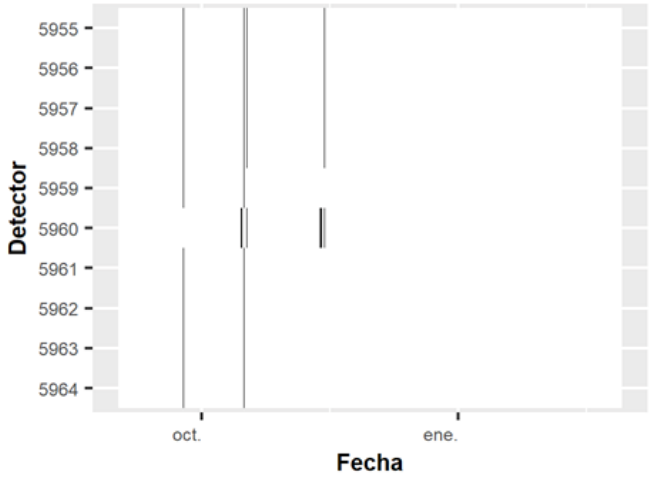


Figura 5-3: Ejemplo de presencia de valores perdidos en escenario.

Se debe definir un umbral del 30% para este indicador que permita filtrar aquellos parámetros con una alta tasa respecto al parámetro de salida

5.2.2.3. Selección basada en relaciones lineales/no lineales

Se ha decidido establecer un criterio de relación simple y de uso generalizado para identificar la intensidad de la relación lineal entre parámetros. Se utiliza el coeficiente de correlación de Pearson o ϱ, de los parámetros del conjunto de datos respecto a d_a.

Este paso es el último que se produce en el preprocesamiento, por lo que define el conjunto de entrada del modelo. El criterio de selección mediante correlación se aplica escogiendo los n parámetros con un valor más alto, entre todos los parámetros del conjunto de datos, siendo n el número máximo de parámetros permitidos en el conjunto de entrada de los modelos. El valor de n se establece en función del número de intervalos incluidos en la ventana temporal.

5.3. Aproximaciones de modelado

En este apartado se detallan las aproximaciones metodológicas elegidas para el

modelado de imputación (sección 5.3.1) y de predicción (sección5.3.2), ofreciéndose una descripción de la aplicación de cada una de ellas a los casos prácticos. Se presentan dos aproximaciones distintas para la imputación y otras dos para la predicción.

Las especificaciones concretas de la estructura interna de cada método se definen dinámicamente en cada situación, mediante la etapa de ajuste de hiperparámetros

5.3.1. Aproximaciones de modelado de imputación

Se ha optado por definir dos alternativas para la imputación, con el objetivo de observar su efecto sobre el proceso de predicción:

5.3.1.1. NI (No imputar)

No se ejecuta la fase de imputación en esta ejecución del marco de predicción sobre el escenario. Implica que la fase de predicción recibe como entrada, directamente, el conjunto de datos ***F***, obtenido en la fase de adquisición de información. En este caso, el conjunto de datos que recibe la fase de predicción contiene una proporción más alta de valores perdidos.

5.3.1.2. DLI (Deep Learning Imputation)

El método DLI se fundamenta en la utilización de una red neuronal profunda, para la imputación de valores perdidos en ***F***. Se trata de una red tipo BPNN con varias capas ocultas (Figura 5-4).

La DLI se diseña siguiendo el proceso de implementación del modelo de imputación detallado en 4.2.3.2, teniendo en cuenta el conjunto de entrada proporcionado por el preprocesamiento de la fase de imputación.

La configuración interna del modelo se elige mediante un proceso de ajuste de hiperparámetros basado en una búsqueda exhaustiva (*Grid Search*), entre una serie de configuraciones que se proporcionan al proceso de ajuste de hiperparámetros.

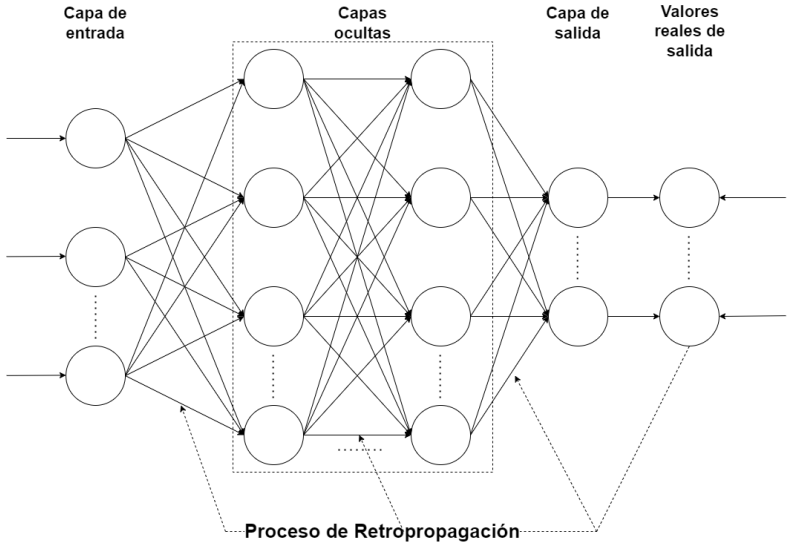


Figura 5-4: Esquema de red neuronal profunda utilizado en la fase de imputación para el método DLI.

En todos los escenarios, la búsqueda exhaustiva se efectúa entre las configuraciones presentadas en la Tabla 5-2. En las configuraciones presentadas solo se varía la estructura de las capas ocultas, definiéndose cinco configuraciones distintas.

Tabla 5-2: Configuraciones probadas en la búsqueda exhaustiva del modelo de imputación.

Configuración	Distribución	Función de pérdida	Nº de pesos/sesgos	Capas (unidades)
1	Gaussiana	Cuadrática	601	3 (4-100-1)
2	Gaussiana	Cuadrática	10.701	4 (4-100-100-1)
3	Gaussiana	Cuadrática	20.801	5 (4-100-100-100-1)
4	Gaussiana	Cuadrática	281	5 (4-10-10-10-1)
5	Gaussiana	Cuadrática	301	3 (4-50-1)

Para todas las configuraciones se realiza un proceso de validación cruzada de 5

iteraciones *(5-fold Cross Validation)*, seleccionándose la configuración que obtiene el mejor índice de rendimiento en este proceso.

En cuanto a la implementación práctica del algoritmo DLI, se ha recurrido a la plataforma de código abierto h2o.ai, en concreto el paquete h2o en su versión para R (The H2O.ai team, 2016) que ofrece una amplia gama de métodos basados en IA y herramientas para el tratamiento de conjuntos de datos. Incluye la función *h2o.deeplearning()*, que crea una red neuronal con los parámetros proporcionados. H2o.ai también dispone de una herramienta para la ejecución automatizada de la búsqueda exhaustiva, en base a un conjunto de entrada, denominada *autoML()*. Esta herramienta proporciona un listado ordenado, de las configuraciones probadas, aportando los valores de diferentes índices de rendimiento relativos a la validación cruzada.

5.3.2. Aproximaciones de modelado de predicción

Se han implementado 2 aproximaciones de modelado de predicción, enmarcadas en la familia de métodos del aprendizaje automático:

i) DLP (Deep Learning Prediction): un método de predicción basado en redes neuronales profundas.

ii) RFP (Random Forest Prediction): un método basado en la técnica de bosque aleatorio.

De cada aproximación, se implementa un modelo individual para cada horizonte de predicción, habiéndose elegido los valores $h_1 = 15'$ y $h_2 = 60'$.

5.3.2.1. DLP (Deep Learning Prediction)

El método DLP se fundamenta en la estructura de una red neuronal profunda Figura 5-5. Se trata de una red tipo BPNN con varias capas ocultas, para predecir los valores de da a corto plazo con un horizonte de predicción definido por h.

La configuración del DLP se define en función de la etapa de ajuste de hiperparámetros correspondiente, teniendo en cuenta el conjunto que recibe como entrada $\boldsymbol{F^{I''}}_a^*$, proporcionado por el preprocesamiento de la fase de predicción.

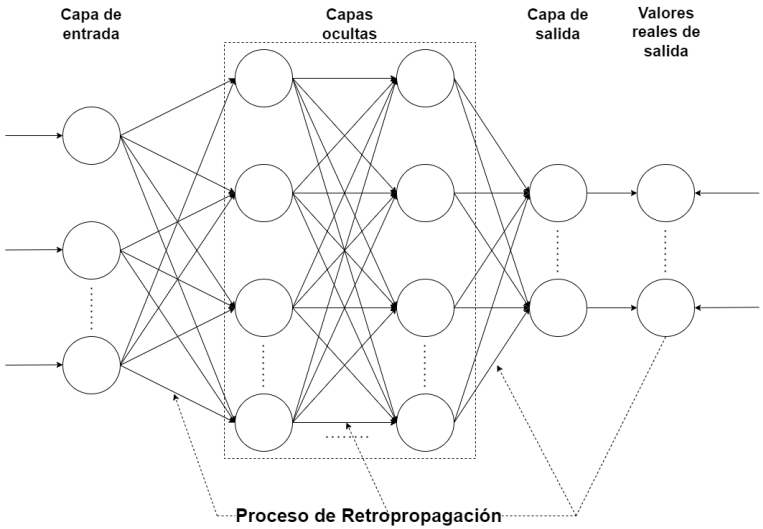


Figura 5-5: Esquema de red neuronal profunda para la definición de DLP.

El ajuste de hiperparámetros se define mediante una búsqueda exhaustiva (Grid Search) entre las configuraciones presentadas en la Tabla 5-3. En las configuraciones presentadas, tanto la distribución como la función de activación del modelo permanecen fijas, mientras que varía la estructura de las capas ocultas, definiéndose cinco configuraciones distintas.

Tabla 5-3: Configuraciones probadas en la búsqueda exhaustiva del modelo de predicción.

Configuración	Distribución	Función de pérdida	Nº de pesos/sesgos	Capas (unidades)
1	Gaussiana	Cuadrática	601	3 (4-100-1)
2	Gaussiana	Cuadrática	10.701	4 (4-100-100-1)
3	Gaussiana	Cuadrática	20.801	5 (4-100-100-100-1)
4	Gaussiana	Cuadrática	281	5 (4-10-10-10-1)
5	Gaussiana	Cuadrática	301	3 (4-50-1)

Para todas las configuraciones se realiza un proceso de validación cruzada de 5

iteraciones (5-fold Cross Validation), seleccionándose la configuración que obtiene el mejor índice de rendimiento.

En cuanto a la implementación práctica del algoritmo DLP, se utilizan las mismas herramientas que las comentadas en las especificaciones de DLI (sección 5.3.1)

5.3.2.2. RFP (Random Forest Prediction DLP) (Deep Learning Prediction)

Se define un modelo de predicción basado en la estructura de Bosque Aleatorio o *Random Forest*, al que se denomina *Random Forest Prediction* (RFP). Este tipo de modelos utiliza una estructura de bosque aleatorio distribuido (Figura 5-6) para modelar el valor de flujo de un detector.

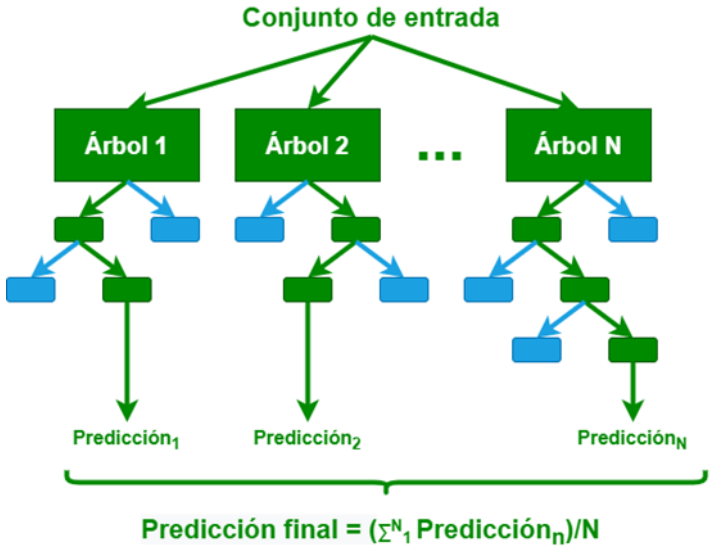


Figura 5-6: Estructura de bosque aleatorio para el método RFP.

El bosque aleatorio distribuido es un algoritmo basado en el aprendizaje automático supervisado que se compone de un conjunto de árboles de decisión, cuyos resultados se agregan mediante el método *bootstrap*, combinando los resultados de distintos métodos de predicción para aumentar la estabilidad y precisión. Se considera una herramienta flexible, fácil de implementar, eficiente y robusta para la clasificación y la regresión.

Dentro de esta familia de modelos, también se encuentran los *Extremely Randomized Trees* o XRT. Se diferencia de DRF en que en la subselección del conjunto de datos que se proporciona a los árboles, se introduce un grado más de aleatoriedad. Suelen presentar un rendimiento equivalente o peor que los DRF, aunque su coste computacional suele ser menor.

El ajuste de hiperparámetros para RFP consiste en realizar una validación cruzada sobre un modelo DRF y otro XRT, durante el proceso de entrenamiento se definen las principales características del bosque aleatorio que, básicamente son:

i) El número de árboles que compone el bosque.

ii) La profundidad mínima en cada árbol.

iii) La profundidad máxima en cada árbol. Se selecciona el modelo con un mejor índice de rendimiento obtenido en el proceso de validación cruzada.

Para RFP se ha recurrido a la plataforma h2o.ai, específicamente el paquete h2o en su versión para R (The H2O.ai team, 2016). Concretamente, se ha utilizado la función h2o.randomforest(), que crea un modelo de tipo bosque aleatorio y la herramienta autoML(), para ejecutar la validación cruzada y selección del modelo con mejor puntuación.

5.4. Evaluación de aplicación del marco

La evaluación de la aplicación del marco a cada escenario consiste en la revisión de una serie de elementos, que pueden ser de tipo cualitativo o cuantitativo, dependiendo de las características de cada fase.

Desde el punto de vista de los esquemas de aplicación, la evaluación consiste en una comparativa de las distintas técnicas aplicadas en una misma fase del marco comparando los índices de rendimiento que generan.

Tabla 5-4: Criterios de evaluación de las fases del marco de predicción.

Etapa	Características evaluadas
Adquisicón	Grado de accesibilidad de la fuente del dato.
	Grado de compatibilTransformaciones necesarias para importar la información al esquema de datos (compatibilidad entre formatos y esquemas de datos).
	Adecuación de la información al proceso de predicción.
	Nivel de detalle de los metadatos.
Análisis	Número de casos en el conjunto de datos.
	Tasa de valores perdidos y distribución, individuales en cada detector y general del escenario.
	Grado de relación entre parámetros.
Imputación	Parámetros relativos a los conjuntos de entrada obtenidos tras el preprocesamiento.
	Tasa valores imputados vs valores perdidos en el conjunto original.
	Índices de rendimiento de modelos de imputación (MAE, MSE, RMSE, R2) en las distintas etapas de implementación del modelo. Comparativa entre valores obtenidos para los distintos detectores.
	Comparación entre el rendimiento de distintas técnicas utilizadas en el método de imputación en un mismo escenario.
Predicción	Parámetros relativos a los conjuntos de entrada obtenidos tras el preprocesamiento.
	Índices de rendimiento de modelos de predicción (MAE, MSE, RMSE, R2) en las distintas etapas de implementación del modelo.
	Comparación entre el rendimiento de distintas técnicas utilizadas en el preprocesamiento para un mismo método de predicción.
	Comparación entre el rendimiento de distintas técnicas utilizadas en el método de predicción en un mismo escenario.

En laTabla 5-4, se definen los elementos que deben ser observados y evaluados en cada una de las fases

5.5. Casos prácticos

A continuación se presentan cuatro casos prácticos sobre los que se ha realizado una aplicación del marco de predicción. A cada uno de los casos presentados se les aplica el marco de predicción con las concreciones presentadas en las secciones 5.1-5.4.

De cada caso práctico se realiza una introducción basada en su ubicación y características. Estableciéndose la escala, el número de detectores que participan y el marco temporal contemplado.

A continuación se presentan los resultados obtenidos en las fase de modelado de forma esquemática, definiéndose el esquema que se aplica y mostrándose la información relativa a los distintos parámetros:

i) Resultados de imputación con h=15'

ii) Resultados de imputación con h=1 hora

iii) Resultados de predicción con h=15'

iv) Resultados de predicción con h=1 hora

Para cada una de estas ejecuciones:

i) En primer lugar se presentan las características del escenario en cuanto a valores perdidos, mediante una tabla resumen con los parámetros relativos a cada detector.

ii) Las restricciones impuestas en el preprocesamiento.

iii) Tabla de resultados presentando las características de los modelos utilizados en cada detector y los índices de rendimiento para los procesos de validación cruzada (RMSEcv) y de test (RMSETest) y de test sobre el conjunto completo, calculando el error sobre agregandolos conjuntos de test de los dos modelos correspondientes a un detector (RMSE cc), para dar mostrar sobre el conjunto de datos completo.

5.5.1. California, PEMS I-405

Este escenario se compone de un tramo de la autovía I-405 de, aproximadamente 2 km de longitud en el que se ubican 8 detectores (Figura 5-7), gestionados por el sistema de monitorización de carreteras PeMS de CalTrans (California Department of Transportation, 2018), la autoridad de carreteras de California.

Figura 5-7: Representación del caso práctico, PEMS I-405

En la Tabla 5-5 se muestran las características básicas de este escenario, presentándose la información sobre los detectores que lo componen, la red de carreteras en la que se localizan, el marco temporal escogido y la tasa general de valores perdidos.

Tabla 5-5: Características del caso práctico PEMS I-405.

Elemento	Descripción
Caso de estudio	PEMS I-405
Detectores	8 detectores situados sobre la I-405 entre los pk 0,37 y 1,34, 4 en cada sentido de circulación. Agregación de parámetros por sentido. 6 carriles por sentido Agregación temporal mínima:5 minutos.
Red	Corredor de autovía en zona interurbana.
Marco temporal	Fecha inicial = 01-10-2018, Fecha final = 31-10-2018
Tasa de valores perdidos	7,68%

5.5.1.1. Análisis de información

En lo relativo al comportamiento del flujo de tráfico diario, se observan perfiles y volúmenes diarios semejantes para todos los detectores (Figura 5-8), a excepción del detector 5, que presenta valores mucho más bajos.

Los valores perdidos se reparten de manera desigual, produciéndose largos periodos en tres detectores.

Las relaciones de coste de desplazamiento entre estaciones muestran unos datos similares entre todos los detectores, destacándose 5 que en el rol de destino presenta valores de coste más altos.

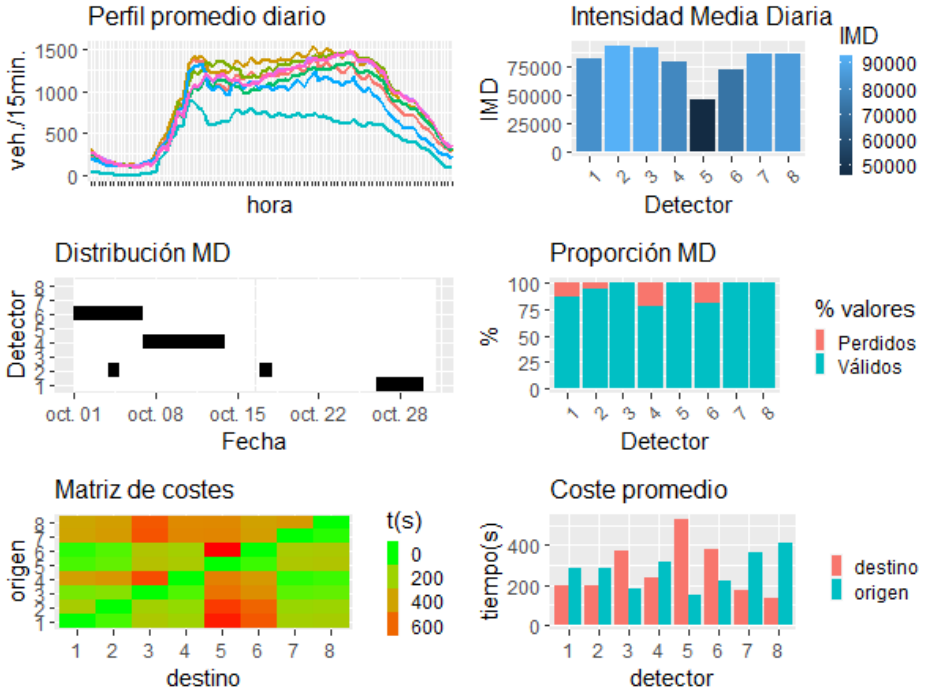


Figura 5-8: Características del caso práctico PEMS-I-405.

A continuación se presentan los resultados obtenidos mediante la aplicación del marco de predicción en las distintas situaciones planteadas.

5.5.1.2. Imputación

En esta sección se realiza una exposición sintética de los resultados obtenidos tras la aplciación de la fase de imputación del marco de predicción al escenario PEMS I-405.

En primer lugar se muestran los resultados de imputación con el modelo DLI, con agregación del conjunto de datos de 15 minutos, por una parte y con agregación de 1 hora, por otra.

Imputación DLI para h=15'

Tabla 5-6: Parámetros relativos a los valores perdidos de cada detector

Detector	Intervalos MD	Tasa MD	Periodos MD	Longitud periodo	Distancia media entre periodos
1	384	12.90%	1	384.00	-
2	193	6.49%	3	64.33	623.50
3	0	0%	0	-	-
4	672	22.58%	1	672.00	-
5	2	0.07%	2	1.00	293.00
6	576	19.35%	1	576.00	-
7	0	0%	0	0.00	-
8	0	0%	0	0.00	-

Tabla 5-7: Restricciones de ejecución de imputación DLI con h=15'.

Máximo nº de parámetros	16
Tamaño de Ventana temporal	8 (2 horas)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-8: PEMS I-405. Resultados de imputación DLI para h=15'

Det.	Grupo	Parám. Entrada	Coste máximo	ϱ mínimo	Configuración Modelo	RMSE cv	RMSE test	RMSE Test cc
1	1	16	226.37	0.927	16-100-100-1	0.0299	0.0304	0.0278
	2	16	226.37	0.969	16-100-100-1	0.0267	0.0214	
2	1	16	225.83	0.952	16-100-100-100-1	0.0556	0.062	0.0602
	2	16	225.83	0.837	16-10-10-10-1	0.0512	0.0559	
3	1	16	421.50	0.953	16-10-10-10-1	0.0251	0.0255	0.0247
	2	16	421.50	0.964	16-10-10-10-1	0.0216	0.0226	
4	1	16	267.39	0.962	16-100-1	0.0276	0.0361	0.0336
	2	16	267.39	0.95	16-10-10-10-1	0.0295	0.0274	
5	1	16	605.36	0.733	16-10-10-10-1	0.0414	0.0404	0.0410
	2	16	605.36	0.863	16-100-1	0.0424	0.0423	
6	1	16	428.56	0.905	16-10-10-10-1	0.0253	0.0245	0.0278
	2	16	428.56	0.937	16-100-100-1	0.0356	0.0358	
7	1	16	200.31	0.946	16-10-10-10-1	0.0167	0.0308	0.0325
	2	16	200.31	0.958	16-100-100-1	0.0314	0.0368	
8	1	16	148.53	0.927	16-10-10-10-1	0.0217	0.0134	0.0141
	2	16	148.53	0.948	16-10-10-10-1	0.0134	0.0157	

Imputación DLI con h=1 hora

Tabla 5-9: Parámetros relativos a los valores perdidos de cada detector.

Detector	Intervalos MD	Tasa MD	Periodos MD	Longitud periodo	Distancia media entre periodos
1	96	12.90%	1	384.00	-
2	48	6.45%	3	64.33	623.50
3	0	0%	0	-	-
4	168	22.58%	1	672.00	-
5	0	0%	0	-	-
6	144	19.35%	1	576.00	-
7	0	0.00%	0	0.00	-
8	0	0.00%	0	0.00	-

Tabla 5-10: Restricciones de ejecución de imputación DLI con h=15'.

Máximo nº de parámetros	4
Tamaño de Ventana temporal	2 (2 horas)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-11: PEMS I-405. Resultados de imputación DLI para h=1 hora'.

Det.	Grupo	Parám. Entrada	Coste Umbral (s)	ϱ Umbral	Configuración Modelo	RMSE cv	RMSE test	RMSE Test cc
1	1	4	226.37	0.9467	4-100-1	0.0697	0.0753	0.0802
	2	4	226.37	0.8971	4-100-100-1	0.1100	0.0923	
2	1	4	225.83	0.9151	4-100-100-1	0.1056	0.1098	0.1027
	2	4	225.83	0.9436	4-50-1	0.0753	0.0853	
3	1	4	421.50	0.9540	4-50-1	0.0579	0.0677	0.0715
	2	4	421.50	0.9433	4-100-100-1	0.0833	0.0807	
4	1	4	267.39	0.9460	4-50-1	0.0611	0.0849	0.0834
	2	4	267.39	0.9310	4-50-1	0.0826	0.0797	
5	1	4	605.36	0.8140	4-10-10-10-1	0.1100	0.1235	0.1124
	2	4	605.36	0.9100	4-50-1	0.0905	0.0853	
6	1	4	428.56	0.8600	4-10-10-10-1	0.0950	0.1054	0.0968
	2	4	428.56	0.8931	4-100-100-1	0.0643	0.0759	
7	1	4	200.31	0.9494	4-10-10-10-1	0.0771	0.0695	0.0828
	2	4	200.31	0.9317	4-100-100-1	0.0896	0.0882	
8	1	4	148.53	0.9515	4-10-10-10-1	0.0835	0.0710	0.0797
	2	4	148.53	0.9321	4-10-10-10-1	0.1080	0.1009	

5.5.1.3. Predicción

En esta sección se realiza una exposición sintética de los resultados obtenidos tras la aplicación de la fase de predicción al escenario PEMS I-405.

En primer lugar se muestran los resultados para los modelos de predicción con h=15′ y posteriormente los de los modelos con h=1 hora.

Predicción con h=15′

Tabla 5-12: PEMS I-405. Restricciones de ejecución de predicción con h=15'.

Máximo nº de parámetros	8
Tamaño de Ventana temporal	5 (1:15h horas)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-13: PEMS I-405. Resultados predicción detector 1, h=15'

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test cc
NI	DLP	1	0.971	8-100-1	0.0344	0.0375	0.0414
		2	0.94	8-10-10-10-1	0.0482	0.0511	
	RFP	1	0.971	XRT-8-32-1	0.0331	0.0395	0.0397
		2	0.94	DRF-8-37-1	0.0403	0.0402	
DLI	DLP	1	0.939	8-10-10-10-1	0.0456	0.0455	0.04613
		2	0.971	8-100-1	0.0379	0.0477	
	RFP	1	0.939	DRF-8-42-1	0.0416	0.0415	0.0412
		2	0.971	XRT-8-30-1	0.0327	0.0406	

Tabla 5-14: PEMS I-405. Resultados predicción detector 4, h=15'

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test cc
NI	DLP	1	0.968	8-100-1	0.0433	0.0416	0.0452
		2	0.957	8-10-10-10-1	0.0528	0.0539	
	RFP	1	0.968	DRF-8-35-1	0.0423	0.0378	0.0398
		2	0.957	XRT-8-39-1	0.0465	0.0449	
DLI	DLP	1	0.968	8-10-10-10-1	0.047	0.0464	0.0469
		2	0.957	8-10-10-10-1	0.0507	0.0481	
	RFP	1	0.968	XRT-8-34-1	0.0429	0.038	0.0390
		2	0.957	DRF-8-38-1	0.0493	0.0416	

Tabla 5-15: PEMS I-405. Resultados predicción detector 7, h=15'

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test cc
NI	DLP	1	0.972	8-100-100-1	0.0416	0.0405	0.0418
		2	0.964	8-10-10-10-1	0.0495	0.0452	
	RFP	1	0.972	XRT-8-34-1	0.0369	0.0381	0.0381
		2	0.964	XRT-8-42-1	0.0467	0.0384	
DLI	DLP	1	0.964	8-10-10-10-1	0.0508	0.0547	0.0514
		2	0.972	8-10-10-10-1	0.0376	0.0434	
	RFP	1	0.964	DRF-8-33-1	0.0475	0.0492	0.0456
		2	0.972	XRT-8-23-1	0.0382	0.0367	

Predicción con h=1h

Tabla 5-16: PEMS I-405. Restricciones de ejecución de predicción con h=1 hora.

Máximo nº de parámetros	4
Tamaño de Ventana temporal	3 (3 horas)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-17: PEMS I-405. Resultados predicción detector 1, h= 1 hora.

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test cc
NI	DLP	1	0.897	4-10-10-10-1	0.1056	0.1025	0.1073
		2	0.947	4-100-100-1	0.1039	0.1191	
	RFP	1	0.897	XRT-4-24-1	0.0678	0.0475	0.0582
		2	0.947	DRF-4-25-1	0.0607	0.0846	
DLI	DLP	1	0.897	4-10-10-10-1	0.1092	0.0936	0.0858
		2	0.947	4-10-10-10-1	0.07	0.0667	
	RFP	1	0.897	XRT-4-31-1	0.0646	0.0602	0.0587
		2	0.947	XRT-4-30-1	0.0644	0.0551	

Tabla 5-18: PEMS I-405. Resultados predicción detector 4, h= 1 hora.

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test cc
NI	DLP	1	0.946	4-50-1	0.0617	0.1024	0.0992
		2	0.931	4-100-100-1	0.0753	0.0917	
	RFP	1	0.946	DRF-4-39-1	0.065	0.0794	0.0774
		2	0.931	XRT-4-41-1	0.0612	0.0726	
DLI	DLP	1	0.946	4-50-1	0.0653	0.0804	0.0827
		2	0.931	4-100-100-1	0.0722	0.0883	
	RFP	1	0.946	XRT-4-30-1	0.0715	0.0676	0.0681
		2	0.931	DRF-4-31-1	0.0606	0.0693	

Tabla 5-19: PEMS I-405. Resultados predicción detector 7, h= 1 hora.

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSEtest cc
NI	DLP	1	0.932	4-100-100-1	0.106	0.0837	0.0859
		2	0.949	4-10-10-10-1	0.0803	0.0913	
	RFP	1	0.932	DRF-4-30-1	0.0676	0.0654	0.0762
		2	0.949	XRT-4-31-1	0.0763	0.1026	
DLI	DLP	1	0.949	4-10-10-10-1	0.0753	0.0632	0.0789
		2	0.932	4-10-10-10-1	0.0976	0.1173	
	RFP	1	0.949	DRF-4-28-1	0.073	0.059	0.0635
		2	0.932	XRT-4-24-1	0.0655	0.0746	

Resumen Predicción

En la Figura 5-9, se muestra un resumen de los resultados obtenidos en la fase de predicción para los modelos de los detectores 1, 4 y 7 del escenario PEMS I-405.

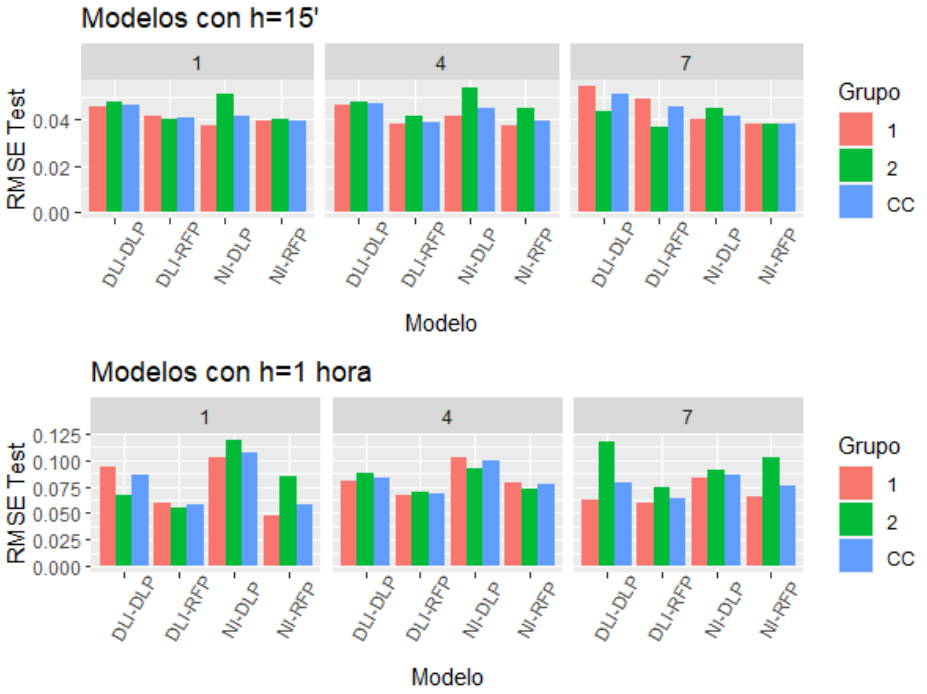


Figura 5-9: Resumen resultados predicción para escenario PEMS I-405

5.5.2. Dublín, M-50

Este caso de estudio se sitúa en el área metropolitana de Dublín, concretamente, en la autovía TMU M-50 (Figura 5-10) que cumple la función de circunvalación de la ciudad. En esta autovía se encuentran instalados 26 detectores de tráfico, gestionados por la *Traffic Infrastructure Ireland* (TII), la autoridad de carreteras de la República de Irlanda.

Se trata de un corredor de alta capacidad que comunica la capital con localidades de su área metropolitana. De igual modo es un nexo entre sudeste y el nordeste de Irlanda. El corredor contemplado consta de 40 km, aproximadamente. Los tramos que lo componen presentan una velocidad máxima de 100 km/h. En este corredor se encuentran instalados 26 detectores, 13 en cada sentido, sobre los que la TII publica series con una agregación temporal de 5 minutos (Tabla 5-20).

Se ha tomado un marco temporal que comprende dos meses completos, desde el 2 de septiembre de 2018, al 1 de noviembre del mismo año.

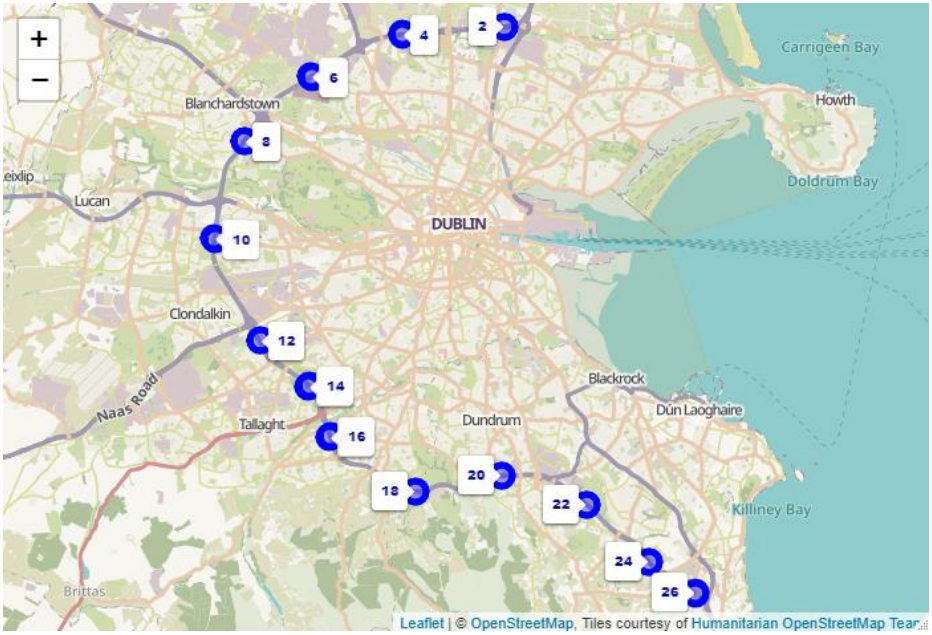


Figura 5-10: Representación del caso práctico, Dublín, M-50

Tabla 5-20: Características del caso práctico Dublín, M-50.

Elemento	Descripción
Caso de estudio	TMU M-50 de Dublín
Detectores	26 detectores situados sobre la M-50 entre los pk 1.7 y 40.
	13 detectores por cada sentido
	Agregación de flujo por sentido.
	Agregación temporal de 5 minutos
Red	Circunvalación de área metropolitana formada por una autovía de 3 carriles en cada sentido.
	Varias incorporaciones.
	Velocidad máxima 100 km/h. (mayor parte del recorrido)
Marco temporal	Fecha inicial = 02-09-2018
	Fecha final = 01-11-2018
Tasa de valores perdidos	0%

5.5.2.1. Análisis de información

En cuanto a las características de los perfiles promedios diarios (Figura 5-11) se observan dos picos de tráfico de manera generalizada en todos los detectores:

i) A primera hora de la mañana, en torno a las 06:00 h

ii) En torno a las 17:00. Con una meseta irregular entre ambas horas, cayendo a niveles mínimos solo entre las 21:00 y las 05:00 horas.

iii) Los detectores con menor volumen presentan un tipo de perfil en el que el pico de la mañana es muy leve, o ni siquiera se produce.

Se observan tres tipos principales de detectores según su IMD,

i) Los situados en arcos en los que el IMD supera los 70.000 veh./día, comprendiendo del 3 al 12.

ii) Aquellos situados en arcos en los que circulan entre 40.000 y 70.000 veh./día, que incluyen a 1 y al 2, por una parte y del 13 al 18, por otra.

iii) Los detectores del 19 al 26, que registran menos de 40.000 veh./día.

El coste promedio entre detectores oscila, en la mayoría de los casos, entre los 500 s y los 800 s debido a la escala física del caso de estudio. En la mayoría de los casos hay varios kms entre un detector y el siguiente del mismo sentido de circulación. Aquellos que se localizan en los extremos norte y sur superan los 800 s de coste promedio; los detectores situados en los extremos (1, 2 ,25 y 26) presentan unos valores especialmente altos.

Los detectores mejor conectados del caso de estudio son aquellos situados en la zona central (del 10 al 16) presentando valores próximos a los 600s, tanto en el rol de origen como en el de destino.

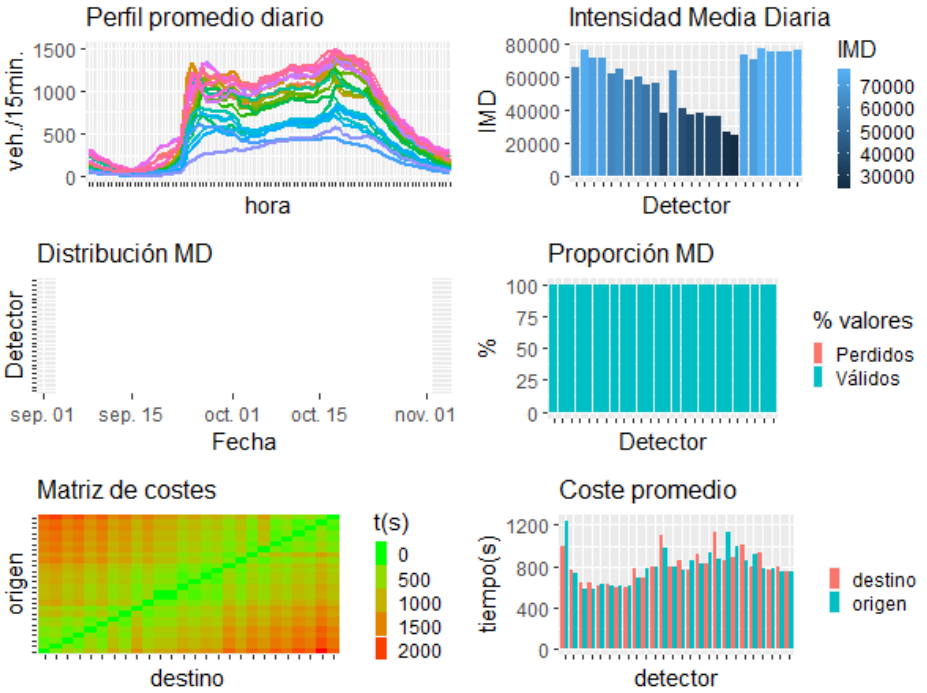


Figura 5-11: Características del caso práctico Dublín, M-50.

A continuación se presentan los resultados obtenidos mediante la aplicación del marco de predicción en las distintas situaciones planteadas.

5.5.2.2. Imputación

En esta sección se realiza una exposición sintética de los resultados obtenidos tras la aplciación de la fase de imputación del marco de predicción al escenario Dublín M-50.

En primer lugar se muestran los resultados de imputación con el modelo DLI, con agregación del conjunto de datos de 15 minutos, por una parte y con agregación de 1 hora, por otra.

Al presentar una tasa de valores perdidos del 0%, no se muestra la tabla con información sobre valores perdidos.

Imputación DLI con h=15'

Tabla 5-21: Restricciones de ejecución de imputación DLI con h=15'.

Máximo nº de parámetros	10
Tamaño de Ventana temporal	5 (1 hora y 15 minutos)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-22: Dublín-M50. Resultados de imputación DLI para h=15'

Det.	Grupo	Parám. Entrada	Coste máximo	ϱ mínimo	Configuración Modelo	RMSE cv	RMSE test	RMSE Test cc
1	1	10	1030.52	0.978	10-10-10-10-1	0.0261	0.0257	0.0300
	2	10	1030.52	0.959	10-10-10-10-1	0.0452	0.0405	
2	1	10	1142.62	0.944	10-10-10-10-1	0.0557	0.0566	0.0495
	2	10	1142.62	0.976	10-10-10-10-1	0.0304	0.032	
3	1	10	913.74	0.968	10-10-10-10-1	0.0298	0.0333	0.0337
	2	10	913.74	0.958	10-10-10-10-1	0.039	0.0347	
4	1	10	1049.52	0.946	10-10-10-10-1	0.0574	0.0557	0.0485

Det.	Grupo	Parám. Entrada	Coste máximo	ϱ mínimo	Configuración Modelo	RMSE cv	RMSE test	RMSE Test cc
	2	10	1049.52	0.984	10-10-10-10-1	0.0309	0.0309	
5	1	10	819.98	0.975	10-10-10-10-1	0.0372	0.0362	0.0372
	2	10	819.98	0.962	10-10-10-10-1	0.0428	0.0395	
6	1	10	974.34	0.948	10-10-10-10-1	0.0526	0.054	0.0477
	2	10	974.34	0.985	10-10-10-10-1	0.0294	0.0322	
7	1	10	791.32	0.985	10-10-10-10-1	0.0305	0.0325	0.0404
	2	10	791.32	0.955	10-10-10-10-1	0.0571	0.0598	
8	1	10	825.07	0.984	10-10-10-10-1	0.0281	0.0295	0.0378
	2	10	825.07	0.955	10-10-10-10-1	0.0556	0.0581	
9	1	10	772.64	0.98	10-10-10-10-1	0.0366	0.0357	0.0380
	2	10	772.64	0.962	10-10-10-10-1	0.0441	0.0437	
10	1	10	794.33	0.987	10-10-10-10-1	0.0292	0.0276	0.0351
	2	10	794.33	0.947	10-10-10-10-1	0.0513	0.0534	
11	1	10	695.79	0.947	10-10-10-10-1	0.0544	0.0544	0.0478
	2	10	695.79	0.987	10-10-10-10-1	0.0285	0.0315	
12	1	10	695.79	0.987	10-10-10-10-1	0.0305	0.0277	0.0357
	2	10	695.79	0.947	10-10-10-10-1	0.0538	0.0554	
13	1	10	631.34	0.978	10-10-10-10-1	0.0324	0.0343	0.0415
	2	10	631.34	0.94	10-10-10-10-1	0.0571	0.0591	
14	1	10	643.72	0.942	10-10-10-10-1	0.0434	0.0427	0.0388
	2	10	643.72	0.987	10-10-10-10-1	0.0309	0.0294	
15	1	10	622.64	0.981	10-100-100-1	0.031	0.0322	0.0359
	2	10	622.64	0.95	10-10-10-10-1	0.044	0.0451	
16	1	10	619.76	0.95	10-10-10-10-1	0.0422	0.0464	0.0421
	2	10	619.76	0.985	10-10-10-10-1	0.0309	0.0314	
17	1	10	803.99	0.984	10-10-10-10-1	0.0288	0.0311	0.0354

Det.	Grupo	Parám. Entrada	Coste máximo	ϱ mínimo	Configuración Modelo	RMSE cv	RMSE test	RMSE Test cc
	2	10	803.99	0.951	10-10-10-10-1	0.044	0.0459	
18	1	10	707.43	0.953	10-10-10-10-1	0.0456	0.0475	0.0433
	2	10	707.43	0.979	10-10-10-10-1	0.0317	0.0329	
19	1	10	819.32	0.979	10-10-10-10-1	0.0295	0.0323	0.0360
	2	10	819.32	0.936	10-10-10-10-1	0.0455	0.0452	
20	1	10	823.62	0.943	10-10-10-10-1	0.0464	0.0459	0.0414
	2	10	823.62	0.979	10-10-10-10-1	0.0316	0.0305	
21	1	10	881.06	0.977	10-10-10-10-1	0.0347	0.0324	0.0339
	2	10	881.06	0.941	10-10-10-10-1	0.0417	0.0374	
22	1	10	785.38	0.982	10-10-10-10-1	0.0372	0.0379	0.0400
	2	10	785.38	0.96	10-10-10-10-1	0.0424	0.0453	
23	1	10	960.42	0.972	10-10-10-10-1	0.0393	0.0434	0.0437
	2	10	960.42	0.938	10-10-10-10-1	0.0399	0.0445	
24	1	10	850.65	0.955	10-10-10-10-1	0.0368	0.0378	0.0374
	2	10	850.65	0.982	10-10-10-10-1	0.0316	0.0363	
25	1	10	1171.44	0.874	10-10-10-10-1	0.0415	0.0435	0.0404
	2	10	1171.44	0.968	10-10-10-10-1	0.032	0.0328	
26	1	10	894.02	0.917	10-10-10-10-1	0.0466	0.0466	0.0460
	2	10	894.02	0.963	10-10-10-10-1	0.0419	0.0447	

Imputación DLI con h=1 hora

Tabla 5-23: Restricciones de ejecución de imputación DLI con h=1hora.

Máximo nº de parámetros	4
Tamaño de Ventana temporal	2 (2 horas)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-24: Dublín-M50. Resultados de imputación DLI para h=1 hora

Det.	Grupo	Parám. Entrada	Coste máximo	ϱ mínimo	Configuración Modelo	RMSE cv	RMSE test	RMSE Test cc
1	1	4	1030.52	0.914	4-100-1	0.0879	0.0949	0.0836
	2	4	1030.52	0.955	4-10-10-10-1	0.0562	0.0561	
2	1	4	1142.62	0.866	4-100-100-1	0.0989	0.0983	0.0885
	2	4	1142.62	0.943	4-10-10-10-1	0.0656	0.0646	
3	1	4	913.74	0.937	4-100-100-1	0.0776	0.0889	0.0852
	2	4	913.74	0.9	4-100-1	0.0824	0.0762	
4	1	4	1049.52	0.904	4-100-100-1	0.1177	0.1154	0.0994
	2	4	1049.52	0.968	4-10-10-10-1	0.0533	0.0603	
5	1	4	819.98	0.91	4-100-100-1	0.0912	0.0746	0.0760
	2	4	819.98	0.946	4-100-100-1	0.0796	0.0794	
6	1	4	974.34	0.904	4-100-100-1	0.1061	0.1057	0.0922
	2	4	974.34	0.967	4-10-10-10-1	0.0593	0.0592	
7	1	4	791.32	0.967	4-10-10-10-1	0.0581	0.0587	0.0719

Det.	Grupo	Parám. Entrada	Coste máximo	ρ mínimo	Configuración Modelo	RMSE cv	RMSE test	RMSE Test cc
	2	4	791.32	0.899	4-100-100-1	0.1067	0.1043	
8	1	4	825.07	0.967	4-10-10-10-1	0.0553	0.0505	0.0659
	2	4	825.07	0.899	4-100-100-1	0.1086	0.1035	
9	1	4	772.64	0.956	4-10-10-10-1	0.0691	0.0615	0.0707
	2	4	772.64	0.925	4-100-100-1	0.0957	0.0932	
10	1	4	794.33	0.886	4-100-100-1	0.1134	0.1234	0.1066
	2	4	794.33	0.966	4-10-10-10-1	0.0556	0.0653	
11	1	4	695.79	0.968	4-10-10-10-1	0.0612	0.0626	0.0792
	2	4	695.79	0.876	4-100-100-1	0.1248	0.1198	
12	1	4	695.79	0.968	4-10-10-10-1	0.0589	0.0583	0.0748
	2	4	695.79	0.876	4-100-100-1	0.1226	0.1153	
13	1	4	631.34	0.948	4-100-1	0.0775	0.0857	0.0949
	2	4	631.34	0.895	4-10-10-10-1	0.1145	0.1174	
14	1	4	643.72	0.97	4-10-10-10-1	0.0716	0.0758	0.0916
	2	4	643.72	0.89	4-100-1	0.1296	0.1304	
15	1	4	622.64	0.901	4-100-1	0.1152	0.1149	0.1021
	2	4	622.64	0.955	4-10-10-10-1	0.0694	0.0709	
16	1	4	619.76	0.964	4-10-10-10-1	0.0839	0.0812	0.0954
	2	4	619.76	0.894	4-10-10-10-1	0.1307	0.1303	
17	1	4	803.99	0.966	4-10-10-10-1	0.0573	0.0628	0.0756
	2	4	803.99	0.899	4-100-100-1	0.099	0.1071	
18	1	4	707.43	0.884	4-100-100-1	0.1261	0.1291	0.1113

Det.	Grupo	Parám. Entrada	Coste máximo	ϱ mínimo	Configuración Modelo	RMSE cv	RMSE test	RMSE Test cc
	2	4	707.43	0.951	4-10-10-10-1	0.0706	0.0677	
19	1	4	819.32	0.954	4-10-10-10-1	0.0729	0.0533	0.0701
	2	4	819.32	0.863	4-100-100-1	0.1113	0.1114	
20	1	4	823.62	0.954	4-10-10-10-1	0.0631	0.0639	0.0857
	2	4	823.62	0.88	4-100-100-100-	0.14	0.1392	
21	1	4	881.06	0.855	4-100-1	0.1077	0.0936	0.0876
	2	4	881.06	0.943	4-10-10-10-1	0.073	0.0729	
22	1	4	785.38	0.93	4-100-1	0.0845	0.0819	0.0843
	2	4	785.38	0.956	4-100-100-1	0.0832	0.0903	
23	1	4	960.42	0.933	4-10-10-10-1	0.0833	0.0796	0.0924
	2	4	960.42	0.838	4-10-10-10-1	0.1075	0.1237	
24	1	4	850.65	0.934	4-100-1	0.0776	0.0712	0.0701
	2	4	850.65	0.964	4-10-10-10-1	0.0584	0.0674	
25	1	4	1171.44	0.933	4-10-10-10-1	0.0626	0.0578	0.0713
	2	4	1171.44	0.757	4-10-10-10-1	0.112	0.1044	
26	1	4	894.02	0.952	4-10-10-10-1	0.0838	0.0738	0.0793
	2	4	894.02	0.908	4-100-1	0.0899	0.0928	

5.5.2.3. Predicción

En esta sección se muestran los resultados de predicción para el escenario Dublín-M50. Se omiten los resultados referentes a los modelos que utilizan DLI en la fase de imputación, porque al ser 0% la tasa de valores perdidos del escenario, no cambia el conjunto de datos respecto a NI.

Predicción con h=15'

Tabla 5-25: Dublín-M50. Restricciones de ejecución de predicción con h=15'.

Máximo nº de parámetros	8
Tamaño de Ventana temporal	5 (1:15h horas)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-26: Dublín-M50. Resultados predicción detector 1, h=15'

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test
NI	DLP	1	0.971	8-100-1	0.0344	0.0375	0.0414
		2	0.94	8-10-10-10-1	0.0482	0.0511	
	RFP	1	0.971	DRF-8-36-1	0.0264	0.0278	0.0397
		2	0.94	XRT-8-37-1	0.0402	0.0404	

Tabla 5-27: Dublín-M50. Resultados predicción detector 14, h=15'

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test cc
NI	DLP	1	0.961	8-10-10-10-1	0.0564	0.0539	0.0469
		2	0.988	8-10-10-10-1	0.0308	0.0297	
	RFP	1	0.961	XRT-8-43-1	0.0403	0.0373	0.0350
		2	0.988	DRF-8-35-1	0.0296	0.0294	

Tabla 5-28: Dublín-M50. Resultados predicción detector 25, h=15'

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test cc
NI	DLP	1	0.971	8-10-10-10-1	0.0319	0.0307	0.0326
		2	0.889	8-10-10-10-1	0.0387	0.0373	
	RFP	1	0.971	DRF-8-38-1	0.0332	0.0288	0.0299
		2	0.889	XRT-8-42-1	0.0354	0.0327	

Predicción con h=1h

Tabla 5-29: Dublín M-50. Restricciones de ejecución de predicción con h=1 hora.

Máximo nº de parámetros	4
Tamaño de Ventana temporal	3 (3 horas)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-30: Dublín-M50. Resultados predicción detector 1, h= 1 hora.

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test cc
NI	DLP	1	0.897	4-10-10-10-1	0.1056	0.1025	0.1073
		2	0.947	4-100-100-1	0.1039	0.1191	
	RFP	1	0.897	XRT-4-24-1	0.0678	0.0475	0.0582
		2	0.947	DRF-4-25-1	0.0607	0.0846	

Tabla 5-31: Dublín-M50. Resultados predicción detector 14, h= 1 hora.

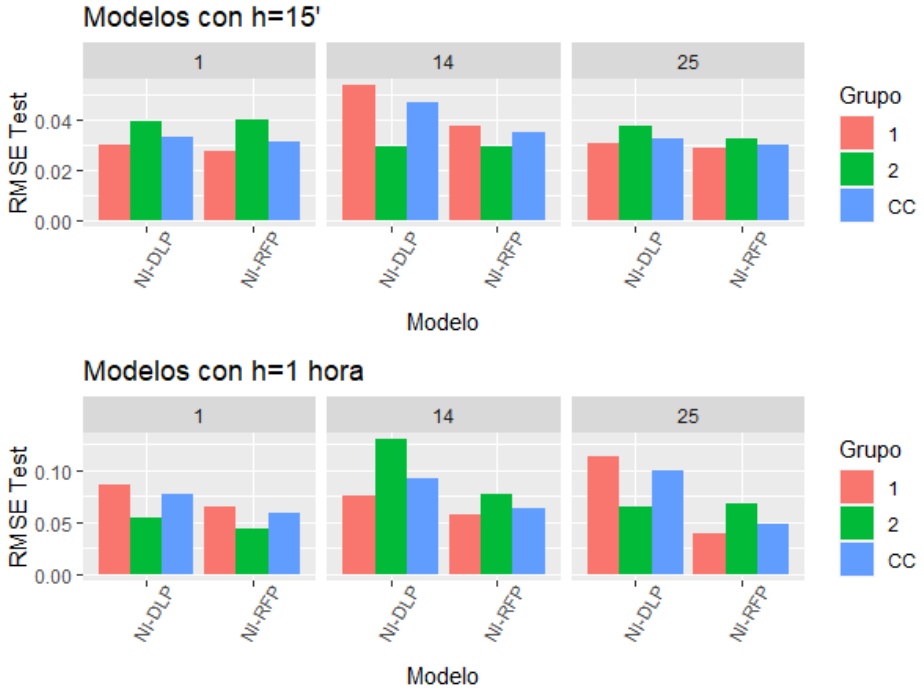
Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test cc
NI	DLP	1	0.97	4-100-100-1	0.0638	0.0756	0.0913
		2	0.89	4-100-100-1	0.1275	0.1298	
	RFP	1	0.97	XRT-4-40-1	0.0537	0.0570	0.0628
		2	0.89	DRF-4-32-1	0.0727	0.0771	

Tabla 5-32: Dublín-M50. Resultados predicción detector 25, h= 1 hora.

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSEtest cc
NI	DLP	1	0.757	4-10-10-10-1	0.1136	0.1135	0.0995
		2	0.933	4-10-10-10-1	0.0596	0.0651	
	RFP	1	0.757	XRT-4-31-1	0.0434	0.0396	0.0476
		2	0.933	XRT-4-26-1	0.0657	0.0673	

Resumen Predicción

Tabla 5-33: Resumen resultados predicción para escenario Dublín-M50



5.5.3. Madrid, M-30 Norte

Este caso de estudio se basa en los datos proporcionados por el portal de datos abiertos del Ayuntamiento de Madrid, que publica la información recopilada por los detectores situados en las vías principales de la ciudad como la M-30. Los datos presentan un nivel de agregación de 15 minutos, suministrando información sobre diferentes parámetros de tráfico a nivel de sentido (Tabla 5-34).

El caso de estudio se compone de la información aportada por 10 detectores localizados en un corredor de 3 km situado al norte de la ciudad (Figura 5-12). En el sentido Este-Oeste se localizan 4 detectores; mientras que en el sentido O-E son 6. El marco temporal es más extenso que en el resto de los casos de estudio, contemplándose datos referentes a 6 meses, desde el 01-09-20016 al 28-02-2017.

Figura 5-12: Representación del caso práctico, Madrid M-30.

En la Tabla 5-34 se muestran las características básicas de este escenario, presentándose la información sobre los detectores que lo componen, la red de carreteras en la que se localizan, el marco temporal escogido y la tasa general de valores perdidos.

Tabla 5-34: Características del caso práctico Madrid-M30.

Elemento	Descripción
Caso de estudio	Madrid, M-30 Norte.
Detectores	10 detectores situados sobre la M-30 entre los pk 26 y 28, 5 por sentido. Agregación por sentido. Intervalo mínimo de registro de datos 15 minutos.
Red	Circunvalación de área metropolitana formada por una autovía de 3 carriles. Varias incorporaciones. Velocidad máxima 90 km/h. (mayor parte del recorrido)
Marco temporal	Fecha inicial = 01-09-2016 Fecha final = 28-02-2017
Tasa de valores perdidos	0.5%

5.5.3.1. Análisis de información

El comportamiento del flujo de tráfico en los arcos contemplados en este caso de estudio se refleja en la Figura 5-13, en la que se observan tres comportamientos bien diferenciados:

i) El mayor IMD del caso de estudio es de unos 150.000 veh/día.

ii) 8 detectores en los que se da un IMD de entre 90.000 veh/día y 120.000 veh/día. mostrando además un perfil muy similar.

iii) Un detector presenta un valor de IMD próximo a los 40.000 veh/día, mucho menor que el del resto, debido a su localización en una vía de incorporación a la principal.

El coste agregado observado arroja tres comportamientos característicos;

i) Un alto valor agregado de coste como origen, en el que se situarían los detectores al final del tramo en el que se ubican.

ii) Un alto coste agregado como destino, que se corresponde con los datos observados para los detectores que se encuentran al inicio de un tramo

iii) Costes agregados similares de origen y destino, con valores cercanos a los 200 s, en el que se colocarían en resto de los detectores.

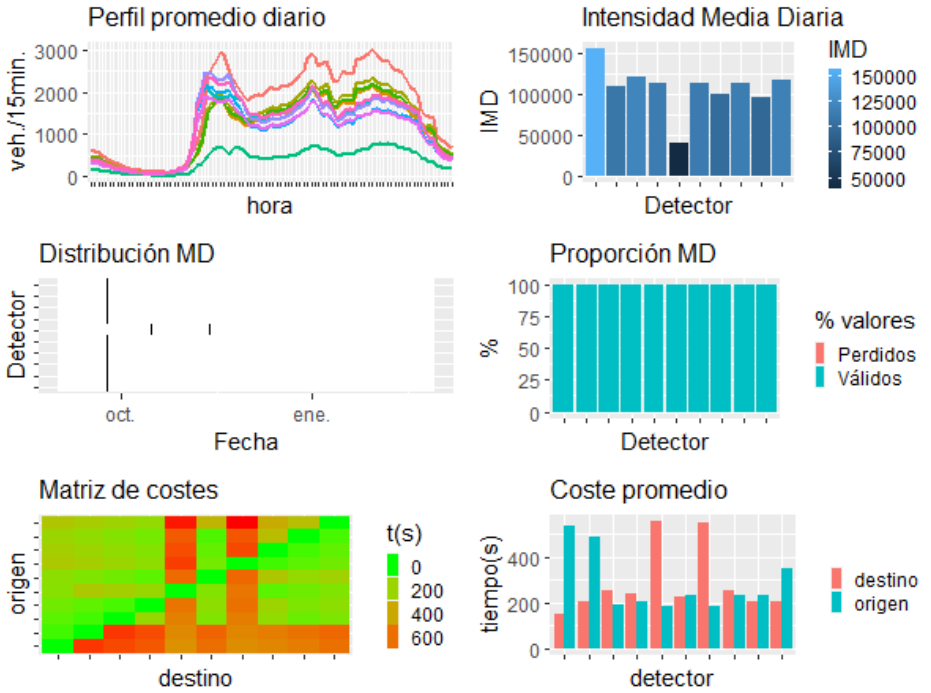


Figura 5-13: Características del caso práctico Madrid-M30.

A continuación se presentan los resultados obtenidos mediante la aplicación del marco de predicción en las distintas situaciones planteadas.

5.5.3.2. Imputación

En esta sección se realiza una exposición sintética de los resultados obtenidos tras la aplciación de la fase de imputación del marco de predicción al Madrid-M30.

En primer lugar se muestran los resultados de imputación con el modelo DLI, con agregación del conjunto de datos de 15 minutos, por una parte y con agregación de 1 hora, por otra.

Imputación DLI para h=15′

Tabla 5-35: Parámetros relativos a los valores perdidos de cada detector.

Detector	Intervalos MD	Tasa MD	Periodos MD	Longitud periodo	Distancia media entre periodos
1	81	0.47%	21	3.86	349.35
2	81	0.47%	21	3.86	349.35
3	84	0.48%	22	3.82	332.71
4	84	0.48%	22	3.82	332.71
5	43	0.25%	11	3.91	698.70
6	148	0.85%	64	2.31	190.02
7	43	0.25%	11	3.91	698.70
8	43	0.25%	11	3.91	698.70
9	43	0.25%	11	3.91	698.70
10	51	0.29%	13	3.92	582.25

Tabla 5-36: Restricciones de ejecución de imputación DLI con h=15'.

Máximo nº de parámetros	10
Tamaño de Ventana temporal	5 (1 hora y 15 minutos)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-37: Madrid-M30. Resultados de imputación DLI para h=15'

Det.	Grupo	Parám. Entrada	Coste máximo	ϱ mínimo	Configuración Modelo	RMSE cv	RMSE test	RMSE Test cc
1	1	5	162.08	0.641	5-10-10-10-1	0.0497	0.0485	0.0547
	2	9	162.08	0.685	9-10-10-10-1	0.0673	0.07	
2	1	5	229.2	0.764	5-10-10-10-1	0.0823	0.0771	0.0715
	2	10	229.2	0.615	10-10-10-10-1	0.0501	0.0579	
3	1	4	279.23	0.784	4-10-10-10-1	0.0874	0.0843	0.0739
	2	10	279.23	0.794	10-10-10-10-1	0.0493	0.0485	
4	1	4	264.84	0.774	4-10-10-10-1	0.0857	0.0837	0.0737
	2	10	264.84	0.787	10-10-10-10-1	0.0471	0.0493	
5	1	5	620.34	0.811	5-10-10-10-1	0.0742	0.0706	0.0613
	2	10	620.34	0.912	10-10-10-10-1	0.0386	0.0387	
6	1	10	247.29	0.653	10-10-10-10-1	0.0271	0.0218	0.0208
	2	10	247.29	0.817	10-10-10-10-1	0.0183	0.0183	
7	1	5	613.64	0.826	5-10-10-10-1	0.0281	0.0272	0.0362
	2	10	613.64	0.755	10-10-10-10-1	0.0566	0.0584	
8	1	5	280.01	0.822	5-10-10-10-1	0.0261	0.0294	0.0392
	2	10	280.01	0.761	10-10-10-10-1	0.0612	0.0632	
9	1	5	224.57	0.698	5-10-10-10-1	0.0592	0.0615	0.0547
	2	10	224.57	0.858	10-10-10-10-1	0.0378	0.0382	
10	1	5	227.03	0.808	5-10-10-10-1	0.0291	0.0284	0.0379
	2	10	227.03	0.79	10-10-10-10-1	0.0627	0.0612	

Imputación DLI con h=1 hora

Tabla 5-38: Parámetros relativos a los valores perdidos de cada detector con h=1h.

Detector	Intervalos MD	Tasa MD	Periodos MD	Longitud periodo	Distancia media entre periodos
1	13	0.96%	7	1.86	291.50
2	13	0.96%	7	1.86	291.50
3	14	1.03%	7	2.00	291.50
4	14	1.03%	7	2.00	291.50
5	8	0.59%	2	4.00	527.00
6	11	0.81%	5	2.20	169.00
7	8	0.59%	2	4.00	527.00
8	8	0.59%	2	4.00	527.00
9	8	0.59%	2	4.00	527.00
10	9	0.66%	3	3.00	347.00

Tabla 5-39: Restricciones de ejecución de imputación DLI con h=15'.

Máximo nº de parámetros	4
Tamaño de Ventana temporal	2 (2 horas)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-40: Madrid-M30. Resultados de imputación DLI para h=1 hora'.

Det.	Grupo	Parám. Entrada	Coste Umbral (s)	ϱ Umbral	Configuración Modelo	RMSE cv	RMSE test	RMSE Test cc
1	**1**	**4**	162.08	0.707	4-10-10-10-1	0.0473	0.0511	0.0594
	2	**4**	162.08	0.607	4-100-1	0.0797	0.0798	
2	**1**	**4**	229.2	0.522	4-100-1	0.0431	0.0395	0.0508
	2	**4**	229.2	0.647	4-100-1	0.0795	0.0786	
3	**1**	**4**	279.23	0.513	4-10-10-10-1	0.0555	0.0673	0.0709
	2	**4**	279.23	0.671	4-10-10-10-1	0.083	0.0797	
4	**1**	**4**	264.84	0.51	4-10-10-10-1	0.053	0.0576	0.0670
	2	**4**	264.84	0.669	4-10-10-10-1	0.076	0.0899	
5	**1**	**4**	620.34	0.603	4-10-10-10-1	0.0787	0.0898	0.0820
	2	**4**	620.34	0.903	4-10-10-10-1	0.0579	0.0629	
6	**1**	**4**	247.29	0.434	4-100-100-1	0.0854	0.0811	0.0693
	2	**4**	247.29	0.773	4-10-10-10-1	0.0407	0.0404	
7	**1**	**4**	613.64	0.76	4-100-100-1	0.0605	0.0582	0.0681
	2	**4**	613.64	0.33	4-100-100-1	0.1264	0.0922	
8	**1**	**4**	280.01	0.333	4-10-10-10-1	0.0712	0.0692	0.0651
	2	**4**	280.01	0.883	4-100-1	0.0601	0.0549	
9	**1**	**4**	224.57	0.38	4-10-10-10-1	0.0717	0.0608	0.0611
	2	**4**	224.57	0.886	4-10-10-10-1	0.0655	0.0617	
10	**1**	**4**	227.03	0.682	4-10-10-10-1	0.0507	0.0466	0.0653
	2	**4**	227.03	0.386	4-100-100-1	0.121	0.111	

5.5.3.3. Predicción

En esta sección se realiza una exposición sintética de los resultados obtenidos tras la aplicación de la fase de predicción al escenario Madrid-M30.

En primer lugar se muestran los resultados para los modelos de predicción con h=15' y posteriormente los de los modelos con h=1 hora.

Predicción con h=15'

Tabla 5-41: Madrid-M30. Restricciones de ejecución de predicción con h=15'.

Máximo nº de parámetros	8
Tamaño de Ventana temporal	5 (1:15h horas)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-42: Madrid-M30. Resultados predicción detector 1, h=15'

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test
NI	DLP	1	0.784	5-10-10-10-1	0.0853	0.0873	0.0784
		2	0.693	8-100-1	0.0518	0.0566	
	RFP	1	0.784	XRT-5-49-1	0.0801	0.0799	0.0728
		2	0.693	XRT-8-38-1	0.0514	0.0554	
DLI	DLP	1	0.901	8-10-10-10-1	0.0404	0.0378	0.0402
		2	0.927	8-10-10-10-1	0.0453	0.0461	
	RFP	1	0.901	8-10-10-10-1	0.0404	0.0378	0.0393
		2	0.927	8-10-10-10-1	0.0453	0.0461	

Tabla 5-43: Madrid-M30. Resultados predicción detector 5, h=15'

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test
NI	DLP	1	0.839	5-100-1	0.0473	0.0431	0.0512
		2	0.865	8-10-10-10-1	0.069	0.0711	
	RFP	1	0.839	DRF-5-40-1	0.0466	0.0399	0.0476
		2	0.865	XRT-8-43-1	0.0686	0.0663	
DLI	DLP	1	0.917	8-10-10-10-1	0.0536	0.0622	0.0572
		2	0.933	8-10-10-10-1	0.0392	0.0451	
	RFP	1	0.917	XRT-8-36-1	0.0534	0.0567	0.0529
		2	0.933	XRT-8-36-1	0.0417	0.0437	

Tabla 5-44: Madrid-M30. Resultados predicción detector 10, h=15'

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test
NI	DLP	1	0.808	5-10-10-10-1	0.0289	0.0285	0.0370
		2	0.793	8-10-10-10-1	0.0611	0.0579	
	RFP	1	0.808	DRF-5-38-1	0.0319	0.0288	0.0362
		2	0.793	DRF-8-45-1	0.0563	0.0542	
DLI	DLP	1	0.935	8-10-10-10-1	0.0331	0.0345	0.0394
		2	0.922	8-10-10-10-1	0.0482	0.0513	
	RFP	1	0.935	XRT-8-32-1	0.0349	0.0339	0.0374
		2	0.922	XRT-8-43-1	0.0443	0.046	

Predicción con h=1h

Tabla 5-45: PEMS I-405. Restricciones de ejecución de predicción con h=1 hora.

Máximo nº de parámetros	4
Tamaño de Ventana temporal	3 (3 horas)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-46: Madrid-M30. Resultados predicción detector 1, h= 1 hora.

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test
NI	DLP	1	0.495	4-10-10-10-1	0.0597	0.0675	0.0692
		2	0.596	4-10-10-10-1	0.0734	0.0735	
	RFP	1	0.495	DRF-4-44-1	0.0632	0.0551	0.0546
		2	0.596	DRF-4-39-1	0.0622	0.0535	
DLI	DLP	1	0.855	4-10-10-10-1	0.0697	0.0653	0.0747
		2	0.852	4-10-10-10-1	0.0934	0.0978	
	RFP	1	0.855	DRF-4-30-1	0.0699	0.0612	0.0668
		2	0.852	DRF-4-33-1	0.0844	0.0804	

Tabla 5-47: Madrid-M30. Resultados predicción detector 5, h= 1 hora.

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test cc
NI	DLP	1	0.342	4-10-10-10-1	0.085	0.0818	0.0721
		2	0.903	4-10-10-10-1	0.0642	0.0671	
	RFP	1	0.342	XRT-4-39-1	0.0654	0.061	0.0556
		2	0.903	DRF-4-31-1	0.068	0.0624	
DLI	DLP	1	0.897	4-10-10-10-1	0.0612	0.0539	0.1296
		2	0.885	4-10-10-10-1	0.1091	0.1145	
	RFP	1	0.897	XRT-4-27-1	0.0677	0.0565	0.0896
		2	0.885	XRT-4-33-1	0.0946	0.0888	

Tabla 5-48: Madrid-M30. Resultados predicción detector 10, h= 1 hora.

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSEtest cc
NI	DLP	1	0.136	4-100-100-1	0.0845	0.0809	0.0775
		2	0.742	4-10-10-10-1	0.0599	0.0507	
	RFP	1	0.136	DRF-4-38-1	0.0553	0.0568	0.0614
		2	0.742	XRT-4-36-1	0.059	0.0528	
DLI	DLP	1	0.809	4-50-1	0.1367	0.152	0.0715
		2	0.881	4-10-10-10-1	0.0588	0.0746	
	RFP	1	0.809	DRF-4-33-1	0.0893	0.096	0.0659
		2	0.881	XRT-4-24-1	0.0608	0.0741	

Resumen Predicción

En la Figura 5-9, se muestra un resumen de los resultados obtenidos en la fase de predicción para los modelos de los detectores 1, 5 y 10 del escenario Madrid-M30.

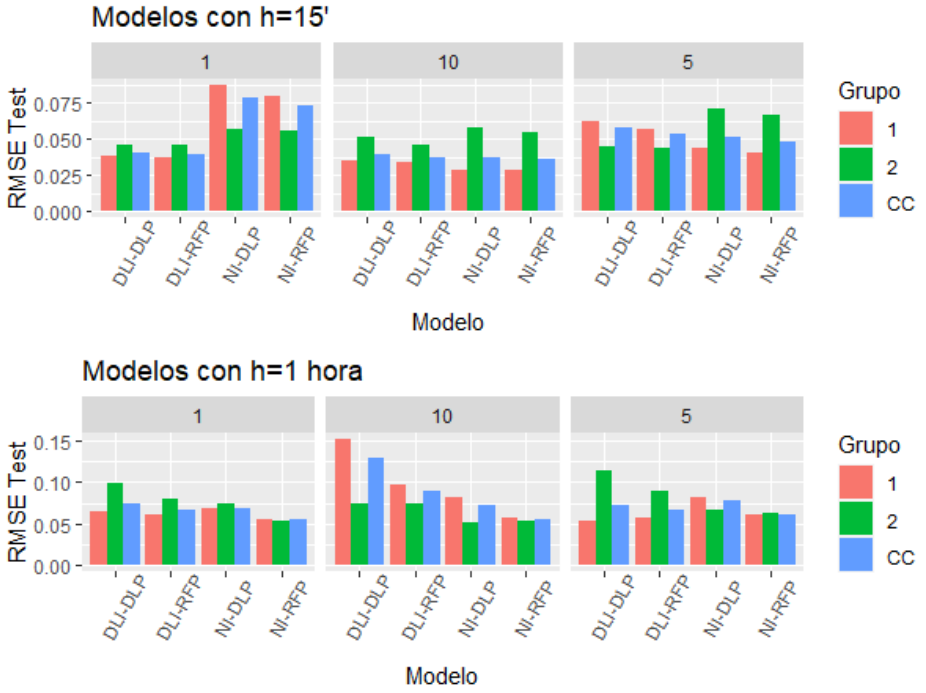


Figura 5-14: Resumen resultados predicción para escenario Madrid-M30

5.5.4. Sevilla-Se30

Este caso práctico se basa en los datos proporcionados por la DGT sobre la carretera SE-30 (Figura 5-15), que funciona como ronda de circunvalación, comunicando la ciudad con su área metropolitana. El trazado presenta un total de 28 km de autovía que rodean completamente la ciudad. En los arcos en los que se localizan los detectores el límite máximo de velocidad oscila entre los entre los 60 y 80 km/h.

Este caso de estudio se compone de los datos recogidos por 27 detectores, distribuidos a lo largo de la circunvalación, 14 en sentido horario, y 13 en el contrario, que toman datos por sentido cada 15 minutos (Tabla 5-49). Se ha decidido utilizar una agregación de los datos de 15 minutos debido a la escala de la zona...

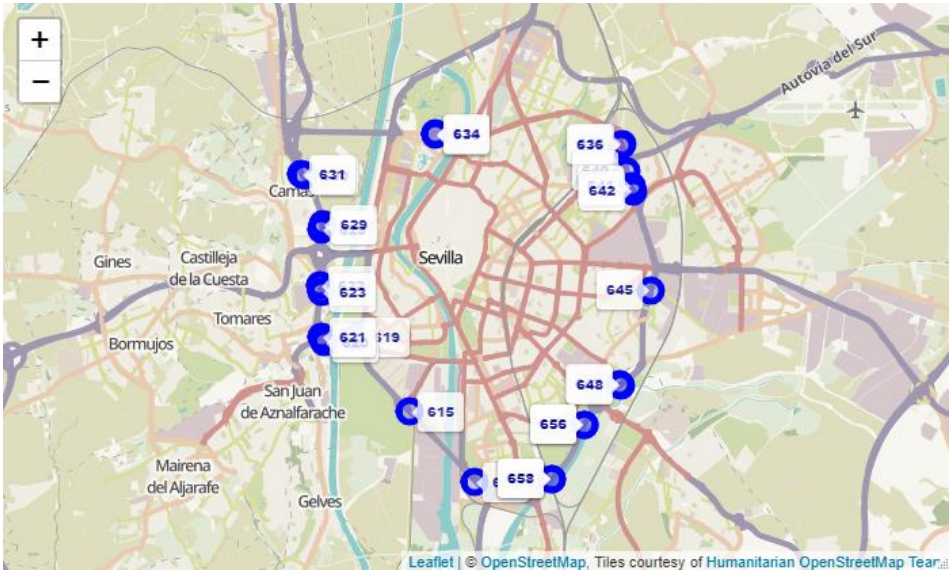


Figura 5-15: Representación del caso práctico, Sevlla SE-30.

En la Tabla 5-49 se muestran las características básicas de este escenario, presentándose la información sobre los detectores que lo componen, la red de carreteras en la que se localizan, el marco temporal escogido y la tasa general de valores perdidos.

Tabla 5-49: Características del caso práctico Sevilla-Se-30.

Elemento	Descripción
Caso de estudio	Sevilla, circunvalación SE-30
Detectores	25 detectores situados sobre toda la Se-30, 13 en sentido ascendente y 12 en el contrario. Agregación por sentido. Intervalo mínimo de registro de datos 15 minutos.
Red	Circunvalación de la ciudad de Sevilla que entre la ciudad y su área metropolitana. Varias incorporaciones. Vía principal con velocidades comprendidas entre los 60 km/h y los 80 km/h. (mayor parte del recorrido), con vías que tienen 3 ó 4 carriles.
Marco temporal	Fecha inicial = 15-01-2022 Fecha final = 15-03-2022
Tasa de valores perdidos	7%

5.5.4.1. Análisis de información

El comportamiento del flujo de tráfico en los arcos contemplados en este caso de estudio se refleja en la Figura 5-16, en la que se observan tres comportamientos bien diferenciados:

i) El mayor IMD del caso de estudio es de unos 50.000 veh/día vehículos localizados en la conexión con el sur del Aljarafe

ii) os detectores que registran entre 35.000 y 50.000 vehículos diarios, que se encuentran junto a incorporaciones de poblaciones más pequeñas.

iii) Los detectores con un IMD inferior a los 35.000 vehículos situados en la zona norte.

El coste agregado observado arroja tres comportamientos característicos;

i) Los detectores situados en la zona suroeste que se caracterizan por presentar un coste agregado como destino bastante superior al que presentan como origen.

ii) Aquellos en los que ambos costes son bastante parejos, y comprenden valores entre los 550 s y los 700 s en los que se incluyen la mayoría de los detectores del caso de estudio.

iii) Aquéllos que presentan un coste como origen notablemente mayor que el coste como destino detectores.

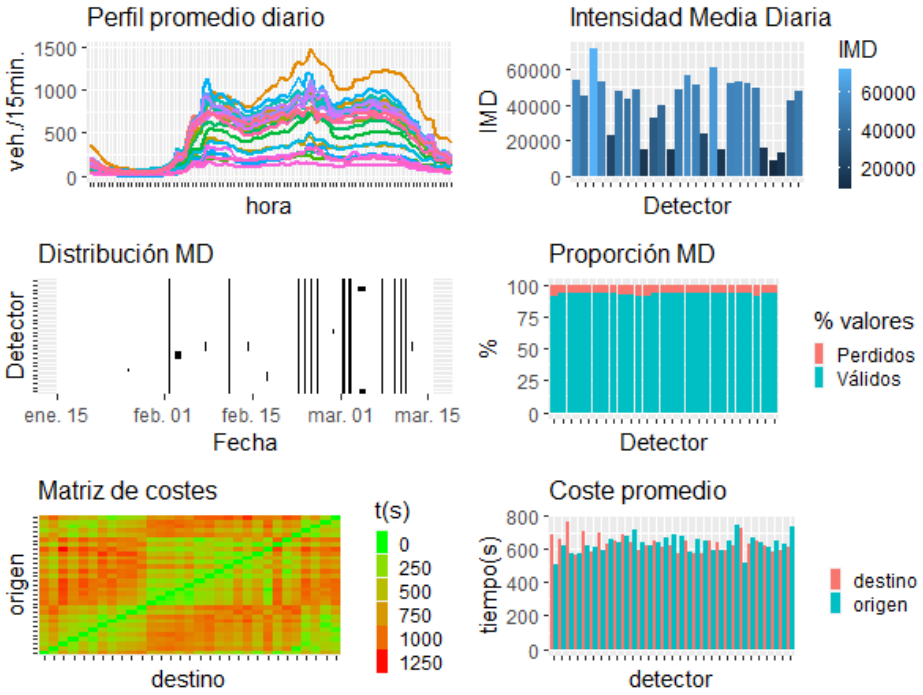


Figura 5-16: Características del caso práctico Sevilla-SE30..

A continuación se presentan los resultados obtenidos mediante la aplicación del marco de predicción en las distintas situaciones planteadas.

5.5.4.2. Imputación

En esta sección se realiza una exposición sintética de los resultados obtenidos tras la aplciación de la fase de imputación del marco de predicción al escenario Sevilla-SE30

En primer lugar se muestran los resultados de imputación con el modelo DLI, con agregación del conjunto de datos de 15 minutos, por una parte y con agregación de 1 hora, por otra.

Imputación DLI para h=15'

Tabla 5-50: Parámetros relativos a los valores perdidos de cada detector. .

Detector	Intervalos MD	Tasa MD	Periodos MD	Longitud periodo	Distancia media entre periodos
1	486	8.44%	27	18.00	204.73
2	389	6.75%	25	15.56	221.79
3	389	6.75%	25	15.56	221.79
4	400	6.94%	26	15.38	212.92
5	400	6.94%	26	15.38	212.92
6	414	7.19%	26	15.92	212.92
7	389	6.75%	25	15.56	221.79
8	389	6.75%	25	15.56	221.79
9	464	8.06%	25	18.56	221.79
10	464	8.06%	25	18.56	221.79
11	488	8.47%	35	13.94	159.53
12	488	8.47%	35	13.94	159.53
13	389	6.75%	25	15.56	221.79
14	389	6.75%	25	15.56	221.79
15	398	6.91%	26	15.31	212.92
16	389	6.75%	25	15.56	221.79
17	389	6.75%	25	15.56	221.79
18	389	6.75%	25	15.56	221.79
19	389	6.75%	25	15.56	221.79
20	389	6.75%	25	15.56	221.79
21	407	7.07%	26	15.65	223.96
22	407	7.07%	26	15.65	223.96
23	389	6.75%	25	15.56	221.79
24	389	6.75%	25	15.56	221.79
25	508	8.82%	27	18.81	204.73
26	395	6.86%	26	15.19	212.92
27	395	6.86%	26	15.19	212.92

Tabla 5-51: Restricciones de ejecución de imputación DLI con h=15'.

Máximo nº de parámetros	10
Tamaño de Ventana temporal	5 (1 hora y 15 minutos)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-52: Sevilla-Se-30. Resultados de imputación DLI para h=15'

Det.	Grupo	Parám. Entrada	Coste máximo	ϱ mínimo	Configuración Modelo	RMSE cv	RMSE test	RMSE Test cc
1	**1**	**10**	681.54	0.796	10-10-10-10-1	0.0742	0.0735	0.0676
	2	**10**	681.54	0.922	10-10-10-10-1	0.0534	0.0531	
2	**1**	**10**	791.29	0.88	10-10-10-10-1	0.0389	0.036	0.0321
	2	**10**	791.29	0.931	10-10-10-10-1	0.022	0.0226	
3	**1**	**10**	578.46	0.928	10-10-10-10-1	0.0485	0.0587	0.0590
	2	**10**	578.46	0.933	10-10-10-10-1	0.0621	0.0596	
4	**1**	**10**	726.51	0.839	10-10-10-10-1	0.0595	0.0588	0.0603
	2	**10**	726.51	0.943	10-10-10-10-1	0.0576	0.0641	
5	**1**	**10**	605.4	0.639	10-50-1	0.1732	0.1784	0.1330
	2	**10**	605.4	0.812	10-10-10-10-1	0.02	0.0219	
6	**1**	**10**	716.82	0.841	10-10-10-10-1	0.0685	0.0711	0.0612
	2	**10**	716.82	0.949	10-10-10-10-1	0.0354	0.0368	
7	**1**	**10**	653.8	0.937	10-10-10-10-1	0.0551	0.0514	0.0476

Det.	Grupo	Parám. Entrada	Coste máximo	ϱ mínimo	Configuración Modelo	RMSE cv	RMSE test	RMSE Test cc
	2	**10**	653.8	0.932	10-10-10-10-1	0.0351	0.0382	
8	**1**	**10**	669.68	0.947	10-10-10-10-1	0.0563	0.0585	0.0535
	2	**10**	669.68	0.955	10-10-10-10-1	0.0389	0.0412	
9	**1**	**10**	707.11	0.882	10-10-10-10-1	0.0791	0.0751	0.0649
	2	**10**	707.11	0.938	10-10-10-10-1	0.0414	0.04	
10	**1**	**10**	660.14	0.911	10-10-10-10-1	0.0599	0.0587	0.0588
	2	**10**	660.14	0.945	10-10-10-10-1	0.0511	0.059	
11	**1**	**10**	616.96	0.927	10-10-10-10-1	0.0722	0.088	0.0874
	2	**10**	616.96	0.884	10-10-10-10-1	0.0631	0.0859	
12	**1**	**10**	638.57	0.87	10-10-10-10-1	0.089	0.0877	0.0778
	2	**10**	638.57	0.812	10-10-10-10-1	0.0477	0.0534	
13	**1**	**10**	671.62	0.94	10-10-10-10-1	0.0663	0.0592	0.0517
	2	**10**	671.62	0.959	10-10-10-10-1	0.0317	0.0332	
14	**1**	**10**	627.74	0.95	10-10-10-10-1	0.0634	0.0571	0.0528
	2	**10**	627.74	0.952	10-10-10-10-1	0.038	0.0423	
15	**1**	**10**	641.47	0.854	10-10-10-10-1	0.0588	0.0605	0.0598
	2	**10**	641.47	0.943	10-100-1	0.063	0.0582	
16	**1**	**10**	594.76	0.768	10-10-10-10-1	0.0398	0.0413	0.0437
	2	**10**	594.76	0.946	10-10-10-10-1	0.0514	0.0496	
17	**1**	**10**	671.46	0.92	10-10-10-10-1	0.0505	0.0537	0.0531
	2	**10**	671.46	0.946	10-10-10-10-1	0.0556	0.0518	
18	**1**	**10**	596.04	0.907	10-10-10-10-1	0.05	0.0498	0.0530

Det.	Grupo	Parám. Entrada	Coste máximo	ϱ mínimo	Configuración Modelo	RMSE cv	RMSE test	RMSE Test cc
	2	**10**	596.04	0.912	10-10-10-10-1	0.0581	0.0607	
19	**1**	**10**	593.01	0.934	10-10-10-10-1	0.0618	0.0524	0.0472
	2	**10**	593.01	0.955	10-10-10-10-1	0.036	0.0343	
20	**1**	**10**	671.05	0.935	10-10-10-10-1	0.0545	0.057	0.0508
	2	**10**	671.05	0.937	10-10-10-10-1	0.0291	0.0357	
21	**1**	**10**	656.53	0.678	10-10-10-10-1	0.0781	0.0769	0.0723
	2	**10**	656.53	0.94	10-10-10-10-1	0.0584	0.0612	
22	**1**	**10**	607.97	0.843	10-100-100-1	0.0696	0.0715	0.0679
	2	**10**	607.97	0.936	10-10-10-10-1	0.0517	0.0592	
23	**1**	**10**	752.81	0.926	10-10-10-10-1	0.0629	0.0617	0.0524
	2	**10**	752.81	0.935	10-10-10-10-1	0.0279	0.0298	
24	**1**	**10**	646.97	0.92	10-10-10-10-1	0.0601	0.0611	0.0652
	2	**10**	646.97	0.915	10-10-10-10-1	0.0702	0.0754	
25	**1**	**10**	666.84	0.715	10-10-10-10-1	0.1221	0.113	0.0974
	2	**10**	666.84	0.818	10-50-1	0.0567	0.0593	
26	**1**	**10**	643.67	0.815	10-10-10-10-1	0.0625	0.0658	0.0639
	2	**10**	643.67	0.928	10-10-10-10-1	0.0603	0.0593	
27	**1**	**10**	600.38	0.937	10-10-10-10-1	0.0601	0.0581	0.0577
	2	**10**	600.38	0.798	10-10-10-10-1	0.0569	0.0568	

Imputación DLI con h=1 hora

Tabla 5-53: Parámetros relativos a los valores perdidos de cada detector con h=1h.

Detector	Intervalos MD	Tasa MD	Periodos MD	Longitud periodo	Distancia media entre periodos
1	107	7.43%	24	4.46	39.39
2	84	5.83%	23	3.65	41.18
3	84	5.83%	23	3.65	41.18
4	86	5.97%	24	3.58	39.39
5	86	5.97%	24	3.58	39.39
6	90	6.25%	24	3.75	46.22
7	84	5.83%	23	3.65	41.18
8	84	5.83%	23	3.65	41.18
9	102	7.08%	23	4.43	41.18
10	102	7.08%	23	4.43	41.18
11	101	7.01%	28	3.61	41.00
12	101	7.01%	28	3.61	41.00
13	84	5.83%	23	3.65	41.18
14	84	5.83%	23	3.65	41.18
15	86	5.97%	24	3.58	39.39
16	84	5.83%	23	3.65	41.18
17	84	5.83%	23	3.65	41.18
18	84	5.83%	23	3.65	41.18
19	84	5.83%	23	3.65	41.18
20	84	5.83%	23	3.65	41.18
21	88	6.11%	24	3.67	42.39
22	88	6.11%	24	3.67	42.39
23	84	5.83%	23	3.65	41.18
24	84	5.83%	23	3.65	41.18
25	112	7.78%	24	4.67	39.39
26	84	5.83%	23	3.65	41.18
27	84	5.83%	23	3.65	41.18

Tabla 5-54: Restricciones de ejecución de imputación DLI con h=15'.

Máximo nº de parámetros	4
Tamaño de Ventana temporal	2 (2 horas)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-55: Sevilla-SE30. Resultados de imputación DLI para h=1 hora'.

Det.	Grupo	Parám. Entrada	Coste máximo	ϱ mínimo	Configuración Modelo	RMSE cv	RMSE test	RMSE Test cc
1	1	10	681.54	0.886	4-100-100-1	0.1185	0.1161	0.1040
	2	10	681.54	0.693	4-10-10-10-1	0.0722	0.0743	
2	1	10	791.29	0.813	4-100-1	0.0879	0.1021	0.0888
	2	10	791.29	0.868	4-10-10-10-1	0.057	0.0562	
3	1	10	578.46	0.935	4-100-1	0.0923	0.0947	0.0950
	2	10	578.46	0.925	4-10-10-10-1	0.0677	0.0959	
4	1	10	726.51	0.694	4-10-10-10-1	0.0935	0.0858	0.0976
	2	10	726.51	0.873	4-10-10-10-1	0.117	0.1264	
5	1	10	605.4	0.512	4-10-10-10-1	0.1796	0.1736	0.1373
	2	10	605.4	0.729	4-50-1	0.0353	0.0483	
6	1	10	716.82	0.694	4-10-10-10-1	0.0884	0.0834	0.0905
	2	10	716.82	0.891	4-10-10-10-1	0.1115	0.1078	
7	1	10	653.8	0.884	4-100-1	0.1134	0.1159	0.1163
	2	10	653.8	0.89	4-10-10-10-1	0.1061	0.1174	
8	1	10	669.68	0.888	4-10-10-10-1	0.1063	0.1133	0.1068

	2	10	669.68	0.909	4-10-10-10-1	0.0948	0.0908	
9	1	10	707.11	0.82	4-100-1	0.1133	0.1173	0.1114
	2	10	707.11	0.909	4-10-10-10-1	0.0809	0.0971	
10	1	10	660.14	0.883	4-10-10-10-1	0.1253	0.1447	0.1286
	2	10	660.14	0.883	4-10-10-10-1	0.0828	0.0891	
11	1	10	616.96	0.855	4-100-100-1	0.1283	0.1565	0.1453
	2	10	616.96	0.713	4-10-10-10-1	0.0882	0.118	
12	1	10	638.57	0.786	4-100-1	0.1004	0.0965	0.1003
	2	10	638.57	0.575	4-10-10-10-1	0.1076	0.1097	
13	1	10	671.62	0.908	4-10-10-10-1	0.1053	0.1302	0.1324
	2	10	671.62	0.898	4-100-100-1	0.1208	0.1378	
14	1	10	627.74	0.918	4-10-10-10-1	0.1373	0.1327	0.1204
	2	10	627.74	0.899	4-100-1	0.0877	0.0902	
15	1	10	641.47	0.898	4-100-100-1	0.1309	0.1274	0.1085
	2	10	641.47	0.741	4-100-1	0.0585	0.0622	
16	1	10	594.76	0.729	4-10-10-10-1	0.0667	0.0578	0.0698
	2	10	594.76	0.915	4-10-10-10-1	0.1062	0.0993	
17	1	10	671.46	0.883	4-10-10-10-1	0.1099	0.108	0.1106
	2	10	671.46	0.912	4-100-100-1	0.1164	0.1171	
18	1	10	596.04	0.875	4-10-10-10-1	0.0987	0.1091	0.1091
	2	10	596.04	0.868	4-10-10-10-1	0.1218	0.109	
19	1	10	593.01	0.884	4-10-10-10-1	0.1247	0.1163	0.1273
	2	10	593.01	0.886	4-100-100-1	0.1311	0.1541	
20	1	10	671.05	0.898	4-100-100-1	0.1151	0.114	0.1037

	2	10	671.05	0.89	4-100-1	0.0667	0.0786	
21	1	10	656.53	0.526	4-10-10-10-1	0.0994	0.1023	0.0986
	2	10	656.53	0.892	4-100-1	0.092	0.0896	
22	1	10	607.97	0.674	4-10-10-10-1	0.0824	0.086	0.0828
	2	10	607.97	0.895	4-100-1	0.0682	0.0749	
23	1	10	752.81	0.878	4-10-10-10-1	0.1033	0.1067	0.1079
	2	10	752.81	0.893	4-100-100-1	0.1179	0.1107	
24	1	10	646.97	0.869	4-100-100-1	0.1334	0.1413	0.1269
	2	10	646.97	0.85	4-10-10-10-1	0.0952	0.0918	
25	1	10	666.84	0.612	4-10-10-10-1	0.1095	0.11	0.0903
	2	10	666.84	0.806	4-100-1	0.048	0.0422	
26	1	10	643.67	0.846	4-100-100-1	0.1386	0.1461	0.1286
	2	10	643.67	0.896	4-100-1	0.0827	0.0856	
27	1	10	600.38	0.9	4-100-100-1	0.1251	0.1283	0.1177
	2	10	600.38	0.857	4-10-10-10-1	0.0963	0.0917	

5.5.4.3. Predicción

En esta sección se realiza una exposición sintética de los resultados obtenidos tras la aplicación de la fase de predicción al escenario Sevilla-SE30.

En primer lugar se muestran los resultados para los modelos de predicción con h=15′ y posteriormente los de los modelos con h=1 hora.

Predicción con h=15′

Tabla 5-56: Sevilla-SE30. Restricciones de ejecución de predicción con h=15'.

Máximo nº de parámetros	8
Tamaño de Ventana temporal	5 (1:15h horas)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-57: Sevilla-SE30. Resultados predicción detector 1, h=15'

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test
NI	DLP	1	0.892	8-10-10-10-1	0.0818	0.0759	0.0733
		2	0.929	8-10-10-10-1	0.0682	0.0668	
	RFP	1	0.892	XRT-8-43-1	0.0792	0.0727	0.0696
		2	0.929	DRF-8-40-1	0.0665	0.0619	
DLI	DLP	1	0.921	8-10-10-10-1	0.0598	0.0549	0.0604
		2	0.92	8-10-10-10-1	0.0718	0.0737	
	RFP	1	0.921	DRF-8-49-1	0.0566	0.0494	0.0545
		2	0.92	XRT-8-46-1	0.0671	0.067	

Tabla 5-58: Sevilla-SE30. Resultados predicción detector 13, h=15'

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test
NI	DLP	1	0.949	8-10-10-10-1	0.051	0.0502	0.0566
		2	0.946	8-10-10-10-1	0.0627	0.0721	
	RFP	1	0.949	XRT-8-42-1	0.0448	0.0463	0.0512
		2	0.946	DRF-8-33-1	0.0516	0.0631	
DLI	DLP	1	0.946	8-10-10-10-1	0.0461	0.0405	0.0475
		2	0.941	8-10-10-10-1	0.0695	0.0647	
	RFP	1	0.946	XRT-8-43-1	0.0454	0.0386	0.0434
		2	0.941	XRT-8-29-1	0.0672	0.0551	

Tabla 5-59: Sevilla-SE30. Resultados predicción detector 26, h=15'

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test
NI	DLP	1	0.829	8-10-10-10-1	0.0606	0.0648	0.0639
		2	0.93	8-10-10-10-1	0.0576	0.0618	
	RFP	1	0.829	XRT-8-38-1	0.0549	0.0568	0.0555
		2	0.93	XRT-8-38-1	0.0526	0.0523	
DLI	DLP	1	0.933	8-100-1	0.0593	0.059	0.0630
		2	0.925	8-10-10-10-1	0.0701	0.0729	
	RFP	1	0.933	DRF-8-42-1	0.0474	0.0404	0.0467
		2	0.925	DRF-8-35-1	0.0652	0.0621	

Predicción con h=1h

Tabla 5-60: Sevilla-SE30. Restricciones de ejecución de predicción con h=1 hora.

Máximo nº de parámetros	4
Tamaño de Ventana temporal	3 (3 horas)
Coste	Promedio de los tiempos de recorrido relacionados con cada detector
Tasa de valores perdidos coincidentes	30%
Grupos en etapa de clasificación	2

Tabla 5-61: Sevilla-SE30. Resultados predicción detector 1, h= 1 hora.

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test cc
NI	DLP	1	0.636	4-100-1	0.0894	0.0878	0.0954
		2	0.893	4-100-1	0.1141	0.1141	
	RFP	1	0.636	DRF-4-30-1	0.0714	0.0709	0.0671
		2	0.893	DRF-4-30-1	0.0844	0.0578	
DLI	DLP	1	0.882	4-50-1	0.12	0.1339	0.1224
		2	0.89	4-100-100-1	0.0857	0.0942	
	RFP	1	0.882	DRF-4-33-1	0.0799	0.092	0.0913
		2	0.89	DRF-4-28-1	0.0862	0.0896	

Tabla 5-62: Sevilla-SE30. Resultados predicción detector 13, h= 1 hora.

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSE test cc
NI	DLP	1	0.9	4-100-100-1	0.1276	0.1185	0.1037
		2	0.914	4-10-10-10-1	0.0761	0.0676	
	RFP	1	0.9	XRT-4-31-1	0.0974	0.0927	0.0856
		2	0.914	XRT-4-31-1	0.0799	0.0682	
DLI	DLP	1	0.903	4-100-1	0.0873	0.1113	0.1204
		2	0.892	4-100-100-1	0.135	0.1426	
	RFP	1	0.903	DRF-4-34-1	0.0845	0.0968	0.1005
		2	0.892	DRF-4-26-1	0.0993	0.1095	

Tabla 5-63: Sevilla-SE30. Resultados predicción detector 26, h= 1 hora.

Imputacion	Prediccion	Grupo	Corr	ConfModelo	RMSEcv	RMSEtest	RMSEtest
NI	DLP	1	0.846	4-100-1	0.1431	0.1551	0.1371
		2	0.896	4-100-100-1	0.085	0.0932	
	RFP	1	0.846	XRT-4-26-1	0.1001	0.1099	0.1046
		2	0.896	XRT-4-39-1	0.0876	0.0917	
DLI	DLP	1	0.844	4-100-100-1	0.1626	0.1844	0.1542
		2	0.896	4-100-1	0.0866	0.0804	
	RFP	1	0.844	XRT-4-33-1	0.1046	0.0997	0.0931
		2	0.896	XRT-4-23-1	0.093	0.0768	

Resumen Predicción

En la Figura 5-9, se muestra un resumen de los resultados obtenidos en la fase de predicción para los modelos de los detectores 1, 13 y 26 del escenario Sevilla SE-30.

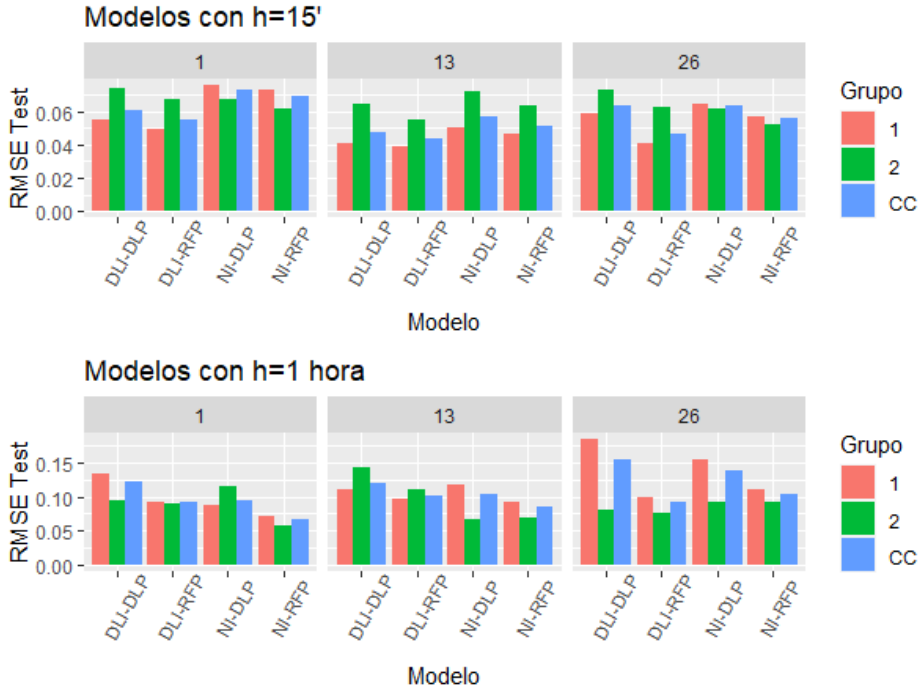


Figura 5-17: Resumen resultados predicción para escenario Sevilla-SE30.

6. CONCLUSIONES Y LÍNEAS FUTURAS

En este trabajo se afronta el problema de la predicción de flujo de tráfico a corto plazo, de una manera integral, contemplando todas las etapas que intervienen, desde la adquisición del conjunto de datos, a la proporción del valor modelado.

La propuesta realizada se basa en un marco predictivo en el que se identifican las fases que intervienen en la predicción de manera totalmente modular y flexible. Mediante el marco definen las funciones que cumple cada fase, mientras que, en cada aplicación de éste, se deben concretar las herramientas utilizadas para para llevar a cabo dichas funciones. De esta forma, se ha diseñado una herramienta que facilita la aplicación de diversas en cada una de sus fases, favoreciendo su comparación.

El marco diseñado cumple con los objetivos planteados al inicio de este trabajo. Durante su desarrollo se ha profundizado en las técnicas de predicción a corto plazo, así como en los procesos que se deben realizar de manera colateral. Del mismo modo, se ha realizado un análisis del estado en el que se encuentra la gestión y la accesibilidad de los conjuntos de datos existentes y su utilización práctica mediante los casos de estudio tratados. Por último, se ha desarrollado un procedimiento que facilita la comparación entre diferentes técnicas en cada una de las fases que intervienen en el proceso de predicción, permitiendo la medición de su influencia en el rendimiento final de la de todo el proceso.

En las siguientes secciones se realiza una evaluación global del trabajo realizado,

separándose en las fases que intervienen en proceso de desarrollo de modelos de predicción de parámetros de tráfico.

6.1. Adquisición, tratamiento y procesamiento de la información

Tras haber aplicado el marco de predicción a 4 casos de estudio diversos, se estima que la fase de armonización de información resulta fundamental dentro del proceso de predicción.

En primer lugar, permite un acercamiento común a las características de diferentes casos de estudio. Dicho acercamiento facilita el procesamiento y análisis de la información de cada caso de estudio, con bases comparativas idénticas, por una parte y, por otra, permite comprender el contexto de cada caso de estudio de una manera simple, observando las características del flujo de tráfico en los diferentes puntos de adquisición de datos, las relaciones espacio-temporales entre éstos y la presencia de valores perdidos.

El aspecto más arduo en este proceso ha resultado la definición del formato homogéneo y del proceso que permite la importación de las distintas fuentes de datos, cuyos formatos originales son totalmente heterogéneos, observándose unidades de medida, tipos de detectores, niveles de agregación de los datos y niveles accesibilidad distintos. Los casos de estudio incluidos en este trabajo presentan características muy variadas en todos los aspectos observados, solo siendo posible su comparación mediante la aplicación del mismo proceso integral de predicción, previa armonización de sus conjuntos originales de datos.

Este paso inicial permite el posterior desarrollo de un proceso de predicción uniforme, aplicable a distintos casos de estudio y que facilita la aplicación de múltiples aproximaciones en cada una de las etapas del proceso de predicción.

La base de comparación homogénea que se ha definido gracias a la elaboración del formato armonizado favorece la comparación de los resultados obtenidos, tras aplicar una misma configuración del marco de predicción entre distintos casos de estudio.

6.2. Imputación

Dentro del proceso global de predicción, la imputación de valores perdidos ha probado su utilidad, contribuyendo a la mejora de la precisión de los valores modelados en la fase de predicción y, sobre todo, ampliando el número de casos para entrenar y testear los modelos. Este proceso es especialmente beneficioso, sobre todo, en aquellos casos de estudio en los que el escenario original presenta una tasa de valores perdidos relativamente alta.

La etapa de preprocesamiento de información en la fase de imputación presenta 2 ventajas fundamentales:

i) Permite obtener medidas de relación de distinta naturaleza entre los parámetros del conjunto de datos y el parámetro de salida del modelo de imputación

ii) Reduce el volumen de información utilizado como entrada, basando dicha reducción en las relaciones observadas, sin que suponga una merma del potencial explicativo sobre el parámetro de salida.

El objetivo de este estudio no es la definición de un método óptimo de selección de parámetros, aunque, el marco de predicción permite establecer comparativas entre metodologías y puede servir como guía para la mejora continua de estas técnicas en función de su influencia sobre el rendimiento de los modelos.

El rendimiento del modelado de imputación se ve influenciado por los parámetros seleccionados en el preprocesamiento, y los criterios elegidos en dicha etapa para componer los conjuntos de entrada de los modelos. Se ha tratado de utilizar un método de selección de parámetros:

i) Fácilmente comprensible, tanto para su implementación como para su inteligibilidad.

ii) Aplicable a distintos casos de estudio de manera genérica.

iii) Que penalice los parámetros para los que se observa una tasa alta de valores perdidos como parámetros de entrada de los modelos.

El método de selección de parámetros ha derivado en comportamientos poco intuitivos del rendimiento de los modelos en casos concretos, observándose que al comparar el rendimiento entre distintas configuraciones de modelado en las que lo

único que cambia es la tasa de valores perdidos, se obtienen mejores índices de rendimiento en escenarios con una tasa mayor de valores perdidos, que en casos en los que la tasa es más baja. Dentro del marco de predicción, se debe incidir en refinar este paso, prestándose un esfuerzo mayor en la fase de preprocesamiento de los conjuntos de entrada de los modelos, tanto de imputación como de predicción.

La posición relativa del detector respecto al resto de detectores de un caso de estudio es determinante para el rendimiento de los modelos ML. La selección de parámetros del conjunto de entrada del modelo se basa en la información del contexto del detector. De esta manera se observan comportamientos más robustos en el rendimiento de imputación, aunque se aumente la tasa de valores perdidos, en detectores que presentan varios detectores aguas arriba. Los detectores más aislados, que se encuentran al principio de un tramo, sin detectores aguas arriba, presentan valores de rendimiento más erráticos, viéndose perjudicados en escenarios con tasas crecientes de valores perdidos.

De la comparativa de los resultados observados para el modelado de imputación entre 4 casos de estudio distintos, se deduce que se puede definir un proceso generalizable y dinámico que, siguiendo exactamente los mismos pasos, se adapte a las características de cada caso de estudio, presentando resultados similares entre éstos.

En este trabajo sólo se han comparado dos aproximaciones para la imputación de valores, en cuanto a la técnica utilizada (NI y DLI), una que consiste en no imputar los datos con el objetivo de comparar su rendimiento con el de la imputación mediante DLI un método basado en una red neuronal BPNN. Para poder extraer conclusiones más definitivas sobre los distintos tipos de técnicas de imputación, se deben comparar aproximaciones de todas las familias de modelado que se utilizan en la actualidad, como las basadas en series temporales y tensores. Se propone introducir técnicas de dicho tipo en aplicaciones futuras del marco a casos de estudio reales.

6.3. Predicción

La utilización del marco de predicción ha permitido la aplicación del proceso de predicción, a los detectores de los 4 casos de estudio, favoreciendo la utilización de múltiples configuraciones de modelado.

El preprocesamiento de la fase de predicción presenta problemas similares al de imputación. Se considera que este paso es de una extrema importancia cuando se utilizan técnicas guiadas por datos. Del mismo modo, se considera que las técnicas elegidas para la concreción del marco en los distintos casos de estudio son optimizables y se debe indagar en este aspecto para conseguir resultados y conclusiones más fiables. Dicho esto, la sencillez de las técnicas utilizadas posibilita su fácil aplicación a los distintos contextos, así como la extracción de información que facilitan la comprensión de características del tráfico en la ubicación de los detectores.

La fase de imputación facilita el cálculo de un mayor número de predicciones sin observarse una merma en la precisión en éstas respecto a las configuraciones de modelado que no introducen dicha fase.

La fase de imputación influye de manera más positiva sobre el rendimiento de predicción con horizontes iguales o menores a 30', observándose los mejores resultados en escenarios con tasas de valores perdidos menores o iguales al 15%.

Se deduce que la imputación ayuda a realizar predicciones con una precisión similar a situaciones en las que el conjunto de entrada del modelo de predicción es completo.

En cuanto a la comparación entre los métodos de predicción, se obtiene una conclusión clara, las configuraciones probadas para el modelo RFP son superiores a las establecidas para el modelo DLP, al menos con las condiciones impuestas

El marco propuesto permite la aplicación de múltiples configuraciones de modelado, en lo referente a las técnicas usadas en cada etapa, de una manera sencilla a distintos detectores y casos de estudio. Esto permite la extracción de índices de rendimiento de cada una de las fases para observar su influencia sobre el proceso completo de modelado, la evolución del rendimiento en función de la tasa de valores perdidos y del horizonte de predicción.

El marco facilita la comparación de distintos aspectos, relativos al proceso de modelado:

1- La aplicación de un mismo método a múltiples detectores, ya sean de un mismo caso de estudio o entre casos de estudio diferentes, permitiendo la comparación de los resultados.

2- La aplicación de distintos métodos en el proceso de modelado de un

mismo detector.

3- La observación de la evolución del rendimiento de los modelos en distintos casos de estudio, con distintas tasas de valores perdidos y horizontes de predicción

Se estima que el marco presentado tiene un claro potencial para la evaluación real de nuevas técnicas de modelado, basadas en ML, sobre mútiples casos de estudio. De esta manera, se considera que se ha generado una herramienta muy útil y práctica para este campo de investigación y favoreciendo la comparación y prestaciones reales de las técnicas más innovadoras.

6.4. Líneas futuras

Se observan diversas vías de mejora de este marco predictivo. Tal y como se ha comentado en las conclusiones, el marco diseñado permite afrontar, de una manera integral y generalista el problema de la predicción a corto plazo, habría ciertos aspectos con posibilidad de ser optimizados en el futuro.

En la Tabla 6-1 se exponen las principales ideas para el desarrollo del marco y de los conceptos que se exponen en ese trabajo, poniendo el foco en las posibilidades de mejora de cada una de las fases del marco, así como del proceso de predicción de manera integral.

Tabla 6-1: Líneas futuras de mejora del marco de predicción.

Etapa	Características evaluadas
Adquisicón	Generación de base de datos común de parámetros de valores perdidos en numerosos casos de estudio.
	Implementar procesos automáticos de importación de conjuntos de datos de principales sistemas de información de tráfico.
	Optimización de la selección de información, mediante la determinación del histórico de información mínimo previo para que se generalice el comportamiento de cada detector con una precisión aceptable.
	Integrar información de coches flotantes y vehículos autónomos y conectados en el modelo de datos.
Preprocesamiento	Optimizar el la selección de parámetros, indagando en las técnicas más adecuadas para cada método., y la selección de los hiperparámetros de los métodos elegidos en cada una de las fases de modelado.
Análisis	Número de casos en el conjunto de datos.
	Tasa de valores perdidos y distribución, individuales en cada detector y general del escenario.
	Grado de relación entre parámetros.
Imputación	Implementación de modelos con familias de técnicas no testadas en este trabajo, como las basadas en series temporales o en tensores. Comprobación de su rendimiento sobre los casos de estudio ya observados.
	Creación de un amplio banco de datos con la aplicación de diversas técnicas a numerosos casos de estudio, y el registro de su rendimiento para comprender el fenómeno de los valores perdidos con una perspectiva más amplia.
	Integración de métodos predictivos de combinación y fusión de modelos para optimizar las predicciones aprovechando las fortalezas de las diferentes técnicas implementadas individuales.
Predicción	Integración de métodos novedosos en el marco para testarlos en diferentes casos de estudio.
	Integración con herramientas de desarrollo de Machine Learning, incorporación autmatizada de modelos más eficientes en el modelado del tráfico.
Marco	Convertir modelado en un proceso iterativo de búsqueda de configuración óptima de las distintas fases del marco.
	Comprobación de validez marco en casos de estudio que contengan detectores situados en distintas escalas desde el punto de vista del tipo de red, mezclando detectores urbanos y periurbanos y en otros tipos de vías extraurbanas

REFERENCIAS

Abdulhai, B., Porwal, H., Recker, W., 1999. Short Term Freeway Traffic Flow Prediction Using Genetically-Optimized Time-Delay-Based Neural Networks. Irvine.

Abirami, U., Sridevi, S., 2017. Traffic flow prophecy with mapreduce job for big data driven. 2016 8th Int. Conf. Adv. Comput. ICoAC 2016 13–18. https://doi.org/10.1109/ICoAC.2017.7951737

Ahmed, M.S., Cook, A.R., 1979. Analysis Of Freeway Traffic Time-Series Data By Using Box-Jenkins Techniques. Transp. Res. Rec.

Asif, M.T., Mitrovic, N., Dauwels, J., Jaillet, P., 2016. Matrix and Tensor Based Methods for Missing Data Estimation in Large Traffic Networks. IEEE Trans. Intell. Transp. Syst. 17, 1816–1825. https://doi.org/10.1109/TITS.2015.2507259

Avazpour, I., Grundy, J., Zhu, L., 2019. Engineering complex data integration, harmonization and visualization systems. J. Ind. Inf. Integr. 16, 100103. https://doi.org/10.1016/j.jii.2019.08.001

Ayuntamiento de Madrid, 2016. Portal de datos abiertos del Ayuntamiento de Madrid [WWW Document]. URL https://datos.madrid.es/portal/site/egob/ (accessed 9.1.19).

Ayuntamiento de Málaga, 2016. Portal de datos abiertos del Ayuntamiento de Málaga [WWW Document]. URL https://datosabiertos.malaga.eu/

Ayuntamiento de Sevilla, 2016. Portal de Datos Abiertos de Sevilla [WWW Document]. URL http://datosabiertos.sevilla.org/data/ (accessed 11.15.18).

Barmpounakis, E., Geroliminis, N., 2020. On the new era of urban traffic monitoring with massive drone data : The pNEUMA large-scale field experiment. Transp. Res. Part C 111, 50–71. https://doi.org/10.1016/j.trc.2019.11.023

Bearn, C., Mingus, C., Watkins, K., 2018. Research in Transportation Business & Management An adaption of the level of traffic stress based on evidence from

the literature and widely available data. Res. Transp. Bus. Manag. 29, 50–62. https://doi.org/10.1016/j.rtbm.2018.12.002

Bellman, R.E., Zadeh, L.A., 1970. Decision-Making in a Fuzzy Environment. Manage. Sci. 17, B-141-B-164. https://doi.org/10.1287/mnsc.17.4.B141

Bengio, Y., Courville, A., Vincent, P., 2012. Representation Learning: A Review and New Perspectives 1–30. https://doi.org/10.1145/1756006.1756025

Benvenuto, N., Piazza, F., 1996. The backpropagation algorithm, in: Neural Networks. pp. 967–969. https://doi.org/10.1109/78.127967

Berends, J., Carrara, W., Vollers, H., Fechner, T., Kleemann, M., 2017. Analytical Report 5: Barriers in working with Open Data.

Berner, J., Hart, T., 2013. Using the Caltrans Performance Measurement System (PeMS) for s TCR's.

Bie, Y., Wang, X., Qiu, T.Z., 2016. Online Method to Impute Missing Loop Detector Data for Urban Freeway Traffic Control. Transp. Res. Rec. J. Transp. Res. Board Volume 259, 37–46. https://doi.org/https://doi.org/10.3141/2593-05

Bing, Q., Gong, B., Yang, Z., Shang, Q., Zhou, X., 2015. Short-Term Traffic Flow Local Prediction Based on Combined Kernel Function Relevance Vector Machine Model. Math. Probl. Eng. 2015, 9. https://doi.org/10.1155/2015/154703

Boquet, G., Morell, A., Serrano, J., Vicario, J.L., 2020. A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection. Transp. Res. Part C Emerg. Technol. 115, 102622. https://doi.org/10.1016/J.TRC.2020.102622

Boukerche, A., Wang, J., 2020. Machine Learning-based traffic prediction models for Intelligent Transportation Systems. Comput. Networks 181, 107530. https://doi.org/10.1016/j.comnet.2020.107530

Bowman, C.N., Miller, J.A., 2016. Modelling traffic flow using simulationa and big data analytics, in: Proceedings of the 2016 Winter Simulation Conference. Washington D.C., pp. 1206–1217.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification And Regression Trees, 1st Editio. ed. Routledge, New York. https://doi.org/10.1201/9781315139470

Caceres, N., Romero, L.M., Benitez, F.G., 2012. Estimating Traffic Flow Profiles According to a Relative Attractiveness Factor. Procedia - Soc. Behav. Sci. 54, 1115–1124. https://doi.org/10.1016/j.sbspro.2012.09.826

Cai, L., Zhang, Z., Yang, J., Yu, Y., Zhou, T., Qin, J., 2019. A noise-immune Kalman filter for short-term traffic flow forecasting. Phys. A Stat. Mech. its Appl. 536, 122601. https://doi.org/10.1016/J.PHYSA.2019.122601

Cai, P., Wang, Y., Lu, G., Chen, P., Ding, C., Sun, J., 2016. A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting. Transp. Res. Part C Emerg. Technol. 62, 21–34. https://doi.org/10.1016/j.trc.2015.11.002

California Department of Transportation, 2018. Caltrans Performance Measurement System (PeMS) [WWW Document]. Perform. Meas. Syst. URL http://pems.dot.ca.gov/ (accessed 9.27.18).

Calvert, S.C., Arem, B. van, Van Lint, J.W.C., 2020. A generic multi-level framework for microscopic traffic simulation with automated vehicles in mixed traffic. Transp. Res. Part C Emerg. Technol. 110, 291–311. https://doi.org/https://doi.org/10.1016/j.trc.2019.11.019

Castro-Neto, M., Jeong, Y.-S., Jeong, M.-K., Han, L.D., 2009. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. Expert Syst. Appl. 36, 6164–6173. https://doi.org/10.1016/j.eswa.2008.07.069

Chen, C., Kwon, J., Rice, J., Skabardonis, A., Varaiya, P., 2003. Detecting Errors and Imputing Missing Data for Single-Loop Surveillance Systems. Transp. Res. Rec. 1855, 160–167. https://doi.org/10.3141/1855-20

Chen, H., Grant-Muller, S., Mussone, L., Montgomery, F., 2001. A study of hybrid neural network approaches and the effects of missing data on traffic forecasting. Neural Comput. Appl. 10, 277–286. https://doi.org/10.1007/s521-001-8054-3

Chen, Y., Guizani, M., Zhang, Y., Wang, L., Crespi, N., Lee, G.M., 2017. When Traffic Flow Prediction Meets Wireless Big Data Analytics. CoRR abs/1709.0,

1–7.

Chowdhury, M., Sadek, A., 2021. WHAT IS ITS? [WWW Document]. (A Guid. Pract. URL https://rno-its.piarc.org/en/intelligent-transport-systems/what-its (accessed 12.6.21).

Consejo Nacional de Información Geográfica de España, 2017. Plan de acción de implementación de la directiva INSPIRE, Plan de Acción del CODIIGE.

Davis, G.A., Nihan, N.L., 1991. Nonparametric Regression and Short-Term Freeway Traffic Forecasting. J. Transp. Eng. 117, 178–188. https://doi.org/10.1061/(ASCE)0733-947X(1991)117:2(178)

Deirdre Lee, D., 2016. Discovering Open Data Standards, in: International Open Data Conference (IODC) 2016. Madrid, pp. 6–9.

Dia, H., 2001. An object-oriented neural network approach to short-term traffic forecasting. Eur. J. Oper. Res. 131, 253–261. https://doi.org/10.1016/S0377-2217(00)00125-9

Dietterich, T.G., 2000. Ensemble Methods in Machine Learning, in: Multiple Classifier Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–15.

Dölger, R., Geißler, T., 2012. DATEX II – The standard for ITS on European Roads, EasyWay.

Dong, C., Shao, C., Richards, S.H., Han, L.D., 2014. Flow rate and time mean speed predictions for the urban freeway network using state space models. Transp. Res. Part C Emerg. Technol. 43, 20–32. https://doi.org/10.1016/j.trc.2014.02.014

Dougherty, M.S., Cobbett, M.R., 1997. Short-term inter-urban traffic forecasts using neural networks. Int. J. Forecast. 13, 21–31. https://doi.org/10.1016/S0169-2070(96)00697-8

Elhenawy, M., Chen, H., Rakha, H.A., 2014. Dynamic travel time prediction using data clustering and genetic programming. Transp. Res. Part C Emerg. Technol. 42, 82–98. https://doi.org/10.1016/j.trc.2014.02.016

European Commission, 2017. Commission outlines next steps towards a European data economy 10–11.

European Commission, 2014. Building infrastructure to strengthen europe's economy.

European Commission, 2013. Mobilising Intelligent Transport Systems for EU cities. https://doi.org/SWD(2013) 527

European Commission, 2007. D2.8.I.7 Data Specification on Transport Networks – Technical Guidelines.

European Parliament, Council of the European Union, 2007. Directive 2007/2/EC of the European Parliament and of the council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), Official Journal of the European Union.

Fan, X., Liu, B., Huang, C., Wen, S., Fu, B., 2021. Utility maximization data scheduling in drone-assisted vehicular networks. Comput. Commun. 175, 68–81. https://doi.org/10.1016/j.comcom.2021.04.033

Gers, F., 2001. Long Short-Term Memory in Recurrent Neural Networks. École polytechnique fédérale de Lausanne.

Ghiasi, A., Li, X., Ma, J., 2019. A mixed traffic speed harmonization model with connected autonomous vehicles. Transp. Res. Part C Emerg. Technol. 104, 210–233. https://doi.org/10.1016/J.TRC.2019.05.005

Ghosh, B., Basu, B., O'Mahony, M., 2009. Multivariate short-term traffic flow forecasting using time-series analysis. IEEE Trans. Intell. Transp. Syst. 10, 246–254. https://doi.org/10.1109/TITS.2009.2021448

Gilmore, J.F., Elibiary, K.J., 1993. AI In Advanced Traffic Management Systems. Work. Notes, AAAI-93 Work. AI Intell. Veh. Highw. Syst. 57–65.

Gilmore, J.F., Elibiary, K.J., Abe, N., 1993. Traffic Management Applications of Neural Networks. Work. Notes, AAAI-93 Work. AI Intell. Veh. Highw. Syst. 85–95.

Gora, P., Katrakazas, C., Drabicki, A., Islam, F., Ostaszewski, P., 2020. Microscopic traffic simulation models for connected and automated vehicles (CAVs) – state-of-the-art. Procedia Comput. Sci. 170, 474–481. https://doi.org/https://doi.org/10.1016/j.procs.2020.03.091

Greenshields, B.D., Thompson, J.T., Dickinson, H.C., Swinton, R.S., 1934. The Photographic Method Of Studying Traffic Behavior. Highw. Res. Board Proc.

Habtemichael, F.G., Cetin, M., 2016. Short-term traffic flow rate forecasting based

on identifying similar traffic patterns. Transp. Res. Part C Emerg. Technol. 66, 61–78. https://doi.org/10.1016/j.trc.2015.08.017

Hamilton, J.D., 1994. State-Space Models, in: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics. Elsevier Science B.V., pp. 3041–3077.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction., Springer Series in Statistics. https://doi.org/10.1007/978-0-387-84858-7

Henrickson, K., Zou, Y., Wang, Y., 2015. Flexible and Robust Method for Missing Loop Detector Data Imputation. Transp. Res. Board 94th Annu. Meet. No. 15-580. https://doi.org/10.3141/2527-04

Herrmann, A., Hempel, U., Jumar, U., 2012. A modular, close-meshed traffic and environmental data acquisition network as a basis for information services, IFAC Proceedings Volumes (IFAC-PapersOnline). IFAC. https://doi.org/10.3182/20120403-3-DE-3010.00061

Hintz, D., 2010. Data Harmonization Principles and Development Approaches as Applied to INSPIRE SDIs 1–8.

Hofleitner, A., Herring, R., Bayen, A., 2012. Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning. Transp. Res. Part B Methodol. 46, 1097–1122. https://doi.org/10.1016/j.trb.2012.03.006

Hoogendoorn, S.P., Bovy, P.H.L., 2001. State-of-the-art of vehicular traffic flow modelling. Proc. Inst. Mech. Eng. Part I J. Syst. Control Eng. 215, 283–303. https://doi.org/10.1177/095965180121500402

Hua Yang, H., Murata, N., Amari, S., 1998. Statistical inference: learning in artificial neural networks. Trends Cogn. Sci. 2, 4–10. https://doi.org/https://doi.org/10.1016/S1364-6613(97)01114-5

Imeers, L.. H., Logghe, S., 2002. Traffic Flow Theory. Kathol. Univ. Leuven 35.

Ishak, S., Al-Deek, H., 2003. Statistical Evaluation of Interstate 4 Traffic Prediction System. Transp. Res. Rec. J. Transp. Res. Board 1856, 16–24. https://doi.org/10.3141/1856-03

Jabari, S.E., Liu, H.X., 2013. A stochastic model of traffic flow: Gaussian

approximation and estimation. Transp. Res. Part B Methodol. 47, 15–41. https://doi.org/10.1016/j.trb.2012.09.004

Jin, K., Wi, J., Lee, E., Kang, S., Kim, S., Kim, Y., 2021. TrafficBERT: Pre-trained model with large-scale data for long-range traffic flow forecasting. Expert Syst. Appl. 186, 115738. https://doi.org/10.1016/j.eswa.2021.115738

Jolliffe, I., 2011. Principal Component Analysis, in: Lovric, M. (Ed.), International Encyclopedia OfStatistical Science. Springer, Berlin, Heidelberg, pp. 1094–1096.

Kamarianakis, Y., Prastacos, P., 2005. Space–time modeling of traffic flow. Comput. Geosci. 31, 119–133. https://doi.org/10.1016/j.cageo.2004.05.012

Kamnik, R., Nekrep Perc, M., Topolšek, D., 2020. Using the scanners and drone for comparison of point cloud accuracy at traffic accident analysis. Accid. Anal. Prev. 135. https://doi.org/10.1016/j.aap.2019.105391

Kawasaki, Y., Hara, Y., Kuwahara, M., 2017. Real-time Monitoring of Dynamic Traffic States by State-Space Model. Transp. Res. Procedia 21, 42–55. https://doi.org/10.1016/j.trpro.2017.03.076

Kirby, H.R., Watson, S.M., Dougherty, M.S., 1997. Should we use neural networks or statistical models for short-term motorway traffic forecasting? Int. J. Forecast. 13, 43–50. https://doi.org/10.1016/S0169-2070(96)00699-1

Kolidakis, S., Botzoris, G., Profillidis, V., Lemonakis, P., 2019. Road traffic forecasting — A hybrid approach combining Artificial Neural Network with Singular Spectrum Analysis. Econ. Anal. Policy 64, 159–171. https://doi.org/10.1016/J.EAP.2019.08.002

Kumar, S.V., Vanajakshi, L., 2015. Short-term traffic flow prediction using seasonal ARIMA model with limited input data. Eur. Transp. Res. Rev. 7, 1–9. https://doi.org/10.1007/s12544-015-0170-8

Kwon, J., Coifman, B., Bickel, P., 2000. Day-to-day travel-time trends and travel-time prediction from loop-detector data. Transp. Res. Rec. 120–129.

Laña, I., Olabarrieta, I. (Iñaki), Vélez, M., Del Ser, J., 2018a. On the imputation of missing data for road traffic forecasting: New insights and novel techniques. Transp. Res. Part C Emerg. Technol. 90, 18–33. https://doi.org/10.1016/j.trc.2018.02.021

Laña, I., Ser, J. Del, Member, S., Vélez, M., Vlahogianni, E.I., 2018b. Road Traffic Forecasting : Recent Advances and New Challenges. IEEE Intell. Transp. Syst. Mag. 93–109. https://doi.org/10.1109/MITS.2018.2806634

Le, Q. V, 2015. A Tutorial on Deep Learning Part 2: Autoencoders, Convolutional Neural Networks and Recurrent Neural Networks. Tutorial 1–20.

Levin, M., Tsao, Y.-D., 1980. On forecasting freeway occupancies and volumes (abridgment). Transp. Res. Rec.

Li, L., Su, X., Zhang, Y., Hu, J., Li, Z., 2014. Traffic prediction, data compression, abnormal data detection and missing data imputation: An integrated study based on the decomposition of traffic time series, in: 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). pp. 282–289. https://doi.org/10.1109/ITSC.2014.6957705

Lippi, M., Bertini, M., Frasconi, P., 2013. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. IEEE Trans. Intell. Transp. Syst. 14, 871–882. https://doi.org/10.1109/TITS.2013.2247040

Lockwood, S., Auer, A., Feese, S., 2016. History of Intelligent Transportation Systems.

Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.Y., 2015. Traffic Flow Prediction with Big Data: A Deep Learning Approach. IEEE Trans. Intell. Transp. Syst. 16, 865–873. https://doi.org/10.1109/TITS.2014.2345663

Ma, X., Yu, H., Wang, Yunpeng, Wang, Yinhai, 2015. Large-Scale Transportation Network Congestion Evolution Prediction Using Deep Learning Theory. PLoS One 10, e0119044. https://doi.org/10.1371/journal.pone.0119044

Mantecchini, L., 2011. Analysis of Traffic Flow Variability on Two-Lane Highways. Contemp. Eng. Sci. 4, 43–53.

McAleer, S.R., Kogut, P., Raes, L., 2018. The Case for Collaborative Policy Experimentation Using Advanced Geospatial Data Analytics and Visualisation, in: Diplaris, S., Satsiou, A., Følstad, A., Vafopoulos, M., Vilarinho, T. (Eds.), Internet Science. Springer International Publishing, Cham, pp. 137–152.

Miglani, A., Kumar, N., 2019. Deep learning models for traffic flow prediction in

autonomous vehicles: A review, solutions, and challenges. Veh. Commun. 20, 100184. https://doi.org/https://doi.org/10.1016/j.vehcom.2019.100184

Min, W., Wynter, L., Amemiya, Y., 2007. Road Traffic Prediction with Spatio-Temporal Correlations, IBM Research Report. New York, New York, USA.

Ministerio de Política Territorial y Función Pública, Ministerio de Economía y Empresa, 2018. datos.gob.es [WWW Document]. URL http://datos.gob.es/es (accessed 11.15.18).

Muspratt, G.R., 2011. Saving money on vehicle detection: Reducing lifetime costs for SCOOT and MOVA via smart vehicle detection 6.

Negrete, J.M., Subirana, J.C., 2012. Trayectoria de la implementación de la Directiva INSPIRE en España, in: III Jornadas Ibéricas de Infraestructura de Datos Espaciales. Madrid.

OECD/ITF, 2015. Big Data and Transport, Oecd/Itf. https://doi.org/10.1002/ajh.23643.

Okutani, I., Stephanedes, Y.J., 1984. Dynamic prediction of traffic volume through Kalman filtering theory. Transp. Res. Part B Methodol. 18, 1–11. https://doi.org/10.1016/0191-2615(84)90002-X

Pamuła, T., 2018. Classification and Prediction of Traffic Flow Based on Real Data Using Neural Networks. IEEE Trans. Intell. Transp. Syst. 24, 1–10. https://doi.org/10.2478/v10174-012-0032-2

Peng, H., Wang, H., Du, B., Bhuiyan, M.Z.A., Ma, H., Liu, J., Wang, L., Yang, Z., Du, L., Wang, S., Yu, P.S., 2020. Spatial temporal incidence dynamic graph neural networks for traffic flow forecasting. Inf. Sci. (Ny). 521, 277–290. https://doi.org/10.1016/J.INS.2020.01.043

Petridis, V., Kehagias, A., Petrou, L., Bakirtzis, A., Kiartzis, S., Panagiotou, H., Maslaris, N., 2001. A Bayesian Multiple Models Combination Method for Time Series Prediction. J. Intell. Robot. Syst. 31, 69–89. https://doi.org/10.1023/A:1012061814242

Polson, N.G., Sokolov, V.O., 2017. Deep learning for short-term traffic flow prediction. Transp. Res. Part C Emerg. Technol. 79, 1–17. https://doi.org/10.1016/j.trc.2017.02.024

Qiu, H., Li, R., Liu, H., 2016. Integrated model for traffic flow forecasting under rainy conditions. J. Adv. Transp. 50, 1754–1769. https://doi.org/10.1002/atr.1427

Qu, L., Li, L., Zhang, Y., Hu, J., 2009. PPCA-based missing data imputation for traffic flow volume: A systematical approach. IEEE Trans. Intell. Transp. Syst. 10, 512–522. https://doi.org/10.1109/TITS.2009.2026312

Roncoli, C., Papamichail, I., Papageorgiou, M., 2016. Hierarchical model predictive control for multi-lane motorways in presence of Vehicle Automation and Communication Systems. Transp. Res. Part C Emerg. Technol. 62, 117–132. https://doi.org/10.1016/j.trc.2015.11.008

Rosenblatt, F., 1957. The Perceptron - A Perceiving and Recognizing Automaton, Report 85, Cornell Aeronautical Laboratory. https://doi.org/85-460-1

Ruiz-Alarcon-Quintero, C., 2016. Harmonization of Transport Data Sources According to INSPIRE Data Specification on Transport Networks. Transp. Res. Procedia 18, 320–327. https://doi.org/10.1016/J.TRPRO.2016.12.043

Russell, S.J., Norvig, P., 2010. Artificial intelligence: a modern approach, 3rd ed. Pearson, Essex. https://doi.org/10.1016/B978-012161964-0/50009-1

Sastre García, D., Torres Arjona, J., Menéndez García, J.M., 2011. Sistemas de adquisición de información de tráfico: Estado actual y futuro (No. 1), 2011. Plataforma Tecnológica Española de la Carretera (PTC)., Madrid.

Schimbinschi, F., Nguyen, X.V., Bailey, J., Leckie, C., Vu, H., Kotagiri, R., 2015. Traffic Forecasting In Complex Urban Networks : Leveraging Big Data and Machine Learning, in: 2015 IEEE International Conference on Big Data (Big Data). pp. 1019–1024.

Shilton, S.J., Anfosso Lédée, F., Van Leeuwen, H., 2015. Conversion of existing road source data to use CNOSSOS-EU, in: Euronoise 2015. p. 6.

Smith, B.L., Demetsky, M.J., 1994. Short-Term Traffic Flow Prediction: Neural Network Approach. Transp. Res. Rec. 98–104.

Smith, B.L., Williams, B.M., Keith Oswald, R., 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. Transp. Res. Part C Emerg. Technol. 10, 303–321. https://doi.org/10.1016/S0968-090X(02)00009-8

Solomatine, D., See, L.M., Abrahart, R.J., 2008. Data-Driven Modelling : Concepts , Approaches and Experiences.

Stathopoulos, A., Dimitriou, L., Tsekeris, T., 2008. Fuzzy Modeling Approach for Combined Forecasting of Urban Traffic Flow. Comput. Civ. Infrastruct. Eng. 23, 521–535. https://doi.org/10.1111/j.1467-8667.2008.00558.x

Stathopoulos, A., Karlaftis, M.G., 2003. A multivariate state space approach for urban traffic flow modeling and prediction. Transp. Res. Part C Emerg. Technol. 11, 121–135. https://doi.org/10.1016/S0968-090X(03)00004-4

Steenberghen, T., Pourbaix, J., Moulin, A., Bamps, C., Keijers, S., 2013. Study on harmonised collection of European data and statistics in the field of urban transport and mobility.

Subdirección General de la Gestión de la Movilidad y Tecnología, 2020. DATEX II. Guía de Utilización. (No. 1.5). Madrid.

Sun, S., Zhang, C., 2007. The Selective Random Subspace Predictor for Traffic Flow Forecasting. IEEE Trans. Intell. Transp. Syst. 8, 367–373. https://doi.org/10.1109/TITS.2006.888603

Sun, S., Zhang, C., Yu, G., 2006. A bayesian network approach to traffic flow forecasting. IEEE Trans. Intell. Transp. Syst. 7, 124–132. https://doi.org/10.1109/TITS.2006.869623

Tampere, C.M.J., Immers, L.H., 2007. An Extended Kalman Filter Application for Traffic State Estimation Using CTM with Implicit Mode Switching and Dynamic Parameters, in: 2007 IEEE Intelligent Transportation Systems Conference. IEEE, pp. 209–216. https://doi.org/10.1109/ITSC.2007.4357755

Tan, P.N., Steinbach, M., Kumar, V., 2005. Chap 8 : Cluster Analysis: Basic Concepts and Algorithms, in: Introduction to Data Mining. p. Chapter 8. https://doi.org/10.1016/0022-4405(81)90007-8

Tarko, A.P., Perez-Cartagena, R.I., 2005. Variability Of A Peak Hour Factor At Intersections, in: Annual Meeting of the Transportation Research Board. Washington D.C., pp. 1–20.

The Highways Agency, The Scottish Office Development Department, The Welsh Office, The Department of Environment for Northern Ireland, 1994. Design Manual for Roads and Bridges, Motorway Incident Detection and Automatic

Signalling (MIDAS).

Tian, Y., Pan, L., 2015. Predicting Short-Term Traffic Flow by Long Short-Term Memory Recurrent Neural Network. 2015 IEEE Int. Conf. Smart City/SocialCom/SustainCom 153–158. https://doi.org/10.1109/SmartCity.2015.63

Tsekeris, T., Stathopoulos, A., 2006. Measuring Variability in Urban Traffic Flow by Use of Principal Component Analysis. J. Transp. Stat. 9, 49–62.

Tselentis, D.I., Vlahogianni, E.I., Karlaftis, M.G., 2015. Improving short-term traffic forecasts: To combine models or not to combine? IET Intell. Transp. Syst. 9, 193–201. https://doi.org/10.1049/iet-its.2013.0191

Tucker, S., Summersgill, I., Fletcher, J., Mustard, D., 2006. Evaluating the benefits of MIDAS automatic queue protection. Traffic Eng. Control 47, 370–373.

Van Der Voort, M., Dougherty, M.S., Watson, S., 1996. Combining kohonen maps with arima time series models to forecast traffic flow. Transp. Res. Part C Emerg. Technol. 4, 307–318. https://doi.org/10.1016/S0968-090X(97)82903-8

van Lint, J.W.C., Hoogendoorn, S.P., van Zuylen, H.J., 2005. Accurate freeway travel time prediction with state-space neural networks under missing data. Transp. Res. Part C Emerg. Technol. 13, 347–369. https://doi.org/10.1016/j.trc.2005.03.001

Vázquez, J.J., Arjona, J., Casanovas-Garcia, J., 2020. A Comparison of Deep Learning Methods for Urban Traffic Forecasting using Floating Car Data. Transp. Res. Procedia 47, 195–202. https://doi.org/10.1016/J.TRPRO.2020.03.079

Veeckman, C., Jedlička, K., De Paepe, D., Kozhukh, D., Kafka, Š., Colpaert, P., Čerba, O., 2017. Geodata interoperability and harmonization in transport: a case study of open transport net. Open Geospatial Data, Softw. Stand. 2, 3. https://doi.org/10.1186/s40965-017-0015-6

Vlahogianni, E.I., Golias, J.C., Karlaftis, M.G., 2004. Short term traffic forecasting: Overview of objectives and methods. Transp. Rev. 24, 533–557. https://doi.org/10.1080/0144164042000195072

Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2015. Recurrence analysis applications to short-term macroscopic and microscopic road traffic 375–397.

https://doi.org/10.1007/978-3-319-07155-8-13

Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2014. Short-term traffic forecasting: Where we are and where we're going. Transp. Res. Part C Emerg. Technol. 43, 3–19. https://doi.org/10.1016/j.trc.2014.01.005

Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2005. Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach. Transp. Res. Part C Emerg. Technol. 13, 211–234. https://doi.org/10.1016/j.trc.2005.04.007

Waibel, Alexander, Hanazawa, T., Hinton, G., Shikano, K., Lang, K.J., 1990. Phoneme Recognition Using Time-Delay Neural Networks, in: Waibel, Alex, Lee, K.-F. (Eds.), Readings in Speech Recognition. Morgan Kaufmann, San Francisco, pp. 393–404. https://doi.org/https://doi.org/10.1016/B978-0-08-051584-7.50037-1

Wang, J., Deng, W., Guo, Y., 2014. New Bayesian combination method for short-term traffic flow forecasting. Transp. Res. Part C Emerg. Technol. 43, 79–94. https://doi.org/10.1016/j.trc.2014.02.005

Whitlock, M.E., Queen, C.M., 2000. Modelling a traffic network with missing data. J. Forecast. 19, 561–574. https://doi.org/https://doi.org/10.1002/1099-131X(200012)19:7<561::AID-FOR785>3.0.CO;2-4

Williams, B., Durvasula, P., Brown, D., 1998. Urban Freeway Traffic Flow Prediction: Application of Seasonal Autoregressive Integrated Moving Average and Exponential Smoothing Models. Transp. Res. Rec. J. Transp. Res. Board 1644, 132–141. https://doi.org/10.3141/1644-14

Williams, B.M., Hoel, L. a., 2003. Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results. J. Transp. Eng. 129, 664–672. https://doi.org/10.1061/(ASCE)0733-947X(2003)129:6(664)

World Customs Organization, 2007. WCO Data Model, Single Window Data Harmonisation.

Wu, Y., Tan, H., 2016. Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework 1–14.

Xiao, X., Chen, Y., Yuan, Y., 2015. Estimation of missing flow at junctions using

control plan and floating car data. Transp. Res. Procedia 10, 113–123. https://doi.org/10.1016/j.trpro.2015.09.061

Yangzhou Chen, Jiang Luo, Wei Li, E.Z. and J.S., 2014. Real Time Traffic Speed Variability Modeling and Prediction. CICTP 2014 Safe, Smart, Sustain. Multimodal Transp. Syst. ASCE 2014 3743–3751.

Yin, H., Wong, S.C., Xu, J., Wong, C.K., 2002. Urban traffic flow prediction using a fuzzy-neural approach. Transp. Res. Part C Emerg. Technol. 10, 85–98. https://doi.org/10.1016/S0968-090X(01)00004-3

Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., Chen, C., 2011. Data-Driven Intelligent Transportation Systems: A Survey. IEEE Trans. Intell. Transp. Syst. 12, 1624–1639. https://doi.org/10.1109/TITS.2011.2158001

Zhang, Y., 2012. How to Provide Accurate and Robust Traffic Forecasts Practically? Intell. Transp. Syst. 19. https://doi.org/10.5772/27741

Zhang, Y., 2011. Hourly traffic forecasts using interacting multiple model (IMM) predictor. IEEE Signal Process. Lett. 18, 607–610. https://doi.org/10.1109/LSP.2011.2165537

Zheng, W., Lee, D.-H., Shi, Q., 2006. Short-Term Freeway Traffic Flow Prediction: Bayesian Combined Neural Network Approach. J. Transp. Eng. 132, 114–121. https://doi.org/10.1061/(ASCE)0733-947X(2006)132:2(114)

Zhong, M., Lingras, P., Sharma, S., 2004. Estimation of missing traffic counts using factor, genetic, neural, and regression techniques. Transp. Res. Part C Emerg. Technol. 12, 139–166. https://doi.org/10.1016/J.TRC.2004.07.006

Zhu, J.Z., Cao, J.X., Zhu, Y., 2014. Traffic volume forecasting based on radial basis function neural network with the consideration of traffic flows at the adjacent intersections. Transp. Res. Part C Emerg. Technol. 47, 139–154. https://doi.org/10.1016/j.trc.2014.06.011

Zhu, Y., Gao, N., Wang, J., Liu, C., 2016. Study on Traffic Flow Patterns Identification of Single Intersection Intelligent Signal Control. Procedia Eng. 137, 452–460. https://doi.org/10.1016/j.proeng.2016.01.280

APÉNDICE I – PUBLICACIONES

Ruiz-Alarcon-Quintero, C., 2016. Harmonization of Transport Data Sources According to INSPIRE Data Specification on Transport Networks. Transp. Res. Procedia 18, 320–327. https://doi.org/10.1016/J.TRPRO.2016.12.043

Nocera, S., Ruiz-Alarcón-Quintero, C., Cavallaro, F., 2018. Assessing carbon emissions from road transport through traffic flow estimators. Transp. Res. Part C Emerg. Technol. 95, 125–148. https://doi.org/10.1016/J.TRC.2018.07.020

APÉNDICE II - GLOSARIO

AADT - Average Annual Daily Traffic.

ANN - Artificial Neural Network.

AI o IA - Artificial Intelligence o Inteligencia Artificial.

API - Application Programming Interface.

ARIMA - Auto Regressive Integrated Moving Average.

ATHENA - Aproximación de modelado híbrido basada en una primera fase de clasificación que utiliza el método k-means para separar la muestra.

AVI - AVerage Imputation.

BCNN - Bayesian Combination Neural Network.

BCP - Bayesian Combination Predictor.

BPNN - Back-Propagation Neural Network.

CAD - Computer-aided design.

CALTrans - CALifornia Department of TRANSportation

CC - Conjunto Completo.

CDE - Conjuntos de Datos Espaciales.

CE - Comisión Europea.

CLTFP - Convolutional Long-Term Neural Network.

CNN - Convolutional Neural Network.

CP-WOPT - Canonical Polyadic - Weighted OPtimization.

DATEX II - DAta EXchange standard.

DGT - Dirección General de Tráfico.

DHV - Design Hour Volume.

DL - Deep Learning.

DLI - Deep Learning Imputation,

DLP - Deep Learning Prediction

FNM - Fuzzy Neural Model.

FRBS - Fuzzy-Rule Based System.

GA - Genetic Algorithms.

GIS - Geographic Information System.

GRNN - General Regression Neural Network.

GRCNN - Graph Recurrent Convolutional Neural Network.

HA - Historical Average.

HALE - Humboldt ALignment Editor.

Html - HyperText Markup Language.

IDE - Infraestructura de Datos Espaciales.

IGN - Instituto Geográfico Nacional.

IMD - Intensidad Media Diaria.

INSPIRE - INfraestructure for SPatial InfoRmation in Europe.

IoT - Internet of Things.

ITS - Intelligent Transport System o Sistema Inteligente de Transporte.

KARIMA - KNN + ARIMA.

KF - Kalman Filter.

KNN - K Nearest Neighbours.

LRTC - Low Range Tensor Decomposition.

LS - Least Squares.

LSTMNN - Long Short-Term Memory Neural Network.

MCR - Missing Completely at Random.

MD - Missing Data.

MIDAS - Motorway Incident and Automatic Signalling.

ML - Machine Learning o Aprendizaje Automático.

MLP - Multi Layer Perceptron.

MNR - Missing Not at Random.

MONICA - Sistema de monitorización de las autovías de los Países bajos.

MR - Missing at Random.

MSTAR - Multivariate State-Space Auto Regressive.

NA - Not A Number.

NN - Neural Network o Red Neuronal.

OL-SVR - On-Line Support Vector Regression.

PCA - Principal Components Analysis.

PDF - Portable Document Format.

PeMS - Performance Measurement System of CalTRANS.

PHP - Hypertext PreProcessor.

PoliVisu - Policy & Data Results Hub.

PPCA - Probabilistic Principal Component Analysis.

RBF - Radial Basis Function.

RBFNN - Radial Basis Function Neural Network.

RCA - Red de Carreteras de Andalucía.

RF - Random Forest o Bosque Aleatorio.

RFI - Random Forest Imputation.

RFP - Random Forest Prediction.

RMSE - Root Mean Square Error.

RW - Random Walk.

SAE - Stacked Auto-Encoder.

SARIMA - Seasonal ARIMA.

SC - SubConjuntos.

SOM - Self Organized Map.

SSA - Singular Spectral Analysis.

SVM - Support Vector Machine.

TDI - Tensor Decomposition Imputation.

TDNN - Time Delay Neural Network.

TII - Traffic Infraestructure of Ireland.

UE - Unión Europea.

VBPCA - Variational Bayesian PCA.

VHT - Vehicle Hours Traveled.

VMT - Vehicle Miles Traveled.

XML - Extensible Markup Language.