CIT2016 – XII Congreso de Ingeniería del Transporte
València, Universitat Politècnica de València, 2016.
DOI: http://dx.doi.org/10.4995/CIT2016.2016.1990

METODOLOGÍA PARA LA ESTIMACIÓN DE LA MOVILIDAD DE VEHÍCULOS DEL PARQUE ESPAÑOL. ESTUDIO PILOTO: AUTOBUSES ARTICULADOS

José M. Mira Mc Williams
Profesor Titular de Universidad. ETSI Industriales. Investigador INSIA. UPM.
josemanuel.mira@upm.es

Blanca Arenas Ramírez
Profesor Contratado Doctor. ETSI Industriales, Directora de la Unidad de Estudios de Transporte e Impacto Medioambiental del Automóvil de INSIA – UPM. España,
barenas@etsii.upm.es

Camino González Fernández
Profesor Titular de Universidad. ETSI Industriales, UPM. España, camino@etsii.upm.es

Arturo Furones Crespo
Investigador INSIA - UPM. España, arturo.furones@ upm.es

Javier Páez Ayuso
Profesor Titular de Universidad. ETSI Industriales, Sub-director de INSIA – UPM.
España, franciscojavier.paez@upm.es

Francisco Aparicio Izquierdo
Profesor Emérito de la UPM. Presidente de INSIA - UPM, España.
francisco.aparicio@upm.es

RESUMEN

España ha intensificado la aplicación de políticas encaminadas a la reducción del número de accidentes y de víctimas, durante la última década, y ha alcanzado un elevado nivel de éxito en esta tarea, situando a nuestro país en un puesto muy destacado entre los países de Europa y del mundo en relación a la seguridad vial. En el año 2013, España ocupó el quinto puesto en el grupo de países de la UE-28 en el indicador muertos por millón de habitante. Un indicador análogo con datos de exposición no existe a nivel europeo, en tanto el denominador (la exposición) es una magnitud de difícil obtención. El logro de nuevas metas exige medidas específicas dirigidas a ámbitos y colectivos con características diferentes, y para ello es necesario un mejor conocimiento de los verdaderos niveles de riesgo de determinados grupos de usuarios, definidos por criterios de género, o edad, o de tipos de vehículos y características constructivas, prestaciones o eficacia de sistemas de seguridad, entre otros factores. En el caso de los vehículos, existen datos de distancias recorridas en diferentes periodos, recogidos en las fichas de inspección técnica, e incorporados, desde hace unos años, en bases de datos gestionadas por las comunidades autónomas y por la DGT, pero estos datos están siendo escasamente explotados.

DOI: http://dx.doi.org/10.4995/CIT2016.2016.1990

En el presente trabajo se presenta una metodología de análisis de los datos obtenidos en las inspecciones técnicas de los vehículos, y de estimación de la exposición (veh-km anual) de diferentes grupos de vehículos (según tipo, edad y potencia, entre otras variables determinantes) que pueden presentar diferencias significativas en sus niveles de exposición. La metodología, que incluye la creación de bases limpias y la formación de conglomerados de vehículos con elevada homogeneidad en el valor de su movilidad, se ha aplicado a una muestra de autobuses articulados a modo de estudio piloto. Con este trabajo se pretende contribuir a un mejor conocimiento de los verdaderos niveles de riesgo de diferentes grupos de vehículos del parque español.

1. Introducción

Durante la última década, España ha alcanzado un notable éxito en la reducción de la accidentalidad en carretera, posicionándose, de acuerdo con el indicador de número de fallecidos al año por habitante, entre los 5 países mejores de Europa.

Sin embargo, con vistas a continuar en este camino de éxito, es necesario profundizar en el conocimiento de los índices de accidentalidad por grupos de vehículos y usuarios. Esto requiere una estimación con precisión razonable de los denominadores de la tasas de accidentalidad, es decir, de la exposición, medida en km por año recorridos.

El problema de la estimación precisa de la exposición es de gran actualidad y objeto de continua investigación (Caroll, P. S. [4], Kweon y Kockelman [5], Keall y Newstead [13], Harrison y Christie [14], Stamatiadis y Deacon [15], pero uno de los problemas que afronta es la escasez de datos abundantes y de calidad.

Con este fin, se ha desarrollado un proyecto de investigación que aprovecha una nueva fuente de datos para medición de la exposición, concretamente los recogidos en las estaciones ITV de España en los últimos años. Esta idea es pionera en Europa.

Se ha aplicado metodología estadística sofisticada, basada en la minería de datos, a una base de datos depurada mediante un procedimiento elaborado, de la movilidad de autobuses articulados, tal y como viene recogida en las estaciones ITV de toda España.

Las técnicas de la minería de datos aplicadas han sido los árboles de regresión de y sus extensiones de randomForests y dynaTree, esta última dentro del enfoque bayesiano. La aplicación es un estudio piloto para la muestra concreta de autobuses articulados. Se ha estimado cómo la exposición (movilidad) de los vehículos depende de una serie de características de los mismos: antigüedad, potencia, peso, cilindrada, número de plazas, entre otras.

Los resultados del análisis estadístico tienen cuatro vertientes: estadística descriptiva univariante y multivariante, estudio de importancia de las variables de entrada sobre la exposición, análisis de bondad de ajuste e identificación de patrones de las variables de

DOI: http://dx.doi.org/10.4995/CIT2016.2016.1990

entrada para los errores más elevados y finalmente predicción puntual y bandas de incertidumbre para algunos ejemplos concretos.

Este trabajo está estructurado como sigue: en el epígrafe 2 se describe la base de datos utilizada para la investigación y el procedimiento de filtrado de datos erróneos y limpieza de la misma. En el epígrafe 3 se describen las técnicas estadísticas de minería de datos que se aplican y en el 4 los resultados de la aplicación a los autobuses articulados.

2. Los datos

2.1. Base de datos

La comunicación a la DGT de los datos del kilometraje de los vehículos, leídos en las ITV's, se realiza de forma obligatoria desde 2011, por lo que en el almacén de datos respectivo existen al menos 4 años con esa información y los vehículos del parque español tienen al menos un registro de la ITV visitada. Para el estudio se han solicitado datos relacionados como la fecha y clase de Matriculación, datos técnicos del vehículo como tipo de combustible, Cilindrada, Potencia Fiscal, Tara, etc., de la Titularidad del vehículo y su historial de ITVs con anotación de kilometraje.

2.2. Depuración de la base de datos y creación de una base operacional estratégica

La base contiene un total de 8785 registros, correspondientes a 650 autobuses articulados. Se ha procedido al cribado de casos y registros contenidos en la base proporcionada para el cumplimiento de los objetivos del estudio. Los criterios de filtrado de los datos se han realizado por dos vías: mediante el uso de programas informáticos (SPSS V.17) y los algoritmos generados en lenguaje R y macros de Excel. En los casos en que no ha sido posible se ha utilizado un método manual analizando "caso a caso". La base de datos "limpia" contiene 484 casos, y un número total de registros de 6419. Sobre esta base "limpia" se ha realizado un filtrado adicional, basado en eliminación de registros con movilidad excesivamente reducida o excesivamente elevada, dando como resultado la base operacional estratégica DWITVAA que contiene 1566 registros de 462 autobuses.

3. Metodología de minería de datos aplicada

Los árboles de decisión o partición se presentan aquí con vistas a su aplicación en tres objetivos: segmentación, estimación de modelos no-paramétricos para predicción y análisis de sensibilidad. Dos referencias interesantes son: Hastie et al. (2008) [10] y Azzalini y Scarpa (2012) [11].

Para todos los cálculos estadísticos se han utilizado las rutinas de los paquetes del software estadístico público R.

Árboles CART ordinarios y condicionales

DOI: http://dx.doi.org/10.4995/CIT2016.2016.1990

Los árboles de partición son una categoría general que incluye los árboles de clasificación y regresión (CART). Los CART son una técnica de aprendizaje supervisado, no paramétrica, es decir, en contraposición a los modelos paramétricos como la regresión lineal en los que se supone que la relación entrada-salida sigue una determinada forma funcional (lineal en este caso), dependiente de unos parámetros que hay que estimar a partir de la muestra. En los modelos no paramétricos como el CART la "forma funcional" viene determinada mucho más directamente por los datos, sin apenas restricciones previas. Su no restricción a forma funcional concreta, que sí tiene la regresión tradicional les confiere flexibilidad para adaptarse a comportamientos locales distintos de la pauta general y a relaciones fuertemente no lineales. Como inconveniente, al no responder a una expresión analítica de la relación entrada-salida, son más difíciles de interpretar y, sobre todo bajo el enfoque bayesiano y requiere recursos computacionales potentes y complejos. El enfoque algorítmico más habitual para obtener la relación entrada salida-final es iterativo: la descomposición binaria recursiva de las variables de entrada, El algoritmo CART escoge la partición en cada nodo de tal forma que se consiga la máxima reducción de variabilidad de la respuesta en el nodo.

Conjuntos de árboles: Random Forests

Los árboles originales ordinarios (CART) son bastante inestables, es decir, son bastante sensibles a pequeños cambios en los datos. Sin embargo los conjuntos de árboles (ensembles of trees) proporcionan un suavizado que resulta en un modelado y predicción más precisas. Las alternativas de suavizado más conocidas, bajo el enfoque frecuencista, son el Bagging, el Boosting y los Random Forests, siendo esta última la más importante y la que se aplicará en este trabajo. En los Random Forests (bosques aleatorios) se replican con bootstrap los datos y se elige al azar un subconjunto de las variables de entrada, ver, Hastie et al. (2008) [10] o, para aplicaciones al análisis de sensibilidad, Grõmping (2009) [9]. Son la técnica de conjuntos de árboles más sofisticados y eficientes dentro del enfoque clásico o frecuencista.

- **Análisis de sensibilidad con Random Forests**

Los Random Forests tienen dos medidas propias de sensibilidad/importancia, la primera se considera más fiable:

  - %MSE, basada en la reducción de pureza, medida como aumento en el error cuadrático medio de predicción dentro de la muestra, cuando se permutan los valores de la variable de entrada en cuestión.
  - Reducción promedio de pureza. Es el promedio de aumento de pureza (reducción de variabilidad) en todos los nodos en los cuales la partición se haga con esa variable.

DOI: http://dx.doi.org/10.4995/CIT2016.2016.1990

Modelos CART bajo enfoque bayesiano: los BART y los Dynatree

Los modelos de árboles bajo enfoque bayesiano son un alternativa muy competitiva a los Random Forests tanto desde el punto de vista del análisis de sensibilidad como de la cuantificación de la incertidumbre. Presentan el inconveniente de ser requerir tiempos de computación mucho mayores, y presentar en ocasiones problemas de convergencia de los algoritmos de estimación de Monte Carlo que aplican. Aunque los primeros modelos CART bajo enfoque bayesiano se basaban en un único árbol, una alternativa más reciente y que proporciona mejores resultados, es el modelo BART (Bayesian Additive RegressionTrees) de Chipman et al. (2010) [7], que se basa en una suma de árboles. Una versión más eficiente computacionalmente son los árboles dinámicos desarrollados por Taddy et al. (2011) [12] basados en los modelos de Monte Carlo secuencial de aprendizaje de partículas (particle learning) de Carvalho et al. (2009) [6]. La aplicación a análisis de sensibilidad es muy reciente, ver Gramacy et al. (2013) [8].

- **Análisis de sensibilidad con dynatree**

Para realizar el análisis de sensibilidad o importancia de las variables de entrada, el Dynatree obtiene a través de un conjunto de árboles bajo el enfoque bayesiano, una aproximación a la relación entrada salida, y a continuación realiza un análisis ANOVA sobre esa aproximación.

4. Resultados de la aplicación a autobuses articulados

En este apartado se presentan resultados de estimación de movilidad, análisis de sensibilidad e importancias y predicción a través de modelos Random Forests aplicados a la base operacional estratégica DWITVAA.

En la Figura 1 se presenta un árbol CART ordinario. La importancia o sensibilidad se puede analizar identificando las variables de entrada involucradas en las particiones de la parte superior del árbol.

Se observa que la variable de entrada involucrada en las particiones superiores es la antigüedad, junto con potencia y número de plazas.

La primera variable “antigüedad de los vehículos” muestra un valor de corte en 10,43 años. Los vehículos con antigüedad mayor que 10,43 años se vuelven a clasificar con la misma variable de entrada y con frontera en 15,26 años. Los autobuses articulados (AA) con antigüedad de más de 15 años tienen una movilidad media de 28.070 km/año. Los que no cumplen con la regla de partición de la antigüedad (con edad inferior a 15), tienen una nueva variable de partición que es la edad ITV, de tal modo que los AA con EDAD-ITV inferior a 13,68 tienen una movilidad media de 44.390 km/año. A partir de esta división los vehículos con EDAD-ITV mayor que 14 tienen una movilidad media de 44.960 km/año y

DOI: http://dx.doi.org/10.4995/CIT2016.2016.1990

los restantes de 80.260 km/año. Los valores tienen orden de magnitud con los valores obtenidos de dos fuentes de datos relacionados con movilidad de autocares y autobuses articulados. En el último Informe del Observatorio de Costes del Transporte de Viajeros en Autocar de más de 55 plazas de longitud de entre 13,5 y 15 m recorren una media anual de 75.000 km. [2] (no hay dato por antigüedad del vehículo). Los datos de movilidad media anual de la flota de 80 autobuses articulados de la Empresa Municipal de Transportes de Madrid fueron provistos por el Director de Material Móvil e Instalaciones para los años 2013 y 2014: 51.910 y 50.694 km anuales y por vehículo respectivamente. Ninguna de las fuentes aporta movilidad en función de la antigüedad del vehículo y la metodología desarrollada constituye una herramienta que permite esta segmentación y otras que se pueden considerar de interés.

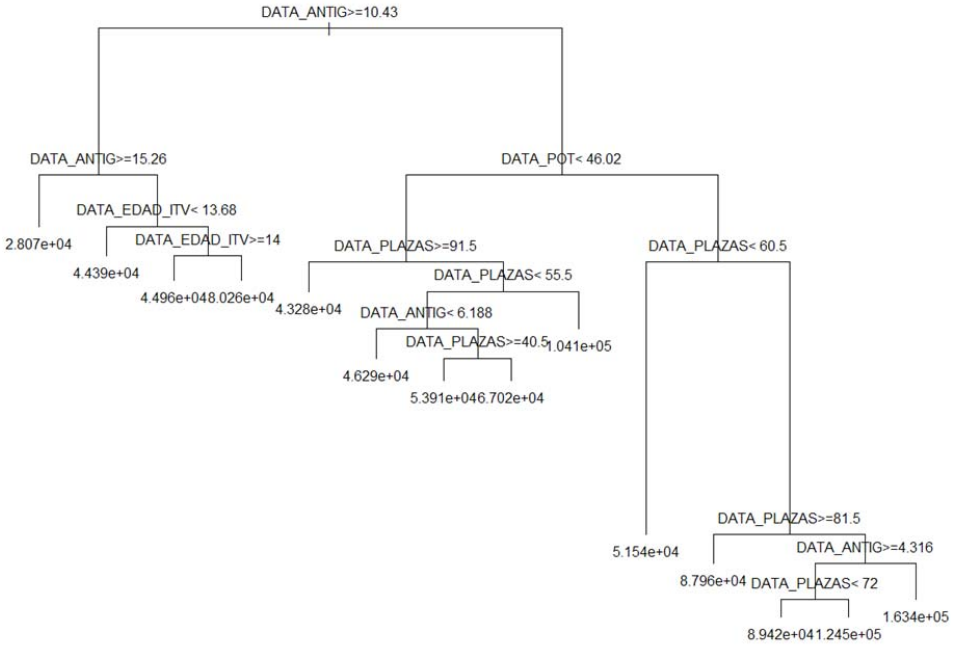

Figura 1. Árbol de regresión CART ordinario para la movilidad con valores promedio para cada nodo final

Un análisis de importancias más potente se ha obtenido a partir de modelos de suma de árboles. En la Tabla 1 se presenta la jerarquía de las variables de entrada por nivel de importancia (normalizada a la suma) según dos criterios: la medida de importancia MSE reduction de randomForest (con ntree= 200 árboles) y la segunda los efectos principales del análisis ANOVA aproximado obtenido con dynaTree (con ntree= 2000 árboles) Se observa que:

DOI: http://dx.doi.org/10.4995/CIT2016.2016.1990

- En primer lugar están permutados las posiciones 1º y 2ª de PLAZAS y ANTIGÜEDAD.
- Le siguen EDAD_ITV y POTENCIA en la ordenación randomForest.
- Para PESO, CIL y AÑO_ITV la ordenación es similar con ambos criterios.
- Las variables EDAD_ITV y POTENCIA no se han introducido en el algoritmo dynaTree por razones de convergencia y de redundancia con ANTIGÜEDAD y CILINDRADA respectivamente.

Tabla 1: Jerarquía de las variables de entrada por nivel de importancia (valores normalizados a la suma).

	Random Forest MSE reduction, ntree = 500 *	Dynatree (ntree=2000) **
DATA_PLAZAS	23,74	17,58
DATA_ANTIG	19,68	36,41
DATA_EDAD_ITV	15,32	-
DATA_POT	15,16	-
DATA_PESO	9,88	15,59
DATA_CIL	8,92	15,20
DATA_ANO_ITV	7,30	15,22
	100,00	100,00

* Efecto total= efectos (individual de la variable+ conjunto con el resto)

La bondad del ajuste de los modelos estimados se ha analizado con la identificación de patrones multivariantes de las variables de entrada que dan como resultado errores grandes observados en los vehículos de menor movilidad correspondiente a autobuses con: ANTIGÜEDAD mayor de 10 años y número de plazas medio-alto.

A continuación se presentan ejemplos de predicción con Random Forests de movilidad de conglomerados homogéneos de autobuses articulados, para el cual se especifican los valores de todas las variables de entrada incluidas en el modelo. En la Tabla 2 se presenta la estimación puntual y el intervalo de confianza de la predicción para cada conglomerado.

Tabla 2: Predicción de movilidad de conglomerados de vehículos homogéneos

Ejemplo	Movilidad (km/año)		
	LI 95%	Predicción puntual	LS 95%
E1.1: AA-9-9-2015-12000-40-30000-75	20.370,50	75.602,80	155.643,40
E1.2: AA-1-1-2015-12000-40-30000-50	20.370,80	58.686,40	137.159,90

DOI: http://dx.doi.org/10.4995/CIT2016.2016.1990

E1.3: AA-17-17-2015-12000-40-30000-75	6.150,00	24.816,30	57.673,00
E1.6: AA-7-4-2012-12000-40-30000-175	37.680,10	75.865,20	137.536,80

Nota: La codificación empleada en la Tabla 2, AA-9-9-2015-12000-40-30000-75 resume los valores especificados para las variables de entrada indicadas de forma correlativa. Así por ejemplo:

- *AA-9-9-2015-12000-40-30000-75 es el caso de un autobús articulado con una antigüedad del vehículo de 9 años, con una edad ITV: 9 años, con año de inspección: 2015, cilindrada: 12000 CC, Potencia: 40 CF, Peso: 30000 kg, y Número de plazas de 75 pasajeros*

Se observan valores de predicción superiores al promedio obtenido para la muestra total de la base para los vehículos de menos de 10 años, y se resalta el patrón relacionado con la antigüedad: los vehículos más antiguos a igualdad del resto de los factores, recorren menos kilómetros que los más modernos.

5. Conclusiones

Se ha realizado una aplicación de metodología sofisticada de criba de datos erróneos en primer lugar, y estudio de importancias y estimación de la movilidad con minería de datos a los datos de movilidad de autobuses articulados obtenidos de las estaciones ITV, a través de un estudio piloto que ha proporcionado resultados de interés para futuras investigaciones con bases de datos más grandes. Los autobuses que recorren más kilómetros son los de menor antigüedad (de menos de 15 años), mientras que los autobuses articulados (AA) con antigüedad de más de 15 años tienen una movilidad media menor de 28.070 km/año. La metodología desarrollada en este trabajo permite la obtención de la movilidad para conglomerados, en función de variables de segmentación de interés (edad, número de plazas, etc.) ya que los mismos pueden presentar diferencias significativas en sus niveles de exposición, de gran interés para las evaluaciones de la seguridad de colectivos.

Agradecimientos

The authors would like to express their gratitude to Dirección General de Tráfico (Spanish Road Traffic Directorate General) for the human and financial resources that it provided for this study (grant reference: SPIP2014-01430).

6. Referencias

DOI: http://dx.doi.org/10.4995/CIT2016.2016.1990

[1] Dirección General de Tráfico, http://www,dgt,es/Galerias/seguridad-vial/estadisticas-e-indicadores/publicaciones/principales-cifras-siniestralidad/Siniestralidad_Vial_2013,pdf

[2] El Observatorio de Costes del Transporte de viajeros en Autocar, Julio de 2015, Consulta en www.fomento.gob.es

[3] Observatorio Social de Transporte por Carretera, Diciembre de 2014, Consulta en www.fomento.gob.es

[4] Carroll, P.S. (1973). Classifications of Driving Exposure and Accident Rates for Highway Safety Analysis, Accident Analysis and Prevention, 5, pp. 71-94.

[5] Kweon Y.J., Kockelman, K. (2003). Overall Injury Risk to Different Drivers: Combining Exposure, Frequency and Severity Models, Accident Analysis and Prevention 35 (3), pp. 414-450.

[6] Carvalho, C., Johannes, M., Lopes, H., Polson, N. (2010). Particle Learning and Smoothing, Statistical Science.

[7] Chipman, H., George, E., McCulloch, R. (2010). BART: Bayesian additive regression trees, The Annals of Applied Statistics, Vol 4(1), pp. 266-298.

[8] Gramacy, R., Taddy, M., Wild, S. (2013). Variable selection and sensitivity analysis via dynamic trees with an application to computer code performance tuning, Annals of Applied Statistics, Vol, 7(1), pp. 51-80.

[9] Grõmping, U., (2009). Variable Importance Assessment in Regression Linear Regression versus random Forests, The American Statistician, Vol, 63(4), pp. 308-319.

[10] Hastie, T., Tibshirani, R., Friedman, J. (2008). The Elements of Statistical Learning: data mining, inference and prediction, Springer.

[11] Azzalini, A., Scarpa, B. (2012). Data Analysis and Data Mining, Oxford.

[12] Taddy, M., Gramacy, R., Polson, R. (2011). Dynamic Trees for Learning and Design, Journal of the American Statistical Association, Vol 106(493), pp.109-123.

[13] Keall M.D., Newstead S. (2009). Selection of comparison crash types for quasi-induced exposure risk estimation, Traffic In., Prev., Vol.10, pp. 23–29.

[14] Harrison, W. A., Christie, R. (2005). Exposure Survey of Motorcyclists in New South Wales, Accident Analysis and Prevention Vol, 37, pp.441-451.

[15] Stamatiadis, N., Deacon J.A. (1997). Quasi-induced exposure: Methodology and insight, Accident Analysis and Prevention, Vol, 29(1), pp. 37–52.