Universidad del País Vasco
Euskal Herriko Unibertsitatea

Máster Universitario en

Ingeniería **C**omputacional y **S**istemas **I**nteligentes

Konputazio Zientziak eta Adimen Artifiziala Saila –
Departamento de Ciencias de la Computación e Inteligencia Artificial

K I S A

I C S I

Tesis de Máster

Análisis de modos de conducción en líneas regulares de autobús hacia una conducción eficiente

Ainara Agundez Arnaez

Tutor(a/es)

Yosu Yurramendi
Departamento de Ciencia de la Computación e Inteligencia Artificial
Facultad de Informática

Iñigo Etxabe
Co-fundador y CEO de Datik

informatika fakultatea
facultad de informática

KZAA /CCIA

Septiembre 2015

Resumen

El presente trabajo analiza la eficiencia de la conducción, comprendida en términos de consumo de combustible y las diferencias entre conductores, medidas en un autobús de la compañía Bizkaibus en Vizcaya. El trabajo se realiza a través del análisis multivariante de datos. Los principales métodos de exploración de datos que se han utilizado han sido la regresión (lineal y por mínimos cuadrados parciales) y los árboles de regresión.

Los datos provienen de un único autobús, contando para ello con un total de 8 variables. Los datos corresponden al periodo comprendido entre el 10 de diciembre de 2014 y el 30 de marzo de 2015. El análisis se realizó en el lenguaje R, haciendo uso de algunas de sus librerías para diferentes propósitos.

Índice

1. Introducción

1.1. Objetivos de la tesis

Actualmente, existen plataformas de gestión de flotas mediante las cuales las compañías pueden sensorizar sus vehículos y conocer, en cada momento, el estado de cada uno de ellos, así como otros muchos datos relacionados con la conducción en dicho instante.

Se quiere añadir y, de alguna forma, completar estas plataformas mediante un sistema consejero, para que, además de poder controlar la situación del vehículo, sea posible conocer las costumbres de conducción de los conductores, e intentar corregir algunos de los malos hábitos que hacen que el consumo se eleve por encima de lo necesario.

Esto requiere un gran trabajo a largo plazo; es importante ser conscientes de que este tema lleva siendo investigado desde hace unas décadas, pero en los últimos años ha habido un aumento en el interés por ello en todo el mundo, como veremos más adelante. Por ello, lo que se busca mediante este trabajo es, además de obtener una visión general, por una parte, evaluar los *pros* y los *contras* de dispositivos ya existentes, y por otra, proponer un nuevo enfoque que sirva para evaluar estas conductas en lo referente al gasto o consumo que conllevan.

1.2. Marco del estudio

El proyecto se ha realizado en la empresa tecnológica Datik en colaboración con la Euskal Herriko Unibertsitatea (UPV/EHU). Datik es una empresa ubicada en el Parke Teknologikoa de Miramón (Donostia) que desarrolla soluciones ITS (Intelligent Transport Systems) destinados a la gestión del transporte, tanto ferroviario como por carretera, y la movilidad ciudadana. Creada en 2008 bajo la denominación Innovate and Transport, pertenece desde el año 2011 al Grupo Irizar.

Desde sus inicios, se han dedicado a la creación de varios dispositivos de seguridad para vehículos (sistemas que detectan peatones u otro tipo de objetos en la calzada, detectores de fatiga, etc), y actualmente dan cobertura a diversas compañías de autobuses, desde algunas en Euskadi y Cataluña, hasta en Polonia, México o Chile.

Además de equipar los autobuses de dichas compañías según la demanda, también se ofrece un servicio de seguimiento y de análisis de datos de los vehículos. Para ello, disponen de iPanel, una plataforma de gestión de flotas que, mediante los datos recogidos en los autobuses y los dispositivos de comunicación, permite conocer el estado del vehículo, la puntualidad con la que se prestan los servicios, el consumo de combustible, la velocidad o las revoluciones por minuto (rpm). Esta plataforma también permite mostrar a los responsables de las flotas, mediante unos umbrales prefijados, las aceleraciones o frenadas bruscas realizadas por sus conductores, así como tiempos excesivos en ralentí o el porcentaje de tiempo conducido a rpm altas.

El siguiente nivel de investigación les lleva a preguntarse si es posible encontrar, entre toda esta cantidad de datos, distintos patrones de conducción, que se clasifiquen por el consumo que generan. En caso afirmativo, podría crearse un modelo de conducción óptima o, al menos, definir unas líneas de hábitos o costumbres que hacen que ésta sea más eficiente. Y de ahí surge la propuesta de esta tesis, la necesidad de un análisis más a fondo de la información disponible, la factibilidad del futuro sistema consejero.

1.3. Distribución de la tesis

El trabajo se ha desarrollado en tres etapas: en la primera, la cual se presenta en la sección 3, se ha realizado un amplio análisis del estado del arte, pues se cree necesario para un conocimiento global de lo existente hasta ahora y sobre todo para la hora de definir las líneas de investigación a seguir.

En la sección 5 se presenta la réplica a un estudio previo a éste, que servirá como iniciativa para la realización de lo que sigue, ya que se han analizado tanto los aspectos positivos como los negativos resultantes, y serán de ayuda para el planteamiento de lo presente.

La tercera etapa corresponde al trabajo realizado puramente con los objetivos mencionados anteriormente. Para terminar, en la sección 8 se presentan las conclusiones y el trabajo futuro que queda por realizar.

2. Motivación

Encontrar patrones en secuencias es una importante área de investigación del descubrimiento del conocimiento (*knowledge discovery)* y el *data-mining*. El CAN genera una cantidad inmensa de datos sobre eventos del motor, pero todavía no se ha abordado el tema de la interpretación de todos estos datos en bruto, por lo que puede ser un área de investigación futura **[1].**

Aunque en el pasado fuera pequeño, en los últimos años ha crecido el interés por la calidad de conducción. Se ha demostrado que reducir la velocidad no es la única, ni la mejor estrategia para el eco-driving, que hay muchos factores que afectan, de alguna u otra manera, en el consumo y las emisiones producidas por los vehículos.

El futuro consejero que se pretende crear debe ser capaz de adaptarse a las características y hábitos de distintos conductores, ya que una vez instalado en un vehículo, será el responsable de dar los consejos oportunos, en tiempo real, a cualquier persona que se siente en él. Sobre todo, se quiere que éste sea un sistema autónomo y funcione sin ninguna influencia externa (como puede ser el establecer unos límites para la hora de tomar de

decisiones), ya que se cree que estas acciones, no contribuyen, sino que perjudican el funcionamiento de los sistemas.

3. Estado del arte

La global crisis financiera ha reducido y, en algunos casos recortado la disponibilidad de recursos. Es por ello que en el sector del transporte, las compañías de servicios públicos, las comunidades y las autoridades públicas han demostrado creciente interés en el uso de diversas tácticas para reducir el uso de autobuses no óptimos, consumos y el número de conductores ineficientes. En lo relativo al consumo de combustible de autobuses, la constante subida de precios de éste hace que sea necesario para las compañías de transporte priorizar en la optimización del mismo, en lo que a consumo y emisiones respecta. Las emisiones relativas al transporte de carretera computan un porcentaje importante del total de contaminantes del aire; en Europa, más del 30% del CO_2 es producido por los sistemas de transporte (Air Quality in Europe AEMA, 2011), por lo que la Agencia Europea de Medio Ambiente (AEMA) marca objetivos a corto, medio y largo plazo **[2].**

Por tanto, podemos decir que el aumento del precio del carburante y de las emisiones de CO_2 y el hecho de que las reservas de petróleo no dejen de disminuir, así como el elevado número de muertes en las carreteras cada año, ha acelerado el desarrollo de tecnologías en el sector de la automoción en el último par de décadas. Hoy en día, mercados automovilísticos de países muy industrializados, Alemania por ejemplo, ya pronuncian expresiones como 'Visión Cero' **[1]**, la cual se identifica con la imagen de un futuro donde no habrá muertos ni heridos graves en accidentes de tráfico.

Pero poco hincapié se ha hecho en la mejora de las costumbres de los conductores para disminuir el consumo de combustible. ¿Cómo se pueden resolver estos problemas, y a cuánto ascienden los beneficios que se pueden obtener? La respuesta está en los cambios a introducir, los cuales se pueden aplicar a tres niveles: gestión del tráfico, vehículo y conductor.

En lo que a la gestión del tráfico respecta, medidas como la reducción del máximo número de pasajeros permitidos en los servicios, o la optimización de autobuses y rutas según la demanda pueden conllevar un aumento en seguridad y reducción en consumo **[3.1]**. Las soluciones aplicadas al vehículo consideran implementar tecnologías alternativas, instalar nuevos equipamientos o cambiar el tipo de combustible, entre otros **[3].** Estas dos grandes familias de aportaciones basadas en la reducción del consumo y de emisiones excesivas necesitan tiempo y dinero para ser adoptadas y llevadas a la práctica por los fabricantes, y además, se estima que el impacto y los beneficios de éstas pueden tardar en llegar hasta 10 años **[4.1].**

A la vez que se siguen desarrollando y mejorando nuevas tecnologías para los vehículos, se está haciendo más y más evidente que uno de los últimos factores influyentes en el ahorro de combustible es el conductor **[5]**. Por ello, durante los últimos años, se ha puesto creciente interés en las conductas de conducción, sobre todo en los efectos de la educación, entrenamiento y *feedback* en el cambio de comportamientos **[3].** Además, centrarse en el comportamiento de los conductores ayuda a la reducción del consumo a corto plazo **[6.1]**, de

una forma mucho más económica, pero queda por determinar la fiabilidad y la calidad de estos métodos. El estudio en **[6]** muestra un ahorro de combustible del 10-15% de media (y hasta un 25-60% de máximo) obtenido mediante la modificación del estilo de conducción, mientras que el dispositivo empleado en **[4]** ahorra un 7.6% de media de combustible (siendo el máximo de 12%).

Una ventaja añadida de algunas de estas técnicas es que la mayor parte de las mejoras obtenidas mediante los cambios de comportamiento será preservada en el tiempo. Y decimos algunas, ya que diversos estudios revelan efectos positivos a largo plazo **[6.2]**, pero otros indican tanto que unos conductores parcialmente recaen en sus viejas costumbres **[6.3]-[6.4]** como que otros lo hacen del todo **[6.5].** Por ello, se están desarrollando distintos métodos para la reducción del consumo del carburante, basados en mejorar tanto el comportamiento como la atención del conductor durante la conducción. En **[7]** y **[8],** por ejemplo, se presentan sistemas automatizados para la detección y predicción de distracciones y faltas de concentración, para mejorar la atención al volante, y Datik también posee un sistema de detección de fatiga para los autobuses, pero aquí nos centraremos únicamente en el análisis de conductas y hábitos.

3.1 Técnicas y dispositivos

Siguiendo la clasificación propuesta en **[9]**, se diferencian tres tipos de técnicas, en tres épocas distintas. Ya en los años setenta, se empezaron a tomar las primeras medidas para reducir los costes de movilidad, como por ejemplo, dispositivos de seguridad pasiva de los pasajeros (*passive passenger safety concepts*), vehículos con bajas emisiones o rutas de tráfico más seguras. Pero para poder disminuir el efecto negativo de las movilizaciones masivas en la economía, medioambiente y sociedad, era necesario idear nuevas estrategias.

El segundo concepto a tratar es la electrificación de vehículos, con el objetivo de reducir costes ecológicos y contrarrestar la cada vez mayor escasez de recursos. A pesar de las soluciones aportadas por los vehículos eléctricos (VE) en respuesta a las desventajas de los vehículos ordinarios, la electrificación en masa está todavía abierta a disputas; se puede tratar como posible a largo plazo, pero no como una opción en un futuro cercano.

De todas formas, todo parece indicar que éstos son una prometedora alternativa a los vehículos de combustión interna. La introducción gradual de vehículos completamente eléctricos en el mercado, nos permite acercarnos a una era de movilidad sostenible, en la que se puede reducir significantemente la dependencia de combustibles fósiles y minimizar las emisiones de tráfico de carretera. Y si la energía empleada para construir y operar los VE es proporcionada por fuentes de energía renovable, se puede decir que la movilidad eléctrica será 100% libre de emisiones **[10]**. Aunque la mayoría de los estudios se hayan centrado en investigar y crear modelos en relación con vehículos de combustión interna, actualmente existen algunos relacionados con los VE **[10]-[12].**

El tercer tipo a tratar es el concepto de asistencia al conductor. Los sistemas de asistencia al conductor (*driver assistance systems o DAS*) son un medio para mejorar, entre otras cosas, la seguridad activa e integrada del vehículo. Hace unos años, estos dispositivos apenas existían y no eran conocidos, por lo que para poder maximizar su contribución a la seguridad del tráfico en general, era necesario incrementar su presencia en el mercado **[9.1],** concienciando a la población de sus beneficios o regulando su implementación por ley. Dos

ejemplos de esto último son la ley que obliga, desde noviembre de 2013, a implantar a todos los nuevos camiones de la Unión Europea un sistema de mitigación de colisiones (EC No 661/2009) y la que exige la presencia de algún sistema de control electrónico de estabilidad (*electronic stability control systems* o ESC) en todos los nuevos vehículos producidos en la Unión Europea, desde noviembre del 2014.

Los primeros DAS se basaron en sensores propioceptivos **[9],** esto es, sensores que miden el estatus interno del vehículo, como pueden ser la velocidad de las ruedas, la aceleración o la velocidad de rotación. Esto permite el control de la dinámica del vehículo, para poder seguir la trayectoria solicitada por el conductor de la mejor manera posible. Uno de los primeros sistemas de asistencia activa basado en sensores propioceptivos, fue el sistema antibloqueo de ruedas (*Anti-lock Braking System* o ABS), con producción en serie desde 1978 (Bosch). El control de tracción (*Traction Control System* o TCS) amplió, más adelante, el sistema. Unos años más tarde, en 1995, la introducción de controles de conducción dinámicos adicionales, como el ESC, marcó un hito en el desarrollo de la asistencia. Gracias al ESC, el giroscopio electrónico se hizo su hueco en el automóvil. Éste no solo sentó las bases del ESC, sino que revelaba todo un rango de posibles aplicaciones futuras. En lo que a seguridad vial se refiere, algunos estudios han demostrado que los controles dinámicos de conducción son los segundos sistemas de seguridad más eficientes para los pasajeros, superados solo por el cinturón de seguridad **[9.2]-[9.3]**, lo cual fue demostrado en el “Moose Test” del Mercedes Clase-A.

El reconocimiento público de las capacidades de los sistemas de control dinámico de conducción en lo que a seguridad se refiere, hizo que la frecuencia de implementación de estos sistemas creciera significantemente. Gracias a ello se han salvado miles de vidas y como ya hemos mencionado, desde noviembre de 2014 es un requerimiento legal para todos los nuevos coches en la UE.

Los sensores exteroceptivos adquieren información externa al vehículo, como por ejemplo los sensores ultrasónicos, radar, lidar o de vídeo, y en cierta medida, también los receptores GNSS (Global Navigation Satellite System). Estos sensores proporcionan información sobre la carretera, la presencia y el estado de otros participantes en el tráfico, o sobre la posición del vehículo en el mundo. En **[9.4]** se expone la evolución internacional del área de los sistemas de navegación en relación con los progresos en las tecnologías de posicionamiento.

La segunda generación de funciones de asistencia para la conducción, introducidos por primera vez sobre 1990 y basados en sensores exteroceptivos, se centra en proveer información y alertas, así como en mejorar el confort durante la conducción. Sustancialmente impulsado por la reducción del coste de las tecnologías de navegación de los dispositivos móviles, el uso del GNSS se ha convertido prevalente en los vehículos de hoy en día. Ayudándolo a orientarse, los sistemas de navegación tienen el potencial para reducir la carga de trabajo del conductor, dejando así más recursos disponibles para dedicarlos a la pura tarea de conducir, reduciendo el peligro de accidentes por falta de atención. Adicionalmente, algunos estudios indican las ventajas de la conducción preventiva y del dirigirse al destino por la ruta óptima en el ahorro de combustible **[9.5]-[9.8].** Con la mejora de precisión de estos sistemas en el futuro, se cree posible concebir más efectos beneficiosos mediante la provisión de datos de localización a los sistemas de soporte, permitiendo que sus funciones se adapten mejor a las condiciones de conducción **[9].**

Los sistemas de asistencia al aparcamiento se introdujeron en el mercado a mediados de los años 90. Los sensores ultrasónicos se emplean para detectar obstáculos en el entorno que los rodea. Inicialmente, estos sistemas disponían simplemente de una función de alerta para intentar evitar colisiones al circular marcha atrás, ya fuera entrando o saliendo de espacios de

aparcamiento. Más adelante, fueron complementados con cámaras para la marcha atrás, para poder ayudar al conductor con información más detallada. Tras hacerse posible la dirección electrónica, los asistentes de aparcamiento obtuvieron la capacidad de relevar por completo al conductor en el aparcamiento paralelo (o en línea), teniendo éste únicamente que acelerar y frenar. Con los años, evolucionaron del estacionamiento en paralelo al perpendicular (o en batería) y los datos de video mejoraron hasta obtener una vista de 360º, pero el mercado no respondió a estos sistemas tan bien como a los ESC y a los sistemas de navegación. De todos modos, hay que tener en cuenta también que son sistemas opcionales, y por tanto, asociados a costes adicionales. En la actualidad, el estado del arte de estos dispositivos está limitado a la dirección automática en un espacio designado por el conductor (y reconocido por el vehículo), pero es posible que pronto haya disponible un dispositivo del tipo aparcacoches, en el que el trabajo del conductor será completamente nulo, de forma que ni tenga que buscar él mismo el espacio para hacerlo.

El desarrollo del control de velocidad crucero adaptativo (*Adaptive Cruise Control* o ACC) marcó otro hito en la historia de los asistentes de conducción. Mediante la implementación de sistemas electrónicos de frenado y accionamiento, y el uso de la previamente muy cara tecnología de radar, la cual se volvió significantemente más asequible, se hizo posible la conducción parcialmente automática. Al igual que el control de velocidad de crucero, el ACC también regula la velocidad de circulación de forma automática. La novedad reside en que con la ayuda de un sistema de radar, controla, también de forma automática, la distancia de circulación con respecto al vehículo precedente, frenando si es necesario para mantener la distancia de seguridad. El ACC visualiza el entorno por delante de nuestro vehículo y si en un momento dado detecta la presencia de otro vehículo circulando a una velocidad inferior, alerta al conductor del peligro por una aproximación excesiva y reduce la velocidad actuando sobre el sistema de frenos. Una vez que el carril por el que circulamos queda libre, el ACC acelera hasta la velocidad que hayamos programado. Cuando se introdujo el ACC en 1999, estas funciones se podían aplicar únicamente a velocidades mayores que 30km/h **[9.9],** pero los sistemas actuales con caja automática, permiten emplearlos a velocidades más pequeñas y, por ejemplo, seguir los demás vehículos automáticamente en atascos **[9.10].**

Actualmente, se emplean sistemas de prevención de colisiones frontales, usando versiones de bajo coste de menor rango y resolución de los sensores lidar, para aplicaciones en velocidades bajas, como “City Safety” de Volvo o “City Stop” de Ford, pero para aplicaciones más avanzadas (por ejemplo, a mayor velocidad), el pequeño rango de detección de dichos sensores es un factor altamente limitador. Aun así, estos sistemas ya han sido presentados **[9.11]**. En ellos, el conductor es alertado de una colisión inminente, y si éste no reacciona, el vehículo frena por sí mismo, para mitigar la intensidad del accidente, una vez que la colisión es ya inevitable **[9.12]**. Han sido analizados en el proyecto Europeo PREVENT (2004-2008), y demostraron su efectividad, sobre todo en grandes vehículos como camiones. Por ello, son obligatorios para nuevos camiones en la UE desde noviembre del 2013.

Diversas extensiones de los actuales DAS están aún por llegar. Podemos mencionar un asistente para evitar colisiones, basado en conducción evasiva (Dang et al. 2012), asistentes para la detección de tráfico y peatones (Enzweiler & Gavrilla 2009) bajo condiciones adversas de visión (mal tiempo) (Roser & Geiger 2009), o asistentes para mejorar la seguridad en los cruces (Hopstock & Klanner 2007). Algunos de estos sistemas requieren un intercambio de datos entre los participantes en el tráfico o con la propia infraestructura de la carretera, lo cual está siendo investigado en diversas pruebas de campo: SIM-TD (2008–2013)(SIM-TD 2013), Ko-FAS (2009–2013) (Ko-FAS 2013), Koline (Saust et al. 2012), DriveC2X (2011–2013) (DRIVEC2X 2013), etc. Este enfoque promete una extensión de los límites de los sistemas en lo que respecta a la disponibilidad de información y la expansión de sus funciones a todo el colectivo de usuarios de la vía, permitiendo así una cooperación asistida o automatizada.

Éstos son solo algunos dispositivos de entre todos los que existen, y no hay duda de que siguen expandiéndose. Además, se pretende que los DAS del futuro sean capaces de llevar a cabo una conducción automática, en cualquier situación concebible, a un nivel de seguridad significativamente superior al de un conductor humano y en cooperación con los demás elementos presentes en la circulación **[9]**. Esto se considera un punto muy importante, ya que el error humano es el responsable del 90% de todos los accidentes **[9.13]**, por lo que es un prerrequisito para una circulación libre de accidentes. Para desarrollar dichos vehículos automáticos y cooperativos, es necesario dividir las tareas de conducción en componentes funcionales básicos, de forma que se puedan implementar técnicamente a un cierto nivel. Por último, queda mencionar que la conducción automática ya ha sido tema de investigación desde el final de los 80, como por ejemplo en los proyectos California PATH (1986-en curso), NAVLAB (1986-en curso) y PROMETHEUS (1987-1995):

Hoy en día, los DAS se distinguen en básicos y avanzados **[13]**. En los básicos, como el sistema antibloqueo de ruedas (ABS) o el control de estabilidad (ESC), el conductor apenas interviene. La mayoría de estos sistemas se activan en caso de emergencia, sin que la persona que conduce tenga que autorizarlo. Por el contrario, los avanzados (advanced *DAS o ADAS*), como los sistemas de advertencia de abandono de carril (*Lane Departure Warning* o LDW), el asistente de mantenimiento de carril (*Lane Keeping Assistance* o LKA) o el control de velocidad crucero (ACC), implican una compleja actuación, y normalmente están limitados por sus costes. Es más, su aplicabilidad está limitada a escenarios específicos, y es necesario que el conductor los supervise. Aunque la integración de ADAS podría ser suficiente para conducir un coche sin intervención humana (hay que tener en cuenta que el objetivo final de todos estos estudios es conseguir una conducción automática), en un sistema como por ejemplo, ACC+LKA, la supervisión del conductor es totalmente necesaria, pues es posible que se den comportamientos incorrectos. En estos casos, si el sistema es capaz de detectar los posibles errores, al solucionarse el ADAS se apaga, y vuelve a la conducción manual, alertando al conductor. Si no, el conductor es el responsable de pasar del modo automático al manual.

Un ADAS puede interactuar tanto con el vehículo como con el conductor, para intervenciones a corto o largo plazo. En las cortas, el sistema corrige una situación peligrosa, y se apaga rápidamente. En este caso, el ADAS trabaja automáticamente. Un ejemplo de este posible escenario es un sistema LDW que corrige o informa al conductor de la salida inminente del carril que está ocupando. En las intervenciones a largo plazo, el problema principal resulta cuando termina la automatización: si el conductor no se ha mantenido a cargo de la conducción, lo que mantiene alerta su atención, podría estar distraído, y el cambio de la conducción automática a la manual puede ser difícil, resultando en una situación gravemente peligrosa.

Actualmente, también existen los denominados EDAS, Eco-DrivingAssistant Systems. El eco-driving es una nueva tecnología emergente, que pretende aprovechar los beneficios de las últimas tecnologías incorporadas en los vehículos para reducir el consumo del combustible y las emisiones de gases de efecto invernadero, y al mismo tiempo mejorar la seguridad vial y reducir la probabilidad de accidentes. En la literatura, esta política es etiquetada como una propuesta en la que todos ganan (*win-win*); tanto el conductor, por el ahorro de costes y mayor seguridad personal, como la sociedad, por la reducción de emisiones de CO_2, gasto de petróleo, resto de contaminantes y fatalidades. Es una nueva forma de abordar el estilo de conducción, que lleva siendo desarrollada desde la mitad de los noventa. No hay una definición formal para ello, pero en términos generales, el eco-driving se refiere usualmente a la modificación del estilo de conducción de brusco a más refinado **[12]**. Hay diversos aspectos relacionados con dicho concepto, y los factores y características de esta técnica de conducción, generalmente se definen y caracterizan de una forma fácil. Las reglas del eco-driving se caracterizan en dos ideas: (1) adoptar un estilo de conducción anticipatorio y (2)

usar el motor de la forma más eficiente posible. Si nos adentramos algo más, los diferentes aspectos de la estrategia son, entre otros, acelerar y decelerar de forma graduada y moderada, cambiar de marchas de forma óptima, mantener un ritmo de conducción constante (conducir en o por debajo del límite de velocidad), eliminar largos tiempos en ralentí, evitar la congestión de las carreteras, anticipar el flujo de tráfico mientras se conduce para evitar paradas o frenadas bruscas, eliminar pesos innecesarios y mantener el vehículo en buen estado (mantener la presión de las ruedas, cambiar regularmente el filtro de aire, etc.) **[9], [14].** La implementación se puede conseguir mediante campañas de publicidad, o educando y entrenando directamente a los conductores.

Las aplicaciones dinámicas de eco-driving proporcionan consejos en tiempo real a los conductores, de forma que puedan modificar su comportamiento o realizar alguna acción para reducir el consumo de energía y las emisiones. Los consejos pueden darse de diversos modos, como pueden ser la velocidad recomendada, aceleración o deceleración óptimos y alertas. En general, las tecnologías dinámicas se categorizan en dos grupos: aplicaciones de autopista y aplicaciones para fuera de autopista. Las primeras, sugieren una velocidad a mantener en cada segmento de la carretera, basada en las características de ésta y las condiciones de tráfico. Las otras, se enfocan más hacia carreteras con señales de tráfico, y proporcionan alertas respecto a la posibilidad de pasar una intersección antes de que el semáforo se ponga en rojo, por ejemplo **[12].**

Los estudios respecto al impacto del eco-driving han llegado a conclusiones optimistas en lo que a su eficiencia respecta, pero sin un testeo estadístico adecuado; aun así, hay un consenso general entre los investigadores de que el esquema del eco-driving tiene, en efecto, el potencial necesario para reducir las emisiones tanto de coches como de autobuses, significantemente. De todos modos, se puede aplicar a cualquier vehículo (nuevo o viejo, grande o pequeño) e implementarlo fácilmente en un corto periodo de tiempo **[9].** Aun así, distintos vehículos pueden requerir distintas estrategias de eco-driving para maximizar su eficiencia. Adicionalmente, el tipo de carretera, su diseño de infraestructuras, el terreno, el tiempo y las condiciones meteorológicas también influyen en el grado de ahorro alcanzable en cada caso **[15]**.

Desde hace unos pocos años, el eco-driving se considera más y más como una solución prometedora para paliar los problemas medioambientales, y como se ha podido deducir de su definición y aplicaciones, es una de las técnicas que centra su interés completamente en la importancia del conductor. Esta forma “ecológica” de conducir tiene dos ventajas principales: se ha demostrado que es la forma más rápida de reducir el consumo de combustible así como la emisión de contaminantes, y su aplicación es prácticamente gratuita. Sin embargo, como ya hemos mencionado, el impacto real de estos sistemas en el sistema de tráfico no ha sido apenas estudiado. El problema consiste en evaluar el impacto del eco-driving en el consumo en una escala global, de acuerdo con la clasificación del conductor en la población **[16]**.

Los primeros estudios estadísticos (Qian y Chung, 2011; Orfila, 2011) han demostrado la posibilidad de modelar el eco-driving en una herramienta de simulación de tráfico, para poder proponer una solución a dicho problema. Pero a estas simulaciones les faltan los principales parámetros, que son las revoluciones por minuto y la relación de transmisión acoplada, por la complejidad de modelación **[16].**

La práctica y la aplicación de las técnicas de eco-driving constan usualmente de dos fases. La primera, es la impartición de cursos o cursillos, donde se presenta a los chóferes el eco-driving y se fijan las políticas a seguir para optimizar la conducción. Esta primera fase no es necesaria ni obligatoria, pero sí muy recomendable, ya que tener una visión más clara de los comportamientos de conducción puede ayudar al conductor no sólo a conocer mejor su

vehículo, sino también a conducir de forma más segura, aportando seguridad en la carretera, y también a reducir el consumo y las emisiones **[3]**. Se ha demostrado que estos programas inducen una reducción de conductas no deseadas, como acelerones y frenazos **[3],[6],[17],** pero quedan aún por analizar y testear la influencia y los efectos de éstos a corto y largo plazo, así como la frecuencia o la regularidad con la que se deben impartir, pues algunos apoyan la idea de que estas charlas sean algo puntual, un curso de iniciación para los nuevos chóferes de una flota, pero también los hay de los que creen que hay que implantarlos como cursos regulares, a los que habría que asistir, por ejemplo, una vez al año. La segunda fase corresponde a la implantación de algún dispositivo (denominados *on board device*-s o *OBD-s*) en el vehículo, que se ocupe de recoger la información del vehículo y de devolver la respuesta oportuna. Estos dispositivos son muy diversos y de distinta complejidad, desde los más simples, que recogen únicamente los datos obtenidos del CAN y generan directrices para conductas más eficientes **[18]**, hasta los que recopilan una gran cantidad de datos como la posición, la velocidad, la aceleración, las rpm o el peso del vehículo, así como los obtenidos mediante sensores exteriores o cámaras y muestran, mediante modelos predictivos y en tiempo real, la velocidad y la aceleración o deceleración óptimas en ese instante **[6]**.

Uno de los temas más importantes a tratar en lo que a los OBD respecta, es la elección de las variables a recoger y analizar. Las más comunes son la velocidad, la aceleración/deceleración, la marcha en la que se está circulando (exceptuando los coches automáticos) y las rpm, además del consumo, claro está. Últimamente también se están analizando con bastante interés los comportamientos de salidas y paradas **[3],[19],[20]**, así como el número de paradas por distancia **[2]**. Los datos de este primer grupo de variables son relativamente fáciles de extraer, gracias a dispositivos como el *CAN bus* **[21]**. CAN son las iniciales de *Controler Area Network* (Red de área de controladores), un sistema desarrollado por Bosch junto con Intel en 1987, especialmente para la industria del automóvil, aunque hoy en día se utiliza también en vehículos industriales. Es un protocolo de comunicaciones en serie para el intercambio de información entre unidades de control electrónicas del automóvil, con un solo cable que recorre el vehículo (se tiene una topología en forma de bus, de ahí el nombre) al que se van conectando los diferentes aparatos electrónicos. Este sistema permite compartir una gran cantidad de información ente las unidades de control (centralitas) abonadas el sistema, lo que provoca una reducción importante tanto del número de sensores utilizados como de la cantidad de cables que componen la instalación eléctrica.

A parte de éstos, hay algunos aspectos sobre los que el conductor no tiene ninguna influencia, como el peso del vehículo o las condiciones meteorológicas, y normalmente éstos factores (la humedad, las precipitaciones, la temperatura…) no suelen ser tan fáciles ni exactos de medir como los anteriores pero eso no significa que no puedan ser variables significativas en el consumo y las emisiones. Y por el contrario, gracias a la gran cantidad de investigaciones que se están llevando a cabo, diversas conclusiones están saliendo a la luz, y algunas de ellas son contrarias a las creencias populares. En **[22]** por ejemplo, se concluye que reducir la velocidad no es la única, ni la más óptima estrategia de eco-driving.

Es necesario presentar, para terminar, los tres puntos de vista que se han empleado **[2]**, **[23]** para crear modelos de estimación de consumo o emisiones o, más bien, para investigar la influencia de la interacciones entre conductor y vehículo. Son los modelos macroscópicos, microscópicos y mesoscópicos. Los primeros se basan en el conocimiento de la velocidad media del vehículo, así como de otros valores agregados, y se emplean para la estimación total del consumo en grandes regiones o para largos periodos de tiempo, pero son de baja precisión, ya que no se toma ninguna información sobre los tiempos de aceleración y deceleración, ni de la potencia específica del vehículo. Los modelos microscópicos se desarrollan normalmente basados en parámetros instantáneos de tráfico (velocidad, aceleración) que se pueden recoger mediante dispositivos GPS. Estos modelos pueden

mejorar sustancialmente la estimación de los anteriores y tienen mayor precisión, pero normalmente se aplican sobre subconjuntos de los trayectos, puesto que necesitan gran cantidad de datos de entrada y es no es fácil hacer una recopilación a gran escala. Para desarrollar estos modelos se emplean tres tipos de métodos: estadísticos, regresiones y derivaciones basadas en el consumo. Finalmente, los mesoscópicos construyen ciclos de conducción, por lo que son una interesante alternativa a los microscópicos en casos en los que no hay datos precisos disponibles de la velocidad y la aceleración.

En **[14]** por ejemplo, se ha desarrollado un algoritmo macroscópico, no-iterativo, para estimar el consumo de combustible de vehículos. Los resultados muestran, para coches manuales, ahorros sustanciales obtenidos gracias a estrategias de cambio de marcha adecuadas. El experimento llevado a cabo demuestra que el cambiar de 3ª a 4ª marcha puede conllevar un 19% de ahorro de combustible, y que esta reducción puede llegar al 25% al cambiar de 4ª a 5ª. En **[2]** se presenta un modelo microscópico, implementado en dos líneas de autobuses. En una de ellas, se ha obtenido un 27% de ahorro, mientras que en la otra no se obtiene nada lo suficiente significativo como para poder aportar resultados aceptables. Estos resultados tan diversos son debidos a que los modelos de consumo están basados en pocos factores o a que los parámetros no son demasiado detallados, por lo que la capacidad de generalización y la precisión de estos modelos están por mejorar **[23]**.

Por lo visto en la literatura, se puede medir la agresividad de los conductores mediante los estudios de consumo **[5]**, o clasificarlos en distintos rangos, como por ejemplo en agresivos, neutros y no agresivos **[20]**. Esta clasificación es el resultado del análisis de los hábitos de conducción, donde los que peores hábitos tienen (en lo que al gasto de combustible o emisiones respecta) son denominados como conductores agresivos y los consejos transmitidos por los OBD más avanzados están basados en las conductas de los no-agresivos (los óptimos), de forma que se busca que los agresivos adopten las conductas de los no-agresivos para conseguir una conducción más eficiente.

3.2 Líneas de investigación relacionadas

Como se ha dicho, ya existen en el mercado sistemas para la detección de distracciones y faltas de concentración al volante, pero a pesar de todo el entusiasmo puesto en los progresos tecnológicos, hay que reconocer que la conducción automática considerable en cualquier situación concebible requiere mucha mayor capacidad cognitiva que la disponible hoy en día. Pero como fase previa, algunos proyectos de investigación están centrados en lo que llaman “semi-automatización”. El factor humano juega un papel importante en estos sistemas, que reparten las tareas y responsabilidades entre en conductor humano y el vehículo semi-automático; incluso hay máquinas que toman el control del vehículo si se da el caso en el que el conductor no pueda, y lo pilotan hasta una posición segura **[9]**.

Pero hasta que la completa automatización sea posible, es deseable apoyar el trabajo del conductor con DAS, de forma que pueda centrar toda su atención en la tarea principal de conducir, aunque estos dispositivos todavía no aportan demasiado a las eficiencia del tráfico, pues es necesario extender sus capacidades hacia una conducción cooperativa **[9]**.

Otra línea emergente de investigación examina los efectos de los OBD y las aplicaciones móviles **[15]** que proporcionan *feedback*s en tiempo real referentes al comportamiento de conducción y eficiencia energética. Diversos estudios han analizado los resultados del uso de

smartphones en la monitorización del conductor **[11],[22],[24]**. Incluso ha habido quien ha tratado de crear una herramienta de entrenamiento basada en los juegos **[25]**; los conductores son puntuados de 0 a 10 (eficiencia), y estas puntuaciones pueden ser comparadas con las obtenidas por otros conductores y también con sus comportamientos para obtener los óptimos, mientras los conductores participan activamente en el "juego", lo que hace que quieran adoptar conductas más eficientes y así evitar que recaigan en sus hábitos originales.

No podemos terminar este apartado sin hacer referencia al Google Car, el proyecto de Google para crear vehículos que se manejen solos, el cual está cada vez más cerca de las calles y carreteras del mundo. Google Car ya ha sido probado en Mountain View (California) y Austin (Texas) y ahora incluye volante, pedales, acelerador y freno, con lo que cumple con las exigencias impuestas por las autoridades de tránsito de esos estados. De esta forma, en caso de una emergencia, la persona dentro del coche podrá detenerlo sin ningún problema.

Y, ¿cómo funciona? Los vehículos tienen sensores diseñados para poder detectar objetos hasta a dos campos de fútbol de distancia en todas las direcciones, incluyendo peatones, ciclistas y vehículos, así como pájaros o pequeñas bolsas de plástico en la carretera. El software procesa toda la información para hacer que el coche conduzca por la carretera de una forma segura, sin cansarse o distraerse. Como ellos mismos explican (http://www.google.com/selfdrivingcar/), el Google Car, como cualquier otro conductor, necesita constantemente responder a las siguientes preguntas: ¿Dónde estoy? El coche procesa dos informaciones (mapa y sensores) para determinar su posición; conoce la calle y el carril exacto en el que se encuentra. ¿Qué tengo alrededor? Los sensores detectan y clasifican los objetos captados por su tamaño, forma y patrón de movimiento. ¿Qué pasará después? El software predice cuál será el próximo movimiento o posición que puedan tener los objetos. ¿Qué debería hacer? Tras la predicción, el software fija una velocidad y trayectoria seguras para dicha situación.

Para todo ello, estos vehículos constan de sensores (láseres, radares y cámaras), baterías eléctricas, forma redondeada (para maximizar el campo de visión de los sensores), un ordenador y sistemas auxiliares (para corregir la dirección, frenar, etc.).

El coche que se conduce solo de Google es un proyecto iniciado en 2009 con la vista puesta en el largo plazo, aunque desde entonces sus dos docenas de Lexus RX450h equipados con sensores ya han circulado y, por tanto, registrado en mapas 3D hasta 2,7 millones de kilómetros, fundamentalmente en autopistas y carreteras. Además, desde hace aproximadamente dos años, los coches de Google también han estado circulando por las calles de Mountain View, donde el gigante tecnológico tiene su sede. Aun así, en caso de salir adelante, el proyecto de Google deberá encontrar un encaje en el actual código de circulación, que en ningún caso contempla la posibilidad de vehículos que circulen sin conductor.

4. Disposición de datos

Los datos disponibles de un vehículo dependen de los dispositivos instalados en ellos. Todos y cada uno de los autobuses llevan el antes presentado CAN, el cual se ocupa de recopilar información de diversas partes del autobús (de las ruedas, las rpm y la velocidad, y

por tanto, la distancia recorrida; del depósito, la cantidad de combustible, etc.) y compartirla con las distintas centralitas (el dato de la velocidad debe llegar al panel de mando, para poder mostrarlo en el tacógrafo).

Para acceder a dichos datos, Datik emplea dos tipos de ordenadores: el MTX y el DCB (Datik Board Computer). La instalación de uno u otro en un vehículo depende de la demanda de datos:

- El MTX es un pequeño computador que recoge, por una parte, la información GPS mediante una antena que debe instalarse en el vehículo, y por otra, los datos GPRS, gracias a una tarjeta SIM que lleva incorporada, para a su vez transmitir la información en tiempo real a iPanel. Al conocer la posición instantánea, se conoce también la velocidad a la que se circula.

- En cambio, el DCB es más completo, puesto que además de obtener los datos de GPS y GPRS, se conecta con el CAN, por lo que la cantidad de variables recogida es mucho mayor.

En las compañías en las que únicamente interesa conocer la puntualidad de los servicios, y conocer la posición de los vehículos para mostrar los tiempos esperados de llegada en los paneles de las paradas, es suficiente equipar los autobuses con el MTX. Si, en cambio, se pretende conocer más a fondo lo que ocurre y hacer, por ejemplo, un seguimiento del combustible consumido, se precisa la instalación del DCB.

Para hacerse una idea, los autobuses con DCB, recopilan, entre otras, las siguientes variables: *fecha, longitud, latitud, velocidad(km/h), rpm, odómetro(km), fuel (L/h), consumo instantáneo(km/l), fuel acumulado, Medición de la resistencia del emisor de nivel de combustible, periodo tacógrafo, status GPS* y *status GPRS.*

Pero además, en el propio iPanel también se hacen diversos cálculos con esos datos primarios que llegan desde los vehículos, obteniendo así variables como *Eficiencia conducción, Velocidad comercial* (velocidad media mientras el vehículo circula con pasajeros), *Paso anticipado*, *Paso retrasado y Paso omitido* (por una parada), *Aceleraciones por distancia* o *Porcentaje de tiempo en ralentí excesivo*. Éstas, como se puede suponer, no son variables medidas a cada instante, sino que se miden por ruta, por servicio o por vehículo, según interese. Con todo esto, un punto importante de todo el estudio corresponderá a la buena selección de variables a tener en cuenta, y su posterior procesamiento.

5. Estudio previo

El primer trabajo realizado ha sido la reconstrucción y revisión de un estudio previo. Este estudio fue llevado a cabo (por una empresa externa) con la intención de hacer un primer análisis de los datos disponibles y su influencia en el consumo, siendo el objetivo final el mismo que el nuestro, encontrar patrones de conducción para ser capaces de dar recomendaciones a los conductores de tal forma que se reduzcan los comportamientos o actitudes que elevan la cantidad de litros de combustible consumidos.

Hay que tener en cuenta que las operaciones realizadas y las decisiones de los siguientes dos apartados (5.1 y 5.2) no se han tomado en referencia a este trabajo, sino que se han seguido los pasos del trabajo que se está replicando, y los resultados que se muestran corresponden a los del trabajo original. En la sección 5.3 se discutirán los lados positivos y los negativos de estas acciones, así como los aspectos en los que, al replicarlo, no se han obtenido resultados semejantes a los esperados.

5.1 Análisis y preproceso de los datos

Como paso previo a la recomendación, es necesario estudiar las variables que caracterizan el problema (variables independientes) y construir un modelo que permita el cálculo o la predicción del consumo de combustible (variable dependiente) a partir del valor de las variables independientes.

Para realizar el estudio se han empleado los datos de las rutas 1673 y 975 contenidos en los ficheros *mezclaServiciosIndicadores1673.csv* y *mezclaServiciosIndicadores975.csv,* obtenidos de iPanel (los datos de estas tablas no son las originales de los vehículos, sino que han sido procesados). Se cuenta con la información de un total de 979 servicios de la ruta 1673 y de 453 de la 975.

Las variables disponibles en estos ficheros se presentan en la Tabla 1.

Id_servicio	%_tiempo_ralenti_exc	Distancia	Frenadas_por_distancia
Cod_servicio	% _ tiempo_rpm_alto	Distancia servicio	rpm_alto
Id_vehiculo	%_tiempo_acel_exc	Distancia total	ralenti_excesivo
Hora inicio	%_tiempo_fren_exc	Velocidad media	rpm_alto__por_distancia
Hora fin	Aceleraciones	Tiempo en servicio	ralenti_excesivo_por_distancia
Id_ruta	Frenadas	Tiempo total	Productividad por tiempo
Id_conductor	Motor on	Rendimiento	Aceleraciones_por _distancia

***Tabla 1**: Variables independientes iniciales.*

Las variables *consignas, seguimiento, seguimiento_cantidad, porc_tiempo_vel_exc, robos_combustible, repostajes_fuera_gasolinera, velocidades_maximas_superadas, velocidades_maximas_por_distancia, combustible_robado, combustible_repostaje_fuera, combustible_robado_litros, combustible_repostaje_fuera_litros, tiempo_en_velocidad_limite_ superada, productividad_distancia* y *tiempo_sin_servicio* no se han tenido en cuenta, pues su valor es 0 para todos los servicios (100 en el caso de la *productividad_distancia*).

De estas variables, se han eliminado aquellas que por lo pronto se creen que no afectan al consumo:

- *Id_servicio*.
- *Cod_servicio.*
- *Id_vehiculo*: el tipo de vehículo sí afectaría al consumo, pero por lo que se puede apreciar en las características de ellos, todos son muy semejantes.
- *Id_ruta*: la eliminamos porque en principio vamos a hacer un modelo por cada ruta.
- *Id_conductor*: el conductor es un factor influyente en la conducción y en el consumo, pero inicialmente, decidimos eliminarlo como variable independiente, ya que estimamos que en el caso de considerarlo sería más coherente hacer un modelo por ruta y por conductor. Este es un aspecto que deberemos valorar en un futuro.

Hora inicio y *hora fin* son fechas junto con una hora, de manera que para poder tratarlas se han pasado a segundos transcurridos desde EPOCH.

Distancia, distancia total y *distancia en servicio* poseen los mismos valores, de forma que solo contaríamos con una de ellas para el análisis. Por otro lado, hay una serie de parámetros que viene expresados en función de la distancia recorrida:

- *Aceleraciones_por_distancia.*
- *Frenadas_por_distancia.*
- *Rpm_alto__por_distancia.*
- *Ralenti_excesivo_por_distancia.*

De forma que en estos parámetros estaría implícita la distancia recorrida, así que el parámetro que representa la *distancia* podríamos obviarlo. De la misma manera, obviaremos el *número de aceleraciones* y *frenadas*, y el número de sucesos de *rpm_alto* y *ralenti_excesivo.*

Tiempo en servicio y *tiempo en total* poseen los mismo valores. Además, la hora de fin menos la hora de inicio, nos proporciona el tiempo en servicio o tiempo total. Por ello, por lo pronto solo incluiremos la hora de inicio y hora de final como mayores indicadores de posible consumo, el cual puede estar relacionado con la franja horaria en la que se realiza el servicio.

La variable *rendimiento* se define como el número de kilómetros recorridos por litro. Se ha estimado, que esta variable se encuentra directamente relacionada con la variable dependiente que queremos predecir, el *consumo medio*. De modo que proporciona información implícita sobre el consumo, y sería equivalente predecir el consumo medio que predecir el rendimiento. Por lo tanto, la variable *rendimiento* no será considerada para el análisis.

Tras esto, el conjunto de variables independientes que estimamos que influyen en el consumo de combustible sería el mostrado en la Tabla 2.

Hora inicio	Velocidad media
Hora fin	Motor on
%_tiempo_ralenti_exc	ralenti_excesivo_por_distancia
% _ tiempo_rpm_alto	rpm_alto__por_distancia
%_tiempo_acel_exc	Aceleraciones_por_distancia
%_tiempo_fren_exc	Frenadas_por_distancia
Productividad por tiempo	

Tabla 2: *Variables independientes finales.*

No existen datos incompletos, pero es necesario un método de detección de outliers, para eliminar valores significativamente distintos del resto, pues suelen ser fruto de errores, e introducen ruido en el modelo predictivo. En este caso, para calcular que valores de consumo medio son outliers, se ha empleado el método de los cuartiles.

Es uno de los métodos más utilizados en estadística para la detección de outliers, y utiliza el concepto de cuartil de un conjunto de datos. Si tenemos un conjunto de datos y lo ordenaremos de menor a mayor, el cuartil 1, llamémosle Q_1, es el valor tal que desde ese valor hacia su izquierda se encuentran la primera cuarta parte de los valores de este conjunto de datos. El cuartil 2, llamémosle Q_2, es el valor tal que desde ese valor hacia su izquierda se encuentran la primera mitad de los valores de este conjunto de datos. Y así sucesivamente. Para detectar valores outliers moderados, tendríamos:

$$LimInf = Q_1 - 1.5 \cdot (Q_3 - Q_1)$$
$$LimSup = Q_3 + 1.5 \cdot (Q_3 - Q_1)$$

Los valores que sean menores que $LimInf$ o mayores que $LimSup$ se consideran valores outliers. Para detectar valores outliers extremos, tendríamos:

$$LimInf = Q_1 - 3 \cdot (Q_3 - Q_1)$$
$$LimSup = Q_3 + 3 \cdot (Q_3 - Q_1)$$

Los valores que sean menores que $LimInf$ o mayores que $LimSup$ se consideran valores outliers.

En este caso, se han eliminado únicamente los valores extremos: 29 (de 979) de la línea 1673 y 7 (de 453) de la línea 975. Tras la extracción de los outliers, se ha procedido a la normalización de las variables.

5.2 Construcción del modelo

Nuestro problema parte de un conjunto de variables independientes que debemos de emplear para ser capaz de predecir el consumo medio de combustible.

Para crear un modelo predictivo inicial se ha empleado la técnica de regresión de mínimos cuadrados parciales (*PLS regression*). La regresión de mínimos cuadrados parciales se utiliza para encontrar las relaciones fundamentales entre dos matrices (X e Y), las cuales se pueden expresar mediante la matriz Z (predicha con el modelo) de la siguiente expresión:

$$\begin{pmatrix} & X_{ij} & \end{pmatrix}_{MxN} \begin{pmatrix} Z_j \end{pmatrix}_{Nx1} = \begin{pmatrix} Y_i \end{pmatrix}_{Mx1}$$

Donde M es el número de muestras que tenemos de cada variable, y N en número de variables independientes. En nuestro caso X estará formada por el conjunto de valores de las variables independientes e Y por los valores de la variable dependiente a predecir, el consumo medio de combustible.

De manera que el modelo PLS nos proporciona el valor de los coeficientes de regresión Z_j, y por lo tanto nos permite obtener la relación entre X_i y Y_i. Los coeficientes de regresión Z_j, indican el valor por el que hay que multiplicar las variables independientes, X_i, para obtener el valor de consumo medio asociado, Y_i. Por tanto, suponiendo que tenemos un número de variables independientes N, tras aplicar el modelo PLS obtenemos la siguiente expresión, donde Z_j son conocidos:

$$X_{i,1}Z_1 + X_{i,2}Z_2 + \cdots + X_{i,N}Z_N = Y_i$$

En este primer estudio, se ha decidido crear un modelo predictivo para cada ruta con el fin de obtener en esta primer aproximación modelos más precisos y de mayor calidad. En ambos casos, se ha empleado el 75% de las muestras para construir el modelo y el 25% restante para testear posteriormente la calidad predictiva de éste. La calidad del modelo se ha medido en términos del Error Cuadrático Medio (ECM) y para determinar cómo de influyente es cada variable independiente en el proceso de predicción del consumo de combustible, se ha aplicado un método de selección de características, el cual proporciona un peso entre 0 y 1 a cada variable independiente, representando 1 el mayor grado de influencia

Ruta 1673

Los coeficientes de la matriz Z, esto es, los asociados a cada una de las variables por la regresión PLS, se muestran en la Tabla 3, el valor del ECM ha sido 0.051 y la Tabla 4 muestra los pesos de cada variable.

Variable	Coeficiente de regresión (Z_j)
Hora inicio	104239.3036983
Hora fin	-104239.017142
%_tiempo_ralentí_exc	0.025838
%_tiempo_rpm_alto	0.058426
%_tiempo_acel_exc	-0.259756
%_tiempo_fren_exc	-0.111073
Motor on	0.0454732
Rpm_alto_por_distancia	0.111195
Ralentí_excesivo_por_distancia	-0.050177
Aceleraciones_por_distancia	0.449269
Frenadas_por_distancia	0.141257
Productividad por tiempo	0.370344
Velocidad media	-0.003794

Tabla 3: *Coeficientes de regresión asociados a cada variable, obtenidos tras crear el modelo predictivo PLS regression para la ruta 1672.*

Variable	Peso
Hora inicio	0.024986
Hora fin	0.025083
%_tiempo_ralentí_exc	0.430681
%_tiempo_rpm_alto	0.278002
%_tiempo_acel_exc	0.591183
%_tiempo_fren_exc	0.132286
Motor on	0.510326
Rpm_alto_por_distancia	1
Ralentí_excesivo_por_distancia	0.539666
Aceleraciones_por_distancia	0.919166
Frenadas_por_distancia	0.189997
Productividad por tiempo	0.517240
Velocidad media	0.799564

Tabla 4: *Peso de cada variable tras aplicar selección de características para la ruta 1672.*

Ruta 975

Los coeficientes de la matriz Z, esto es, los asociados a cada una de las variables por la regresión PLS, se muestran en la Tabla 5, el valor del ECM ha sido 0.063 y la Tabla 6 muestra los pesos de cada variable.

Variable	Coeficiente de regresión (Z_j)
Hora inicio	164305.079675
Hora fin	-164302.960297
%_tiempo_ralentí_exc	-0.046432
%_tiempo_rpm_alto	0.116718
%_tiempo_acel_exc	-0.049174
%_tiempo_fren_exc	0.190257
Motor on	0.205109
Rpm_alto_por_distancia	0.066642
Ralentí_excesivo_por_distancia	-0.015872
Aceleraciones_por_distancia	0.171833
Frenadas_por_distancia	-0.234502
Productividad por tiempo	1.716774
Velocidad media	-0.184684

Tabla 5: *Coeficientes de regresión asociados a cada variable, obtenidos tras crear el modelo predictivo PLS regression para la ruta 975.*

Variable	Peso
Hora inicio	0.008514
Hora fin	0.008510
%_tiempo_ralentí_exc	0.074582
%_tiempo_rpm_alto	0.735909
%_tiempo_acel_exc	0.508936
%_tiempo_fren_exc	0.133734
Motor on	0.051441
Rpm_alto_por_distancia	1
Ralentí_excesivo_por_distancia	0.034044
Aceleraciones_por_distancia	0.779896
Frenadas_por_distancia	0.089246
Productividad por tiempo	0.005199
Velocidad media	0.104254

Tabla 6: *Peso de cada variable tras aplicar selección de características para la ruta 975.*

5.3 Estudio y conclusiones

El estudio se ha aplicado sobre las rutas 1673 y 975. No se conoce el software con el que se ha realizado, pero para la réplica se ha trabajado con R (http://www.r-project.org/) (ver apartado 6.3).

El primer punto a analizar es la decisión de crear el modelo por servicio, lo que significa que se recopila un dato de cada variable por cada servicio (o recorrido) de la ruta. Teniendo en cuenta que el objetivo final es estudiar las conductas de conducción, se considera que tener un único dato representando todo lo sucedido en el rango de tiempo en el que se ha realizado el servicio implica una importante pérdida de información.

El segundo punto de interés es el método de selección de variables. No es recomendable hacer la selección manualmente, pues puede llevar, como ha sido el caso, a introducir variables insignificantes y empeorar el modelo. En esta ocasión, como se puede observar en las tablas 5 y 6, las variables *Hora inicio* y *Hora fin* introducidas en el modelo de regresión no aportan apenas nada en lo que a la previsión del consumo respecta; en una escala de 0 a 1, ninguna de las dos supera el 0.0086. Y aunque nos puedan engañar los coeficientes obtenidos por el PLS (ambos son muy grandes comparados con el resto de variables), si tenemos en cuenta que tienen signo contrario, y que el valor de las dos variables para un mismo servicio es muy parecido, debemos darnos cuenta de que a la hora de estimar el consumo, una resta el efecto de la otra, por lo que la significancia es casi nula.

Es remarcable también la presencia de dos variables (“porcentaje de tiempo” y “por distancia”) por cada concepto (*ralentí excesivo, rpm altas, aceleración excesiva* y *frenada excesiva*), puesto que aunque no contengan exactamente los mismos datos, debe haber una gran correlación entre las dos variables, por lo que, seguramente, sería suficiente introducir una de cada pareja. Por ello, para hacer la selección de variables se requiere emplear algún método estadístico, empezando por las correlaciones, o las regresiones simples, hasta modelos de regresión múltiple o algoritmos de selección de variables.

Otro aspecto a tratar es la extracción de outliers, la cual no debería aplicarse únicamente sobre la variable de interés, sino también sobre todas las demás. No hacerlo también puede conllevar errores en el modelo y, por consiguiente, su empeoramiento.

Con todo ello, si evaluamos el modelo obtenido para la ruta 1673 (recordemos que el modelo se construyó con el 75% de los datos originales, por lo que tenemos el 25% restante para el testeo) resulta que solo re presenta sólo el 45% de la variabilidad del consumo, lo cual se traduce en que no es un modelo demasiado preciso. Esto se puede observar en la Figura 1, donde se muestran en el eje Y los valores observados (reales) de los datos de test, y en el eje X los valores de consumo predichos por el modelo. En el caso en el que todos los puntos cayesen sobre la recta (*predichos=obesrvados*), el modelo seria exacto, pero vemos que hay muchos puntos que quedan bastante lejos de ella.

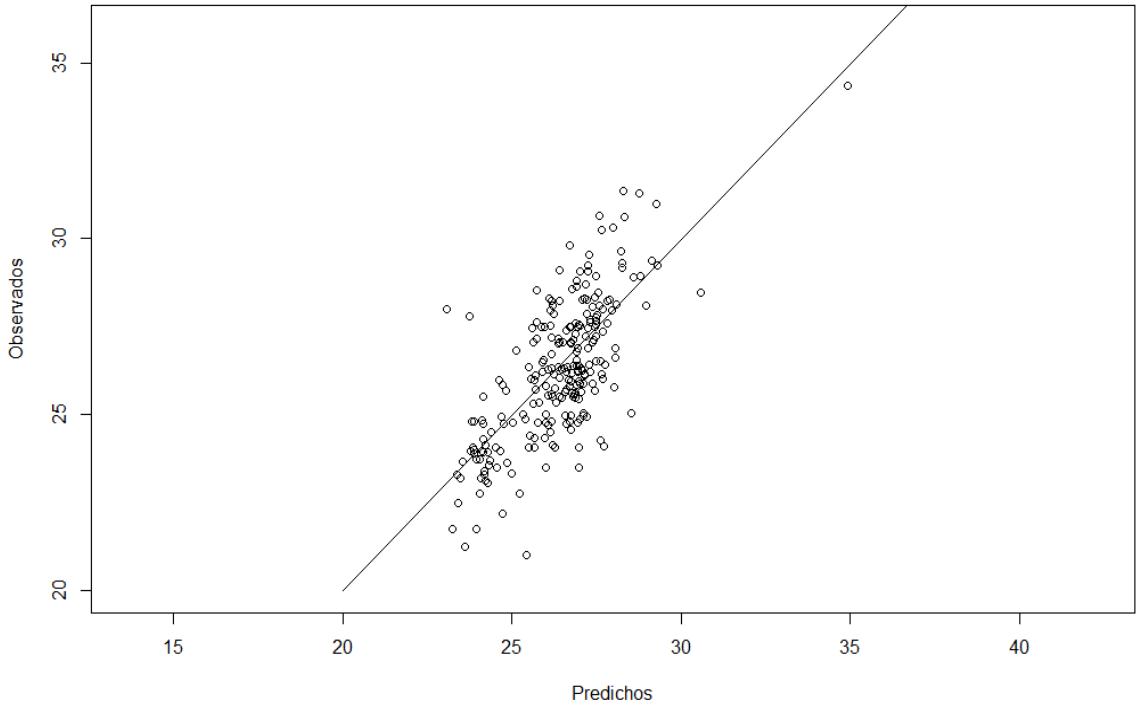

Figura 1: *Valores de consumo medio predichos por el modelo vs. valores observados.*

Otra de las causas de la imprecisión de la estimación del consumo, es la presencia de variables subjetivas. Ocho de las trece variables del modelo (todas, excepto *Hora inicio, Hora fin, Motor on, Productividad por tiempo* y *Velocidad media*) no son variables reales, sino que han sido creadas a partir de los datos que se reciben del vehículo. En el caso de las revoluciones por minuto, por ejemplo, se recibe el dato de las rpm, pongamos, cada segundo. Hay un umbral prefijado (a mano, claro) que delimita las rpm altas, por lo que si el valor que se recibe supera el umbral, cuenta como rpm altas, y si no, no; lo que significa que si un conductor hipotéticamente conduce siempre justo por debajo del umbral, su valor de *%_tiempo_ranlenti_exc* (y lo mismo ocurre con *ralentí_exc_por_distancia*) será 0. En cambio, otro que conduzca justo por encima del umbral tendrá valor 100. Esto significa que para dos conductas muy parecidas, en lo que a rpm respecta al menos (pueden tener valores de rpm muy parecidos en todo el trayecto, pero uno algo por encima y el otro por debajo del umbral) tendremos valores de la variable *%_tiempo_ranlenti_exc* completamente opuestos, mientras que es posible que apenas haya diferencia entre el consumo de ambos.

Esta subjetividad inducida por la implicación 'humana' hace que también se pierda información, por lo que no es nada conveniente. En caso de querer discretizar las variables *ralentí, rpm, aceleración* y *frenada* convendría aplicar, por ejemplo, un algoritmo de clustering al consumo, y analizar si hay diferencias respecto a los valores de la variable que se pretende discretizar.

Por último, en lo que a diferencia de resultados respecta, hay que subrayar el Error Cuadrático Medio. El ECM mide la diferencia entre el estimador y lo que se estima, esto es, la precisión del modelo. Los valores del ECM obtenidos para ambas rutas son muy pequeños (0.051 y 0.063 para las rutas 1673 y 975 respectivamente), lo que debería significar que la estimación es muy buena, pero ya hemos visto que no es así (véase Figura 1). En cambio, y teniendo en cuenta que el resto de resultados de este estudio prácticamente han coincidido

con los del original, llama la atención las diferencias en los valores de los EMC: en nuestro caso han sido 0.0968 y 0.152.

En el próximo apartado se pretende corregir estos problemas, y construir, a su vez, un modelo que se espera que trabaje de un modo más eficiente, pues una vez realizado este primer análisis, estamos en disposición de evitar y no volver a caer en los mismos errores.

6. Trabajo actual

Como ya se ha comentado, se cree que con la creciente precisión de las tecnologías de navegación del futuro, es posible concebir más efectos beneficiosos mediante la provisión de datos de la localización, de forma que se permita que las funciones se adapten a las condiciones locales, pero para llegar a conseguir esos objetivos, hay que empezar a construir con lo que se tiene.

6.1 Propuesta del modelo

Hemos visto que un modelo por servicios implica la pérdida de la mayoría de la información del viaje, y por tanto, la falta de datos para observar los hábitos de los conductores. Por otra parte, trabajar con todos los datos disponibles de un servicio (en un servicio de dos horas de duración, en el que se reciben los datos a cada segundo, tendríamos 7200 observaciones de cada una de las variables disponibles) puede resultar un tanto engorroso, además de ser posible que en gran parte del transcurso del recorrido, no haya ningún cambio en la conducción, como puede ocurrir en casos en los que la ruta incluya trayectos por autopista, por ejemplo.

Para intentar reducir la cantidad de datos, sin perder demasiada información de las conductas, nos centraremos, como ya lo hacen algunos de los estudios más recientes **[3],[41],[47],** en las conductas de parada y salida de los conductores, tomando para ello, los datos correspondientes a los segundos anteriores y posteriores a una parada.

Con dichos datos, se puede construir un modelo por cada parada, y observar parecidos y diferencias entre modelos, de donde puede que sea posible extraer o identificar algunas conductas de aceleración o frenado que induzcan altos consumos de combustible.

Para terminar, se analizará el combustible consumido en paradas y salidas de distintos conductores, para ver si hay diferencias significativas y si es posible clasificarlos por el consumo.

Por otra parte, con los datos de todas las paradas de un mismo servicio, también podremos analizar si lo ocurrido en las paradas es realmente significativo en el consumo total del viaje o, o por lo contrario, es necesario tener en cuenta otros aspectos del trayecto.

6.2 Obtención y tratamiento de datos

6.2.1 Datos internos

Los datos recopilados por los dispositivos instalados en los autobuses pueden ser extraídos fácilmente con una memoria USB. De entre los autobuses disponibles, se ha seleccionado el 3255 de la compañía Bizkaibus, ya que es uno de los que repite una misma ruta una y otra vez.

El vehículo está equipado con un DCB, y disponemos de los datos recopilados desde el 10 de diciembre de 2014 hasta el 30 de marzo de 2015, con las siguientes catorce variables:

Fecha	Odómetro (km)
Longitud	Fuel (L/h)
Latitud	Fuel acumulado (L)
Nº de satélite GPS	Periodo tacógrafo
Velocidad (km/h)	Status GPS
rpm	Status GPRS
Medición de la resistencia del emisor de nivel de combustible (MRC)	Consumo instantáneo (km/l)

Tabla 7: *Variables disponibles del autobús 3255.*

De estas variables se han eliminado, en un principio, *nº de satélite GPS, periodo tacógrafo, status GPS, status GPRS* y *MRC*, puesto que no nos interesan, no aportan información al modelo a construir.

La medición de cada una de estas variables se realiza cada segundo, siempre y cuando el vehículo se encuentra arrancado. Una vez apagado el motor, no se guarda ninguna información.

6.2.2 Datos externos

Para completar los datos, se ha pedido ayuda a Bizkaibus, desde donde se ha recibido la información del gráfico y conductor. No se han proporcionado los datos al completo, solo disponemos de los conductores de algunos servicios de enero y marzo (ninguno de diciembre y febrero) y del gráfico de casi todos los días laborables y algunos fines de semana.

Cada conductor se identifica por un número y el gráfico es un número que se asigna diariamente a cada vehículo, el cual designa las rutas y servicios a realizar en ese día, así como los horarios de los mismos. En nuestro caso, en los 111 días en los que tenemos datos, el autobús tiene diversos gráficos, pero hay uno que se repite con mucha frecuencia. Los servicios realizados los días con dicho gráfico son los siguientes:

Gráfico: 1201
06:20 Vac Le/Cochera Eibar
07:00 **A3912** Ord Eibar - Durango - Amore - Lemona - H - Bilbao
09:30 **A3912** Ord Bilbao - H - Lemona - Amorebieta - Durango - Eibar
12:00 **A3912** Ord Eibar - Durango - Amore - Lemona - H - Bilbao
14:30 **A3912** Rel Bilbao - H - Lemona - Amorebieta - Durango - Eibar
Releva con Gráfico **1202** en Le/Cochera a las 15:10
15:10 Cab Le/Cochera Le/Cochera

Gráfico: 1202
15:05 Cab Le/Cochera Le/Cochera
15:10 **A3912** Rel Bilbao - H - Lemona - Amorebieta - Durango - Eibar
Releva al Gráfico **1201** en Le/Cochera
17:00 **A3912** Ord Eibar - Durango - Amore - Lemona - H - Bilbao
19:30 **A3912** Ord Bilbao - H - Lemona - Amorebieta - Durango - Eibar
21:25 Vac Eibar Le/Cochera

Figura 2: *Indicaciones para el gráfico 1201.*

Como se puede observar, cada día con este gráfico se realizan tres servicios de ida (09:30, 14:30 y 19:30) y otros tres de vuelta (07:00, 12:00 y 17:00) de la línea 3912, que recorre los municipios de Bilbao, Etxebarri, Galdakao, Bedia, Lemoa, Amorebieta-Etxano, Iurreta, Durango, Abadiño, Berriz, Zaldibar, Mallabia, Ermua y Eibar.

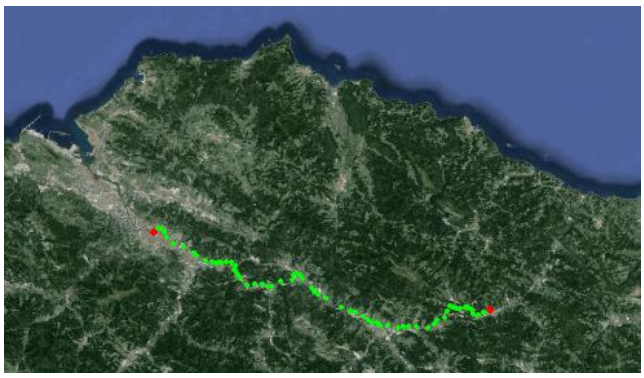

Figura 3: *Recorrido de la línea 3912 (izda.: Bilbao, dcha.: Eibar).*

Cada vez que se hace un relevo de gráfico (en este caso, a las 15:10), el autobús continua haciendo los servicios del gráfico con el que se ha relevado, y también se procede al cambio de conductor. En algunos casos, hay un único relevo (y por tanto, cambio de conductor) y en otros dos:

Gráfico: 1121
06:20 Vac Le/Cochera - Eibar
07:00 **A3912** Rel Eibar - Durango - Amorebieta - Lemona - H - Bilbao
Releva con Gráfico 1201 en Durango a las 07:40
08:10 A3924 Ord Durango - Leioa (Rfzo)(Grf 73)
08:45 Vac Leioa Le/Cochera

Gráfico: 1201
07:20 Vac Le/Cochera Durango (Autopista)
07:40 A3912 Re l Durango - Amore - Lemona - H - Bilbao
Releva Al Gráfico 1121 en Durango
09:30 **A3912** Ord Bilbao - H - Lemona - Amorebieta - Durango - Eibar
12:00 **A3912** Rel Eibar - Durango - Amore - Lemona - H - Bilbao
Releva con Gráfico 1202 en Le/Cochera a las 13:05
13:30 Vac Le/Cochera Bi/Termibus

Gráfico: 1202
13:00 Cab Le/Cochera Le/Cochera
13:05 A3912 Re l Lemona - H - Bilbao
Releva al Gráfico 1201 en Le/Cochera
14:30 **A3912** Ord Bilbao - H - Lemona - Amorebieta - Durango - Eibar
17:00 **A3912** Ord Eibar - Durango - Amore - Lemona - H - Bilbao
19:30 **A3912** Ord Bilbao - H - Lemona - Amorebieta - Durango - Eibar
21:25 Vac Eibar Le/Cochera

Figura 4: *Indicaciones para el gráfico 1121.*

Aunque este recorrido parezca en un principio distinto del anterior, resulta que la única diferencia existente es que en este segundo hay dos cambios de conductor. Por lo demás, la línea es la misma, y también coinciden las horas y el sentido de los servicios prestados.

6.2.3 Preproceso de los datos

Para empezar, se ha llevado a cabo una limpieza de datos corruptos: eliminamos filas con símbolos o mezclas de letras y números sin sentido, así como las líneas de 'NULL'-s, que aparecen cuando hay saltos en el tiempo (esto es, cada vez que se enciende el vehículo). Éstas han sido eliminadas a mano, puesto que no había un criterio de identificación de las mismas. Tras ello, la dimensión de los datos es de 5031035 filas por 10 columnas. Se han eliminado otros 18 casos en los que la fila estaba vacía o solo había parte de los datos, y se han modificado otras dos fechas, que aunque no eran correctas, eran evidentes:

22014-12-23 18:29:42 → 2014-12-23 18:29:42
2014-11-27 132014-12-26 11:57:33 → 2014-12-26 11:57:33

Siendo conocedores de los horarios de los servicios prestados por el autobús, se ha descubierto que la variable *fecha* lleva una hora de retraso, por lo que es necesario modificarla. También se han observado saltos atrás en el tiempo en la variable fecha, haciendo que haya dos líneas distintas con la misma fecha.

Se ha creado la variable *seg*, pasando la *fecha* (contiene fecha y hora) a tiempo Epoch. Fijado como origen por convenio el 1-1-1970 a las 00:00, esta variable mide los segundos transcurridos desde el origen hasta dicho instante. Gracias a ella, se ha podido adelantar una hora los datos de la variable *fecha*, así como eliminar los saltos hacia atrás y valores repetidos, un total de 23528, por lo que nos hemos quedado con 5007489 observaciones.

A continuación, se han añadido las variables *conductor*, *línea*, *sentido* (ida=1 o vuelta=2) y *gráfico*.

La siguiente variable analizada ha sido *odómetro*. Esta variable cuenta los kilómetros recorridos por el vehículo, por lo que se supone que su trazo debe ser ascendente. Pero al observar saltos raros en sus valores, y graficarlo, se ha obtenido lo siguiente:

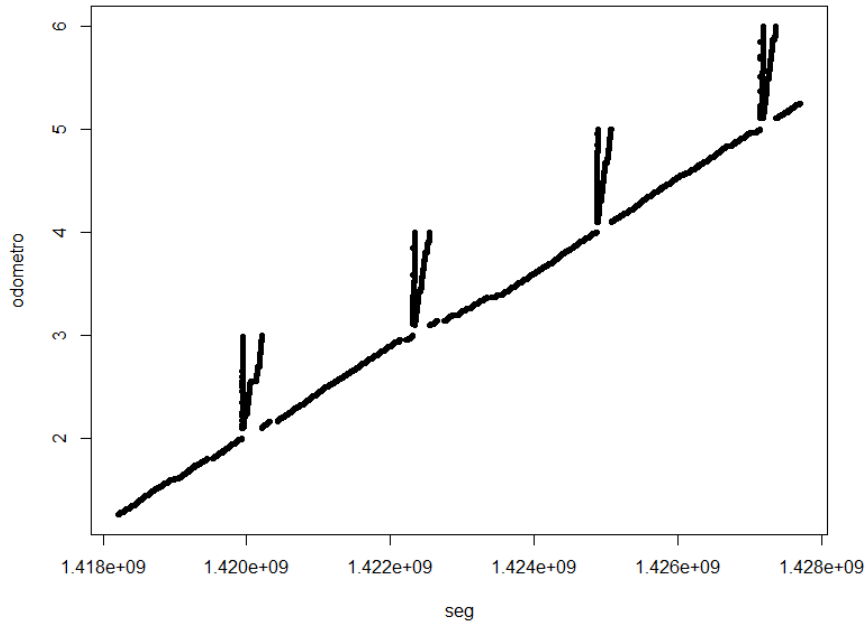

Figura 5: *Valores de la variable odómetro respecto a la fecha.*

Lo cual no tiene sentido, ni por sus valores, comprendidos entre 1.26063 y 5.9999, ni por los saltos. El problema ha resultado ser una cuestión de escritura de decimales en la recogida de datos, por lo que haciendo los cambios oportunos hemos conseguido corregir los errores y obtener los valores correctos de la variable. En la siguiente imagen se pueden observar las dos variables (los valores de la corregida se han dividido por 10000, para que en escala coincidieran con los originales, en realidad toma valores ente 12606.3 y 52483.5).

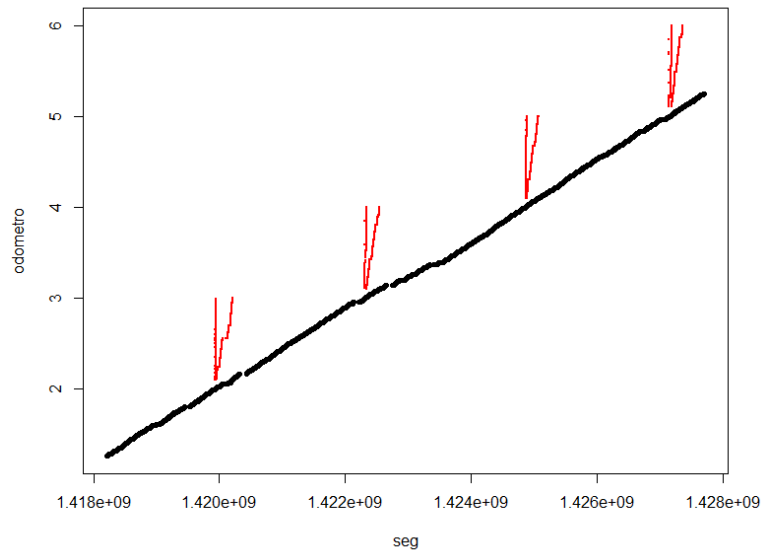

Figura 6: *Valores originales de odómetro (rojo) y valores de la variable corregida (negro)*

La variable *fuel* toma el valor de error por defecto, '-5.-5', en todos los casos, por lo que ha sido eliminada, y las variables *fuel acumulado* y *velocidad* no han dado ningún problema.

El análisis de las *rpm* ha resultado caótico, pues se ha observado que no hay ninguna lógica entre los datos, y que al contrario que el odómetro, en este caso no hay modificación alguna que solucione el problema. Por ello, se ha tenido que prescindir de dicha variable.

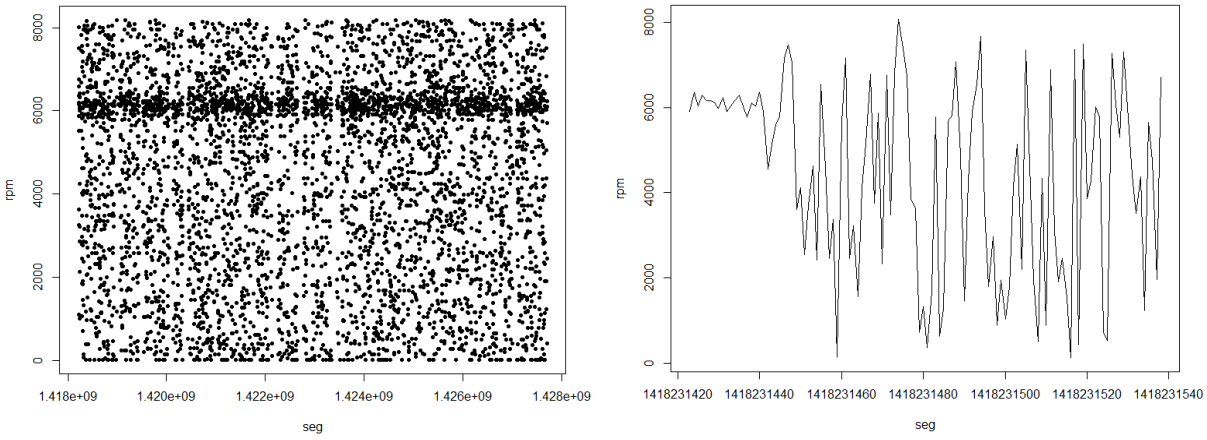

Figura 7: *Valores de rpm (izda) y valores de rpm de los primeros 2 minutos tras una parada (dcha).*

Se ha intentado cambiar de autobús, ya que las revoluciones son un dato bastante relevante y de gran interés en el estudio del consumo, pero en ninguno de los autobuses disponibles se recogen las rpm adecuadamente, por lo que se ha seguido trabajando con el 3255.

Por último, se ha analizado el *consumo instantáneo*, nuestra variable de interés. Es importante fijarnos en que está medida en kilómetros por litro, por lo que valores grandes significan consumos pequeños, y viceversa. Para conocer el consumo en litros de un tramo, hay que dividir la distancia recorrida entre el consumo instantáneo. Llama la atención el valor 125.498, cuya frecuencia es mayor que el 20% (aparece más de un millón de veces). Pero nos informan que este valor está fijado como límite inferior de la variable, cuando el consumo es casi nulo; a eso se debe su alta frecuencia de aparición. Hay que mencionar también que el valor 0 forma otro 15% de los datos, y aunque no tiene sentido como tal (puesto que valores pequeños indican consumos grandes, el valor 0 significaría consumo 'infinito'), resulta que en el 98% de los casos, la velocidad es también 0, por lo que se cree que el valor 0 realmente significa consumo 0.

Para terminar, se han creado las variables *distancia* (mide la distancia, en metros, recorrida por el vehículo en cada segundo), *aceleración* (m/s^2) y *fuel consumido* (L). Para un segundo dado, i, la *distancia* y la *aceleración* se obtienen del siguiente modo (prestando siempre especial atención a los valores perdidos):

$$distancia[i] = velocidad[i] \cdot tiempo(1\ seg)$$

$$aceleración[i] = \frac{velocidad[i] - velocidad[i-1]}{\Delta tiempo\ (1\ seg)}$$

El *fuel consumido*, en cambio, es una variable aditiva, como *odómetro*, por lo que se obtiene recursivamente:

$$fuel\ consumido[i] = fuel\ consumido[i-1] + \frac{distancia[i]}{consumo\ instantáneo[i]}$$

A continuación se puede observar que las variables *distancia* y *fuel consumido* se han construido correctamente:

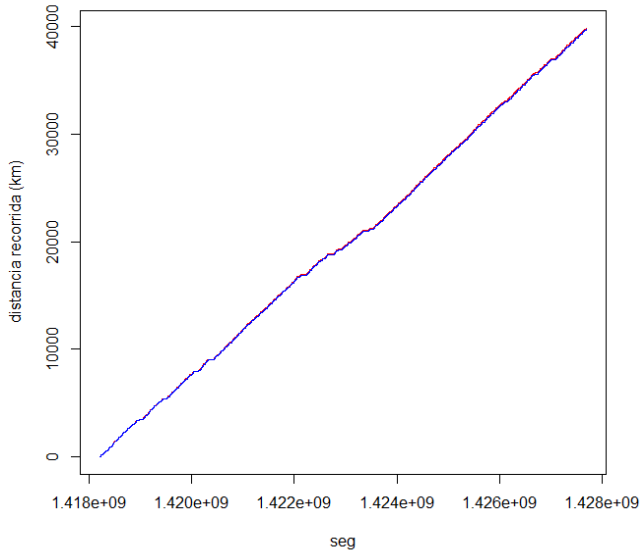

Figura 8: *Gráfico de odómetro (rojo) y la suma de distancias (azul).*

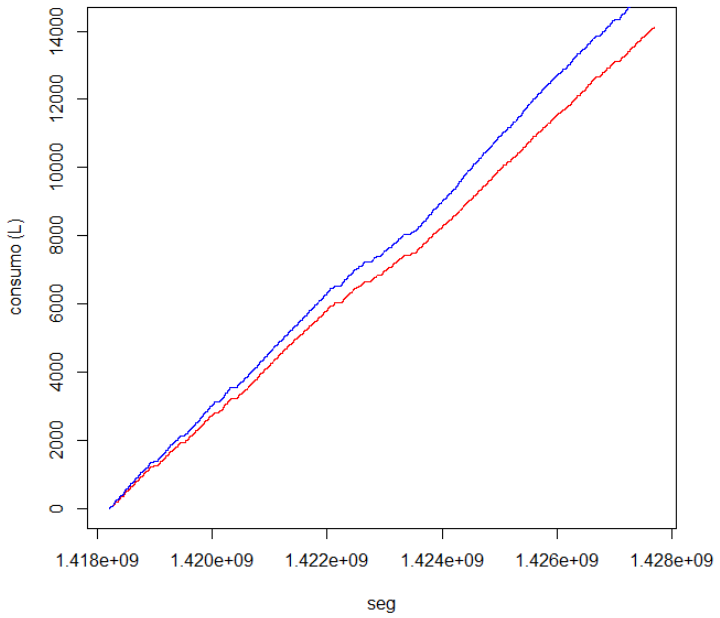

Figura 9: *Gráfico de fuel acumulado (rojo) y fuel consumido (azul).*

La suma de las *distancias* (que se ha dividido por 1000 para adecuar las unidades, pues el *odómetro* esta medido en kilómetros) coincide perfectamente con el *odómetro* (corregido) *(Figura 9)* y el *fuel consumido* se acerca bastante al *fuel acumulado (Figura 9)*. Aun así, siendo el primer caso tan exacto, las diferencias en el segundo nos hacen sospechar sobre los valores y la medición de la variable *consumo instantáneo*.

Llegados a este punto, nuestros datos consisten en 5007489 observaciones de un total de 15 variables: *fecha, seg, longitud, latitud, velocidad, odómetro, consumo instantáneo, fuel acumulado, gráfico, línea, sentido, conductor, aceleración, distancia* y *fuel consumido*.

6.2.4 Métodos de detección y extracción de outliers

Un valor atípico es una observación que se desvía tanto de otras observaciones, que despierta la sospecha de que se generó por un mecanismo diferente. Para proteger los resultados de estas posibles observaciones atípicas, se aplican los procedimientos denominados métodos de detección de outliers.

La mayoría de las bases de datos reales contienen valores atípicos o outliers que toman valores raramente mayores o menores en comparación con el resto del conjunto. Estos valores pueden causar efectos negativos en diversos análisis de datos, como regresiones o ANOVA, basados en hipótesis de distribuciones, pero también pueden dar información importante sobre los datos, o la forma en la que han sido tomados. Por ello, es importante realizar la detección y posterior extracción de outliers en nuestra base de datos.

Existen diversos métodos de detección de outliers **[26]**, clasificados generalmente como formales (hacen uso de test estadísticos) e informales (también denominados técnicas de etiquetado de outliers).

Teniendo en cuenta la magnitud de nuestros datos, se ha utilizado el mismo método que en el primer estudio, el de los cuartiles, para hacer una primera limpieza de los datos. Según nos vayamos adentrando en ellos, puede que sea necesaria una segunda revisión.

Antes de detectar outliers por el método seleccionado, puede ser de ayuda visualizar las variables (solo se representan algunas, pues el resto de histogramas no se creen necesarios):

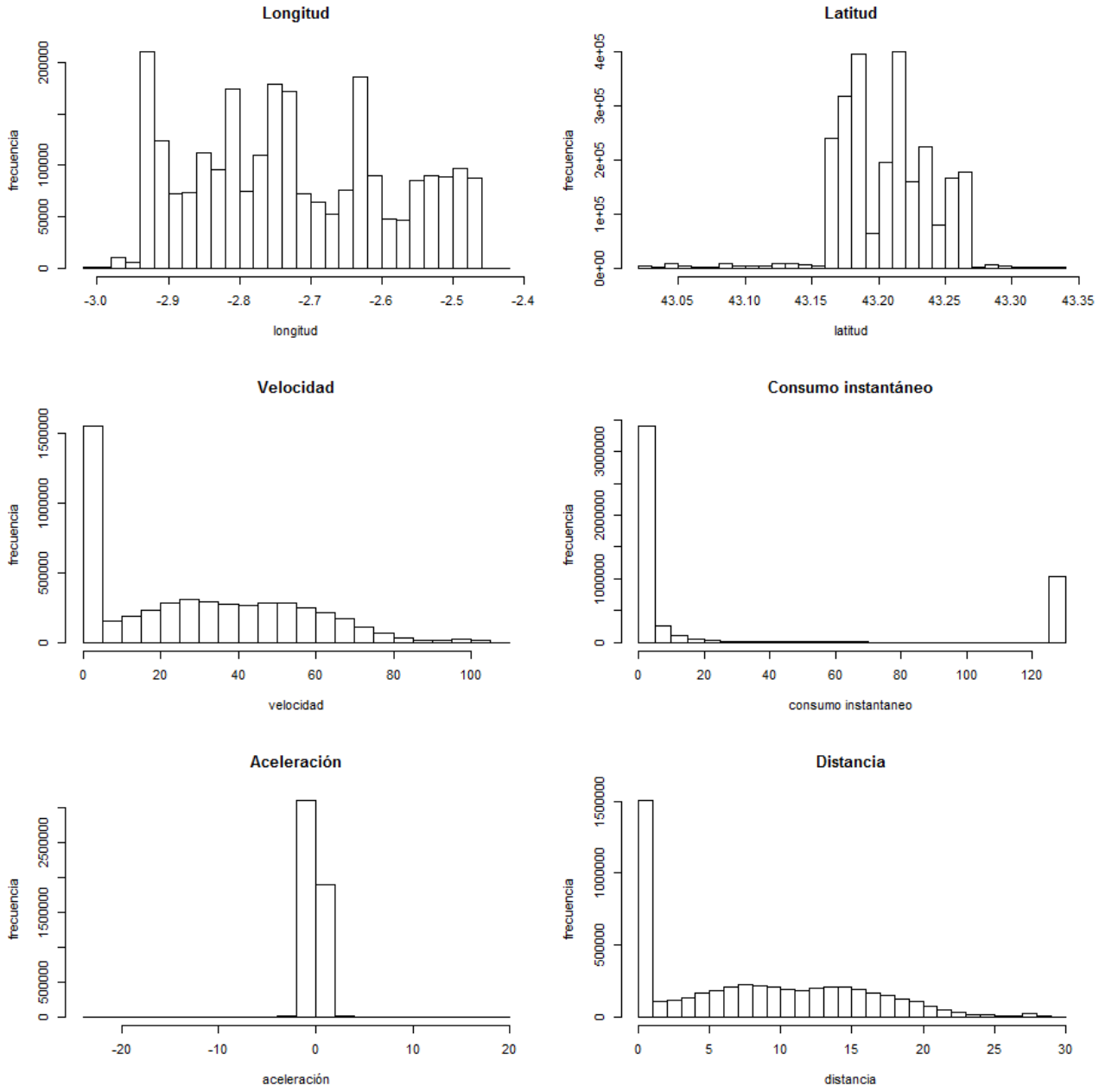

Figura 10: *Histogramas de las variables que nos pueden interesar.*

A continuación, se ha procedido a la detección de outliers en cada variable (solo se han considerado los valores outlier extremos):

- *latitud*: Hay 3083 valores por debajo del límite inferior, pero al graficar dichos datos resulta que corresponden a posiciones de una ruta distinta de las demás, por lo que no se van a eliminar.

- *consumo instantáneo*: Como se ha mencionado en el prepoceso y se puede observar en el histograma, el valor 125.498 aparece muchas veces, pero se encuentra muy alejado del resto, por lo que todas sus apariciones (1040529) y otras observaciones con valores parecidos (1077246 en total) quedan por encima del límite superior considerado por el método en este caso. Pero como conocemos nuestra base de datos, podemos decir que no corresponden a observaciones erróneas, por lo que no serán eliminadas.

- *aceleración*: Hay 179392 valores que se escapan de los límites considerados. La mayor parte de ellos son resultado de los saltos en el tiempo producidos cada vez que se apaga el autobús o, en la mayoría de los casos, por la pérdida de datos durante unos pocos segundos. Se ha intentado corregir estos errores mediante funciones de suavizado, pero teniendo en cuenta que los datos de aceleración que nos van a interesar son más bien pocos (ya que vamos a analizar las paradas y arranques), y que el resto de variables son correctas, se ha decidido no modificarlos ni eliminarlos.

Las variables *longitud*, *velocidad*, *odómetro*, *fuel acumulado, distancia* y *fuel consumido* no han presentado ningún valor considerable como outlier. El resto (*fecha, seg, gráfico*, *línea*, *sentido* y *conductor*) no se ha considerado en este análisis, puesto que no tiene sentido realizarlo.

6.3 Software y métodos

En este apartado se presentan el software y los métodos utilizados, tanto de análisis y procesamiento de datos, como del resto de cuestiones que se han tenido que abordar durante en estudio.

6.3.1 Software empleado: R-Project y RStudio

R (http://www.r-project.org/) es un lenguaje orientado a objetos para la manipulación, análisis estadístico y representación de datos y creación de gráficos. Se trata de un proyecto de software libre, resultado de la implementación GNU del lenguaje S.

R es uno de los lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy popular en el campo de la investigación biomédica, la bioinformática y las matemáticas financieras.

En este trabajo, se han utilizado las siguientes herramientas y paquetes de R:

1) **R** de 64 bits para Windows (v.3.2.1) (http://cran.rstudio.com/)

2) Librería **pls**: Paquete que permite la implementación de la regresión de mínimos cuadrados parciales (PLSR) y la regresión de componentes principales (PCR). (https://cran.r-project.org/web/packages/pls/index.html)

3) Libraría **geosphere**: Paquete que implementa funciones de trigonometría esférica para aplicaciones geográficas.(https://cran.r-project.org/web/packages/geosphere/index.html)

4) Librería **ggmap**: Colección de funciones para visualizar datos y modelos espaciales mediante diversas fuentes online, como Google Maps y Stamen Maps. (https://cran.r-project.org/web/packages/ggmap/index.html)

5) **Librería *rpart***: Paquete que permite la partición recursiva para árboles de clasificación y regresión. (http://cran.r-project.org/web/packages/rpart/index.html)

6) **Librería *rpart.plot*:** Paquete que permite la representación gráfica de árboles generados por rpart. (http://cran.r-project.org/web/packages/rpart.plot/index.html)

RStudio (http://www.rstudio.com/) es una interfaz gráfica de usuario (*Graphical User interface* o GUI) para R que permite al usuario ejecutar R en un entorno amigable. Al igual que R, es un entorno libre y de código abierto y se puede ejecutar en el escritorio (Windows, Mac o Linux) o incluso a través de Internet mediante el servidor RStudio.

6.3.2 Medición de distancias en la Tierra

Sean $a = (lat1, long1)$ y $b = (lat1, long1)$ dos puntos situados en la superficie de la Tierra, y R el radio de la misma. Para calcular la distancia entre los puntos a y b en términos de su latitud y longitud es necesario adoptar cierto nivel de abstracción, puesto que es imposible obtener el valor exacto, debido a la inmensa cantidad de irregularidades en la superficie terrestre.

Comúnmente, son tres los puntos de vista desde los que de afronta este problema:

1) La Tierra como superficie plana

   Se considera una aproximación en plano (2D) de la Tierra, de forma que el camino más corto entre dos puntos, es una línea. En estos casos, la base de las fórmulas reside en el teorema de Pitágoras. Como se puede intuir, esta aproximación se vuelve imprecisa a la vez que la separación entre los puntos crece o cuanto más cerca de los polos se encuentran.

   Éste método es conocido como **Aproximación equirrectangular**, y la distancia entre a y b se obtiene de la siguiente forma:

$$d = \sqrt{\left((lon2 - lon1) \cdot cos\left(\frac{lat1 + lat2}{2}\right)\right)^2 + (lat2 - lat1)^2} \cdot R$$

   Es muy funcional, sobre todo si la precisión no es demasiado importante, pero sí el rendimiento, como veremos a continuación.

2) La Tierra como superficie esférica

Asumir que la Tierra es un esferoide (en vez de un elipsoide) simplifica bastante los cálculos a la hora de trabajar, pero ello conlleva a poder cometer un posible error del 0.5%, lo que en la mayoría de los casos seguirá siendo una aproximación acertada. La distancia más corta entre dos puntos sobre la superficie de una esfera, se calcula sobre el círculo máximo (*great-circle*) que contiene ambos puntos. Esta distancia es conocida como distancia ortodrómica (*great-circle distance*), y existen diversas formas de calcularla.

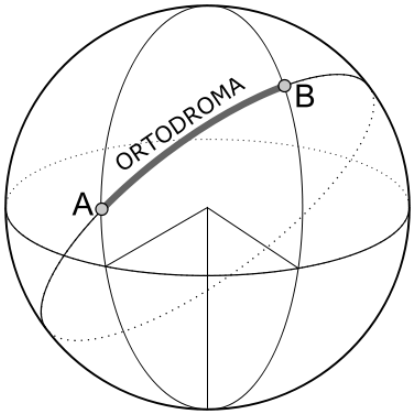

Figura 11: *Distancia ortodrómica entre dos puntos a lo largo de un círculo máximo sobre la superficie de una esfera.*

Ley Esférica del Coseno:

$$d = \operatorname{acos}(\sin(lat1) \cdot \sin(lat1) + \cos(lat1) \cdot \cos(lat1) \cdot \cos(lon2 - lon1)) \cdot R$$

Esta es seguramente la forma más simple de calcular la distancia ortodrómica, pero el mayor problema de esta aproximación es que en pequeñas distancias (a menos de un kilómetro de distancia, más o menos), está muy condicionado a errores de redondeo.

Método de Haversine:

$$d = R \cdot 2 \cdot \operatorname{asin}\left(\min\left\{ 1, \sqrt{\sin\left(\frac{lat1 + lat2}{2}\right)^2 + \cos(lat1) * \cos(lat2)} \right\} \right)$$

Ésta es una formulación alternativa a la anterior, y es más robusta ante cálculos de pequeñas distancias.

3) La Tierra como elipsoide

Una elipsoide aproxima la superficie de la Tierra mucho mejor que una esfera o que una superficie plana. La distancia más corta entre dos puntos de una elipsoide se calcula sobre la geodésica que los une.

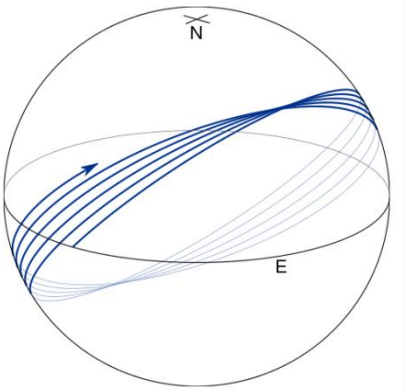

Figura 12 *Una geodésica sobre una elipsoide ovalada.*

En estos métodos, en lugar de definir el radio R de la Tierra, hay que decidir qué elipsoide simula mejor la superficie terrestre (hoy en día lo más empleado es el WGS84, un sistema de coordenadas geográficas mundial estándar en geodesia, cartografía y navegación, que data de 1984). Las fórmulas de Vincenty se basan en esta perspectiva para el cálculo de las distancias.

Método de Vincenty Elipsoide:

La distancia entre dos puntos se calcula mediante un algoritmo iterativo, que es ampliamente utilizado por la buena precisión que obtiene.

a, b = semiejes mayor y menor del elipsoide
f = achatamiento del elipsoide $(a-b)/a$
φ_1, φ_2 = latitud geodética
L = diferencia de longiud
$U_1 = \text{atan}((1-f).\tan\varphi_1)$ *(U es 'latitud reducida')*
$U_2 = \text{atan}((1-f).\tan\varphi_2)$
$\lambda = L$, $\lambda' = 2\pi$
while $\text{abs}(\lambda-\lambda') > 10^{-12}$ { *(implica un error<0.06mm)*
  $\sin\sigma = \sqrt{[\ (\cos U_2.\sin\lambda)^2 + (\cos U_1.\sin U_2 - \sin U_1.\cos U_2.\cos\lambda)^2\ ]}$
  $\cos\sigma = \sin U_1.\sin U_2 + \cos U_1.\cos U_2.\cos\lambda$
  $\sigma = \text{atan2}(\sin\sigma, \cos\sigma)$
  $\sin\alpha = \cos U_1.\cos U_2.\sin\lambda / \sin\sigma$
  $\cos^2\alpha = 1 - \sin^2\alpha$
  $\cos 2\sigma_m = \cos\sigma - 2.\sin U_1.\sin U_2/\cos^2\alpha$
  $C = f/16.\cos^2\alpha.[4+f.(4-3.\cos^2\alpha)]$
  $\lambda' = \lambda$

$\lambda = L + (1-C).f.\sin\alpha.\{\sigma+C.\sin\sigma.[\cos2\sigma_m+C.\cos\sigma.(-1+2.\cos^2 2\sigma_m)]\}$
}
$u^2 = \cos^2\alpha.(a^2-b^2)/b^2$
$A = 1+u^2/16384.\{4096+u^2.[-768+u^2.(320-175.u^2)]\}$
$B = u^2/1024.\{256+u^2.[-128+u^2.(74-47.u^2)]\}$
$\Delta\sigma = B.\sin\sigma.\{\cos2\sigma_m+B/4.[\cos\sigma.(-1+2.\cos^2 2\sigma_m) -$
$B/6.\cos2\sigma_m.(-3+4.\sin^2\sigma).(-3+4.\cos^2 2\sigma_m)]\}$
$d = b.A.(\sigma-\Delta\sigma)$

Figura 13 *Pseudocódigo para el cálculo de la distancia de Vincenty.*

La fórmula puede no tener solución para dos puntos casi antipodales, por lo que se requiere limitar el número de iteraciones del algoritmo para evitar ese caso.

Hay mucho más por analizar respecto a los distintos métodos de cálculo (como por ejemplo, si se tiene o no en cuenta la altitud a la que se encuentran los puntos, esto es, la irregularidad de la superficie terrestre, o si se decide fijar el radio R dependiendo de la latitud media en la que se sitúan), pero centrándonos en nuestro caso particular, teniendo en cuenta que las distancias de nuestro interés no van a pasar de los 200m, se ha decidido emplear la aproximación equirrectangular.

Las dos razones principales han sido la precisión de este método en pequeñas distancias, y sobre todo, la simplicidad de los cálculos a realizar, ya que se va a tener que calcular miles de veces.

Para terminar, queda por definir el radio R. Realmente, solo hay que decidir si tomar un radio general, o hacerlo dependiente de la latitud de los puntos. Tras haber realizado pruebas de ambos casos, apenas ha habido diferencias entre las distancias computadas, por lo que una vez más, y poniendo especial énfasis en la sencillez del método, el radio R queda fijo, como el valor medio de la Tierra, considerado $R = 6371km$.

6.3.3 Métodos de análisis de datos

Los métodos empleados para el análisis y la construcción de los modelos han sido la regresión lineal y los árboles de regresión.

6.3.3.1 Regresión lineal

El análisis de regresión lineal es una técnica estadística utilizada para estudiar la relación entre variables. Tanto en el caso de dos variables (regresión simple) como en el de más de dos (regresión múltiple), la regresión lineal puede utilizarse para explorar y cuantificar la relación entre una variable dependiente (Y) y una o más variables llamadas independientes o predictoras ($X_1, X_2, \dots, X_N$). En el caso de una única variable independiente, la ecuación de regresión define una recta en plano (Figura 14), mientras que en la regresión múltiple, define un hiperplano en un espacio multidimensional (Figura 15).

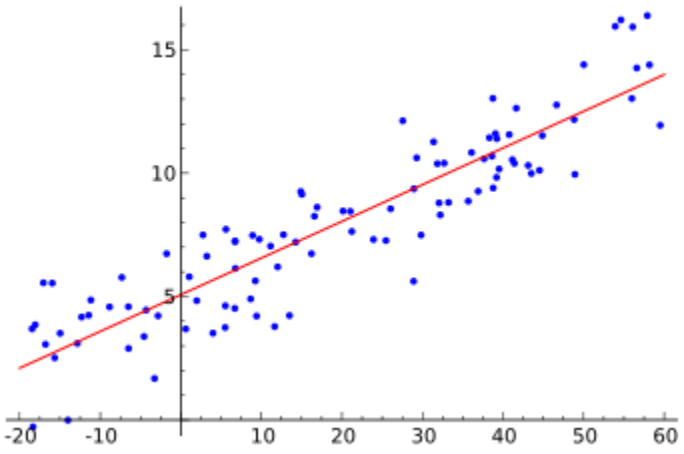

Figura 14: *Regresión lineal con una variable independiente y una variable dependiente y la recta de regresión (en rojo).*

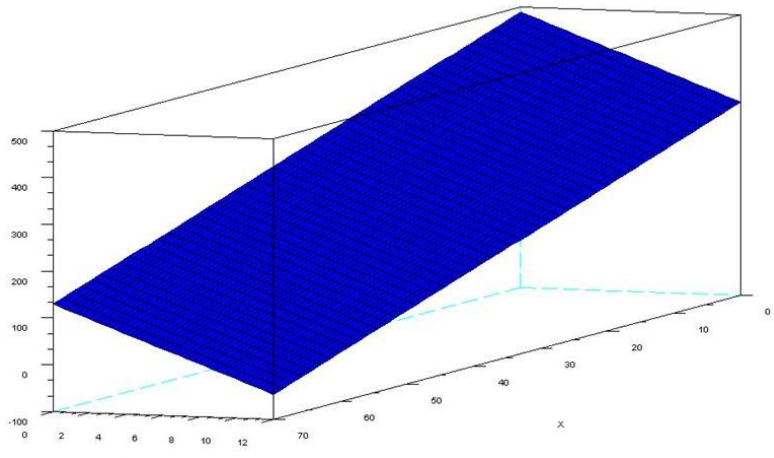

Figura 15: *Regresión múltiple con dos variables independientes y el hiperplano (en azul).*

El algoritmo empleado en el estudio anterior, la regresión de mínimos cuadrados parciales (PLS), en lugar de encontrar hiperplanos de mínima varianza entre la variable de respuesta y las variables independientes, encuentra una regresión lineal mediante la proyección de las variables de predicción y las variables observables a un nuevo espacio. Es especialmente adecuada cuando la matriz de predictores tiene más variables que observaciones, y cuando hay correlación entre los valores de ella.

En nuestro caso, las observaciones superan, por mucho, las variables y el tema de la correlación entre variables se puede solucionar haciendo un buen análisis y selección de los datos a introducir para la regresión, por lo que nos hemos decantado por la Regresión lineal.

6.3.3.2 Árboles de regresión

Los métodos basados en árboles (o árboles de decisión) son bastante populares en el *data mining*, pudiéndose usar para clasificación y regresión. Son útiles para la exploración inicial de datos y apropiados cuando hay un número elevado de datos, y existe incertidumbre sobre la manera en que las variables explicativas deberían introducirse en el modelo.

Un árbol de decisión tiene unas entradas las cuales pueden ser un objeto o una situación descrita por medio de un conjunto de atributos y a partir de esto devuelve una respuesta la cual en últimas es una decisión que es tomada a partir de las entradas. Los valores que pueden tomar las entradas y las salidas pueden ser valores discretos o continuos.

Los árboles de decisión se han adaptado para tareas como la regresión, el agrupamiento o la estimación de probabilidades. Un árbol de regresión se construye de arriba hacia abajo y la estrategia que se utiliza en la construcción del árbol es la segmentación de la muestra en grupos homogéneos respecto a la variable respuesta. Esto es, se parte de un nodo raíz añadiendo particiones (que constituyen los hijos del nodo partido). En cada partición los ejemplos se van dividiendo entre los hijos. Finalmente, se llega al punto en que todos los ejemplos que caen en los nodos inferiores son de la misma clase y esa rama ya no sigue creciendo. En un árbol de regresión, la función aprendida tiene dominio real (en los de clasificación el dominio es discreto), y los nodos hoja del árbol se etiquetan con valores reales, de tal forma que una cierta medida de calidad se maximice, en este caso, obtener la mínima varianza de los ejemplos que caen en ese nodo respecto al valor asignado.

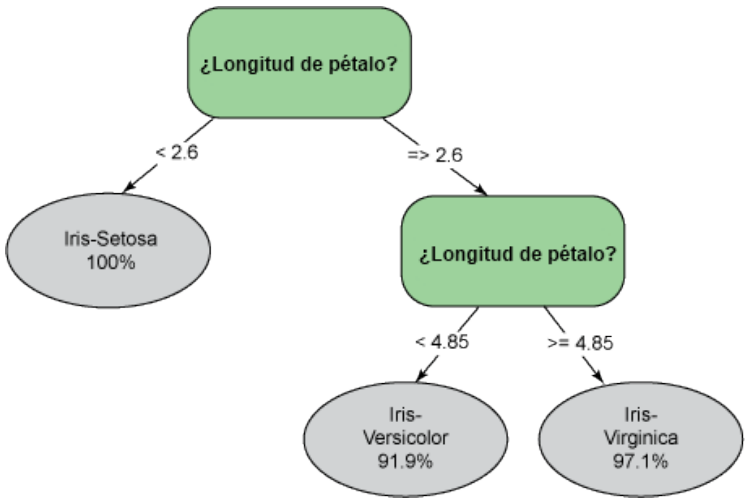

Figura 16: *Árbol de regresión.*

La partición se obtiene de forma recursiva mediante una serie de cuestiones binarias (sí/no), expresadas en términos de una única variable independiente en cada momento. Inicialmente se asigna a todos los individuos a un nodo inicial. Se determina la variable que mejor divide los datos en 2 grupos, según los criterios de partición o medida de calidad a mejorar. Los datos se separan y el mismo proceso se aplica separadamente a cada subgrupo. Este proceso de partición se repite hasta que los subgrupos contienen un número mínimo de datos o hasta que no puede realizarse ninguna mejora en la homogeneidad de los subgrupos. La estructura que representa las diferentes particiones aplicadas de forma recursiva da lugar a

un árbol binario y al final del proceso el conjunto de datos resulta dividido en diversas clases: los nodos terminales. A cada nodo terminal se le asigna una clase, en el caso de que la variable respuesta sea categórica, o se le predice un valor de la variable respuesta en el caso de que sea continua. Estos árboles de clasificación y regresión son conocidos como CART (*Classification and Regression Trees*).

La librería *rpart* de R sirve para construir modelos de clasificación o de regresión de una estructura muy general usando el proceso de construcción de árboles visto anteriormente. El número mínimo de observaciones, *nmin*, que debe tener un nodo para dividirlo es, de manera predeterminada, 20 y el número mínimo de observaciones que debe tener un nodo termina es *nmin/3*. Para construir árboles de regresión emplea el método *anova*, esto es, elegir la división que maximice la suma de cuadrados entre grupos en un simple análisis de varianza **[27]**.

Un parámetro importante de estos métodos es el parámetro de complejidad, *cp*, el cual tiene una interpretación bastante directa: si una partición no mejora el R^2 del modelo por lo menos en *cp*, se considera irrelevante. La principal función de este parámetro es ahorrar tiempo de cálculo mediante la poda de divisiones que no valen la pena.

Para terminar, queda podar el árbol. En general, cada partición de más significa menor error de clasificación en el grupo de entrenamiento, pero se corre el peligro de sobreajustar el modelo. La función *plotcp* muestra el punto óptimo para ejecutar la poda, que por convenio se toma en el árbol con mayor error de validación cruzada, que sea menor que la suma del menor error de validación cruzada y el error estándar en dicho árbol.

6.4 Construcción del modelo

6.4.1 Detección de pasos por paradas y detenciones

Como la idea es analizar las conductas de parada y arranque por paradas, de los datos obtenidos tras el preproceso de los mismos se han seleccionado los correspondientes a los días en los que se ha trabajado con gráficos equivalentes, que como hemos visto anteriormente, corresponden a los mismos itinerarios, coincidiendo líneas, servicios y horarios.

Tras darnos cuenta que en muchos de los casos el dato del *gráfico* no coincide con los servicios que luego se observan en los datos, se ha hecho un análisis manual, día a día, y nos hemos quedado con el patrón que más se cumple, que es equivalente a los gráficos 1201 y 1121 anteriormente presentados, en el que se realizan 3 servicios de ida y otros 3 de vuelta de la línea 3912.

De los 111 días de los que disponemos datos, nos quedamos con 80: 16 días de diciembre, 24 de enero, 24 de febrero y 16 de marzo (de los 31 restantes, 9 corresponden a días de revisión o en taller y 22 a distintos gráficos). Gracias al análisis manual realizado, se ha podido comprobar que algunos días no hay datos a partir de una hora (por ejemplo, el 24 y 31 de diciembre el último servicio es el de las 17:30), que apenas hay datos de algunos servicios concretos o que ha habido una desviación de la ruta esperada (puede ser porque haya habido un accidente, presencia de obras, etc.), por lo que se han eliminado los datos correspondientes a dichos servicios, para quedarnos solo con los completos (o casi completos).

Cuando decimos que no hay datos, nos referimos a dos tipos de situaciones. Por una parte, lo que ocurre en algunos casos (como pueden ser las vísperas de los festivos navideños) es que el autobús deja de emitir datos a partir de un momento, por lo que suponemos que se han terminado los servicios y el autobús está parado. Pero lo que ocurre la mayoría de las veces es que desde que se arranca el vehículo (al comienzo de un servicio, por ejemplo) hasta que se empiezan a recibir los datos de posición *(latitud, longitud)*, transcurre bastante tiempo. Y aunque el resto de variables se recibe correctamente, para poder analizar lo que ocurre en las paradas del recorrido, nos resulta indispensable contar con la información de posición. Teniendo en cuenta que los servicios de la línea 3912 tienen una duración aproximada de hora y media, no tener datos de los primeros 30 minutos implica la pérdida de un tercio de la ruta.

En total, faltan 84 servicios (44 de ida y 40 de vuelta) y se han eliminado los datos de 13 servicios (10 de ida y 3 de vuelta), por lo que de los 240 servicios de ida y 240 de vuelta teóricos, se estima que contamos con unos 196 de ida y 200 de vuelta, que conforman 386813 observaciones de los datos. Los servicios con muy pocos datos no se han eliminado, ya que la poca información de la que disponen nos puede ser de ayuda.

	Servicios sin datos		**Servicios a eliminar**		**Total servicios eliminados**		**Total por mes**	**Servicios con muy pocos datos (<1/3)**
	Ida	Vuelta	Ida	vuelta	Ida	Vuelta		
Dic.	2	0	2	1	4	1	5	0
Ene.	2	6	0	0	2	6	8	2 ida
Feb.	20	7	6	1	26	8	34	5 ida, 1vuelta
Mar.	20	27	2	1	22	28	50	1 ida
Total	44	40	10	3	54	43	**97**	8 ida, 1 vuelta
	84		**13**		**97**			

Tabla 8: *Número de servicios prescindidos por mes.*

Para la detección de pasos por paradas, lo primero es identificar las paradas. La línea 3912 consta de 66 paradas de ida (Bilbao→Eibar) y 67 paradas de vuelta (Eibar→Bilbao).

Figura 17.1: *Paradas de ida de la línea 3912.*

Figura 17.2: *Paradas de vuelta de la línea 3912.*

Aun contando con la ayuda de Google Maps (https://www.google.es/maps) y la página web de Bizkaibus (http://apli.bizkaia.net/apps/danok/tq/index.html?Idioma=ES), donde aparecen las paradas dibujadas geográficamente, no ha sido fácil la tarea de posicionarlas, pues las carreteras experimentan continuos cambios, lo que hace que los mapas de las dos webs que se han consultado en diversos casos no coincidan entre ellas, ni con la trayectoria realizada por el autobús (Figura 18), y no quede clara la posición exacta de la parada. Tras un primer posicionamiento, viendo que en algunas paradas no se detectaba ningún paso, éstas se han modificado de acuerdo con lo observado en los datos (se ha marcado el punto medio en el que el autobús paraba en los distintos servicios como el lugar de parada).

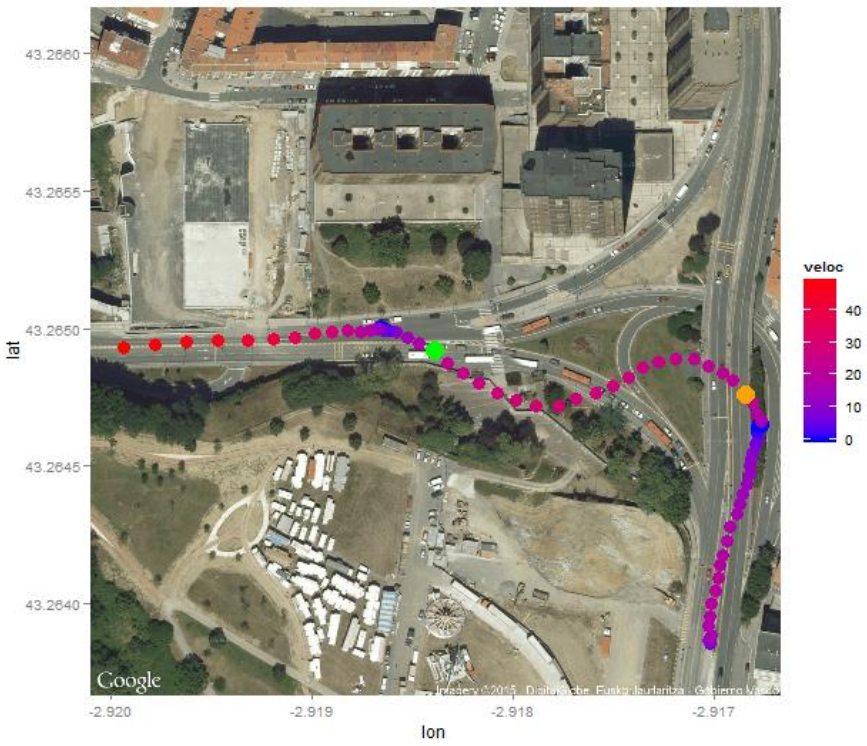

Figura 18: *Parada de bus (verde) y trayectoria del vehículo.*

Tras obtener la posición más o menos exacta de las paradas, queda calcular el número de veces que nuestro vehículo pasa por cada una de ellas. Para identificar los pasos, se ha fijado un radio r y se considera que el autobús ha pasado por la parada P si pasa por algún punto dentro de la circunferencia de centro P y radio r, $C(P,r)$.

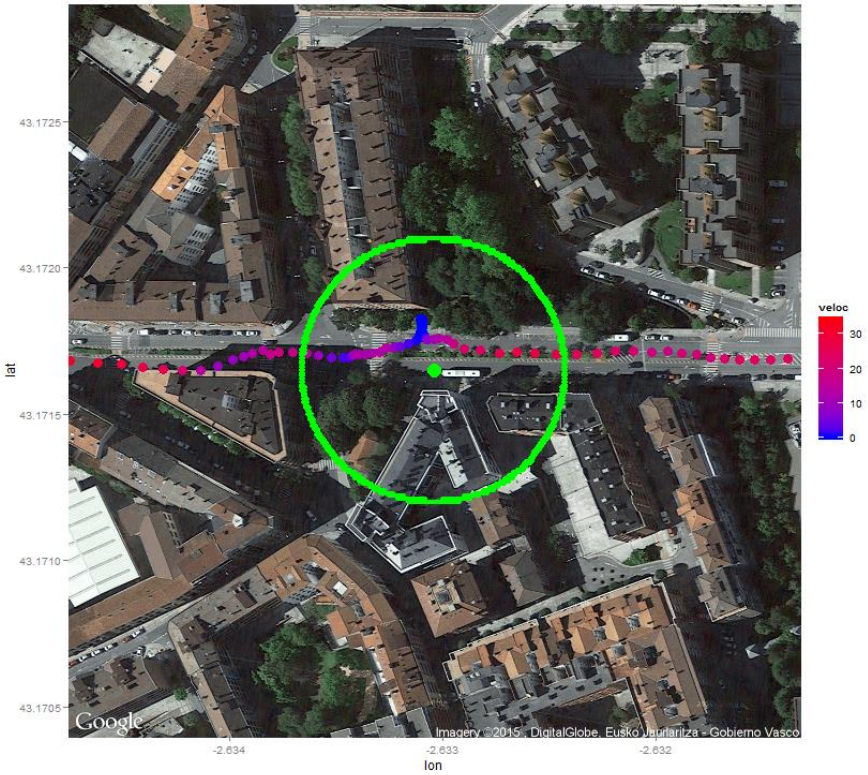

Figura 19: *Parada de bus (verde) y circunferencia a considerar con $r = 50m$.*

Para fijar el radio, se han realizado pruebas con tres medidas: 10m, 25m y 50m. En ocasiones, los datos de posicionamiento son bastante inexactos, y el vehículo parece estar circulando en dirección contraria, incluso por fuera de la calzada. Por ello, con un radio de 10m, muchos pasos por parada no son detectados. Por otra parte, en las pruebas con $r = 50m$, se ha observado que dependiendo de la trayectoria de la carretera, se detectan pasos por paradas que en realidad no lo son (Figura 20). También tenemos en cuenta que en un círculo de esas dimensiones (hay 100 metros de distancia entre los puntos más alejados), cuando vayamos a detectar si el vehículo ha parado o no, puede que para distintos servicios se detecten distintos puntos de parada dentro del espacio considerado. En la Figura 18, por ejemplo, obviando que el recorrido no coincide con el trazo de las vías, el punto naranja indica la presencia de un semáforo; además, unos metros antes de la parada hay un paso de peatones, y otro después.

Por tanto, y viendo los resultados obtenidos, se ha decidido fijar el radio en 25m, que es el que menos problemas ocasiona.

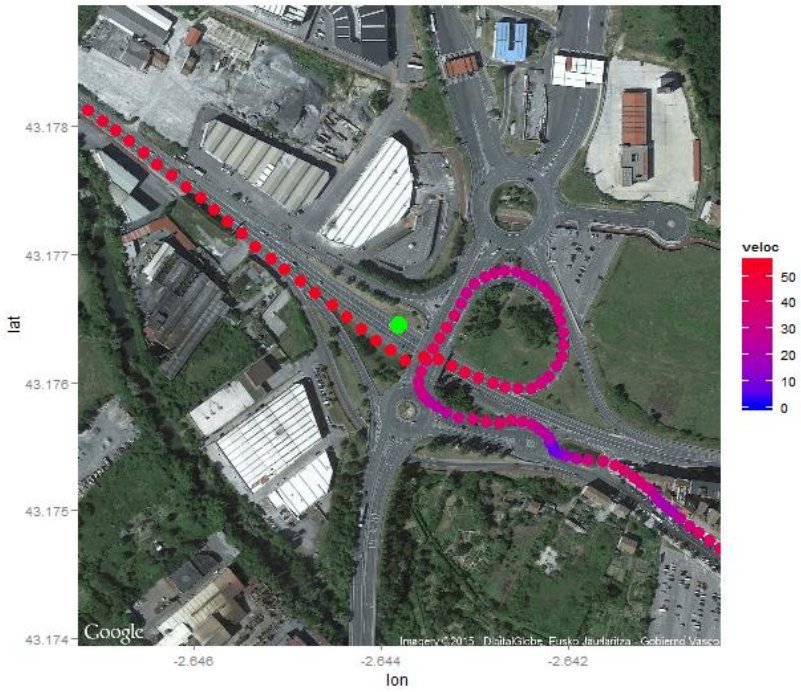

Figura 20: *Parada de bus (verde) y trayectoria del vehículo.*

No se han tenido en cuenta la primera ni la última parada de los servicios, porque antes de la primera no hay frenada, ni tampoco aceleración tras la última, y además la primera de la ida es la última de la vuelta, y viceversa, lo que ocasiona problemas de diferenciación de servicios. En total, se han detectado 18267 pasos por las 129 paradas, 8827 en las de ida y 9440 en las de vuelta, con las siguientes frecuencias:

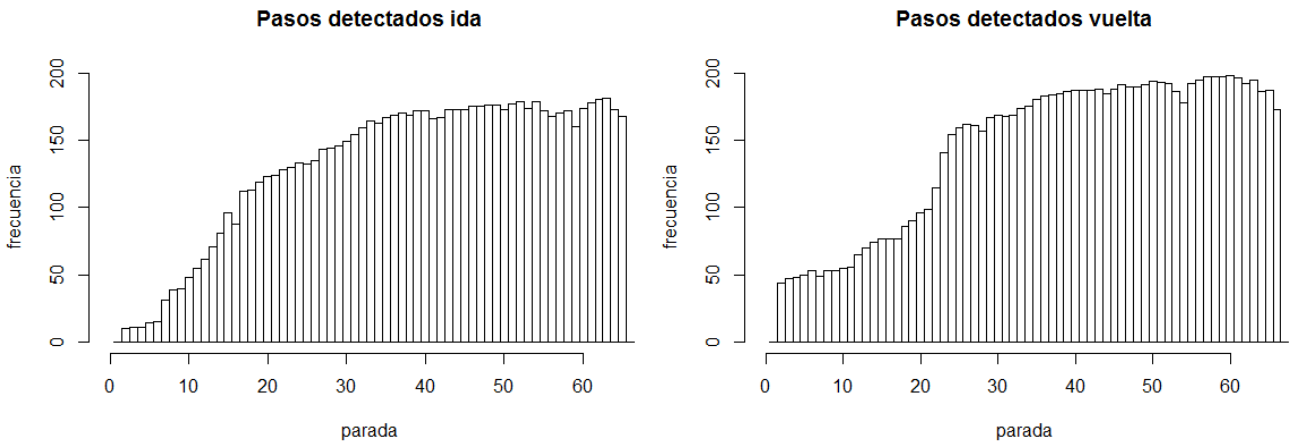

Figura 21: *Frecuencia de pasos por paradas de ida y de vuelta.*

El siguiente paso consiste en determinar si en los pasos de parada el vehículo ha parado o no. Para ello, se parte del punto más cercano a la parada detectado y nos vamos alejando, hacia ambos lados de la marcha alternamente, hasta un máximo de 30 metros. Si se encuentra un punto con *velocidad=0*, se para y se guarda como parada. En caso contrario, pero si la velocidad llega a reducirse por debajo de los 10km/h, se guarda como reducción, pero no parada.

En total, de los 18267 pasos de parada localizados, se han detectado 9657 detenciones, de entre las cuales 7018 han sido detenciones por completo y 1592 han reducido la velocidad por debajo de los 10km/h, pero no se han detenido del todo.

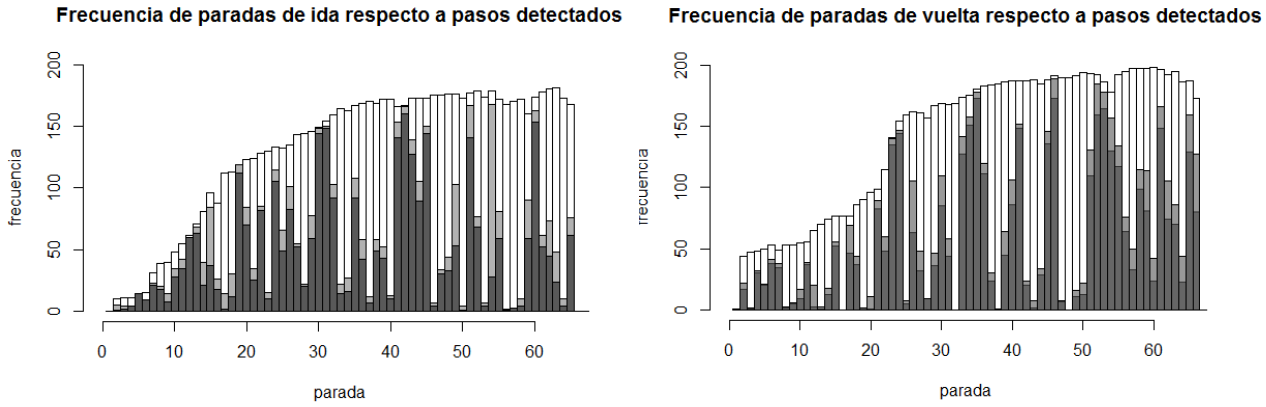

Figura 22: *Frecuencia de paradas (gris oscuro), reducciones (gris) y no paradas (blanco) de los pasos detectados.*

Tras analizar los datos obtenidos, se ha decidido no contar con los casos en los que el vehículo no llega a detenerse, ya que la información de la que prescindimos es pequeña (16.5%), la frenada anterior y la aceleración posterior no son claras en la mayoría de los casos y los datos distan bastante de los casos en los que sí se detiene por completo.

6.4.2 Análisis del consumo

Para realizar el análisis del consumo de combustible se han seleccionado dos de las paradas con más detenciones detectadas: la parada San Miguel en Amorebieta (Figura 23), la 31º del sentido Bilbao→Eibar (ida) con un total de 148 detenciones, a la que denominamos *i31* y la parada 46º del sentido Eibar→Bilbao (vuelta) en el Hospital de Galdakao (Figura 24) con173, a la que denominamos *v46*.

Figura 23: *Parada San Miguel en Amorebieta y parada del Hospital de Galdakao.*

Para cada parada se han medido las siguientes 17 variables:

consumoA	consumoF
velocM	FvelocM
velocm	Fvelocm
velocmedia	Fvelocmedia
acelM	FacelM
acelm	Facelm
acelmedia	Facelmedia
tiempoAcel	tiempoFren
conductor_p	

Tabla 9: *Variables creadas para el análisis del consumo en las paradas.*

Donde *conductor_p* es el conductor en dicha parada, *consumoA* es el combustible consumido, *velocM* la velocidad máxima, *velocm* la mínima y *velocmedia* la media, *acelM* la aceleración máxima, *acelm* la mínima y *acelmedia* la media y *tiempoAcel* el tiempo (en segundos) que ha tardado, medidas en la aceleración. *consumoF*, *FvelocM*, *Fvelocm*, *Fvelocmedia*, *FcelM, Facelm*, *Facelmedia* y *tiempoFren* son las equivalentes a las anteriores, pero referentes a la frenada.

Pero quedan por definir las propias 'aceleración' y 'frenada'. Un punto que se ha querido mantener durante todo el trabajo, es la objetividad de los datos, y muy especialmente el no mezclar ni comparar datos que no correspondan a mismas situaciones. Si tomásemos como 'aceleración' los primeros x segundos tras el arranque, podría pasar que en un caso se recorriera cierta distancia, y en otro el doble, o la mitad, depende de la velocidad y aceleración con las que salga el chófer. Por tanto, en vez de plantearlas en función del tiempo, se han definido en función de la distancia recorrida.

Una vez más, para fijar esta distancia, se han hecho pruebas con distintos valores. Esta vez han sido 30m, 50m y 100m. Para elegir la mejor opción, se ha creado un modelo de regresión lineal para *consumoA* por cada una, en las paradas que vamos a analizar, incluyendo en un principio todas las variables excepto el conductor y *consumoF*. El consumo en la frenada no se ha tenido en consideración en este primer análisis, puesto que se cree que está condicionado al modo de circulación de los segundos previos a la propia frenada. Por ello, en este caso entenderemos como *consumo* (nuestra variable de interés a analizar en este trabajo) lo consumido en la aceleración.

Todo el proceso (respectivo a la construcción del modelo lineal) que se va a presentar a continuación, se ha realizado con las tres medidas, pero se ha visto que tomando 100m, suponemos que por la gran cantidad de datos, no se consigue representar nada bien el consumo en función de las demás variables. En el caso de la parada de Amorebieta, los resultados con 30m y 50m han sido parecidos, y en la de Galdako, algo mejores con 30m. Por tanto, y sin perder de vista el objetivo de construir un modelo equilibrado en cuanto a sencillez y calidad, se ha decidido considerar como parada y frenada, los 30 metros anteriores y posteriores a la detención, respectivamente.

En las próximas líneas encontramos el trabajo realizado y los resultados obtenidos con dicho criterio.

Para continuar el estudio, se han creado dos nuevos ficheros de datos, correspondientes a las dos primeras paradas que van a ser analizadas, con las 17 variables presentadas en la Tabla 9. El primero consta de 148 observaciones, y el segundo de 173, el número de paradas detectadas en cada una de ellas.

Parada San Miguel (Amorebieta)

El primer paso es la visualización de las variables (Figura 24).

Tras analizar los histogramas y dibujar los diagramas de dispersión dos a dos (con todas las variables, excepto con el conductor), se han detectado dos puntos raros, pertenecientes a dos servicios en los que ha habido problemas al paso de esa parada (en el primer caso, un fallo de medición de la velocidad y en el segundo, la falta de datos durante 7 segundos), por lo que han sido eliminados. Los diagramas resultantes se encuentran en la Figura 25.

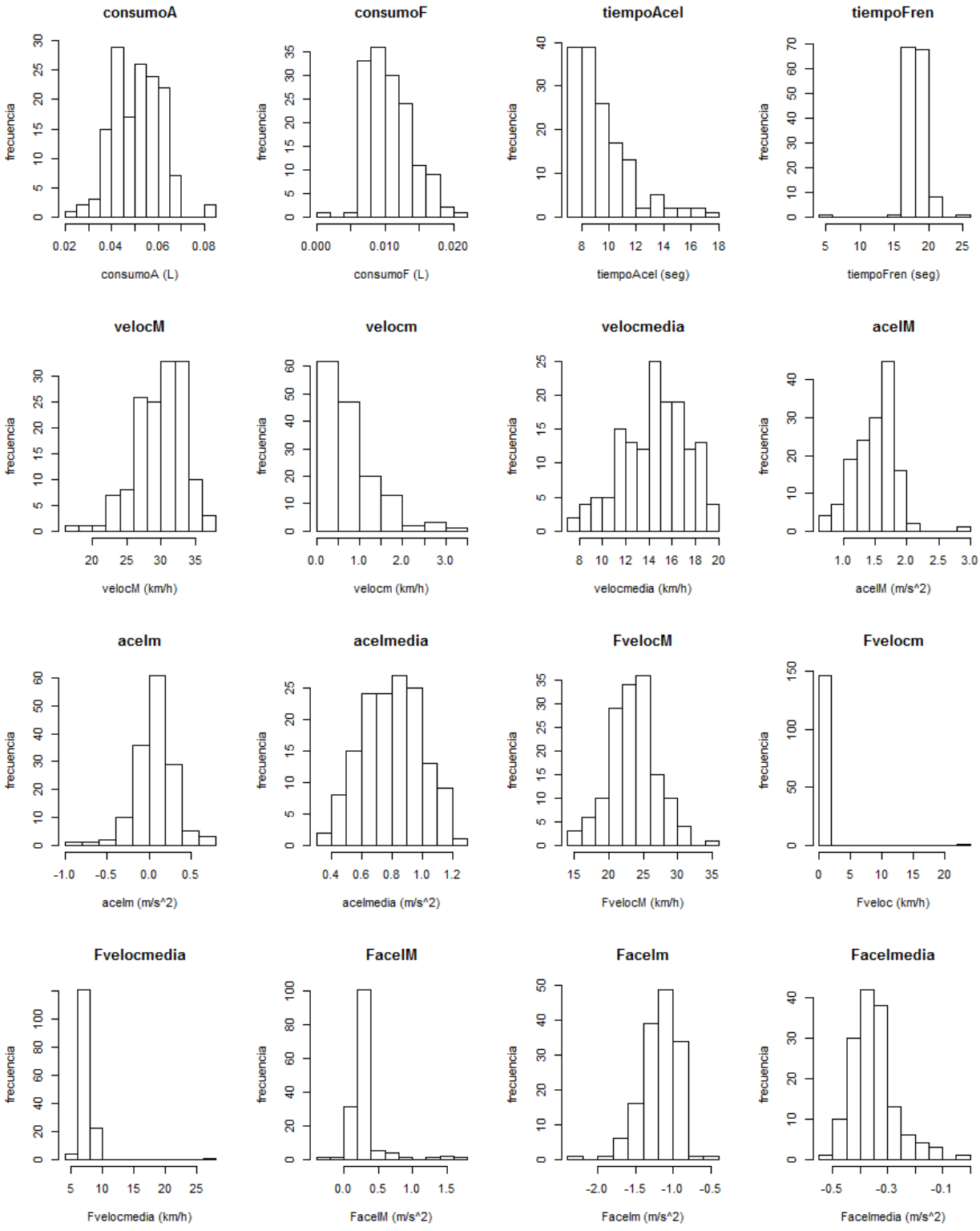

Figura 24: *Histogramas de las variables de i31.*

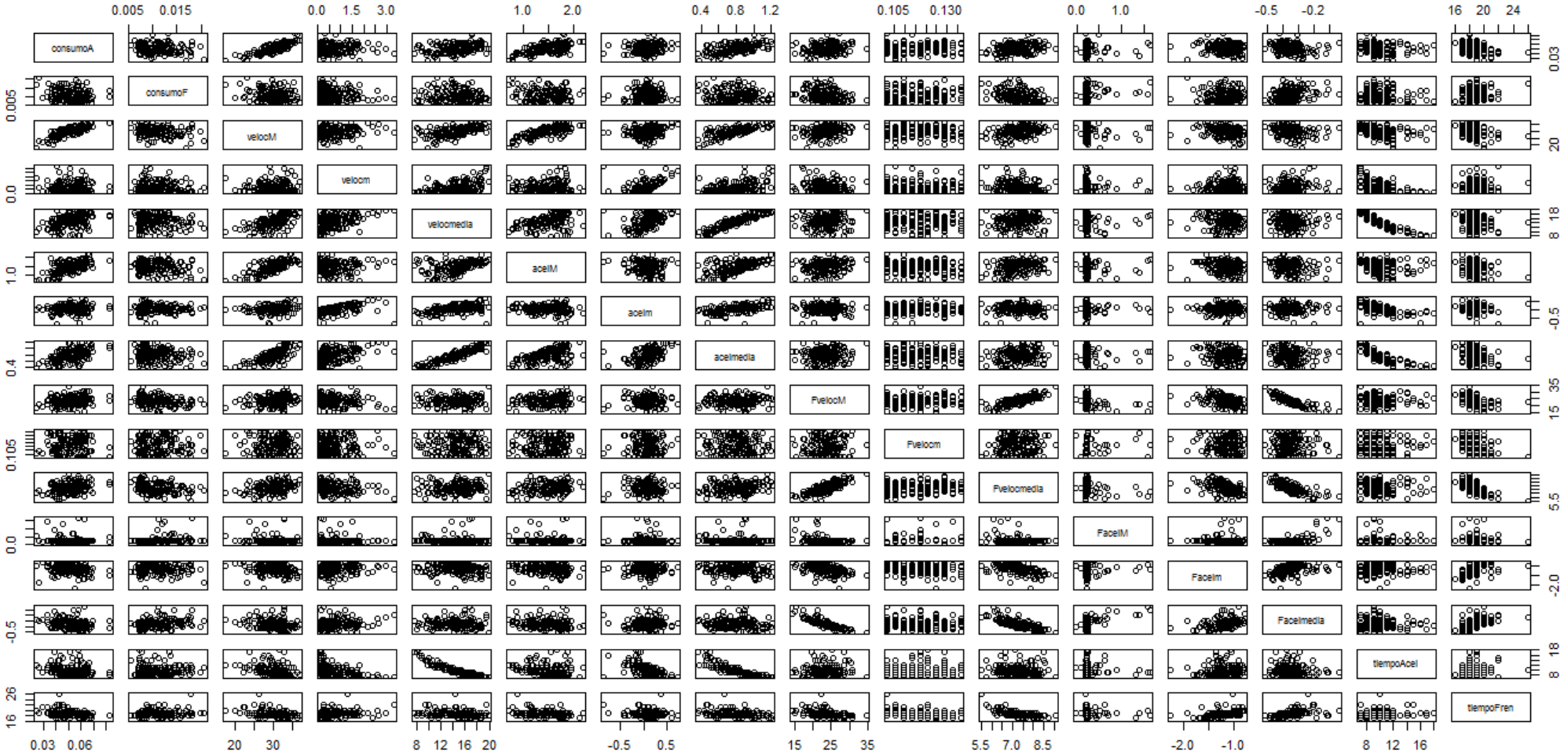

Figura 25: *Diagramas de dispersión de las variables de i31.*

En los diagramas se visualizan variables correlacionadas (tanto positiva como negativamente) con el consumo, y también entre ellas. Para cuantificar estas relaciones, y ver el verdadero peso que tienen en la predicción del consumo, se ha realizado una regresión lineal. Como ya se ha comentado, el modelo inicial contenía todas las variables excepto el *conductor* y *consumoF*, pero tras ir modificándolo para mejorarlo (eliminar variables, análisis de residuos y cerciorarnos de que no hay elementos influyentes, basándonos en los *hat values* y las distancias de Cook), el modelo final es el siguiente (teniendo en cuenta que el modelo se ha creado con las variables normalizadas):

$$\widehat{consumoA} = 0.92 \cdot velocM - 0.16 \cdot velocmedia + 0.157 \cdot FvelocM + 0.14 \cdot Facelmedia$$

Siendo las cuatro variables estadísticamente significativas y con un valor de $R^2 = 0.7111$, el cual mide el porcentaje de variabilidad del consumo explicada por las demás variables, esto es, cómo de bien predecimos el consumo conociendo el valor del resto, podemos decir que es un modelo de predicción bastante bueno: explicamos el 71.11% de la variabilidad. En la Figura 26 se observan los valores predichos por el modelo, frente a los valores reales de la variable *consumoA*. Parece que, exceptuando los dos casos de mayor consumo, que han quedado más lejos, en el resto de casos la predicción se acerca bastante a la realidad.

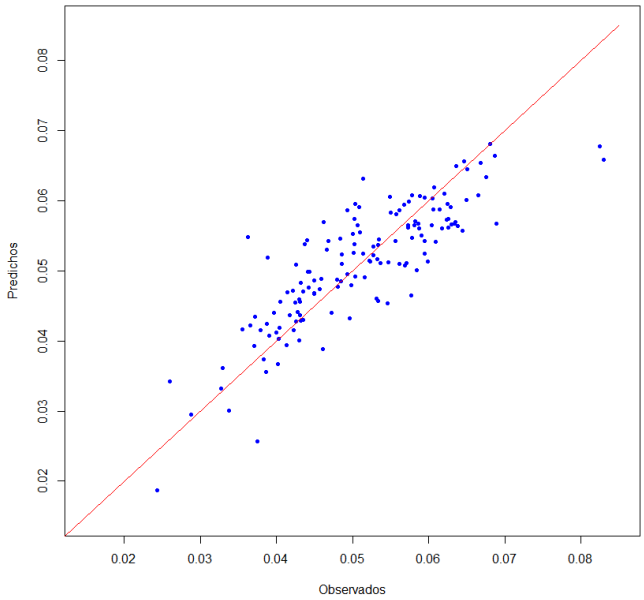

Figura 26: *Valores de consumo predichos por el modelo vs. valores observados.*

De esta forma hemos observado que sí existe una relación directa entre las variables y el consumo, pero nuestro objetivo es clasificar conductores, o al menos, detectar conductas que eleven el consumo. Una forma más clara de visualizar estas relaciones es construir un árbol de regresión.

En este caso también, el árbol genérico para la predicción del consumo en el arranque se ha construido incluyendo todas las variables excepto el conductor y el consumo en la frenada. Tras aplicar el criterio de corte óptimo (ver apartado 6.3.3.2), se ha podado el árbol al valor de $cp = 0.015$. El árbol resultante es el siguiente:

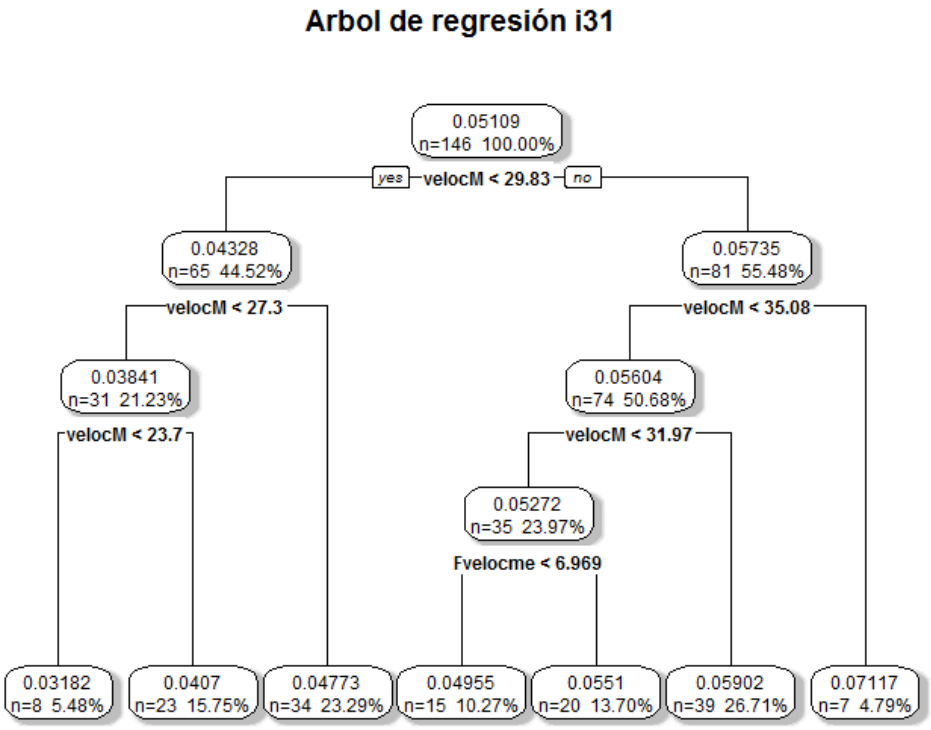

Figura 27: *Árbol de regresión para la parada i31.*

La calidad de los árboles también se puede medir en función del parámetro R-cuadrado. En este caso, hemos obtenido un valor de $R^2 = 0.7277$, muy parecido al del modelo de regresión lineal. Podemos visualizar también los valores predichos por ambos modelos, frente a los valores reales del consumo (Figura 28).

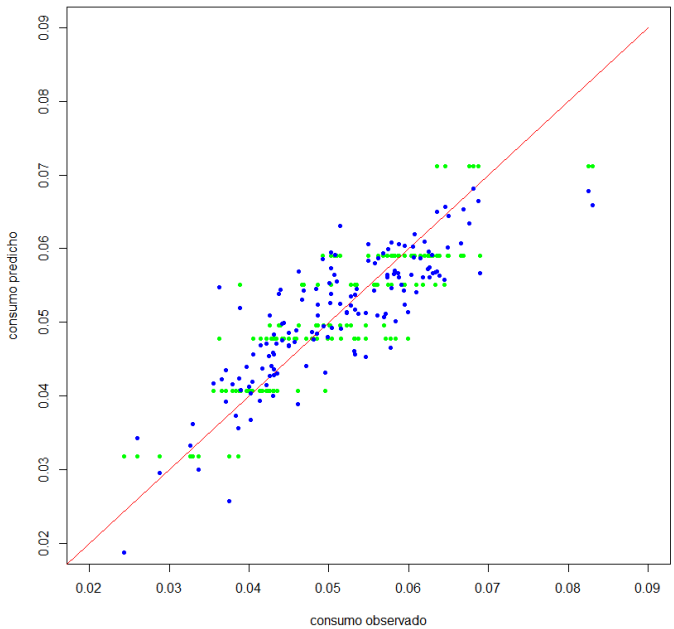

Figura 28: *Valores de consumo predichos por la regresión (azul) y el árbol (verde) vs. valores observados.*

Al crear dos modelos distintos para la estimación del consumo, es de esperar que ambos modelos posean una estructura parecida. En el modelo de regresión, cuatro variables son las responsables de modelar el consumo, mientras que en el árbol solo son dos. Esto tiene una explicación bastante sencilla. Si analizamos el peso de las variables del primer modelo, podemos observar que la velocidad máxima en el arranque aporta casi seis veces más información que cualquiera de las otras tres, es la más influente con mucha diferencia. Y esto se muestra en el árbol, pues al realizar la poda para no sobreajustar el modelo, se ha quedado con las que más le aportan, que son *velocM* y la velocidad media de frenada. Pero el primer árbol generado por R lo completaban las variables *velocM, Fvelocmedia, FacelM, Fvelocm* y *Facelm*. Cuanto menos, curioso, pues cuatro de estas cinco (todas excepto velocM, que ya hemos visto que es la más significativa) ni siquiera aparecen en el modelo de regresión.

En cuanto a la predicción, no se puede obviar la mayor diferencia entre ambos modelos: regresión calcula valores del consumo en la recta real, mientras que el árbol devuelve siempre un valor de entre los 7 de los nodos finales. Aun así, parece que las diferencias de estimación no son tan grandes como cabría esperar.

Para terminar, vamos a analizar si hay diferencias en el consumo entre los conductores. De las 146 detenciones localizadas (recordemos que hemos eliminado dos observaciones al inicio), disponemos de la información del conductor en únicamente 58 casos, y existen 11 conductores distintos:

Conductor	Frecuencia
155	5
207	4
210	13
220	7
237	3
246	1
258	11
278	1
347	9
381	1
711	3
Desconocido	88
TOTAL	**146**

Tabla 10: *Frecuencia con la que aparecen los conductores en i31.*

Empezamos visualizando las variables de i31, pero solo mostramos las observaciones con conductor conocido. Tras la primera prueba y la presencia de demasiados colores, se ha decidido no mostrar tampoco los conductores cuya frecuencia es menor que 5. Los diagramas de dispersión de las variables se muestran en la Figura 29. Aun habiendo recortado la cantidad de individuos, no es posible clasificar o distinguir a los conductores por los valores que toman en las diversas variables.

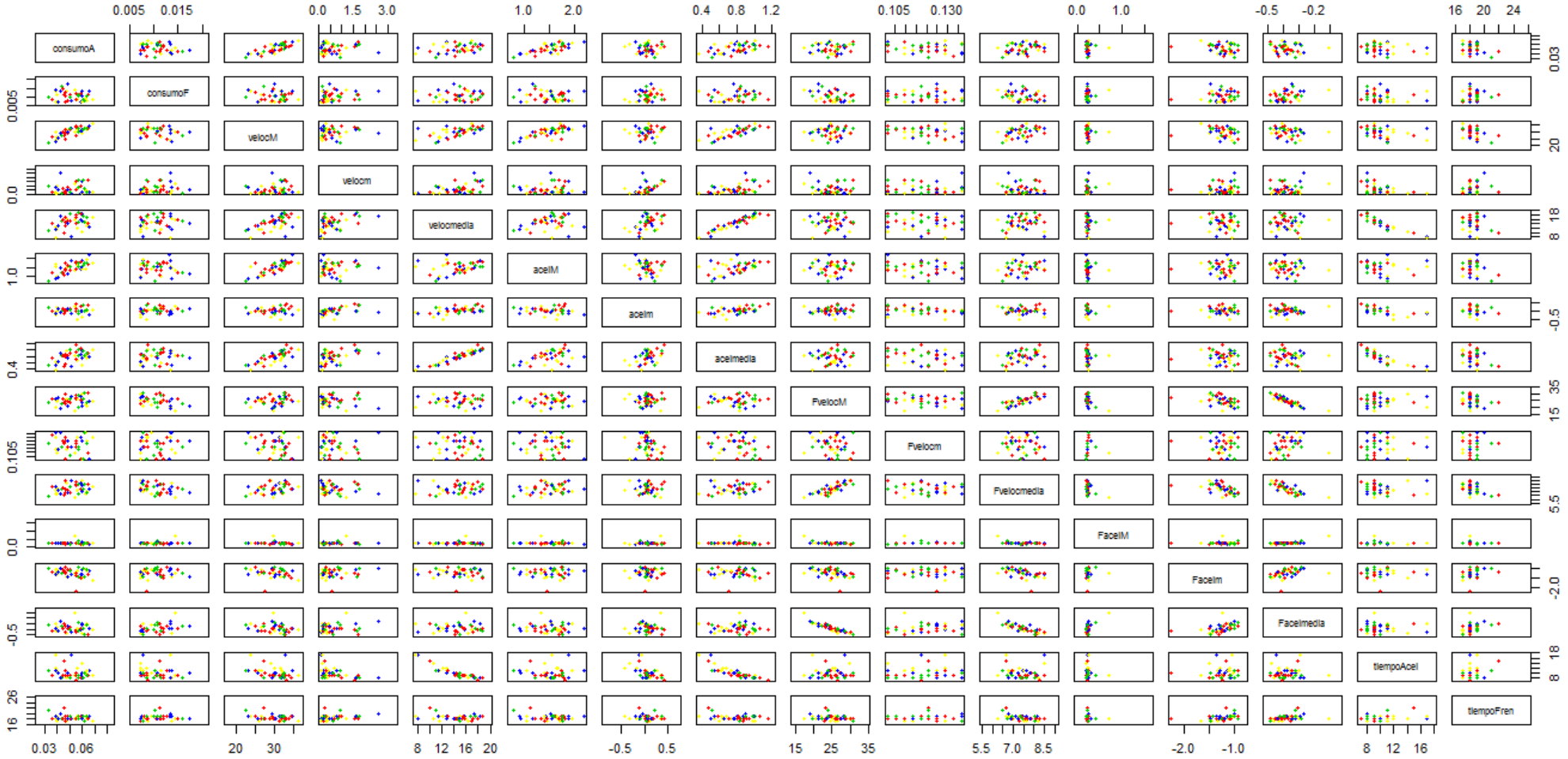

Figura 29: *Diagramas de dispersión de las variables de i31. Conductores por colores.*

Como nuestro interés se centra en el estudio del consumo, veamos qué ocurre si mostramos únicamente el consumo en función del conductor. En este caso se han incluido los 11 conductores, así como los valores de consumo para el resto de los casos, cuando el conductor no es conocido (*conductor=0*).

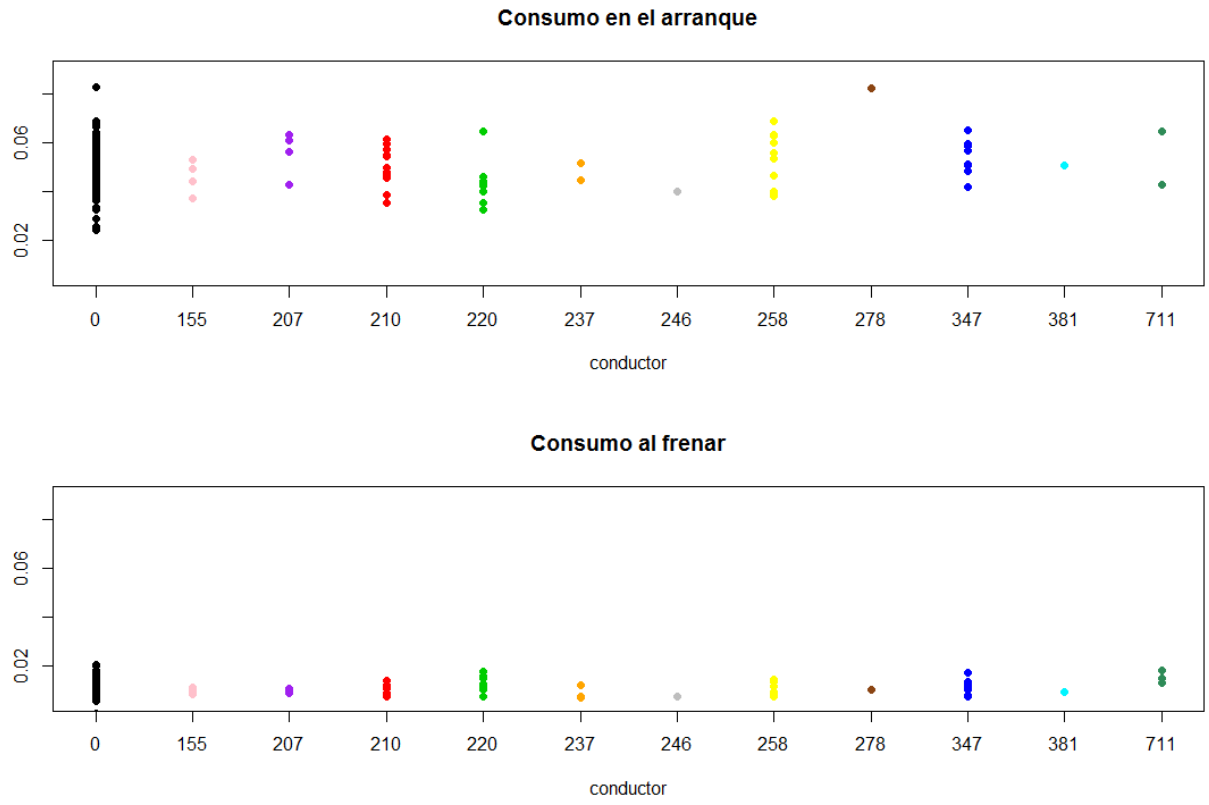

Figura 30: *Consumo de combustible en el arranque y frenada de los conductores en la parada i31.*

Como es de esperar, el consumo es mayor en el arranque que en la frenada, pero no parece haber claras distinciones entre unos conductores y otros. Aunque en algunos casos da la sensación de que unos tienen un rango más amplio de consumos que otros, que parece que hacen un consumo casi fijo, la razón es tan simple como que se disponen más datos de unos que de otros.

A continuación, se ha repetido este mismo proceso con la parada v46. Veamos si los resultados obtenidos se asemejan a estos primeros, o por lo contrario, los contradicen.

Parada Hospital (Galdakao)

Empezamos visualizando las variables, donde diversos valores llaman nuestra atención en los histogramas. En total, se han eliminado 7 observaciones, 2 de ellas por errores en la recolecta de datos (la velocidad tiene un valor constante) y las otras 5 por ser muy distintas del resto (arranques muy muy lentos, muy largo tiempo de frenada, etc.). En la Figura 32 se muestran los diagramas de las variables tras esta primera limpieza.

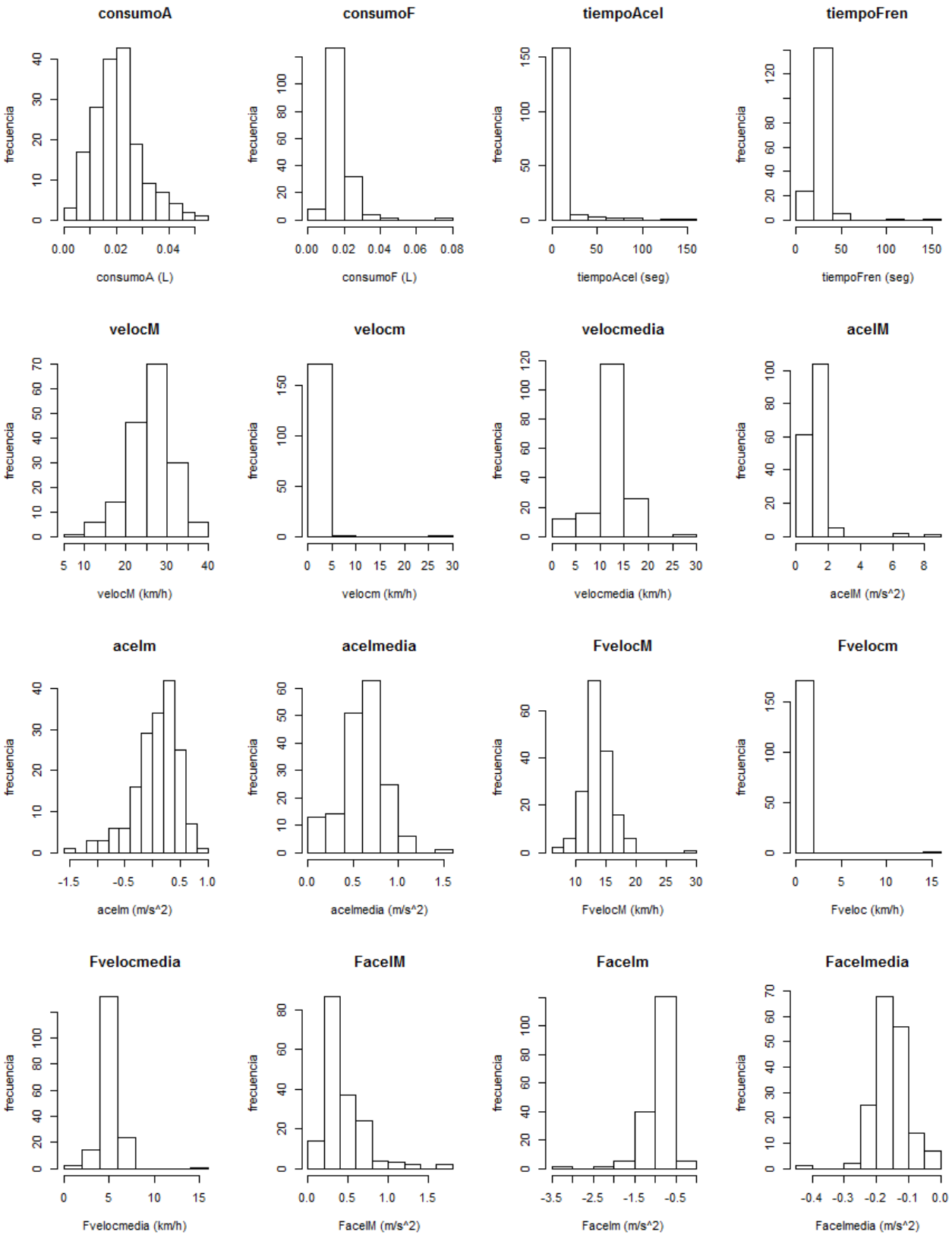

Figura 31: *Histogramas de las variables de v46.*

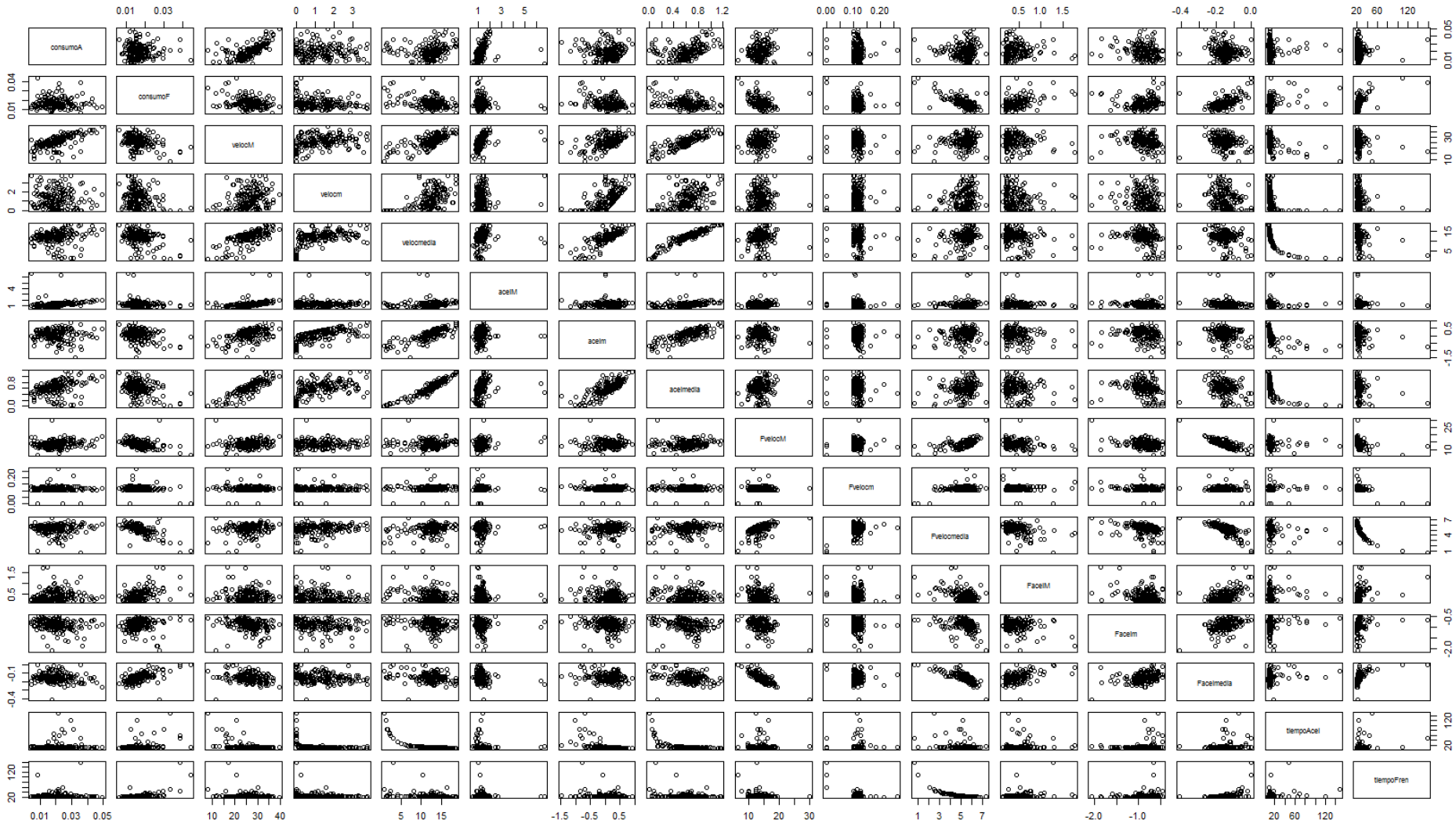

Figura 32: *Diagramas de dispersión de las variables de v46.*

En este caso también se observa la correlación entre diversas variables, lo que nos da un impulso positivo. El mejor modelo de regresión obtenido ha sido el siguiente:

$$\widehat{consumoA} = 1.016 \cdot velocM - 0.3 \cdot acelm + 0.137 \cdot Fvelocm + \\ +0.3 \cdot Fvelocmedia + 0.388 \cdot tiempoAcel + 0.31 \cdot tiempoFren$$

Con todas las variables estadísticamente significativas y con un valor de $R^2 = 0.7446$, podemos afirmar que nos encontramos ante un buen modelo de predicción: explicamos el 74.46% de la variabilidad del consumo. En la Figura 33 se observan los valores predichos por el modelo, frente a los valores reales de la variable *consumoA*.

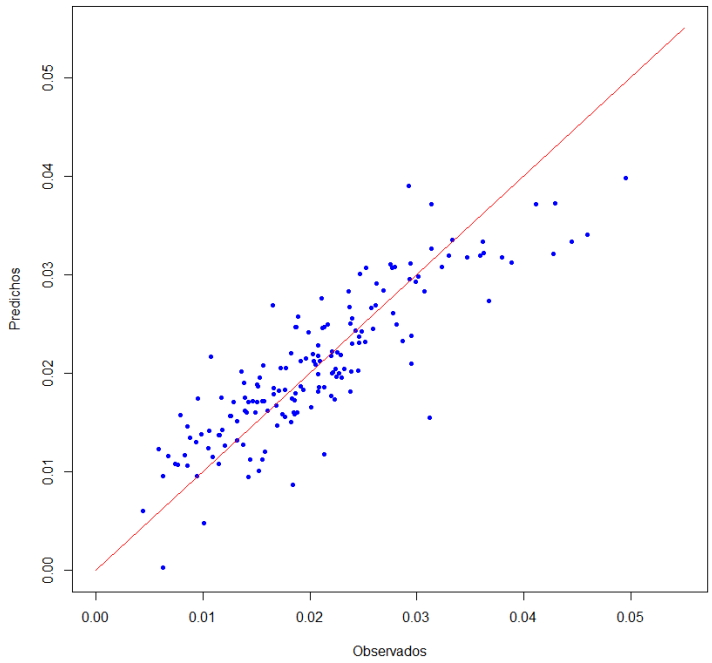

Figura 33: *Valores de consumo predichos por el modelo vs. valores observados.*

Parece que el modelo se ajusta peor a las observaciones cuyo consumo está en los valores máximos; los valores centrales están mejor aproximados que los extremos superiores, en los cuales se obtiene una aproximación a la baja.

Al igual que para i31, se ha construido un árbol de regresión para esta parada, que se ha podado en el valor de $cp = 0.015$, obteniendo justo el mismo número de nodos que para i31. En este caso, el valor de R^2 es 0.7471, también muy parecido al obtenido con el modelo lineal. El árbol resultante es el que se muestra en la Figura 34.

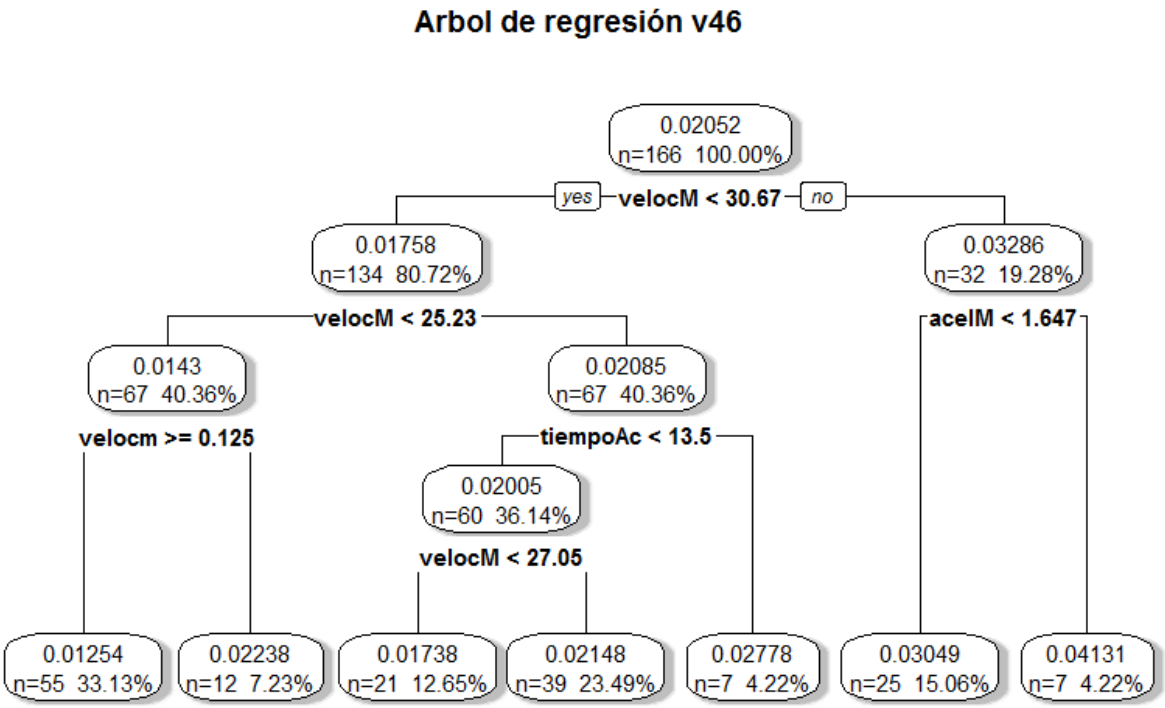

Figura 34: *Árbol de regresión de la parada i31.*

Tenemos aqui la comparación estre ambos modelos de prediccion:

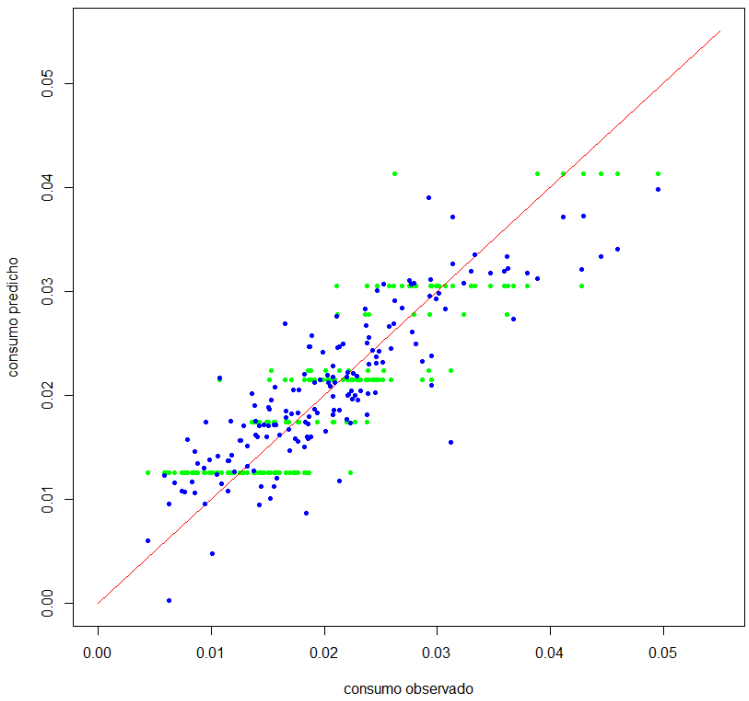

Figura 35: *Valores de consumo predichos por ambos modelos vs. valores observados.*

Esta vez los resultados tampoco nos dejan nada claro. El árbol contempla la velocidad y la aceleración máximas, el tiempo y la velocidad minima en el arranque como predictoras del consumo, mientras que en el modelo de regresión, con un total de 6 variables incluidas, solo se repiten la velocidad máxima y el tiempo de arranque.

De todas formas, queda claro en ambos casos que la información de la velocidad máxima en el arranque es crucial para la estimación del consumo.

Veamos por último qué sucede en esta parada con el consumo de aceleracióny frenada, si lo analizamos por conductores. De las 166 detenciones localizadas desconocemos el conductor en 103 de ellas. El resto está compuesto por 10 de los 11 conductores identificados en la parada i31.

Conductor	Frecuencia
155	8
207	6
210	14
220	10
237	2
258	10
278	1
347	9
381	1
711	2
Desconocido	103
TOTAL	**166**

Tabla 11: *Frecuencia con la que aparecen los conductores en v46.*

Para una primera visión de los datos, se han dibujado los diagramas de dispersion de las variables medidas, mostrando solo los datos de las observaciones cuyo conductor conocemos. Como en el análisis anterior, para evitar la abundancia de colores, se muestran solo los casos de los conductores que aparecen con una frecuencia mayor que 5 (Figura 36).

Sin mayor éxito, procedemos a la visualización únicamente del consumo, por condutores (Figura 37). En este caso sí, se han incluido los 10 conductores identificados, ya que aportan información general del consumo, aunque suponemos que nada específico podremos deducir sobre ellos y sus conductas.

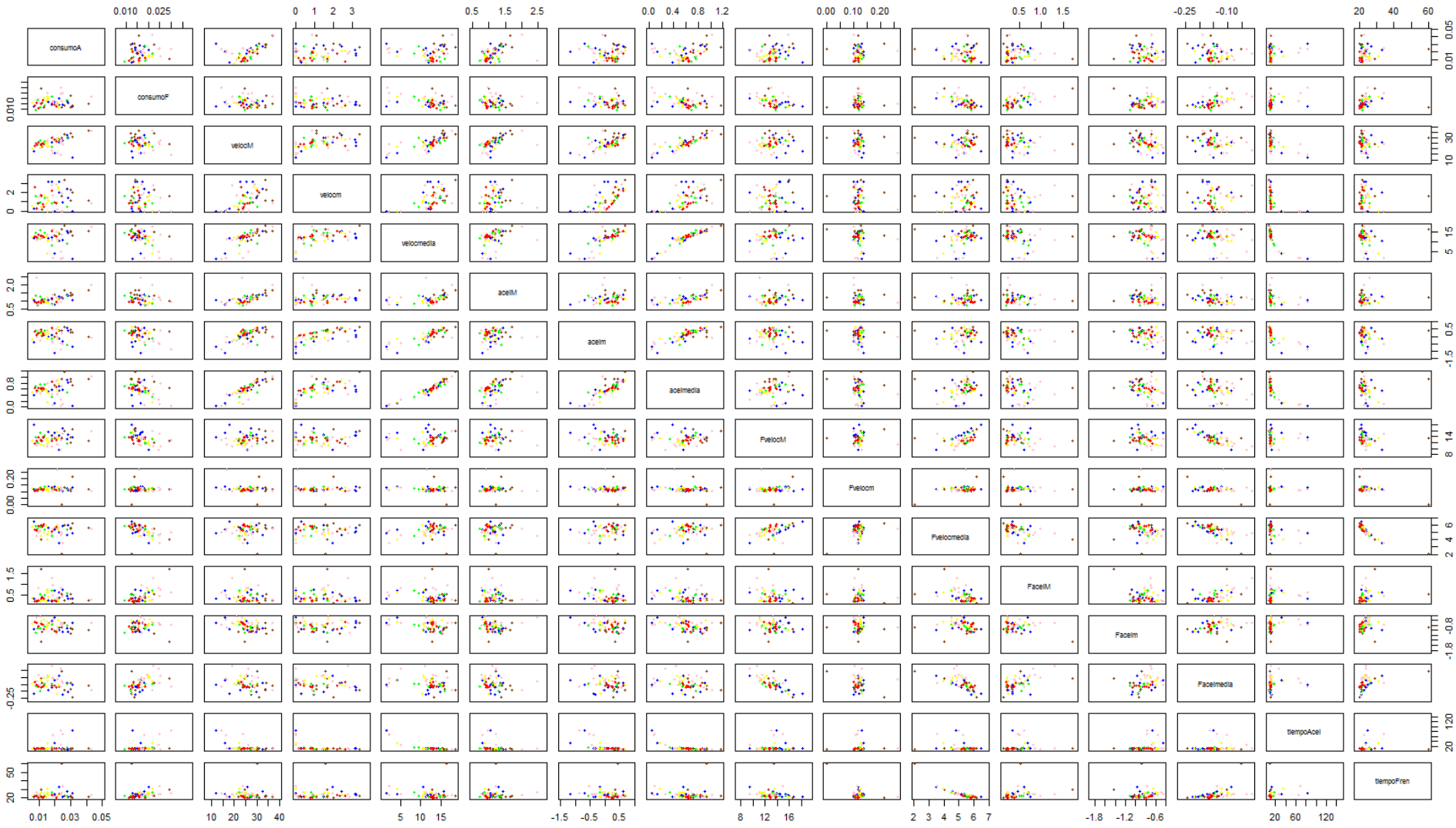

Figura 36: *Diagramas de dispersión de las variables de v46. Conductores por colores.*

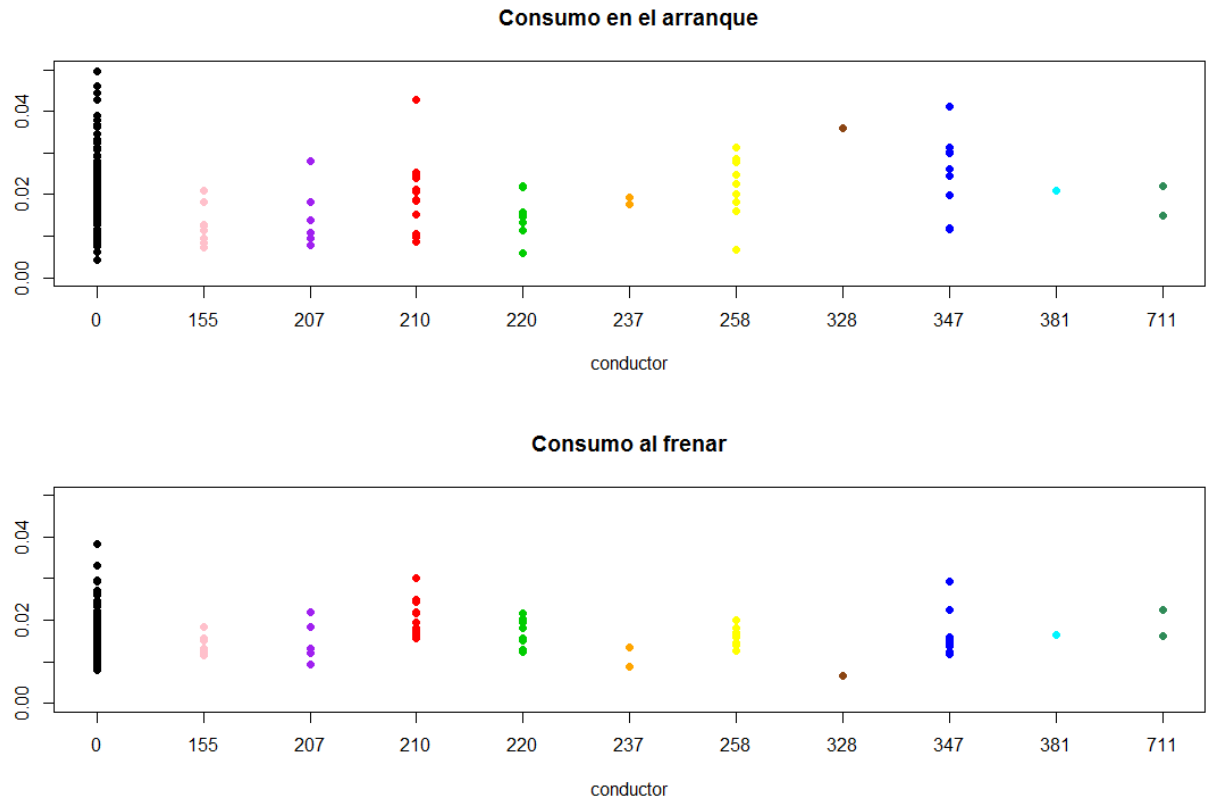

Figura 37: *Consumo en el arranque y frenada de los conductores en la parada v46.*

Aquí parece que se ve algo. A diferencia del caso anterior, esta vez sí que se observan diferencias entre los conductores, sobre todo si comparamos los datos de frenado y aceleración. Los conductores 155 y 220 obtienen valores parecidos en ambos casos. El 207 está algo por encima y el 210 un poco más. El 258 parece situarse entre medias de éstos dos últimos, y el 347, a pesar de aparecer solo en 9 ocasiones, parece que toma valores en un amplio rango de valores, por lo que no podemos decir nada sobre él.

7. Experimentos y resultados

La construcción de los cuatro primeros modelos de estimación de consumo (dos para una parada del servicio de ida, y otros dos para una de vuelta) no ha despejado demasiadas dudas. Lo único que hemos podido sacar en claro ha sido la importancia de la velocidad máxima obtenida en el arranque, puesto que los cuatro modelos difieren en el resto de variables a considerar.

A pesar de todo, parece que los modelos funcionan correctamente en la estimación del consumo de sus paradas correspondientes, por lo que nos preguntamos si también lo harán en las demás. Para testearlo, se han seleccionado entre las paradas con más detenciones 5 de ida y 5 de vuelta, y se ha tratado de predecir el consumo mediante los modelos de

regresión lineal de i31 (Figura 38) y v46 (Figura 39). Las paradas seleccionadas han sido la 19, 26, 30,44 y 49 de ida y la 24, 35, 52, 61 y 65 de vuelta (i19, i26, i30, i44, i49 y v24, v35, v52, v61 y v66 respectivamente). A continuación, mostramos los valores reales frente a los estimados por ambos modelos:

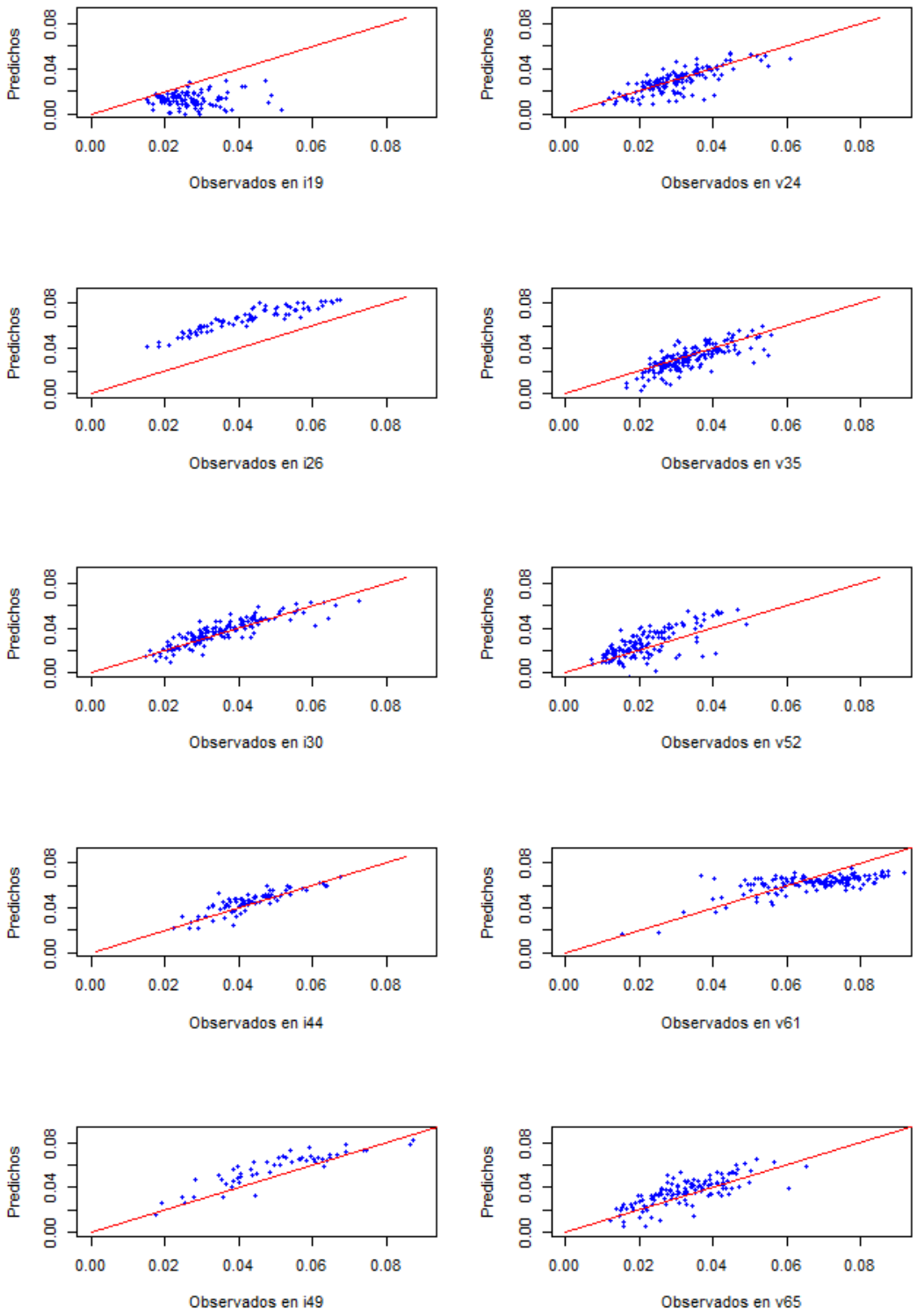

Figura 38: *Consumos predichos por el modelo de regresión lineal de i31.*

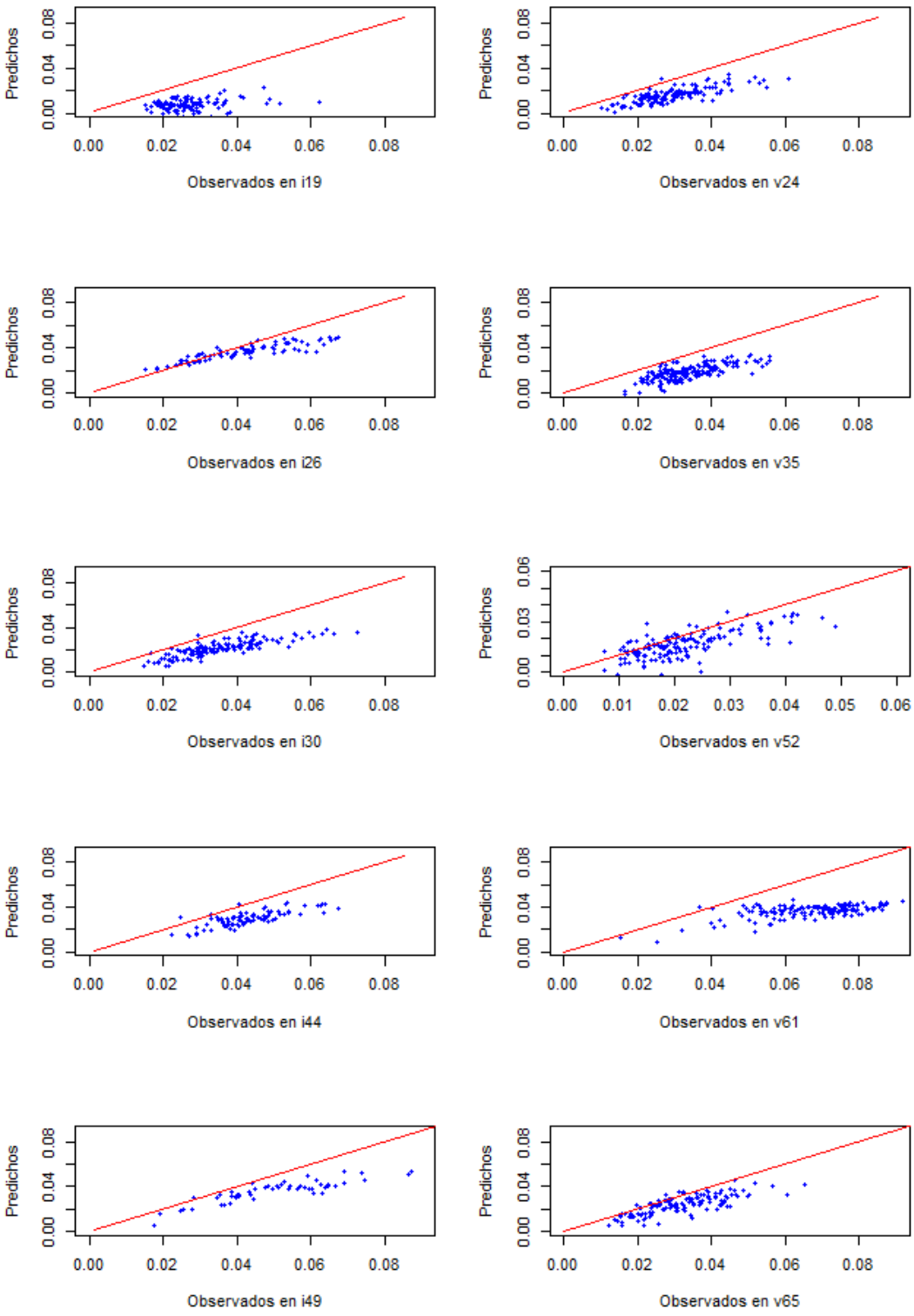

Figura 39: *Consumos predichos por el modelo de regresión lineal de v46.*

Aunque en un principio el modelo de v46 parecía mejor (tomaba en cuenta más variables y el valor de R^2 era algo mayor), queda claro que no sirve para modelar el resto de paradas. Tampoco tenemos noticias demasiado buenas sobre el i31. Consigue aproximaciones

bastante buenas en algunos casos, pero dista mucho de ser el modelo general (si es que existe) que querríamos encontrar.

Aun así, de este pequeño análisis podemos sacar un dato importante a efectos de futuras investigaciones. En los gráficos se observan grandes distinciones en los valores del consumo dependiendo de la parada, teniendo en cuenta que en todas ellas el consumo está medido en los 30 primeros metros de recorrido tras el arranque. Así, en la parada v65, los consumos observados son, en general, menores que los de la parada v61, lo que nos puede hacer pensar que aparte de las variables que se puedan medir en el vehículo, el consumo también puede verse afectado por las características de la carretera (en este caso, del lugar de parada).

Para terminar, la idea era verificar si de verdad lo ocurrido en las paradas es significativo del consumo realizado en todo un servicio, o es necesario tomar otros puntos de referencia o incluso cambiar de perspectiva. Pero la falta de datos ha imposibilitado este último análisis, puesto que no tenemos casi ningún servicio del que dispongamos de la información de todas las paradas; es más, podemos afirmar que en la mayoría de los servicios identificados falta al menos un tercio de los datos.

8. Conclusiones y trabajo futuro

Diversos dispositivos y sistemas han sido creados y testeados en la misma línea de trabajo de esta propuesta, y en muchos de ellos se han obtenido ahorros considerables en el consumo de combustible.

Ante la magnitud de este problema, hemos decidido entrar en la raíz y hacernos con lo investigado hasta ahora, para así poder realizar un estudio desde lo más profundo de los datos. No ha sido tarea fácil recopilar y aunar la información de los cientos de estudios llevados a cabo, sobre todo porque cada día aparecen nuevos artículos, y cada vez más; es un ámbito que se encuentra en constante cambio, no es posible estar completamente al día de todos los resultados.

A pesar de ello, con todo lo recopilado y estudiado, hemos tomado la decisión de analizar dos de los aspectos que están tomando cada vez más fuerza en el mundo de la automoción. Por un lado, la influencia del conductor en el consumo y por otro, el análisis del consumo en las frenadas y arranques.

Los resultados no han sido tan exitosos como quisiéramos esperar, pero podemos afirmar que hemos sacado algunas ideas a la luz. Por un lado, ha resaltado el peso de la velocidad en el consumo, pero no ha quedado claro el papel del resto de las variables observadas. Los modelos construidos funcionan correctamente para los datos para los que han sido creados, pero no tanto para el resto de situaciones, lo que nos hace plantearnos la inexistencia de un buen modelo de predicción global, válido para toda clase de situaciones y capaz de reconocer conductas generales que impulsen consumos elevados. Eso sí, queda por verificar si el punto de vista desde el que se ha trabajado es bueno realmente.

Por otro lado, parece que el conductor, o más bien, sus conductas, sí influyen en el combustible consumido. Para poder reafirmar esto, es necesario también realizar más estudios con distintos conductores y en distintas situaciones, pero da la impresión de que esta investigación va por buen camino; las técnicas de eco-driving, presentados al comienzo de este trabajo, podrían realmente tener la capacidad de mejorar los hábitos de los conductores, logrando así la reducción deseada.

Por último, ha llamado nuestra atención las diferencias obtenidas en las distintas paradas, en lo referente al consumo, que como hemos comentado, abren las puertas a nuevos aspectos que analizar.

Han sido muchos los motivos por los que los resultados obtenidos no han sido tan satisfactorios como hubiéramos querido. A pesar de disponer, en un principio, de una gran cantidad de datos (más de 5 millones de observaciones), la falta de información de posicionamiento del vehículo (indispensable para nuestro estudio) y la errónea información de las rpm han hecho mella en los resultados. Las fechas repetidas (algunos segundos, en ocasiones minutos enteros), y pequeños, pero frecuentes, saltos hacia adelante y hacia atrás en el tiempo también han perjudicado a variables creadas *a posteriori* como la aceleración, o las matrices de datos con las que se han creado los modelos de las paradas. Hemos visto, por ejemplo, que en el modelo de i31 un par de observaciones han tenido que ser eliminadas por errores en los datos originales. Y por último, la medición del propio consumo, que no acaba de quedar claro. No hace falta decir que es indispensable que los valores que estamos analizando sean lo más exactos y certeros posibles.

Con todo ello, podemos decir que la investigación debe seguir adelante, tanteando nuevas y diversas perspectivas desde las cuales abordar el tema del análisis del consumo, con más tiempo y más y mejores datos.

8.1 Posibles futuras líneas de trabajo

Queda mucho trabajo por realizar, pero más que el puro análisis de datos, toca impulsar el trabajo de campo, aportar nuevas variables a medir, encontrar factores (no tan obvios como los hechos que ocurren en el propio vehículo) que puedan afectar el consumo. Hablamos de las características de la carretera, el peso que lleva el vehículo o las condiciones meteorológicas. También podríamos considerar el tiempo que lleva conduciendo el chófer o medir de alguna manera su nivel de fatiga o cansancio.

Pero la realidad es que medir y cuantificar estos factores no es tan sencillo como instalar un nuevo dispositivo y guardar un valor cada segundo. Por ello, queda un largo pero, en mi opinión, esperanzador camino por recorrer.

9. Referencias

[1] José de Almeida, João C. Ferreira: BUS Public Transportation System Fuel Efficiency Patterns. 2nd International Conference on Machine Learning and Computer Science (IMLCS'2013) August 25-26, 2013 Kuala Lumpur (Malaysia).

[2] Stefano Carrese, Andrea Gemma, Simone La Spada: Impacts of driving behaviors, slope and vehicle load factor on bus fuel consumption and emissions: a real case study in the city of Rome. Procedia - Social and Behavioral Sciences 87 (2013) 211 – 221.

[3] Catarina Rolim, Patrícia Baptista, Gonçalo Duarte, Tiago Farias, Yoram Shiftan: Quantification of the impacts of eco-driving training and real-time feedback on urban buses driver's behaviour. 17th Meeting of the EURO Working Group on Transportation, EWGT2014, 2-4 July 2014, Sevilla, Spain.

[3.1] WSP (2012). Energy Efficiency Measures for Buses and Bus Transport - Possibilities and Experiences from Other Actors.

[4] Deepak Hari, Christian J. Brace, Christopher Vagg, John Poxon, Lloyd Ash: Analysis of a driver behaviour improvement tool to reduce fuel consumption. 2012 International Conference on Connected Vehicles and Expo.

[4.1] J. Alson, B. Ellies, D. Ganss: Interim Report: New Powertrain Technologies and Their Projected Costs. U.S. Environmental Protection Agency EPA420-R-05-012, 2005.

[5] Jonathan Seth Stichter: Investigation of vehicle and driver aggressivity and relation to fuel economy testing. University of Iowa, 2012.

[6] Thomas J. Daun, Daniel G. Braun, Christopher Frank, Stephan Haug, Markus Lienkamp: Evaluation of driving behavior and the efficacy of a predictive eco-driving assistance system for heavy commercial vehicles in a driving simulator experiment. Proceedings of the 16th International IEEE Annual Conference on Intelligent Transportation Systems (ITSC 2013), The Hague, The Netherlands, October 6-9, 2013.

[6.1] A. E. Wahlberg, J. Gothe: Fuel Wasting Behaviors of Truck Drivers. In Industrial Psychology Research Trends, I. M. Pearl, Ed., 2007, ch. 4, pp. 73-87.

[6.2] M. A. Symmons, G. Rose, G. H. Y. Doom: The effectiveness of an ecodrive course for heavy vehicle drivers. In 2008 Australasian Road Safety Research Policing and Education Conference, no. November, Adelaide, Australia, 2008, pp. 187-194.

[6.3] M. Zarkadoula, G. Zoidis, E. Tritopoulou: Training urban bus drivers to promote smart driving: A note on a Greek eco-driving pilot program. Transportation Research Part D: Transport and Environment, vol. 12, no. 6, pp. 449-451, Aug. 2007.

[6.4] A. E. Wahlberg: Long-term effects of training in economical driving: Fuel consumption, accidents, driver acceleration behavior and technical feedback. International Journal of Industrial Ergonomics, vol. 37, no. 4, pp. 333-343, Apr. 2007.

[6.5] B. Beusen, S. Broekx, T. Denys, C. Beckx, B. Degraeuwe, M. Gijsbers, K Scheepers, L. Govaerts, R. Torfs, L. I. Panis: Using on-board logging devices to study the longer-term impact of an eco-driving course. Transportation Research Part D: Transport and Environment, vol. 14, no. 7, pp. 514-520, Oct. 2009.

[7] Belhassen Akrout, Walid Mahdi: Spatio-temporal features for the automatic control of driver drowsiness state and lack of concentration. Machine Vision and Applications (2015) 26:1–13.

[8] Shouyi Wang, Yiqi Zhang, Changxu Wu, Felix Darvas, Wanpracha Art Chaovalitwongse: Online Prediction of Driver Distraction Based on Brain Activity Patterns. IEEE Transactions On Intelligent Transportation Systems, Vol. 16, No. 1, February 2015.

[9] Klaus Bengler, Klaus Dietmayer, Berthold Färber, Markus Maurer, Christoph Stiller, Hermann Winner: Three Decades of Driver Assistance Systems - Review and Future Perspectives. IEEE Intelligent transportation systems magazine. Winter 2014.

[9.1] M. Lu: Modelling the effects of road traffic safety measures. Accid. Anal. Prev., vol. 38, no. 3, pp. 507–517, May 2006.

[9.2] M. Aga, A. Ogada: Analysis of vehicle stability control effectiveness from accident data. In Proc. 18th Int. Enhanced SafetyVehicles-Conf., Nagoya, AI, 2003.

[9.3] R. Sferco, Y. Page, J. Y. LeCoz, P. Fay: Potential effectiveness of the electronic stability programs–What European field studies tell us. In Proc. 17th Int. Enhanced Safety Vechicles Conf., Amsterdam, The Netherlands, 2001.

[9.4] M. Akamatsu, P. Green, K. Bengler: Automotive technology and human factors research: Past, present, and future. Int. J. Veh. Technol, 2003

[9.5] T. Bär, R. Kohlhaas, J. M. Zöllner, K. Scholl: Anticipatory driving assistance for energy efficient driving. In Proc. IEEE Forum Integrated Sustainable Transportation System, 2011, pp. 1–6.

[9.6] K. Boriboonsomsin, M. J. Barth, W. Zhu, A. Vu: Eco-routing navigation system based on multisource historical and real-time traffic information. IEEE Trans. Intell. Transp. Syst., vol. 13, no. 4, pp. 1694–1704, 2012.

[9.7] B. Dornieden, L. Junge, P. Pascheka: Anticipatory energy-efficient longitudinal vehicle control. In ATZ worldwide. 2012, pp. 24-29.

[9.8] D. Popiv, M. Rakic, F. Laquai, M. Duschl, K. Bengler: Reduction of fuel consumption by early anticipation and assistance of deceleration phases. In Proc. World Automotive Congr. Int. Federation Automotive Engineering Societies, Budapest, Hungary, June 30, 2010.

[9.9] W. D. Jones: Keeping cars from crashing. IEEE Spectrum, vol. 38, no. 9, pp. 40–45, Sept. 2001.

[9.10] Intelligent Transport Systems - Full Speed Range Adaptive Cruise Control Systems - Performance Requirements and Test Procedures. International Standard Organization 22179 TC204/WG14, 2009.

[9.11] K. Kodaka, M. Otabe, Y. Urai, H. Koike: Rear-end collision velocity reduction system. In Proc. SAE WorldCongr., Detroit, MI, 2003.

[9.12] M. Maurer: Forward collision warning and avoidance. In Handbook of Intelligent Vehicles, A. Eskandarian, Ed., London, U.K.: Springer- Verlag, 2012.

[9.13] Volvo Trucks European Accident Research and Safety Report, Volvo, Gothenburg, Sweden, 2013.

[10] Magnus Helmbrecht, Cristina Olaverri Monreal, Klaus Bengler, Roman Vilimek, Andreas Keinath: How Electric Vehicles Affect Driving Behavioral Patterns. Institute of Ergonomics, Technische Universität München, Germany, 2014.

[11] Alberto Díaz Álvarez, Francisco Serradilla García, José Eugenio Naranjo, José Javier Anaya, Felipe Jiménez: Modeling the Driving Behavior of Electric Vehicles Using Smartphones and Neural Networks. University Institute for Automobile Research (INSIA), Madrid, Spain, 2014.

[12] Sabreena Anowar, Amir Zahabi: Evaluation of energy reduction in transportation, impact of eco-driving and hybrid-electric vehicle technologies in urban Québec. 2013.

[13] Alberto Broggi, LucaMazzei, Pier Paolo Porta: Car-driver cooperation in future vehicles. In Procs. Intl. Conf. On Models and Technologies for Intelligent Transportation Systems, Rome, Italy, June 2009.

[14] Imed Ben Dhaou: Fuel Estimation Model for ECO-Driving and ECO-Routing. College of Engineering, Al Jouf University, Kingdom of Saudi Arabia. 2011.

[15] Kristin Lovejoy, Susan Handy, Marlon G. Boarnet: Impacts of Eco-driving on Passenger Vehicle Use and Greenhouse Gas Emissions. University of California and University of Southern California. 2013.

[16] Olivier Orfila, Guillaume Saint Pierre, Cindie Andrieu: Gear Shifting Behavior Model for Ecodriving Simulations Based on experimental data. EWGT 2012.

[17] Modeling the Relation Between Driving Behavior and Fuel Consumption. CGI GROUP.

[18] Tianyi Guan, Christian W. Frey: Model adaptive driver assistance system to increase fuel savings. 2012 IEEE International Conference on Vehicular Electronics and Safety. July 24-27, 2012. Istanbul, Turkey.

[19] William Frith, Peter Cenek: Standar Metrics for Transport and Driver Safety and Fuel Economy. Opus International Consultants Central Laboratories. November 2012.

[20] Munzilah Md Rohani: Bus Driving Behaviour and Fuel Consumption. FACULTY OF ENGINEERING, SCIENCE & MATHEMATICS, UNIVERSITY OF SOUTHAMPTON. November 2012.

[21] Wei Lun Ng, Chee Kyun Ng, Borhanuddin Mohd. Ali, Nor Kamariah Noordin, Fakhrul Zaman Rokhani: Review Of Researches In Controller Area Network. Department of Computer and Communication Systems, Faculty of Engineering, Universiti Putra Malaysia, Malasya.

[22] Stewart A. Birrell, Mark Fowkes, Paul A. Jennings: Effect of Using an In-Vehicle Smart Driving Aid on Real-World Driver Performance. IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 15, NO. 4, AUGUST 2014.

[23] Xiaohua Zhou, Jian Huang, Weifeng Lv, Dapeng Li: Fuel Consumption Estimates Based on Driving Pattern Recognition. 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing.

[24] German Castignani, Thierry Derrmann, Raphaël Frank, Thomas Engel: Driver Behavior Profiling Using Smartphones: A Low-Cost Platform for Driver Monitoring. Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg.

[25] V. Corcoba Magaña, M. Muñoz-Organero: GAFU: Using a Gamification Tool to Save Fuel. Departament of Telematic Engineering, University Carlos III, Leganés, Spain.

[26] Songwon Seo: A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets. BS, Kyunghee University, 2002.

[27] Therneau, T.M., E.J., Atkinson: An Introduction to Recursive Partitioning Using the Rpart Routine. En: Technical Report 61, Mayo Clinic, Section of Statistics, 1997.