

## II Jornadas Internacionales de Estadística Aplicada 5 y 6 de Diciembre de 2019

### Aplicación de algoritmos de Machine Learning en estudios medioambientales

Autores: Angélica Arenas<sup>1</sup>, Jorge Ruiz Porro<sup>2</sup>, Gema Cereceda<sup>2</sup>, Blanca Arenas Ramírez<sup>3</sup>

<sup>1</sup> Facultad de Ingeniería. IIDISa. Universidad Nacional de Salta. Argentina (UNSa), <sup>2</sup> Escuela de Ingenieros Industriales de Madrid, Universidad Politécnica de Madrid (ETSII-UPM). España, <sup>3</sup> Instituto Universitario de Investigación del Automóvil "Francisco Aparicio Izquierdo". Universidad Politécnica de Madrid (INSIA-UPM). España

*Datos de contacto: aarena@unsa.edu.ar.*

#### RESUMEN.

En este trabajo se presentan algunas aplicaciones de algoritmos de Machine Learning en estudios medioambientales, de especial interés por la contaminación del aire en las grandes ciudades debido al elevado número de vehículos que circulan por ellas a diario.

Los modelos desarrollados son variados como son las aplicaciones y muestran el potencial para el cálculo de emisiones contaminantes. La aplicación de este trabajo consiste a la modelización de ciclos de conducción de autobuses a partir de datos obtenidos en condiciones reales de circulación en entorno urbano y desarrollo de modelos para el cálculo de emisiones contaminantes de una flota de autobuses en la ciudad de Madrid. Los datos se han obtenido en dos campañas de medición en el marco de proyectos de investigación realizadas en el año 2008 y en 2019 respectivamente. Si bien los objetivos de sendos proyectos fueron distintos, los datos obtenidos en la primera campaña han servido de entrenamiento y ajuste de modelos mediante algoritmos del campo de Machine Learning. Los de la segunda campaña sirvieron para validación de las herramientas desarrolladas y también para el desarrollo de otros modelos estadísticos avanzados de contaminantes de interés particular. Aquí se presentan algunas aplicaciones de modelos que muestran la potencia de los algoritmos hasta ahora probados por el equipo investigador y colaboradores.

**Palabras Claves:** algoritmos de Machine Learning, estudios medioambientales, emisiones contaminantes, autobuses, ciclos de conducción

#### INTRODUCCION

Los algoritmos de Machine Learning destacan por la gran capacidad para identificar patrones complejos y descubrir tendencias a partir de grandes volúmenes de datos, creando modelos propios que sirven para la predicción del comportamiento de parámetros y variables de interés. Las aplicaciones son extensas y dentro de la algoritmia se incluyen los modelos de decisión como son los modelos de Árboles de clasificación y regresión (CART), modelos de conjuntos multimodelos "Ensemble models" como son los modelos de bosques aleatorios "Random Forest", modelos de redes neuronales como "Perceptron" y modelos de Clústers tipo "K-means" entre otros. Debido a la extensión del trabajo, aquí se presentan aplicaciones de modelos de árboles CART y Random Forest desarrolladas para el cálculo de emisiones de la flota de autobuses de transporte urbano de la ciudad de Madrid.

## METODOLOGIA

### Planteamiento del trabajo

En este trabajo se presenta:

- Creación de una herramienta capaz de generar ciclos de conducción urbana de autobuses, a partir de los registros de velocidades reales obtenidos mediante los equipos embarcados en el autobús, partiendo de la creación de bases de datos “*ad hoc*”.
- Creación de modelos para el cálculo de emisiones contaminantes en los ciclos de conducción de los autobuses.

El objetivo final es la identificación de los patrones de conducción que mejor se adaptan a las condiciones de explotación de vehículos para el transporte colectivo de personas y a las exigencias medioambientales de calidad del aire en grandes ciudades como Madrid.

### Creación de bases de datos

Los valores de las variables medidas con equipos embarcados en un autobús modelo NL 273-F de MAN y que cumple con la normativa Euro IV de emisiones de la Empresa Municipal de Transportes (EMT) de Madrid, han sido proporcionadas por el Instituto Universitario de Investigación del Automóvil “Francisco Aparicio Izquierdo” de la Universidad Politécnica de Madrid (INSIA-UPM), España. Los ensayos fueron realizados con equipos embarcados que han registrado variables cinemáticas (tiempos, velocidades, etc...) y niveles de emisiones contaminantes. El equipo PEMS –Portable Emissions Measurement System–, fue embarcado en el vehículo para la circulación por los trayectos habituales de algunas de las líneas regulares de la EMT de Madrid tanto en sentido de ida como en el de vuelta, con registros en intervalos de 1 segundo. Los ensayos se han realizado en diferentes momentos del día para que las condiciones de realización de los ensayos cubran la mayor variabilidad posible.

En la Tabla 1 se muestran los ensayos realizados según la línea de recorrido del autobús, el itinerario (ida o vuelta), el combustible utilizado (biodiesel B100 o gasóleo) y el nivel de carga (vacío, media carga y lleno):

Tabla 1: Ensayos realizados

		COMBUSTIBLE UTILIZADO						
		BIO DIESEL 100% (B100)			GASÓLEO			
Línea	Itinerario	Vacio	Media Carga	Plena Carga	Vacio	Plena Carga	Total por itinerario y línea	Total por línea
C1 (Circular 1)	Ida	2	4	1	4	2	13	26
	Vuelta	2	3	2	4	2	13	
27	Ida	2	2	2	4	2	12	24
	Vuelta	2	2	2	4	2	12	
63	Ida			2			2	4
	Vuelta			2			2	
145	Ida	2	2		4	2	10	20
	Vuelta	2	2		4	2	10	
Total por nivel de carga y combustible		12	15	11	24	12	74	
Total por combustible		38			36			

Para el modelado de ciclos, se generaron cinco bases de datos correspondientes al número de combinaciones posibles de combustible y nivel de carga a partir de los datos de partida.

### Identificación de tipos de ciclos de conducción

Se ha creado una herramienta para generar ciclos de conducción urbana de autobuses, a partir de los registros de velocidades reales obtenidos mediante los equipos embarcados en el autobús,

y el correspondiente tratamiento de los bases de datos.

La metodología seguida para la obtención de los ciclos suavizados de conducción urbanos de autobuses comprende dos etapas:

- En primer lugar, se ha desarrollado un algoritmo de programación que permite dividir el ciclo completo de conducción en los diferentes microciclos de movimiento (tramos del ciclo de conducción con velocidad no nula) y de parada (tramos del ciclo de conducción con velocidad nula). Los microciclos de movimiento y de parada aparecen intercalados entre sí.
- A continuación se ha generado un código de programación en MATLAB que permite suavizar cada microciclo de movimiento de tal modo que se consigue que la evolución temporal de la velocidad sufra menos cambios de pendiente y que estos cambios sean menos bruscos en comparación con los asociados a los valores de velocidad reales obtenidos mediante los equipos embarcados utilizados en los ensayos. El resultado son los microciclos de movimiento suavizados a partir de los datos de velocidad reales.

Un microciclo de movimiento estará constituido por los puntos comprendidos entre un punto de inicio de microciclo de movimiento y el punto de fin del siguiente. Un punto de comienzo de microciclo de movimiento tiene asociada una velocidad nula mientras que la velocidad posterior no sea nula. Un microciclo de parada contendrá todos los puntos que se encuentren en el intervalo formado por un punto de fin de microciclo de movimiento y el punto de inicio de microciclo de movimiento siguiente, y los puntos iniciales y finales del mismo tienen velocidad nula. De este modo, pueden identificarse los microciclos de movimiento y de parada que forman el ciclo de conducción completo. Se realizó la formación de conjuntos ordenados de microciclos, formados por cada microciclo de movimiento y sus microciclos de parada anterior y posterior. Los microciclos de movimiento y de parada ordenados en un archivo de tipo texto (.txt) pueden ser utilizados en las aplicaciones dedicadas al desarrollo de algoritmos de programación.

A continuación, se desarrolló un código de programación en MATLAB para suavizar cada microciclo de movimiento de tal modo que se consigue que la evolución temporal de la velocidad sufra menos cambios de pendiente y que estos cambios sean menos bruscos en comparación con los asociados a los valores de velocidad reales obtenidos mediante los equipos embarcados utilizados en los ensayos.

El suavizado se basa en un procedimiento iterativo de comparación de pendientes entre conjuntos de cinco puntos consecutivos de valores de velocidad y los 5 valores del siguiente conjunto.

Si la pendiente de los dos conjuntos es parecida, (se adopta un valor límite para el error relativo entre ambas pendientes), ambos conjuntos se almacenan en un mismo vector y se concatenan estos valores de velocidad a los anteriores si las pendientes continúan siendo similares.

Si la pendiente cambia sustancialmente, se genera un nuevo vector en el que se almacenan los valores de velocidad utilizados para el cálculo de la pendiente que resulta ser diferente. A este nuevo vector seguirán concatenándose valores de velocidad que se asocien a pendientes parecidas hasta que vuelva a producirse un nuevo cambio en la pendiente que conllevará la generación de un nuevo vector.

Este proceso de comparación de pendientes va repitiéndose hasta que se han recorrido todos los puntos pertenecientes al microciclo de movimiento en cuestión y se hace extensivo al resto de microciclos de movimiento pertenecientes al ciclo de conducción completo del autobús.

Para los casos en que los microciclos de movimiento estén compuestos por menos de ocho puntos, no se pueden formar dos conjuntos de cinco valores de velocidad consecutivos.

Por ello, en estos casos, el microciclo de movimiento se divide por la mitad de tal modo que la primera mitad de valores sirven para calcular la pendiente inicial y con la segunda mitad se obtiene la pendiente final. Si ambas pendientes son similares se almacenan en un mismo vector todas las velocidades y, si no lo son, se generan dos vectores almacenando en el primero las velocidades empleadas para calcular la pendiente inicial y en el segundo las asociadas a la pendiente final.

A continuación se generan los microciclos de movimiento aproximados mediante rectas de unión entre los puntos inicial y final de cada uno de los vectores obtenidos anteriormente (denominados teóricos o "idealizados") y calculando valores de velocidad cada segundo (para poder comparar

fácilmente los valores de velocidad reales cuya frecuencia de toma de datos es de un segundo). A continuación, se calcula la diferencia puntual existente entre las velocidades reales (microciclos reales) y las de microciclos de movimiento aproximados (teóricos o idealizados) empleando la siguiente fórmula:

$$\text{Diferencia puntual} = (v_{\text{real}} - v_{\text{teórica}})^2$$

Entonces, se ha obtenido una medida que refleja lo alejado que se encuentra cada microciclo de movimiento aproximados con respecto a las velocidades reales sumando las diferencias puntuales de los puntos pertenecientes a cada microciclo de movimiento.

Para los microciclos de movimiento que no se aproximan lo suficiente a las velocidades reales, se repite el proceso de cálculo de pendientes y generación de vectores con velocidades que dan lugar a pendientes similares pero permitiendo un error límite prefijado menor que el utilizado en el paso anterior. Con ello se consigue que estos microciclos se aproximen más a las de velocidades reales consiguiendo una mayor bondad de ajuste.

Para determinar el grado de cercanía entre las velocidades respectivas se calcula el error cuadrático medio (MSE) con;

$$MSE = \sqrt{\frac{\sum_{i=1}^n (v_{\text{real}} - v_{\text{teórica}})^2}{n}}$$

Se ha comprobado que el MSE con los microciclos aproximados de movimiento alejados de la realidad sin corregir es mayor que tras proceder a su corrección. En la Figura 1 se muestra un esquema de los pasos seguidos.



Figura 1: Proceso de suavizado de microciclos de movimiento

El procedimiento adoptado confiere cierto grado de libertad al investigador, ya que éste puede modificar el nivel de error admisible entre el ciclo suavizado y el real así como el error relativo límite entre las pendientes calculadas entre dos conjuntos de valores de velocidad consecutivos. En el proceso, el investigador asume un nivel de compromiso del ajuste entre el ciclo real y el modelizado. Los microciclos suavizados o modelizados se concatenan en el orden adecuado (los de movimiento con los de parada) para representar el ciclo completo de conducción urbana del autobús y se someten a un proceso de verificación del nivel de alejamiento entre el real y el modelizado mediante la medida del error MAPE (*Mean Absolute Percentage Error*).

En la Figura 2 se muestra la representación gráfica del ciclo de conducción real completo y los modelizados (suavizados e idealizados) del autobús en una de las pruebas experimentales (test070720080850\_L145i) que permite su comparación visual.

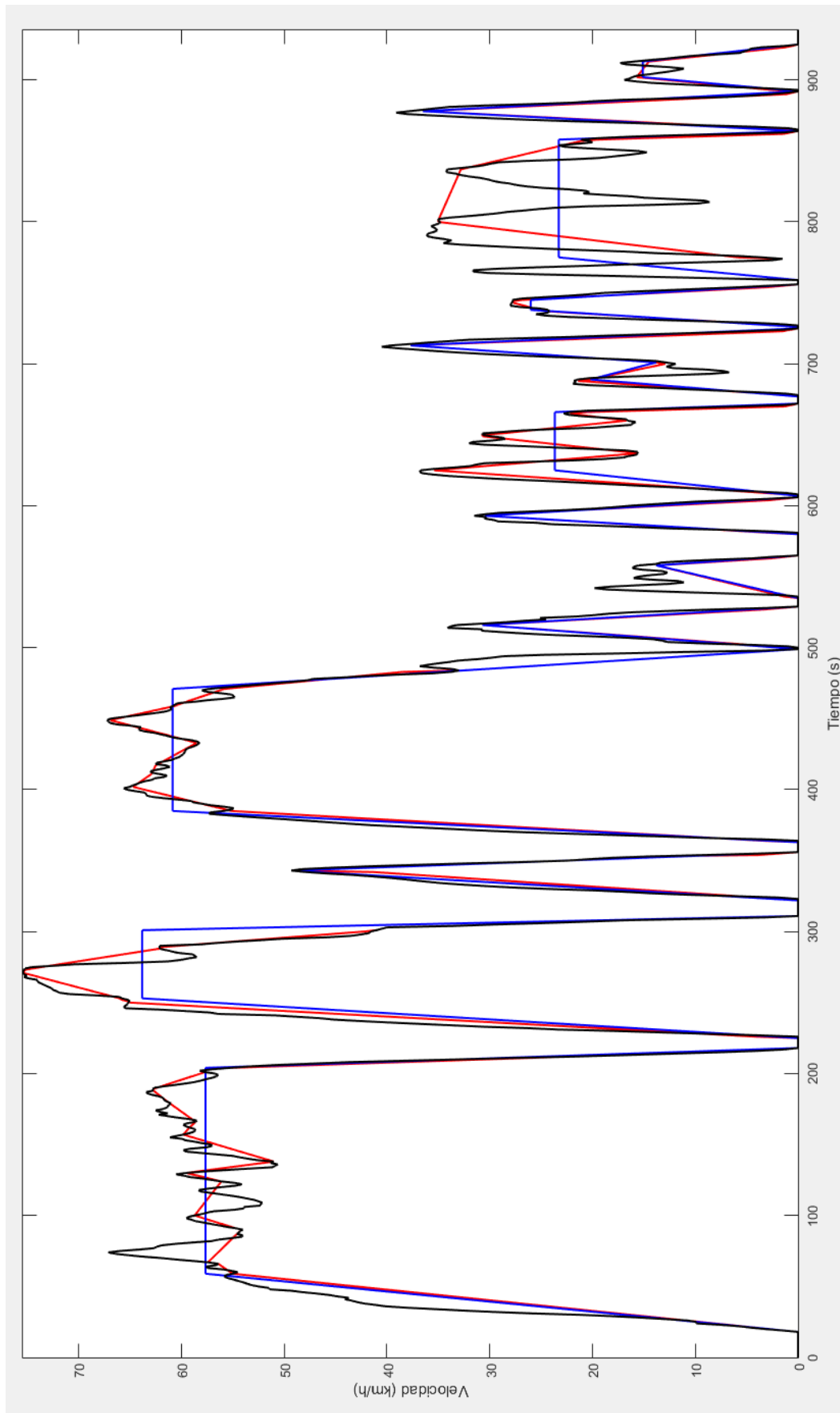


Figura 2: Ciclo de conducción real, suavizado e idealizado para el test070720080850\_L145i. Salida del Programa MATLAB

Los errores obtenidos para el caso concreto del ensayo test070720080850\_L145i, se muestran en la siguiente tabla:

Tabla 2: Errores de aproximación de los ciclos suavizado e idealizado para el ensayo test070720080850\_L145i. Elaboración propia

CICLOS	MAPE (%)
Real - Suavizado	15,09
Suavizado - Idealizado	20,59
Real - Idealizado	21,52

↑ 42,61%

Como se puede observar en la tabla anterior, el MAPE entre el ciclo de conducción suavizado y el ciclo real resulta ser del 15,09%; entre el ciclo suavizado y el idealizado es de 20,59% y, finalmente, entre el ciclo real y el idealizado, el MAPE es del 21,52%. Por tanto, la conclusión final es que el ciclo de conducción idealizado se aleja un 42,61% más que el ciclo suavizado con respecto al ciclo real (el MAPE es un 42,61% mayor).

En el proceso se ha identificado y modelizado los 6 tipos de microciclos de movimiento idealizados:

Tabla 3: Tipología considerada para los microciclos de movimiento idealizados. Elaboración propia

Tipo	1	2	3	4	5	6
Morfología						

## Generación de modelos Random Forest para el cálculo de emisiones en ciclos de conducción

En la Tabla 4 se muestran las variables obtenidas en la campaña de mediciones realizadas en los ensayos indicados. Además, se señalan las variables explicativas para la generación de los modelos explicativos y predictivos de emisiones contaminantes. Los dos bloques finales contienen los nombres de las variables respuesta de los modelos.

Los microciclos suavizados o modelizados se han aplicado al cálculo de las emisiones de contaminantes de CO, CO<sub>2</sub>, NO<sub>x</sub> y de partículas en el intervalo de tiempo correspondiente a los microciclos, con modelos Random Forest del campo de Machine Learning.

Los algoritmos para la generación de modelos de bosques aleatorios de árboles de regresión y de clasificación (Random Forest- RF-) se han escrito para el software estadístico de uso público R y permite disminuir la correlación entre árboles, partiendo de dos fuentes de aleatoriedad: la de los datos y la de los clasificadores seleccionados para el proceso iterativo de los modelos RF.

Se ha utilizado la técnica de la validación cruzada dividiendo el conjunto completo de observaciones iniciales en dos partes: la primera el "training set" con un 90% de las observaciones sirve para el ajuste de los modelos RF y la segunda constituye el conjunto de comprobación o "test". Los clasificadores son las variables de velocidad, aceleración, temperatura exterior, presión, humedad, y batería; con los datos de la velocidad y de la aceleración teórica obtenidas del suavizado del ciclo, sin modificar.

Para observar la influencia de ambos parámetros cinemáticos, se realiza el estudio con un retardo de 1 segundo tanto en la velocidad como en la aceleración; y, por último, se realiza con un retardo de 3 segundos. Para las distintas estrategias de modelado se obtuvo la variabilidad explicada con

la expresión:

$$VE = 1 - \frac{MSE_{OOB}}{\sigma_y^2}$$

Tabla 4: Variables contenidas en las bases de datos

Datos de la base de datos	Tipo de variable	Variable	Modelo
Datos propios del ensayo	Vehículo	Datos del autobús	
	Equipo de medida	Valores de calibración	
	Combustible	Composición	
	Gases de escape	Composición	
Variables medidas cada segundo	Variables temporales	Hora exacta	
		Tiempo transcurrido desde el comienzo del ensayo	X
	Posición del autobús	Latitud	
		Longitud	
		Altitud (m)	
	Valores cinemáticos	Velocidad instantánea (km/h)	X
	Condiciones ambientales	Temperatura ambiente (°C)	X
		Presión ambiental (kPa)	X
		Humedad relativa (%)	X
	Condiciones de los gases de escape	Temperatura (°C)	
		Presión (kPa)	
	Emisiones contaminantes	CO (%w/v y g/s)	X
		CO <sub>2</sub> (%w/v y g/s)	X
		NO <sub>x</sub> (ppm y g/s)	X
		PM (mg/m <sup>3</sup> y g/s)	X
		HC (ppm y g/s)	X
		H <sub>2</sub> O (%w/v y g/s)	
	Otras	Consumo de combustible (g/s)	
		Relación aire/combustible	
		Batería (V)	X

## DESARROLLO

Los resultados de las distintas estrategias de modelado para los distintos contaminantes se muestran en las Tablas 5, 6 y 7.

Por último, dentro del análisis de la emisión de contaminantes, se realizó un estudio de la variación de la predicción de los contaminantes en función de las velocidades reales o suavizadas en el caso de un microciclo y un contaminante concretos para un retardo de tres segundos en las variables cinemáticas. Además, se consideran tres posibilidades en cuanto al retardo en las emisiones: que no haya retardo, que el retardo sea de un segundo o que sea de tres.

Tabla 5: Variabilidad explicada de ensayos sin retardo

Contaminante	sin retardo
CO <sub>2</sub>	86.07%
NO <sub>x</sub>	81.33%
Partículas	93.05%
CO	66.16%

Tabla 6: Variabilidad explicada de ensayos con un retardo de 1 segundo

Contaminante	retardo de 1 s
CO <sub>2</sub>	86.07%
NO <sub>x</sub>	77.99%

Partículas	92.27%
CO	66.16%

Tabla 7: Variabilidad explicada de ensayos con un retardo de 3 segundos

Contaminante	retardo de 3 s
CO <sub>2</sub>	87.13%
NO <sub>x</sub>	80.23%
Partículas	92.22%
CO	67.85%

Se puede observar que se consiguen mejores resultados sin aplicar retardos para el NO<sub>x</sub> y para las partículas mientras que, para los compuestos con carbono la mejor situación se alcanza con tres segundos de retardo. La mínima variabilidad se obtiene para el CO, dándose un dato muy similar para las tres situaciones.

Un resultado de mucho interés de los modelos RF, es la importancia de las variables seleccionadas para la explicación y predicción de las emisiones. Los estadísticos MSE y el de impureza del nodo de Gini, coinciden en el orden de las variables explicativas del modelo pero el MSE discrimina de forma muy clara la importancia que las variables cinemáticas tienen frente al resto.

Como se obtienen resultados similares para todos los contaminantes, se indica como ejemplo uno de ellos, concretamente para las emisiones de NO<sub>x</sub>. Las variables velocidad y aceleración se sitúan por delante del resto, indicando la relevancia que las variables cinemáticas representan para el modelo de emisiones de NO<sub>x</sub>, seguidas por las ambientales (Figura 3).

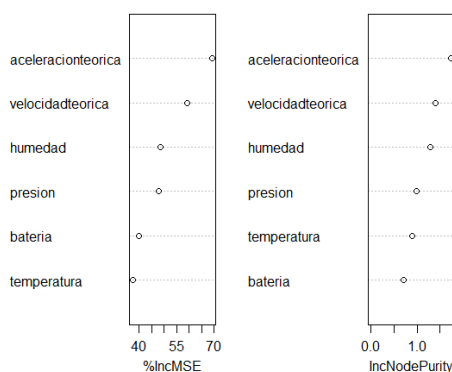


Figura 3: Importancia de las variables para el NO<sub>x</sub>

## CONCLUSIONES

Sobre el procedimiento de identificación de ciclos:

Se han desarrollado herramientas de identificación y segmentación de ciclos de movimiento y de parada, de forma automatizada.

Los códigos y la algoritmia implementada se basan en la búsqueda de patrones de forma automática, en base a comparación de niveles y valores y concatenación de valores próximos.

Asimismo se ha desarrollado una herramienta de suavizado que produce nuevos ciclos de movimiento suavizados que reducen el nivel de ruido, lo que permite el desarrollo de nuevos modelos de emisiones más precisos.

Con esta herramienta se pueden generar ciclos idealizados, y cuyas emisiones totales puedan ser determinadas dentro de márgenes de errores acotados a valores aceptables, y que tienen características de repetibilidad y normalización para su recomendación de ejecución con ajuste a las exigencias medioambientales de las ciudades.

Sobre los modelos de emisiones contaminantes:

Se han desarrollado, modelos de minería de datos del campo de Machine Learning, para la identificación de las variables con influencia en las emisiones con distintas estrategias de modelado, en la búsqueda de los mejores modelos de predicción para los diferentes contaminantes.

En cuanto a la importancia de las variables empleadas en la generación de los modelos estadísticos de emisiones, se concluye que las más influyentes son las cinemáticas (en primer lugar la aceleración y a continuación la velocidad), seguido de la humedad, la presión y finalizando en la temperatura y la batería, y que concuerdan con otros trabajos reflejados en la bibliografía técnico – científica consultada.

Analizando la variabilidad explicada de las variables, se puede observar que tanto para el NO<sub>x</sub> como para las partículas se consiguen mejores resultados sin aplicar retardos. Sin embargo, para el CO<sub>2</sub> y el CO el valor óptimo se obtiene con un retardo de tres segundos. En el caso de este último la variabilidad explicada es muy similar en las tres situaciones, detectándose que es relativamente menor que para el resto de contaminantes.

Las aplicaciones de los modelos muestran la potencia de los algoritmos hasta ahora probados por el equipo investigador y colaboradores.

## AGRADECIMIENTOS

Este trabajo se ha realizado en el proyecto de investigación CÍCLOPE (Sistema de optimización de ciclos urbanos de conducción. Aplicación a la generación de patrones adaptados a exigencias medioambientales y situaciones de explotación de flotas de vehículos) financiado por el Plan Nacional 2016-2018 del Ministerio de Economía y Competitividad- Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad. TRA2015-68803-R. Además, los autores agradecen la financiación parcial de este trabajo a la Comunidad de Madrid, que mediante el programa SEGVAUTO TRIES, ha contribuido a su desarrollo.

## BIBLIOGRAFIA

Román, A., “Metodología para la asignación de vehículos de una flota a rutas preestablecidas”, Tesis Doctoral, Escuela Técnica Superior de Ingenieros Industriales. Universidad Politécnica de Madrid, España, 2014.

Jorge Ruiz Porro, Gema Cereceda, José M. Mira McWilliams, Blanca Arenas Ramírez, Francisco Aparicio Izquierdo. (2017). Modelos de identificación de ciclos de conducción urbana de autobuses. 13º Congreso Iberoamericano de Ingeniería Mecánica (CIBIM 2017). 23 al 26 de octubre de 2017. Lisboa, Portugal

García, Natalia Fonseca, José M. Mira, and Zamir Mera. (2019). Modelling urban bus fleet emissions with machine learning boosting methods: City of Madrid. 23rd International Transport and Air Pollution Conference. 15-17 May 2019, Thessaloniki, Greece.

Edinalva Gomes Bastos, Víctor Pita González-Campos, Blanca Arenas-Ramírez, José M. Mira, Francisco Aparicio-Izquierdo. (2019). Modelado de emisiones de partículas de autobuses urbanos. 2do CONGRESO sobre MEDIOS de TRANSPORTES y sus TECNOLOGÍAS APLICADAS. Gral. Pacheco, 11 – 13 de Setiembre de 2019.

Zarkadoula, M., Zoidis, G. & Tritopolou, E., “Training urban bus drivers to promote smart driving: A note on a Greek eco-driving pilot program”, Transportation Research Part D, 12:449–451, 2007.

Carrese, S., Gemma, A. & La Spada, S., “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. 87:211–221, 2013.

José María López, Felipe Jiménez, Javier Páez, M. Nuria Flores, Angélica N. Arenas, Blanca Arenas-Ramirez, Francisco Aparicio. (2017). Modelling the fuel consumption and pollutant emissions of the urban bus fleet of the city of Madrid. (2017). Transportation Research Part D 52: Transport and Environment Journal. pp. 112-127. Editorial: Elsevier. ISSN: 1361-9209. U.K. DOI: 10.1016/j.trd.2017.02.016.

Horiba, 2017. OBS-2200. [En línea]  
Available at: <http://www.horiba.com/de/automotive-test-systems/products/emission-measurement-systems/portable-emission-measurement-systems/details/obs-2200-877/>  
[Último acceso: 17 Marzo 2017].

Instituto de Investigación del Automóvil (INSIA), 2016. Proyecto CÍCLOPE. [En línea]  
Available at: <http://insia-upm.es/portfolio-items/proyecto-ciclope/>  
[Último acceso: 25 noviembre 2019].

The R Foundation, 2017. Programa R. [En línea]  
Available at: <https://www.r-project.org/>  
[Último acceso: 17 octubre 2019].