

# Manejo de “Big Data” Aplicado a Estudios de Plazas de Peaje Utilizando un Simulador de Conducción

Didier Valdés Díaz, Ph.D., Benjamín Colucci Ríos, Ph.D., P.E., Enid Colón Torres, Ricardo García Rosario, Johnathan Ruiz González, B.S.C.E., Bryan Ruiz Cruz, B.S.C.E.  
Universidad de Puerto Rico Recinto de Mayagüez

*Abstracto— La ingeniería y las ciencias continúan experimentado grandes cambios que se incorporan y modifican los sistemas existentes. Para utilizar, evaluar y mejorar los sistemas innovadores que se van desarrollando en diversas disciplinas se dispone de un gran número de datos que es necesario analizar. Uno de los campos de estudio que trabaja con sistemas que generan volúmenes significativos de datos a una alta velocidad de insumo (“Big Data”) es la ingeniería de transportación. Esta ponencia presenta una metodología para el análisis de datos de simuladores de conducción utilizando el primer simulador de conducción de la Universidad de Puerto Rico Recinto de Mayagüez (UPRM) con énfasis en autopistas con plazas de peaje en Puerto Rico. Esta investigación considera dos configuraciones de señalizaciones en la Plaza de Peaje Caguas Sur con el propósito de evaluar cuál configuración tiene potencial para reducir choques en la plaza de peaje. Los resultados generados por la simulación pueden ser catalogados como “Big Data” debido a la gran cantidad y complejidad de los mismos. Los métodos de análisis tradicionales no son suficientes para procesar y analizar todos los resultados. La metodología presentada utiliza como plataforma el programa R® que facilita el manejo y procesamiento de los datos para sus respectivos análisis e ilustraciones gráficas de los resultados. El procedimiento utilizado permite estudiar todos los datos en el espacio, ofreciendo una mejor visualización de lo que ocurre con cada conductor a lo largo de la simulación. La metodología presentada elimina la limitación de tener que seleccionar segmentos de análisis a priori mediante el uso de las técnicas de “Big Data.” Por lo tanto, los análisis se pueden concentrar en los datos que mejor representan el comportamiento de los conductores.*

**Palabras Claves—“Big Data”; R®; Simulador de Conducción; Plazas de Peaje; Análisis de Datos**

## I INTRODUCCIÓN

El término “Big Data” se utiliza para describir un conjunto de datos complejos que son difíciles de analizar con las herramientas estadísticas tradicionales [1]. La mayoría de las investigaciones indican que el término “Big Data” se divide en tres elementos importantes: Volumen, Velocidad y Variedad de datos [2]. Sin embargo, otros estudios involucran otros dos elementos: Valor y Veracidad [3]. En los últimos cinco años, se ha observado un crecimiento en las investigaciones para resolver los problemas asociados al almacenamiento, la velocidad de análisis y el entendimiento de la variabilidad de los datos. Una de las disciplinas en la que se está trabajando activamente con la implantación y

desarrollo de tecnologías para el uso de “Big Data” es la ingeniería de transportación. Existen factores relacionados a la transportación como son las redes de transporte, los itinerarios de transporte colectivo, la operación de puertos y los datos de velocidad y volumen de vehículos en segmentos de carreteras entre otros, que se han beneficiado por las nuevas técnicas de interpretación de “Big Data” en particular en lo relacionado a la medición y análisis de variables de transporte en tiempo real. Sin embargo, no se ha encontrado en la literatura investigaciones sobre el procesamiento y análisis de “Big Data” asociado a la recopilación y análisis de datos de un simulador de conducción.

Los simuladores de conducción son una herramienta costo efectiva que brinda información esencial a los investigadores sobre temas de interés relacionados a la transportación como son: la seguridad vial, factores humanos y la implementación de nuevas tecnologías, entre otros. En el año 2013, la Universidad de Puerto Rico, Recinto de Mayagüez (UPRM) se unió a otras cuatro universidades de los Estados Unidos en el Centro Universitario de Investigación en Transporte SAFER-SIM, con el fin de investigar temas relacionados a la seguridad vial haciendo uso de simuladores de conducción. La primera investigación realizada con el simulador de conducción de la UPRM consistió en evaluar dos configuraciones de señalizaciones en la Plaza de Peaje de Caguas Sur, localizada en el municipio de Caguas, Puerto Rico. Se evaluó cuál configuración tiene el menor potencial para choques en la plaza de peaje. Se analizaron 12 escenarios para cada configuración, alterando las condiciones del tráfico, el carril donde comenzaba la simulación, el carril por donde debía pasar el sujeto en la plaza de peaje y las condiciones de día y de noche. El estudio consistió en tomar la información de 20 sujetos, entre las edades de 18 a 70 años, para las dos configuraciones [4]. Se analizaron tres variables dependientes, desviación estándar de la posición del vehículo en la carretera, velocidad promedio y ruido de aceleración (desviación estándar de la aceleración) [5]. El simulador de conducción de la UPRM toma datos aproximadamente cada 0.02 segundos para 35 variables. Esto significa que cada minuto el simulador toma aproximadamente 105,000 datos en total.

Para la investigación de la Plaza de Peaje de Caguas Sur, se utilizaron 12 de las 35 variables provistas por el simulador.

Estas fueron: número de sujeto, edad, género, configuración, escenario, zona de interés, tiempo, velocidad, aceleración y las coordenadas de la posición del vehículo (X, Y, Z). Los escenarios simulados para esta investigación tenían una duración promedio de 5 minutos lo cual representa en promedio 180,000 datos para un solo sujeto. Teniendo en cuenta que la investigación incluye 12 escenarios y un total de 20 sujetos, se tiene un total de 32,578,780 datos. Ante la cantidad de datos que obtenemos de las simulaciones las herramientas comerciales comunes, como Excel®, entre otras, no tienen la capacidad para procesar y realizar los análisis necesarios. Por esta razón se requiere el uso de otras herramientas comerciales especializadas para realizar análisis estadísticos para grandes cantidades de datos, como lo es el programa R®.

R® es un programa de código abierto (“open source” como se le conoce en inglés) que se utiliza para realizar análisis estadísticos. Esta herramienta permite la creación de códigos para generar gráficos y realizar análisis sin restricción de cantidad de datos. Utilizando la plataforma R® se creó un programa, ReadData Toll Plaza, que une los archivos de todos los sujetos y configuraciones y analiza las variables dependientes en sus respectivas zonas de interés sin restricciones de cantidad de datos.

## II. OBJETIVOS

El objetivo de esta ponencia es presentar la metodología utilizada en la UPRM para analizar los datos recopilados por el simulador de conducción. A la vez se presentan los resultados del uso de la metodología desarrollada en un estudio de caso correspondiente al análisis de la seguridad vial en una plaza de peaje.

## III. REVISIÓN DE LITERATURA

La revisión de literatura se concentra en tres aspectos principales que se entrelazan en esta ponencia: “Big Data”, seguridad vial y simuladores de conducción. Inicialmente se presenta la literatura relevante a “Big Data” y se resumen los conceptos principales y su aplicación en el campo de la transportación. La aplicación particular de la metodología desarrollada tiene como objetivo mejorar la seguridad vial que es el segundo aspecto presentado mientras que el tercer aspecto es el relacionado con simuladores de conducción que es la fuente de los datos para la cual se desarrolla la metodología de análisis.

### A. “Big Data”

El término “Big Data” es un concepto que se ha mencionado mucho durante la pasada década. Según el departamento de transportación de los Estados Unidos (U.S. DOT por sus siglas en inglés), “Big Data” es el proceso en el cual se genera

información a gran escala, incluyendo la recopilación, manejo y análisis de datos [6]. Este concepto surgió de forma espontánea y fue aceptado por la industria, tanto pública como privada, y posteriormente incorporado en investigaciones y publicaciones por parte de la academia [3]. No fue hasta el 2011 que el concepto de “Big Data” comenzó a investigarse a nivel multidisciplinario, generando diferentes significados donde algunos estudios se enfocan en definir que es el término “Big Data” mientras otros estudian cómo se puede utilizar [7].

Una de las características que definen “Big Data” es el gran tamaño de los datos. Sin embargo, varias publicaciones indican que existen tres características que definen el concepto “Big Data”: volumen, variedad y velocidad [8]. El volumen se define como magnitud de los datos. El volumen de datos necesario para considerarse como “Big Data” varía de acuerdo al tipo de dato. Además, lo que hoy en día se considera “Big Data” puede ser distinto a lo que se considere “Big Data” en años futuros debido a la rápida evolución de dispositivos utilizados para almacenar y procesar datos, con lo cual se genera mayor capacidad [7]. La segunda característica, variedad, se define como la heterogeneidad entre la estructura de los datos. Emani et al. (2015) indican que la variedad se puede representar como estructurado (i.e. datos procesados), semi-estructurado o no estructurado (i.e. datos no procesados) [8]. La velocidad se define como la razón a la cual se generan y deben ser procesados los datos. Otros estudios añaden dos características adicionales: valor y veracidad [2]. El valor se define como el proceso de extraer la información valiosa de una gran cantidad de datos mientras que la veracidad se define como la precisión y exactitud de los datos. Hashem et al. (2015) indican que “Big Data” se puede clasificar en diferentes categorías para poder tener un mejor entendimiento de las características debido a la gran escala de los datos [3]. Las categorías que definen los investigadores son: fuente, formato, almacenaje, “*staging*” y la forma de procesar los datos.

El análisis de “Big Data” se ha utilizado en diversos campos como redes sociales, aplicaciones médicas e ingeniería, entre otros [9, 10]. Una de las áreas de la ingeniería civil que ha estado trabajando con la aplicación de las metodologías desarrolladas en el campo de “Big Data” en los pasados años es la transportación. Nuevas tecnologías en el campo de la transportación, como son los Sistemas Inteligentes de Transportación (ITS por sus siglas en inglés), recopilan una inmensa cantidad de datos en forma continua a lo largo de una extensa zona de estudio durante un periodo de tiempo que dura desde días hasta años. Algunas compañías utilizan sensores y Sistemas de Posicionamiento Global (GPS, por sus siglas en inglés) para recopilar datos como la velocidad, aceleración y coordenadas a intervalos de tiempo de 5 a 10

segundos con el propósito de extraer información valiosa que ayude a maximizar la productividad del sistema mientras se minimizan los costos de operación [11]. El problema es que los métodos de análisis tradicionales no son suficientes para manejar la cantidad de datos, especialmente cuando deben ser procesados lo más rápido posible para proveer soluciones a tiempo real [12]. Sin embargo, las agencias que operan sistemas de transportación deben invertir en profesionales y programas para mantenerse al día con la rápida y constante evolución de herramientas que se utilizan para trabajar con “Big Data” [13]. Es por esta razón que se necesita desarrollar nuevas metodologías que permitan el procesamiento de “Big Data” de una forma rápida, eficiente y efectiva. Esto permitiría mejorar la operación y eficiencia de diversos sistemas de transportación para los cuales se puede recopilar información a tiempo real.

### B. Seguridad en Plazas de Peaje

La seguridad en las carreteras es uno de los asuntos de mayor importancia a nivel global entre las agencias que operan sistemas de transportación. Según la Organización de la Salud (WHO por sus siglas en inglés), ocurren 1.25 millones de fatalidades como consecuencia de choques en las carreteras, donde los más afectados son las personas entre las edades de 15 a 29 años [14]. Las plazas de peaje constituyen uno de los sistemas que se han visto afectados por el incremento en choques fatales y heridas severas. Durante las pasadas décadas, las plazas de peaje se han diseñado y construido sin utilizar un mismo estándar que mantenga uniformidad en los mensajes, colores y localización de los dispositivos de control de tráfico [15]. Además, la evolución de nuevas tecnologías en los sistemas de peaje, como son los sistemas de cobro automático (ETC por sus siglas en inglés), ha generado un cambio en el comportamiento del conductor. La uniformidad de los dispositivos de control de tráfico es un factor que afecta positivamente la seguridad en la carretera debido a que reduce los puntos de conflicto, facilitando la toma de decisiones del conductor [16]. A pesar de que se incorporó una nueva sección dedicada a plazas de peajes en la última revisión del Manual de Dispositivos de Control de Tráfico (MUTCD por sus siglas en inglés), aún existen sistemas de plazas de peaje que no han actualizado debidamente las señalizaciones [17]. Por lo tanto, el efecto de las señalizaciones en las plazas de peaje debe ser estudiado para entender como el comportamiento del conductor se ve afectado y evaluar las posibles alternativas que beneficien la seguridad en la carretera.

### C. Simulador de Conducción

Una de las herramientas que se han utilizado en los pasados años para estudiar el comportamiento del conductor y su relación con seguridad vial de forma costo-efectiva son los simuladores de conducción. Este equipo permite evaluar

condiciones existentes y futuras de un sistema de transportación de una forma segura, donde el participante puede ser expuesto a una situación de alto riesgo si sufrir daños físicos. Diferentes investigaciones han utilizado simuladores de conducción para evaluar la efectividad de señalizaciones con mensajes variable y sistemas de amortiguadores de choques entre otros dispositivos de control de tráfico [18-20]. De igual forma, los simuladores de conducción se pueden utilizar para analizar factores que afecten la seguridad, como la configuración de señalizaciones, en sistemas de carreteras que contengan plazas de peajes [4-5].

Los simuladores de conducción permiten recopilar datos de variables como velocidad, aceleración y posición del vehículo en un intervalo de tiempo definido a lo largo de la simulación. La cantidad de datos varía de acuerdo al número de variables, escenarios, participantes e intervalo de colección de datos establecido por el diseño de experimento del estudio. Por lo tanto, los resultados obtenidos de la simulación pueden ser catalogados como “Big Data” debido al gran volumen y variedad de los datos. Esto brinda la oportunidad para el desarrollo de una metodología que permita procesar los datos no estructurados que se obtienen de los resultados de la simulación.

## IV. METODOLOGÍA UTILIZADA PARA EL ANÁLISIS DE DATOS

El simulador nos provee varios archivos, los cuales contienen información de más de 35 variables. Para el estudio se analizaron las variables de velocidad, posición y aceleración de las cuales tenemos datos cada 0.02 segundos. Estos archivos son tipo .PLT y pueden ser leídos utilizando programas como Excel, Notes, Gnuplot, MatLAB y R<sup>®</sup> entre otros. Cada uno de los archivos del simulador es de aproximadamente 17 kilobytes, dependiendo del tiempo que dure cada experimento y la distancia recorrida. Para analizar los datos obtenidos, se deben agrupar todos en un solo archivo. Este proceso es limitado por algunos programas como Excel<sup>®</sup> el cual tiene un límite de 1,048,576 filas. Debido a estas limitaciones se decidió crear un código utilizando la plataforma R<sup>®</sup>, el cual nos permite crear una base de datos ilimitada con toda la información de los sujetos, escenarios, configuraciones y las variables pre-determinadas que se utilizarán para el análisis. El procedimiento llevado a cabo para evaluar y analizar los datos provenientes del simulador son los consiste de los siguientes ocho pasos:

- Recopilación de los archivos de datos.
- Creación de una base de datos.
- Graficar variables de interés.
- Identificar zonas de interés.
- Creación del modelo lineal.

- Analizar el modelo lineal.
- Aplicar la Prueba ANOVA con corrección de Tukey.
- Interpretación de los resultados.

Para mostrar el procedimiento, utilizaremos el caso de estudio de la Plaza de Peaje de Caguas Sur, Puerto Rico. Ver Fig. 1.



Fig. 1. Imagen de la Plaza de Peaje simulada de Caguas Sur, Puerto Rico

#### A. Procedimiento para recopilación de archivos y creación de base de datos.

Se comienza con la recopilación de los archivos de datos provenientes de SimCreator/SimVista. Estos son generados por el simulador a medida que cada sujeto recorre los escenarios bajo estudio. Luego de obtener todos los archivos de los sujetos del estudio se agrupan por sus extensiones, en este caso utilizamos los archivos con la extensión .PLT. Al estar agrupados podemos utilizar el código en el programa R<sup>®</sup> para unir todos los datos y además seleccionar las variables deseadas. Si el código no ejecuta ningún error, se completa el proceso creando un archivo con extensión .CSV el cual contiene una matriz NxM donde N es la cantidad de filas de datos y M es la cantidad de columnas que representa las variables de interés. Este proceso es explicado en la Fig. 2. En el caso de estudio de la Plaza de Peaje Caguas Sur se creó un archivo con extensión .CSV el cual contiene una matriz de 2,506,060 filas con 13 columnas.

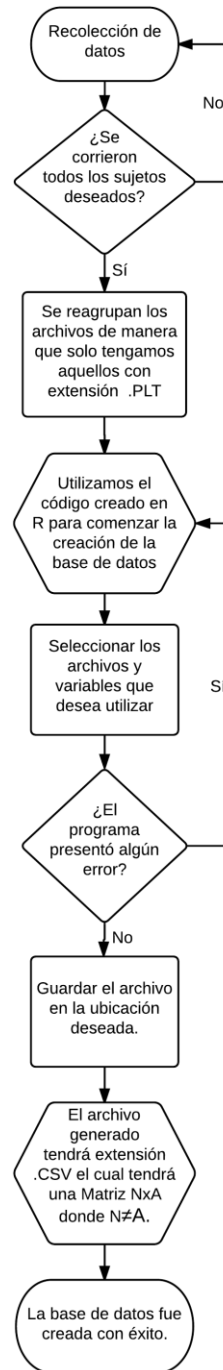


Fig. 2. Procedimiento para recopilación de archivos y creación de base de datos.

#### B. Graficar variables e identificar zonas de interés

Luego de agrupar los datos, se procede a determinar las zonas de interés de estudio. Estas varían dependiendo del tipo de estudio que se esté realizando y la variable que se desee analizar. En forma de ejemplo se presentará la variable velocidad para ilustrar la metodología utilizada. Al ilustrar la

velocidad en función de la distancia podemos identificar diferentes áreas de interés. Este método nos permite agilizar el proceso de selección de zonas a analizar, ya que en la mayoría de los casos se pueden observar patrones de interés. Los cambios de velocidad se dan mayormente en las áreas donde se encuentran las señalizaciones de tránsito. La Fig. 3 ilustra una gráfica de velocidad vs distancia con una posible zona de interés a estudiar.

Esta imagen muestra que existe una diferencia en las zonas denotadas con círculos rojos. En la Configuración 1, la velocidad máxima alcanzada por un conductor es de 75 mph y en la Configuración 2 es de 65 mph, esta zona es de interés ya que hay una diferencia notable de 10 mph.

La Tabla 4D-2 del MUTCD del 2009 se utilizó para definir el tamaño de las zonas de interés [16]. De acuerdo a esta tabla las distancias de visibilidad dependen de la velocidad rotulada en el segmento de carretera. La Tabla 1 muestra la distancia mínima para visualizar los letreros según el MUTCD 2009 adaptada a sistema internacional.

Tabla 1. Distancia mínima de visibilidad de los letreros.

Velocidad de 85 Percentil (mph)	Distancia Mínima de Visibilidad (metros)
20	53.34
25	65.53
30	82.29
35	99.06
40	118.87
45	140.20
55	164.59
60	190.5
65	217.93

Nota: Las distancias establecidas están derivada de la distancia de visibilidad de parada más un largo de cola asumido para largos de ciclo cortos de 60 a 75 segundos. Tabla adaptada a metro utilizando factor de convención de 1 metro = 3.281 pies.

Esta metodología presenta una ventaja sobre las metodologías tradicionales para la recopilación y análisis de datos de simuladores de conducción. En la metodología tradicional, Se determinan a priori los segmentos de la carretera en los cuales se va a tomar la información. Luego, se utilizan detectores virtuales para identificar los tramos de tal manera que se identifiquen los datos correspondientes. Finalmente, los datos identificados se extraen de la base de datos global para ser estudiados en detalle utilizando un programa comercial como Excel®. Es necesario hacer esto para que los datos obtenidos puedan ser manejados por el programa.

Cuando se aplica la metodología de análisis de datos presentada en esta ponencia, no es necesario definir a priori los segmentos en los cuales se va a tomar la información. Se

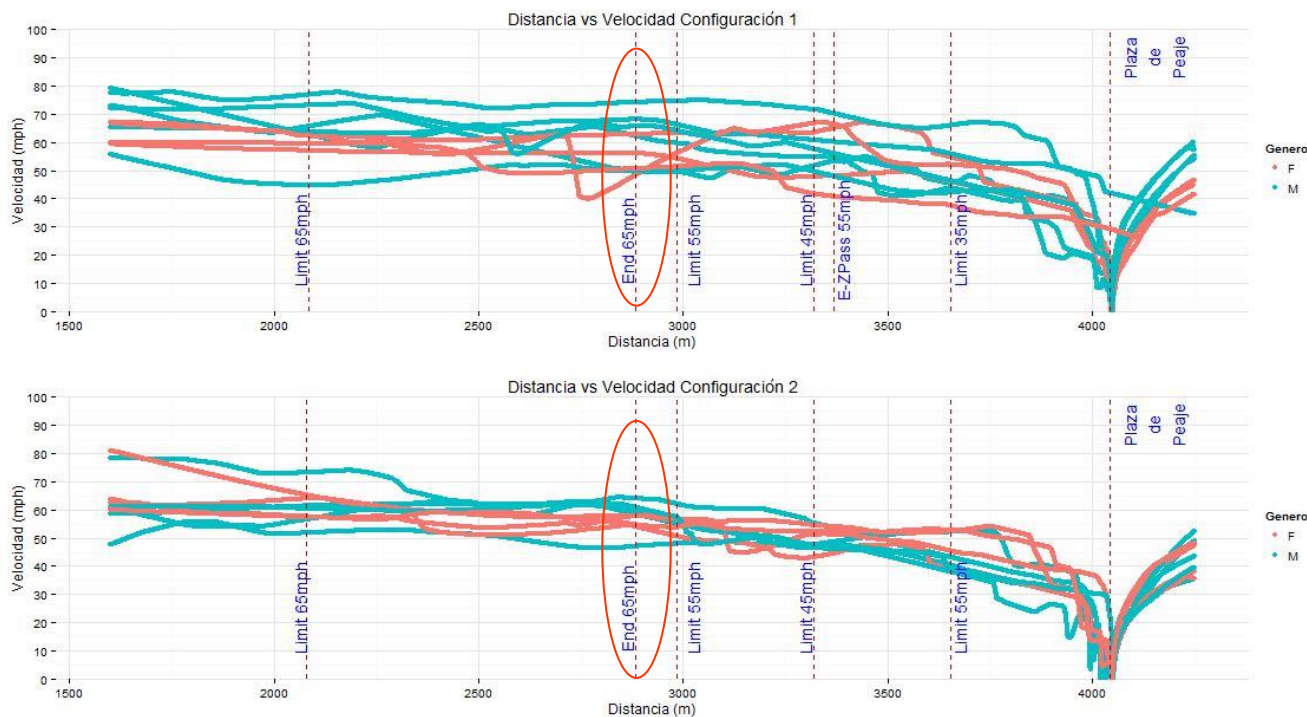


Fig. 3. Gráfica de velocidad con respecto a la distancia.

analizan todos los datos. Los segmentos a analizar resultan de visualizar en forma gráfica de todos los datos que se producen en el simulador de tal manera que los segmentos en los cuales se deben tomar los datos se definen a posteriori. Mediante el análisis de las variables consideradas en el estudio, en el proceso de análisis gráfico es posible investigar en forma detallada las diferencias que se observan y visualizar los segmentos donde los conductores realizan cambios que pueden tener un impacto en la seguridad vial. De esta manera se aprovecha toda la información disponible y los resultados permiten explorar el comportamiento de los conductores a todo lo largo de los experimentos antes de definir los segmentos específicos que son de interés para llevar a cabo los análisis de seguridad vial.

Al determinar las áreas de interés, al rango de datos que se encuentra en esa área se le asigna un número de zona, el cual crea en el archivo .CSV una nueva variable llamada zona. Al completar este proceso se obtiene la base de datos completa para realizar las pruebas estadísticas deseadas.

### C. Análisis estadístico

El análisis estadístico de la variable velocidad, consistía de realizar un modelo lineal mixto con múltiples variables, el cual toma en consideración la zona, la configuración y los sujetos como un efecto aleatorio. Para realizar las pruebas de hipótesis utilizamos una prueba ANOVA a la cual se le aplica una corrección Tukey la cual nos permite eliminar la posibilidad de que ocurra el “Family Wise Error Rate” (FWER, por sus siglas en inglés). El FWER es la probabilidad de al menos obtener un falso positivo al realizarse múltiples pruebas de hipótesis, lo que en estadísticas se conoce como Error de Tipo I. Esta prueba nos provee información para determinar si hubo una diferencia significativa entre las configuraciones estudiadas. Los resultados de este análisis se muestran en la Tabla 2. Todos los procesos son ejecutados en el programa R® y luego se exportan los resultados a tablas y gráficas en las cuales se pueden presentar los resultados del análisis.

Tabla 2. Comparación entre Configuración 1 y Configuración 2 para la variable de Velocidad.

Es ^	Zona 1		Zona 2		Zona 3		Plaza de Peaje	
	Configuración		Configuración		Configuración		Configuración	
	1	2	1	2	1	2	1	2
1	62.13	59.98	54.08	55.63	52.16	51.78	51.83	47.33
2	61.05	59.84	53.15	55.50	50.92	52.43	51.57	48.98
3	61.53	58.33	56.18	53.97	54.44	51.93	52.09	50.12
4	59.30	59.00	53.77	53.63	53.44	51.74	51.10	49.28
5	63.36	59.46	54.01	52.43	45.61	40.63	21.08	18.46
6	61.17	59.52	51.30	52.01	41.73	42.69	20.03	20.50
7	56.48	57.08	51.17	49.90	45.61	40.70	18.99	19.79
8	57.46	54.71	54.10*	49.73*	47.70	43.44	22.45*	16.20*
9	60.46	61.71	55.34	54.42	52.19	51.36	49.79	49.26
10	61.66	57.52	55.38	52.82	51.64	51.56	48.04	49.35
11	61.34	57.22	54.98	52.86	53.68	52.58	51.99	52.36
12	60.47	58.69	56.75	54.36	55.16	50.95	51.72	48.83

^Escenario  
\* Valor-P<0.05

## V. CONCLUSIONES

Esta ponencia presenta una metodología desarrollada y utilizada para recolectar, manejar y analizar grandes cantidades de datos generadas por un simulador de conducción. Se presenta como ejemplo un caso de estudio donde un grupo de sujetos experimenta con dos escenarios de señalización en plazas de peaje en una autopista en Puerto Rico. La metodología desarrollada utilizando el programa R® permite:

- Recolectar y manejar cantidades ilimitadas de datos para la creación de bases de datos.
- Generar gráficas que permiten visualizar los segmentos específicos en los que se realizan análisis estadísticos más detallados.
- Analizar el comportamiento de los conductores utilizando todos los datos provenientes de la simulación.
- Aplicar modelos y pruebas estadísticas detalladas para interpretar los datos provenientes de la simulación.

Mediante el análisis de estos resultados, se llega a conclusiones específicas para mejorar la seguridad vial en el sistema bajo estudio indicando cuál de los escenarios estudiados presenta mejores indicadores relacionados con la seguridad vial. Los métodos existentes presentan la limitación de tener que seleccionar los segmentos de análisis a priori. La metodología presentada elimina esta limitación mediante el estudio de todos los datos utilizando las técnicas de “Big Data” presentadas en esta ponencia.

## VI. RECOMENDACIONES

Se recomienda que para los procesos de análisis de “Big Data” generados por simuladores de conducción se continúe automatizando la creación y análisis de los datos generados por los programas. A corto plazo se recomienda crear una interface genérica de tal manera que se definan las variables de interés y se puedan observar las gráficas iniciales para introducir los segmentos en los cuales se desea realizar los análisis estadísticos detallados. El propósito de la

automatización es mejorar la efectividad del programa reduciendo el tiempo de computación y mejorar la interface para los usuarios.

#### AGRADECIMIENTOS

Los autores reconocen la aportación económica de “Research Innovative Transportation Administration” (RITA) para financiar los estudios del Centro Universitario de Investigación en Transporte (UTC, por sus siglas en inglés) y de la Universidad de Puerto Rico, Recinto de Mayagüez (UPRM). En adición se quiere reconocer la aportación del Dr. Raúl Macchiavelli y el Dr. Pedro Torres, Catedráticos de la Universidad de Puerto Rico Recinto de Mayagüez y a la estudiante graduada del Departamento de Matemáticas Hilda Calderón en el análisis estadístico.

#### REFERENCIAS

- [1] C. Snijders, U. Matzat, and U. Reips. “Big Data: Big Gaps of Knowledge in the Field of Internet Science,” *International Journal of Internet Science*, 7(1), pp. 1-5, 2012.
- [2] G. Bello-Orgaz, J. Jung and D. Camacho, “Social big data: Recent achievements and new challenges,” *Information Fusion*, 28 (2016) 45-59.
- [3] I. Hashem, I. Yaqooba, N. Anuara, S. Mokhtara, A. Gania and S. Khanb, “The rise of big data on cloud computing: review and open research issues,” *Information Systems*, 47 (2015), pp. 98-115.
- [4] D. Valdés, B. Colucci, D. Fisher, E. Colón, R. García and J. Rivera, “Driving Simulation in the Safety and Operation Performance of the Freeway Toll Plaza,” *Advances in Transportation Studies*, 2016.
- [5] D. Valdés, B. Colucci, D. Fisher, E. Colón, R. García and J. Rodríguez, “Operational and Safety-Based Analysis of Toll Plaza Signage using Driving Simulation,” *Transportation Research Board*, 2016.
- [6] M. Burt, M. Cuddy and M. Razo “Big Data’s Implications for Transportation Operations: An Exploration,” U.S. Department of Transportation, December 2014.
- [7] A. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,” *International Journal of Information Management*, 35 (2015), pp. 137-144.
- [8] C. Emani, N. Cullot and C. Nicolle, “Understandable Big Data: A survey,” *Computer Science Review*, 17 (2015), pp. 70-81.
- [9] F. Batarseh and E. Latif, “Assessing the Quality of Service Using Big Data Analytics With Application to Healthcare,” *Big Data Research* (2015), <http://dx.doi.org/10.1016/j.bdr.2015.10.001>.
- [10] A. Vij and K. Shankari, “When is big data enough? Implications of using GPS-based surveys for travel demand analysis,” *Transportation Research Part C*, 56 (2015), pp. 446-462.
- [11] A. Ayed, M. Halima, and A. Alimi, “Big Data Analytics for Logistics and Transportation,” 4<sup>th</sup> IEEE International Conference on Advanced Logistics and Transport, 2015.
- [12] C. Wang, X. Li, X. Zhou, A. Wang and N. Nedjah, “Soft computing in big data intelligent transportation systems,” *Applied Soft Computing*, 38 (2016), pp. 1099-1108.
- [13] S. Buckley and D. Lightman, “Ready or Not, Big Data is Coming to a City (Transportation Agency) Near You,” Presentation in Transportation Research Board 94<sup>th</sup> Annual Meeting Compendium of Papers, Transportation Research Board, (2015), Washington D.C.
- [14] Global Status Report on Road Safety 2015, Geneva. World Health Organization, 2015.
- [15] A. Shaufler, “Toll Plaza Design,” National Cooperative Highway Research Program, Synthesis of Highway Practice 240, 1997.
- [16] American Association of State Highway and Transportation Officials, “A Policy on Geometric Design of Highways and Streets,” AASHTO, 2011.
- [17] Manual on Uniform Traffic Control Devices, Washington, D.C., U.S. Dept. of Transportation, Federal Highway Administration, 2009.
- [18] G. Watson, Y. Papelis, and O. Ahmad, “Design of Simulator Scenarios to Study Effectiveness of Electronic Stability Control Systems,” *Transportation Research Record*, (2006), Vol. 1980, pp. 79-86.
- [19] K. Fitzpatrick, S. Chrysler, A. Nelson and V. Iragavarapu, “Driving Simulator Study of Signing for Complex Interchanges,” In TRB 92<sup>nd</sup> Annual Meeting Compendium of Papers. Transportation Research Board, (2013), Washington, D.C.
- [20] M. Jeihani, B. Spell, and A. Ardeshiri, “Speed Pattern Analysis in the Proximity of Dynamic Message Signs Using a Driving Simulator,” In Transportation Research Board 93<sup>rd</sup> Annual Meeting Compendium of Papers. Transportation Research Board, (2014), Washington D.C.