Metodología para explorar datos abiertos de accidentalidad vial usando Ciencia de Datos: Caso Medellín

Methodology to explore open data of road crashes using Data Science: Case Medellín

Jorge Pérez-Rave[1]* Juan Carlos Correa Morales[2] Favián González Echavarría[3]

Recibido 19 de febrero de 2018, aceptado 20 de agosto de 2018
Received: February 19, 2018 Accepted: August 20, 2018

RESUMEN

La accidentalidad vial es un serio problema de salud pública en el mundo. El estudio de datos abiertos sobre este tema puede estimular decisiones más oportunas e informadas. El objetivo es proponer una metodología para estudiar datos abiertos sobre accidentalidad vial (caso Medellín) usando Ciencia de Datos, considerando desde la planificación del estudio hasta la visualización web. La metodología consta de cuatro macroprocesos: 1. Planificación, 2. Preparación de datos, 3. Análisis automático y 4. Visualización de datos (aplicación web). Estos constan de una o más etapas, desagregadas en 15 subetapas con alcances univariado, bivariado y multivariado. Los macroprocesos 2-4 fueron automatizados en lenguaje R. Como resultado, el analista puede familiarizarse con el tema (descriptivos), explorar relaciones entre variables, localizar sucesos, inducir patrones de agrupación e identificar algunos factores asociados con los eventos de accidentalidad. Todos estos, combinando variables para una segmentación más detallada. El caso de estudio también tiene valor para otros ámbitos, ya que la accidentalidad vial genera mayores efectos en países en desarrollo, lo cual está atrayendo el interés de los investigadores.

Palabras clave: Accidentalidad vial, ciencia de datos, visualización web, metodología de análisis, programación en R.

ABSTRACT

Road crashes is a serious public health problem in the world. The study of open data on this subject can stimulate more timely and informed decisions. The objective is to propose a methodology to study open data on road accident (Medellín case) using Data Science, considering from the planning of the study to the web visualization. The methodology consists of four macroprocesses: 1. Planning, 2. Data preparation, 3. Automatic analysis and 4. Data visualization (web application). These consist of one or more stages, disaggregated into 15 sub-stages with univariate, bivariate and multivariate scopes. Macroprocesses 2-4 were automated in R language. As a result, the analyst can become familiar with the topic (descriptive), explore relationships between variables, locate events, induce patterns of grouping and identify some factors associated with the events of accidents. All these, combining variables for a more detailed segmentation. The case study also has value for other areas, since road accidents generate greater effects in developing countries, which is attracting the interest of researchers.

Keywords: Road crashes, data science, web visualization, analysis methodology, R programming.

[1] Grupo de investigación IDINNOV. IDINNOV S.A.S. Medellín, Colombia. E-mail: investigacion@idinnov.com
[2] Escuela de Estadística. Universidad Nacional de Colombia. Medellín, Colombia. E-mail: jccorrea@unal.edu.co
[3] Departamento de Ingeniería Industrial. Universidad de Antioquia. Medellín, Colombia. E-mail: favian.gonzalez@udea.edu.co
* Autor de correspondencia: investigacion@idinnov.com

INTRODUCCIÓN

Los accidentes viales representan un serio problema de salud pública. La Organización Mundial de la Salud expresa que, en el mundo, más de 1.25 millones de personas fallece cada año en las vías y, además, los accidentes generan un alto costo para las naciones. De hecho, este tema se encuentra en la agenda 2030 para el desarrollo sostenible, buscando reducir tanto el número de muertes como lesionados a causa de accidentes viales a nivel mundial [1].

Hay varias formas de hacerle frente a dicha problemática, entre ellas los programas de educación y respeto por las normas de tránsito y la disposición de la tecnología para el mejoramiento preventivo. Esto último comprende la mejora de los vehículos, pero también, el aprovechamiento de los abundantes conjuntos de datos que diariamente producen los ecosistemas de transporte, a fin de estimular decisiones más oportunas e informadas.

La alternativa basada en datos viene en auge, a causa de la era de ciudades digitales, Big Data, etc., que provee cada vez más datos abiertos disponibles, y Colombia no es la excepción, tal como lo promueve el Ministerio de Tecnologías de la Información y las Comunicaciones (MINTIC), en función del desarrollo y/o la adaptación de soluciones basadas en datos en variedad de sectores económicos [2].

Una de las principales disciplinas para abordar el fenómeno de los datos abiertos es la Ciencia de Datos. Se trata de un área emergente que incorpora conocimientos de estadística, matemática, programación/algoritmia, teoría de visualización y el contexto de aplicación propiamente [3-4], para estudiar todo el ciclo de vida del dato en cada campo de aplicación: planificación, captura, estructuración, limpieza, procesamiento, análisis y visualización. Su objetivo, dependiendo del alcance, restricciones y recursos, es proveer hallazgos de valor, de modo que se contribuya con la generación de conocimiento y con mejores procesos de toma de decisiones. Dichos hallazgos bien pueden ser de naturaleza descriptiva, comparativa, evolutiva, correlacional, predictiva o explicativa, según las teorías de soporte, la comprensión del fenómeno y los datos con los que se cuente.

Esta disciplina naciente aún tiene mucho por aportar y, para ello, las naciones deben superar barreras que limitan sus aplicaciones. [5-6] recalcan que las pocas organizaciones que están aprovechando la Ciencia de Datos están dándole un alcance limitado, por razones de deficiencia de competencias profesionales en el tema y poca calidad de las fuentes de información. [7] advierte sobre el uso desarticulado de métodos o paquetes computacionales sin un propósito global que los una, o sin explotación de tareas complejas y determinantes como la captura y la preparación de datos (limpieza, estructuración, creación de nuevas variables, etc.). Asimismo, señala la dependencia de procedimientos manuales, debido al uso de software basado en botones o de instrucciones de código básicas o no generalizables.

En Colombia es marcada la escasez de contribuciones en Ciencia de Datos. De hecho, Colciencias recientemente se ha manifestado sobre el déficit de profesionales relacionados con TIC, y ha lanzado una "Convocatoria para la formación de ciudadanos en ciencia de datos" [8].

En el marco de la transformación digital del país, la Alcaldía de Medellín viene publicando una serie de datos a través de portales como GeoMedellin y datos.gov.co. Allí puede encontrarse variables de accidentalidad vial, proyecciones de población, Encuesta de Calidad de Vida, entre otras.

A la fecha, diversos estamentos en Colombia ejecutan esfuerzos individuales y en pocos casos se articulan datos de diferentes fuentes para enriquecer las perspectivas de análisis. Además, en la mayoría de los casos los análisis son de naturaleza descriptiva - univariada o de zonificación (mapas), dejando oportunidades para explorar enfoques bivariados y multivariados que complementen las visualizaciones existentes. Por ejemplo, uno de los principales sitios de visualización de datos de accidentalidad de Medellín (https://geomedellin-m-medellin.opendata.arcgis.com/) presenta varias oportunidades de mejora en cuanto a: 1) posibilidad de crear nuevas variables secundarias informativas, a partir de las variables originales. 2) Trascender los conteos de eventos (frecuencias absolutas) o mapas, adicionando reportes de estadísticos de localización, de dispersión y de tendencia central (cuando apliquen), así como de combinación entre frecuencias absolutas y relativas. 3) Adicionar visualizaciones que complementen los diagramas de barras simples, adicionando campos para análisis de

segmentación, etc. 4) No todas las visualizaciones de las variables están en funcionamiento. 5) El alcance de los análisis es en su mayoría univariado, dejando oportunidades para complementar con enfoques bivariados y multivariados.

Dichas oportunidades de mejora no solo se presentan en Colombia, sino que son un desafío declarado en estudios recientes internacionales, los cuales expresan la necesidad de contar con sistemas que trasciendan la mera agregación univariada de estadísticas de accidentalidad [9]. Esto ayudaría a la identificación de hallazgos de mayor valor para la toma de decisiones o para guiar nuevos estudios relacionados con análisis de causas.

Recientemente, [10] desarrolló la aplicación web "Pasa-Segura-Medellín", que toma como insumos datos abiertos de accidentalidad vial en Medellín, abarca los tres alcances estadísticos (univariado, bivariado y multivariado) desde un enfoque de Ciencia de Datos, y ayuda a saldar varias de las oportunidades descritas.

No obstante, a la fecha se carece de un procedimiento documentado en revistas académicas que guíe sobre cómo comenzar a abordar los datos abiertos disponibles sobre accidentalidad vial en el país. Si bien se cuenta con abundantes libros de texto que exponen con detalle métodos, técnicas y herramientas estadísticas o de inteligencia artificial, el mero uso de estos recursos no garantiza per sé la generación de ventajas competitivas para los tomadores de decisiones. Más aún, considerando la teoría de recursos y capacidades [11], vale recalcar la importancia de la "orquestación de los recursos" [12]. En este contexto, ello se traduce en formas lógicas y efectivas de combinar secuencialmente los métodos, los datos abiertos y los demás recursos disponibles, a fin de generar capacidades analíticas competitivas [13].

Así, haciendo una retrospectiva de la vivencia en el diseño, desarrollo y puesta a punto de la aplicación "Pasa-Segura-Medellín", y tomando en cuenta algunas recomendaciones de la literatura, así como el conocimiento y la experiencia de los autores, este trabajo tiene como objetivo proponer una metodología para estudiar datos abiertos sobre accidentalidad vial usando Ciencia de Datos (caso Medellín), abarcando desde la planificación del estudio hasta la visualización bajo un entorno web.

El caso Medellín también tiene valor para otros ámbitos, ya que, a la fecha, estudios como [14] advierten que la accidentalidad vial es un fenómeno que está generando mayores consecuencias negativas en países de ingresos medios y bajos, a causa del acelerado crecimiento poblacional, la urbanización y los problemas de inversión para afrontar el tema. Por consiguiente, este tipo de escenarios está atrayendo la atención de dirigentes, académicos e investigadores y puede ayudar en otros contextos geográficos.

Este trabajo tiene tres implicaciones prácticas: 1) documenta un marco procedimental que otros investigadores pueden considerar para abordar datos abiertos disponibles sobre accidentalidad vial, e incluso para generar nuevas propuestas metodológicas; 2) estimula la generación de nuevos desarrollos en otros campos, a través de la Ciencia de Datos, en sincronía con lo que están demandando organismos como Colciencias y MINTIC; 3) contribuye a una mejor comprensión de la accidentalidad vial en un contexto de ingresos medios-bajos, lo cual está despertando cada vez más interés en los investigadores.

El resto del documento está organizado así: la sección 2 expone la metodología empleada, la sección 3 los resultados (tomados de la aplicación web) acompañados de la discusión. La sección 4 lista las conclusiones generales.

METODOLOGÍA PROPUESTA

La metodología tiene como fin proveer de forma automática información estadística (exploratoria) sobre métricas de accidentalidad vial desde un enfoque de Ciencia de Datos, usando datos abiertos disponibles. Está compuesta por cuatro macroprocesos con un orden lógico: 1. Planificación del estudio, 2. Preparación de datos, 3. Análisis automático y 4. Visualización de datos. Estos constan de una o más etapas, las cuales, a su vez, se desagregan en 15 subetapas en total. Los macroprocesos 2-4 fueron integrados secuencialmente y automatizados en lenguaje R [15], bajo el entorno RStudio [16]. A continuación, se describe cada uno de los macroprocesos.

Macroproceso de planificación del estudio

El objetivo de este macroproceso es definir las preguntas de analítica, considerando las necesidades de los usuarios, del analista, etc., así como tomando en cuenta los datos abiertos disponibles. Asimismo,

precisar los detalles técnico-metodológicos del estudio. A continuación, se exponen las preguntas de analítica resultantes:

P.1 ¿Qué características han presentado los eventos de accidentalidad en Medellín, en términos de: gravedad (día, mes, año), horas, comunas, barrios, fines de semana, quincena, jornada, día de la semana, día calendario del mes, diseño, mes y clase?

P.2 ¿Cuál es el mapa de accidentalidad vial de en Medellín? Y ¿Cuáles son las calles y carreras que describen dicha accidentalidad?

P.3 ¿Cómo ha evolucionado la accidentalidad vial en Medellín considerando los periodos 2014 – 2017_1?

P.4 ¿Cómo se caracterizan las comunas considerando las métricas disponibles de población, inversión y calidad de vida? Y ¿Cómo se relacionan los eventos de accidentalidad con dichas métricas?

P.5 ¿Qué patrones de agrupación subyacen en las comunas de Medellín? Y ¿Qué indicadores latentes los describen?

P.6 ¿Qué posibles factores (latentes u observables) se relacionan de forma significativa con los eventos de accidentalidad en las comunas?

En la Tabla 1 se describen los aspectos técnico – metodológicos del caso de aplicación.

Macroproceso de preparación de datos

En este macroproceso se diseñan y desarrollan algoritmos que ponen a punto los datos objeto de estudio. Esto comprende una sola etapa, con el mismo nombre del macroproceso en cuestión, e incluye dos subetapas, las cuales se describen en la Tabla 2.

En la Figura 1 se esquematiza el procedimiento empleado en la preparación de los datos, mostrando cómo se va pasando de los conjuntos de datos a la extracción de variables originales y, de estas, a la creación de variables secundarias. Unos grupos de variables son empleados para describir los eventos de accidentalidad, en cambio, otros grupos (integración de varios conjuntos de datos) son empleados para explorar relaciones y patrones de agrupación.

Este macroproceso es vital, considerando que la mayoría de los programas de formación en Colombia, sobre Ciencia de Datos o tópicos relacionados, se enfocan básicamente en los métodos de análisis y poco o nada cubren estos procedimientos previos. Sin embargo, cerca del 70% del tiempo de un proyecto

Tabla 1. Ficha de aspectos técnico = metodológicos del caso de estudio.

Criterios	Descripción
Muestras y tamaños	Eventos de accidentalidad vial en Medellín, reportados por GeoMedellin. Periodos: 2014 (41594 obs.), 2015 (42080 obs.), 2016 (42841 obs.), 2017_1 (24352 obs.).
Unidades de análisis	1) Evento de accidentalidad (150867 obs); 2) Comuna (20 obs. previa agregación).
Conjuntos de datos	1. Accidentalidad 2014 – 2017_1 (GeoMedellin) [17]; 2. Inversión por comuna y corregimiento (datos.gov.co, 2016) [18]; 3. Encuesta de calidad de vida (medellin.gov.co, 2016) [19]; 4. Proyecciones de población (datos.gov.co, 2016) [20]. Datos proveídos por Alcaldía de Medellín a dichos medios.
Diseño	Observacional, de tipo descriptivo y relacional
Métodos	Métodos de lectura, limpieza, procesado e integración de datos. Métodos de análisis: descripción univariada, correlaciones Kendall, Análisis Clúster, Análisis de Componentes Principales, Regresión usando Componentes Principales. Análisis de Pareto, Métodos de visualización interactiva de la información.

Tabla 2. Descripción del macroproceso de preparación de datos.

Etapa	Subetapa	Descripción
Preparar datos	1. Lectura, limpieza y procesado	Leer datos, unificar textos, homogenizar formatos, eliminar valores anómalos, creación de variables secundarias a partir de los datos disponibles.
	2. Integración de conjuntos de datos	Cuatro conjuntos de datos (accidentes, comunas, encuesta calidad de vida, e inversión por comunas). Véase Figura 1.

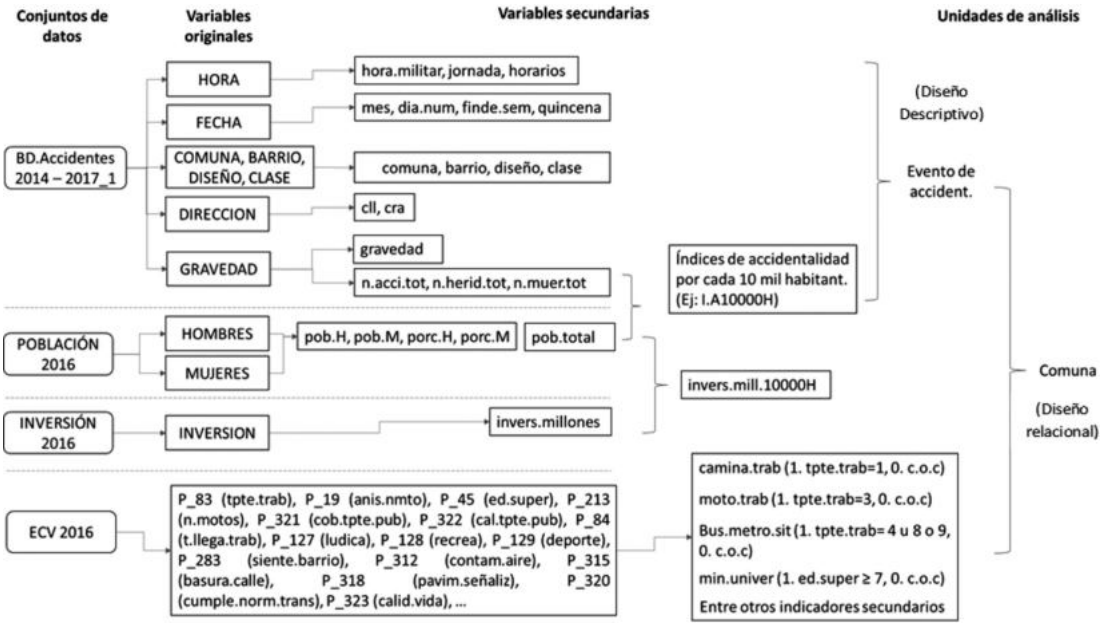

Figura 1. Esquema de tareas de creación de variables secundarias e integración de conjuntos de datos. Fuente: elaboración propia.

en dicha disciplina es consumido por tareas de captura y preparación de los datos, a las cuales no solo se les atribuye ciencia, sino también gran parte de arte que se va dominando con la práctica [21].

Macroproceso de análisis automático

El objetivo de este macroproceso es someter los datos (ya preparados) a una serie de pruebas estadísticas y consta de tres etapas. La primera toma como unidad de análisis el evento de accidentalidad. Su finalidad es retratar el estado del tema en periodos específicos, considerando frecuencias, tendencia central, evolución y localización de eventos (mapas y diagramas de dispersión). En la Tabla 3 se describen las seis subetapas que conforman esta etapa descriptiva.

La segunda etapa toma como unidad de análisis la comuna (20 en total) y su finalidad es describir

Tabla 3. Descripción del macroproceso de análisis automático - etapa descriptiva.

Etapa	Subetapa	Descripción
Análisis descriptivo tomando como unidad de análisis el evento de accidentalidad (periodos: 2014 - 2017_1)	1. Gravedad de accidentes	Medias, totales y frecuencias relativas de eventos de accidentalidad según gravedad (herido, muerto, solo daños) con opciones por día, mes y año. Gráfico de barras verticales.
	2. Segmentación de accidentes	Frecuencias absolutas y relativas, o valores medios, según factores (horas, comuna, barrio, fin de sem, quincena, jornada, dia, sem, dia num, diseño, mes, clase). Gráfico de barras horizontales.
	3. Mapa de accidentalidad	Mapa interactivo de Medellín con eventos de accidentalidad, según gravedad y proporción de casos elegidos.
	4. Calles y carreras de accidentalidad	Diagrama de dispersión (X: Calle, Y: Carrera), con visualizaciones según comuna, gravedad y zoom.
	5. Evolución gráfica de la accidentalidad	Gráfico de serie interactivo para eventos de accidentalidad al mes para cada periodo (2014 - 2017_1).
	6. Evolución tabular de la accidentalidad	Resumen estadístico de la evolución de la accidentalidad (Mín, Máx, Med, DE, Median, Q1, Q3, n. datos) para cada año.

relaciones y patrones de agrupación, a través de variables provenientes de los diversos conjuntos de datos (en este caso, 4 conjuntos). En esta se abarcan los tres alcances estadísticos: univariado, bivariado y multivariado. La tercera etapa parte de las exploraciones realizadas en las primeras dos etapas y provee resultados de posibles factores relacionados con los eventos de accidentalidad. Así, la primera subetapa toma como unidad de análisis la comuna (integrando datos de 2016) y la segunda el evento de accidentalidad (2014 – 2017_1). En la Tabla 4 se describen las subetapas que componen esta etapa relacional.

Macroproceso de visualización de datos

Este macroproceso se encarga de transmitir al usuario los resultados de los análisis previos, de modo que este pueda encontrar en ellos significados e interpretaciones, convirtiéndolos en información de utilidad para la toma de decisiones. Para ello, se recurre a tecnologías de visualización de datos en forma de aplicación web, usando Shiny [22]. Este macroproceso consta de una etapa y tres subetapas, expuestas en la Tabla 5.

La validación de la aplicación web se va haciendo con la ejecución de cada función y algoritmo programado. No obstante, la prueba final de la consistencia se realizó por medio de una comparativa con visualizaciones de algún medio oficial disponible. Dicho medio debe proveer reportes para al menos una de las variables observadas. En este caso, se cuenta con visualizaciones de la API de GeoMedellín, que provee resultados descriptivos para cuatro de los

Tabla 4. Descripción del macroproceso de análisis automático - etapas relacionales.

Etapa	Subetapa	Descripción
Análisis de relaciones y patrones de agrupación integrando conjuntos de datos (periodo: 2016)	1. Métricas sobre comunas	Descripción univariada en formato de diagrama de barras horizontales, tipo ranking, para indicadores de la base de datos integrada (accidentes, población, inversión, ECV).
	2. Relaciones bivariadas	Diagrama de dispersión con histograma y coeficiente de correlación Kendall para pares de variables, valor-p para explorar la significancia, y con opción de con y sin comuna “La Candelaria”.
	3. Patrones de agrupación de comunas	Dendrograma derivado de análisis clúster para comunas, con opciones de: variables de accidentalidad (solo eventos de accidentalidad) o tipos de población (métricas de caracterización sociodemografica, etc., de comunas), método de distancia (Euclideana, Máxima, Manhattan, Canberra) y tipo de enlace (Promedio, Vecino lejano, Ward).
	4. Indicadores latentes de tipos de población	Dos nuevos indicadores latentes derivados de Análisis de Componentes Principales, considerando métricas sociodemográficas, etc., de comunas. Gráfico de Componentes Principales por individuos (comunas) y Boxplot para los Scores de las componentes 1 y 2.
Posibles factores asociados a eventos de accidentalidad	1. Factores latentes (integrando conjuntos de datos de 2016)	Modelos de regresión múltiple usando como predictores los Scores de las dos componentes principales de la subetapa anterior, con opciones de tipo de gravedad. Resumen de las regresiones (completo, sin La Candelaria, sin Palmitas, y sin ambos) y gráfico de normalidad para los residuales de cada modelo.
	2. Jornada y horarios (periodos: 2014 – 2017_1)	Análisis de segmentación bajo gráficos tipo Pareto, desagregando por clase de evento (Atropello, Choque, Otros), jornada (AM, PM) y horarios.

Tabla 5. Descripción del macroproceso de visualización web.

Etapa	Subetapa	Descripción
Aplicación web	1. Diseño	Visualizaciones considerando las preguntas de analítica y las etapas y subetapas descritas del análisis automático, donde además se definieron los tipos de gráficos. Uso de Shiny bajo entorno tipo dashboard.
	2. Desarrollo	Funciones y los algoritmos que conectan los análisis con la visualización automática.
	3. Validación	Comparativa con visualizaciones de referencia. En este caso: visualizaciones disponibles en GeoMedellín para el conjunto de datos “Accidentalidad 2016” [19] para Medellín.

campos del conjunto de datos "Accidentalidad 2016" (Alcaldía de Medellín). En el anexo 1 se muestra evidencias de consistencia entre las visualizaciones.

RESULTADOS Y DISCUSIÓN

La visualización de los resultados de la metodología propuesta puede encontrarse en https://idinnov.shinyapps.io/pasa-segura-medellin/

A continuación, se presentan los tipos de visualizaciones para cada subetapa ya descrita, tomando como referencia las preguntas de analítica previamente definidas.

-Características de los eventos de accidentalidad en Medellín:
En la Figura 2 se ilustra el formato de visualización desarrollado, que permite dar respuestas a la pregunta en consideración.

Nótese que el usuario tiene la posibilidad de interactuar con las visualizaciones, indicando el periodo de observación (2014-2017_1), así como eligiendo reporte por día, mes o año (gráfico de barras verticales, izquierdo). En el caso de la gráfica de barras de la derecha, puede elegir la variable de segmentación (Ej: horas, comuna, barrio, día de la semana,…) y el tipo de gravedad del evento (herido, muerto, solo daños o todos). A modo de ejemplo, puede verse que durante el periodo 2016 los eventos de accidentalidad en Medellín reportaron, en promedio al mes, 1999.5 casos de heridos (56%), 1551 casos de solo daños (43.4%) y 19.6 casos de muertes (0.6%). También a modo ilustrativo, puede decirse que, en cuanto a las horas de mayor reporte de accidentalidad, se encuentran entre 7 am - 7:59 am, 11 am - 11:59am, 12 pm - 12:59 pm y 5 pm - 5:59 pm. Es decir, se puede ver la influencia de las horas pico (de inicio de la jornada laboral/académica, de pausa para almorzar y de terminación de la jornada laboral/académica), como es de esperar. Vale recordar que el usuario puede explorar otras 10 variables de segmentación, diferente a la hora del evento del día y elegir un tipo específico de evento (herido, muerto, solo daños).

Mapa de accidentalidad vial de en Medellín y calles - carreras que describen dicha accidentalidad:
En la Figura 3 se presentan dos gráficos. El de la izquierda provee al usuario un mapa interactivo en el que puede reconocer las zonas de la ciudad donde fue reportada la ubicación de los accidentes viales. El usuario puede elegir el reporte para cada uno de los tres tipos de gravedad (herido, muerte y solo daños), los cuales tienen diferente color asignado (Ej: puntos rojos para casos de muerte). Además, se le permite al usuario especificar la proporción de casos que desea visualizar, de modo que se obtenga una visualización más amigable ante abundantes datos para determinados eventos que dificulten la interpretación. Para ello se genera de forma automática una muestra aleatoria del tamaño especificado, a partir del total de la población reportada. Este mapa también ofrece al usuario la posibilidad de aumentar o reducir el zoom, en

Fuente: Pasa-Segura-Medellín [10].

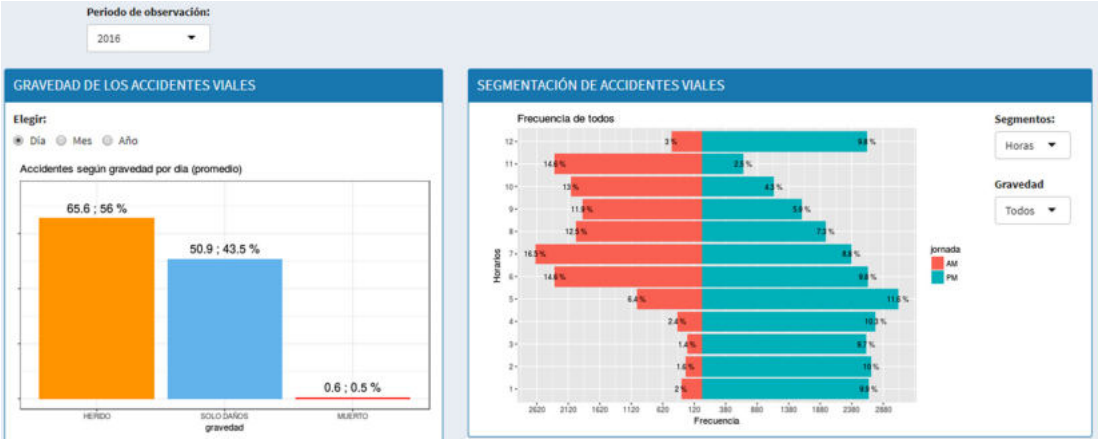

Figura 2. Descripción de gravedad (izq.) y otros atributos (der.).

busca de más o menos detalles de la zona, según se desee. La gráfica de la derecha (Figura 3) se deriva de un procesamiento y análisis especializado de los datos, en el que fue necesario usar métodos de procesamiento de texto, pues la variable de partida fue la dirección en que se produjo el evento.

Esta variable es tipo carácter, siendo necesario el uso de procedimientos de homogenización, limpieza y extracción adicionales. Así, fue posible crear dos variables secundarias (véase esquema de la figura 1), una de ellas representa la calle en que se produjo el evento y la otra la carrera. En la gráfica de dispersión (derecha, Figura 3), la calle se retrata en el eje de la abscisa y la carrera en el de la ordenada. (nota: se trata de un enfoque general, se omiten letras de las direcciones. Ej: Cll 52 A, se toma como parte de la calle 52). Nótese lo asertiva que es esta visualización para hacerse a una idea de las zonas de mayor concentración de eventos de accidentalidad, recurriendo a uno de los gráficos más famosos de la estadística, pero esta vez no para retratar relaciones entre variables, sino para otro uso: describir zonas geográficas de mayor concentración. Esto es aún más útil, en la medida en que el usuario puede especificar la comuna (entre las 20 objeto de estudio), el tipo de gravedad e incluso aumentar o disminuir el zoom. A modo de ilustración, vale decir que en la comuna "La Candelaria" los eventos de accidentalidad con reporte de solo heridos en el periodo de 2016, tienden a concentrarse alrededor de la calle 45 con la carrera 50.

Evolución de la accidentalidad vial en Medellín entre 2014 – 2017_1:

En la Figura 4 se presenta la visualización para describir la evolución de la accidentalidad vial mes a mes para diversos años. Esta consta de dos partes. La gráfica de serie (izquierda, Figura 4) y la representación tabular (derecha, Figura 4).

En la gráfica de serie (izquierda, Figura 4) el usuario puede comparar cómo se ha venido comportando los eventos de accidentalidad mes a mes. Es decir, puede establecer comparaciones (visuales - exploratorias) entre meses (Ej: enero tiende a presentar menores eventos de accidentalidad con reporte de heridos, en comparación con los demás meses). Pero también, permite realizar comparaciones (visuales - exploratorias) entre los años. A su vez, la visualización de la derecha (Figura 4) ofrece un reporte tabular donde el usuario encuentra ocho estadísticos descriptivos de los eventos de accidentalidad.

Ambas gráficas se actualizan según la combinación comuna - gravedad, que el usuario elija, y también según se active (o no) la casilla de "Cada 10 mil habitantes". Dicha opción está habilitada solo para la categoría de comuna "Todas" y resulta útil para una comparación más razonable entre los diferentes periodos, pues es de esperar que el denominador poblacional cambie con los años.

Caracterización de comunas considerando métricas disponibles de población, inversión y

Fuente: Pasa-Segura-Medellín [10].

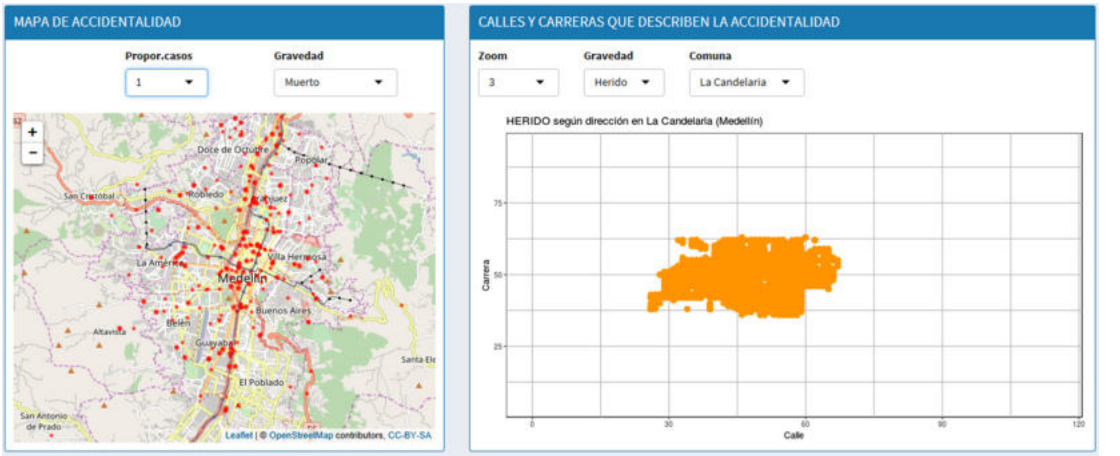

Figura 3. Mapa de accidentalidad (izq.) y diagrama calle-cra (der.).

Fuente: Pasa-Segura-Medellín [10].

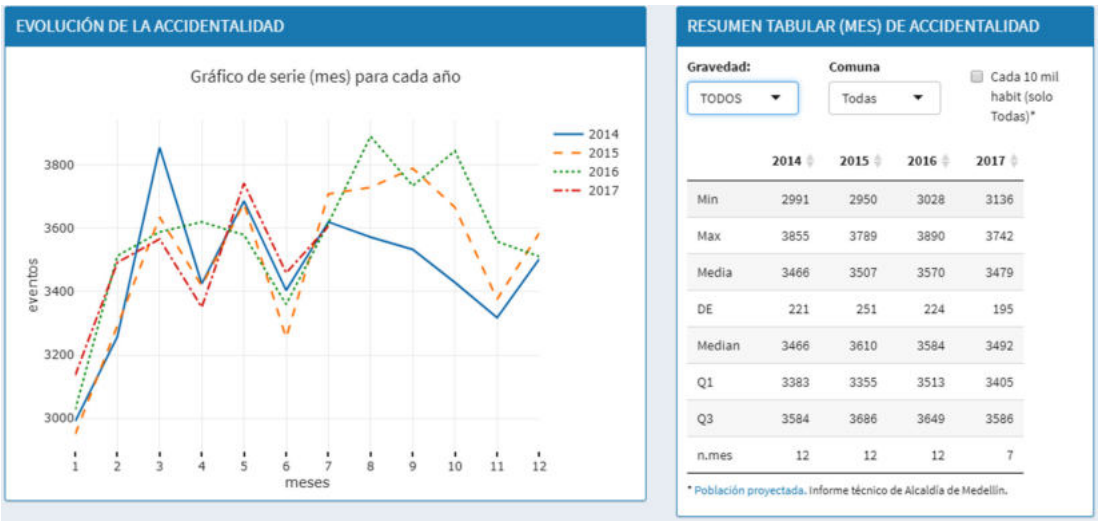

Figura 4. Gráfico de serie múltiple (izq.) y resumen tabular (der.).

calidad de vida, y relaciones entre estas y los eventos de accidentalidad:
A partir de este momento, mediante procedimientos de agregación de datos tomando esta vez como unidad de análisis a la comuna (20 observaciones en total), se integran variables de cuatro conjuntos de datos, lo cual permite, entre otros, explorar relaciones entre las variables. En la Figura 5 se presenta la visualización para la pregunta de analítica de caracterización de comunas y relaciones bivariadas.

En la Figura 5, gráfica izquierda, el usuario obtiene la caracterización de las diferentes comunas. Así, según la variable que se elija, se provee un diagrama de barras horizontales, ordenadas de mayor a menor valor reflejando una jerarquía y variando los colores para mejor discriminación. Cada que se elige una variable el aplicativo reporta la descripción narrativa de la variable y su escala. En la gráfica derecha (Figura 5) se muestran cuatro elementos para pares de variables elegidos: histogramas, dispersión, coeficiente de correlación Kendall (dada su pertinencia para pocos datos y/o desviaciones de la normalidad), así como el valor-p alusivo para explorar si la relación resulta o no significativa, siendo costumbre usar una significancia de 0.05.

Fuente: Pasa-Segura-Medellín [10].

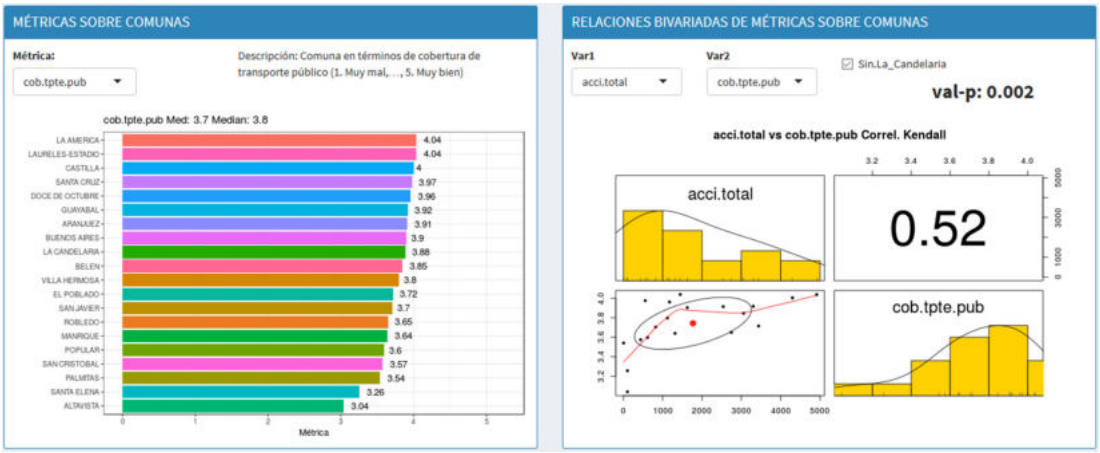

Figura 5. Barras (izq.) y diagrama de dispersión enriquecido (der.).

Así, un valor-p inferior a esta cifra deduce una relación significativa entre las variables. En esta visualización (derecha, Figura 5), el usuario puede explorar los resultados con y sin una comuna que muestra un comportamiento notablemente más elevado en términos de los eventos de accidentalidad (La Candelaria). A modo de ilustración, nótese que entre las variables número total de accidentes por comuna y percepción de la cobertura de transporte público por comuna se muestra una relación positiva (0.52) y significativa (valor-p: 0.002). Cuando se incluye La Candelaria, esta relación sigue siendo significativa pero con un menor coeficiente de correlación Kendall (0.48; valor-p: 0.003).

Patrones de agrupación subyacentes en las comunas de Medellín e indicadores latentes que los describen:

En la Figura 6 se muestran las visualizaciones de pasar al mundo multivariado, considerando grupos de variables a la vez. Estas permiten encontrar hallazgos relacionados con patrones de agrupación entre comunas, así como indicadores latentes.

En la gráfica izquierda (Figura 6) se provee el dendrograma resultante del método no supervisado "Análisis Clúster". Allí puede verse diversos grupos dependiendo del método de cálculo de las distancias (euclideana,…) y del tipo de enlace (vecino lejano, promedio…). Nótese, a modo de ilustración, que en la Figura 6, La Candelaria se muestra separada de las demás comunas. Es decir, en un nivel superior se podría hablar de con y sin La Candelaria. A medida que se desciende en el dendrograma, se van detallando mejor otros grupos. Por ejemplo, Aranjuez, Guayabal, El poblado, Robledo, Belén y La América se muestran en el mismo grupo (usando distancias Euclidianas y enlace vecino más lejano). El usuario puede explorar nuevos patrones de agrupación, efectuando el análisis no por accidentalidad, sino para indicadores sociodemográficos, etc., de la población. En la visualización derecha (Figura 6) se resumen los resultados de efectuar Análisis de Componentes Principales. En concreto, se muestran las cargas de las variables en cada factor, las cuales, al estar estandarizadas, representan la correlación entre cada variable y el factor. También se provee las dos variables latentes (Dim1 y Dim2), que justo corresponden a las componentes principales 1 y 2. Cada componente principal induce un tipo de población. Por ejemplo, los scores de la primera componente (Dim1) están centrados en cero, representan un continuo bipolar y aumentan (se hacen más positivos) en la medida en que aumentan los valores de las siguientes variables: percepción de cobertura de transporte público, edad, proporción de hogares con al menos una persona con formación mínima universitaria. De igual forma, los scores de dicha componente (Dim1) disminuyen a medida que las variables descritas se hacen menores, así como si las siguientes variables aumentan su valor: proporción de hogares que tienden a usar bus, o metro o sistema integrado de transporte (es decir, no taxi ni vehículo propio, etc.), proporción

Fuente: Pasa-Segura-Medellín [10].

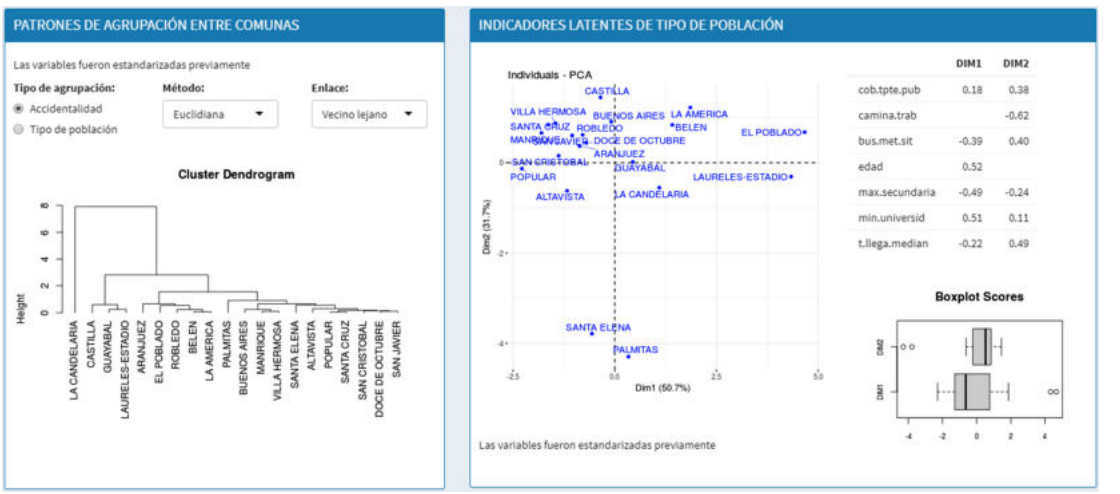

Figura 6. Dendrograma (izq.) y resultados de ACP (der.).

de hogares con miembro de máximo formación secundaria, y tiempo (mediana) en llegar al sitio de trabajo. En otras palabras, ambos indicadores latentes (Dim1 y Dim2) puede inducir dos tipos de población según condiciones socioeconómicas, donde a mayores scores, las condiciones se muestran más facilitadoras que limitadoras, y viceversa. Tomando esto en consideración y al combinar de forma gráfica dichas componentes principales, es posible ver las distancias entre las comunas (Figura 6, derecha). Nótese que Palmitas y Santa Elena se muestran más cercanas en comparación con los demás. Lo mismo entre el par Poblado y Laureles - Estadio.

Posibles factores relacionados con los eventos de accidentalidad:

Factores latentes usando regresión:
Considerando los indicadores latentes inducidos para el caso anterior (sección 3.5, Figura 6), en la Figura 7 se ofrecen visualizaciones que resumen posibles relaciones funcionales entre estos (predictores) y variables de accidentalidad (respuesta). Pueden verse cuatro modelos de regresión, (1) con todas las comunas, (2) sin La Candelaria, (3) sin Palmitas y (4) sin ambas comunas. El análisis de sensibilidad variando tales comunas se hace por sus distancias pragmáticas y analíticas con respecto a los comportamientos de las demás.

Nótese, en la Figura 7, que con solo esas dos componentes principales (Dim1 y Dim2) el porcentaje de explicación de los modelos (R^2) osciló entre 53% y 76.1% para el logaritmo del número total de eventos de accidentalidad. Estos porcentajes también tendieron a ser razonables cuando se varía el tipo de gravedad (heridos, muerte,…). Del mismo modo, en todos los cuatro modelos dichos predictores (Dim1 y Dim2) resultaron estadísticamente significativos al 0.05 (o menos en varios casos), lo cual deduce la robustez de los mismos. Al explorar los supuestos clásicos de la regresión, no hay razones para invalidar el supuesto de normalidad, lo cual puede verse en la gráfica QQnorm expuesta en la parte derecha de la Figura 3. Tampoco se encontró patrones en los gráficos de los residuales, que llevarán a invalidar la homocedasticidad.

Las relaciones positivas - significativas entre las dos componentes principales (indicadores latentes Dim1 y Dim2) y las métricas de accidentalidad, pueden quizás deberse a menor probabilidad de desplazamientos a pie o en transportes públicos (más vehículos particulares), personas con domicilios o lugares de trabajo ubicados en sitios de alta afluencia vehicular, cargos de mayor responsabilidad intelectual que tal vez incrementan el afán (por llegar a reuniones, etc.) o el estrés (no se logran desconectar de compromisos laborales o académicos durante los desplazamientos), entre otros.

Jornada y horarios bajo enfoque Pareto:
En la Figura 8 se muestra la segmentación de eventos de accidentalidad para la combinación de

Fuente: Pasa-Segura-Medellín [10].

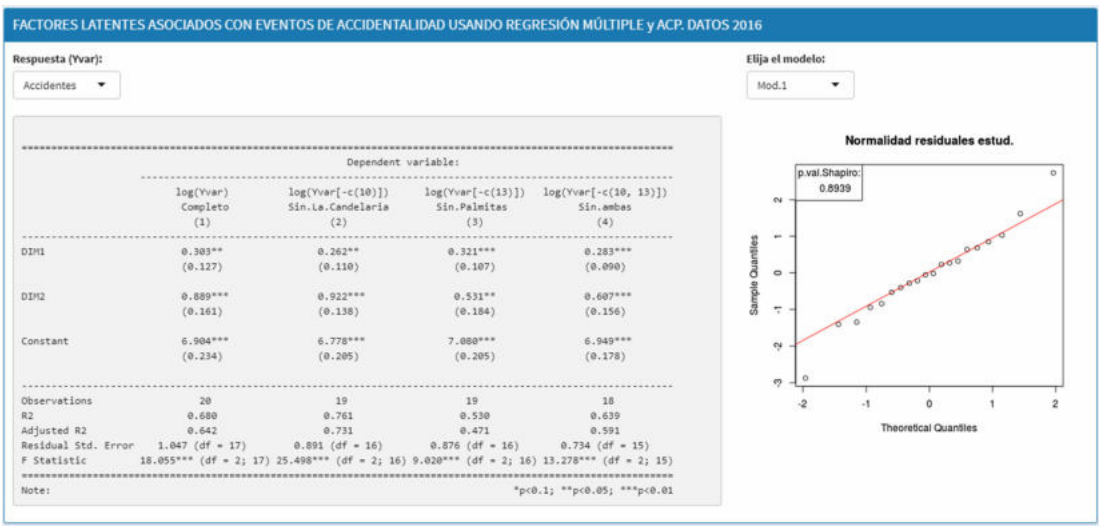

Figura 7. Resumen de regresiones (izq.) y normalidad (der.).

Fuente: Pasa-Segura-Medellín [10].

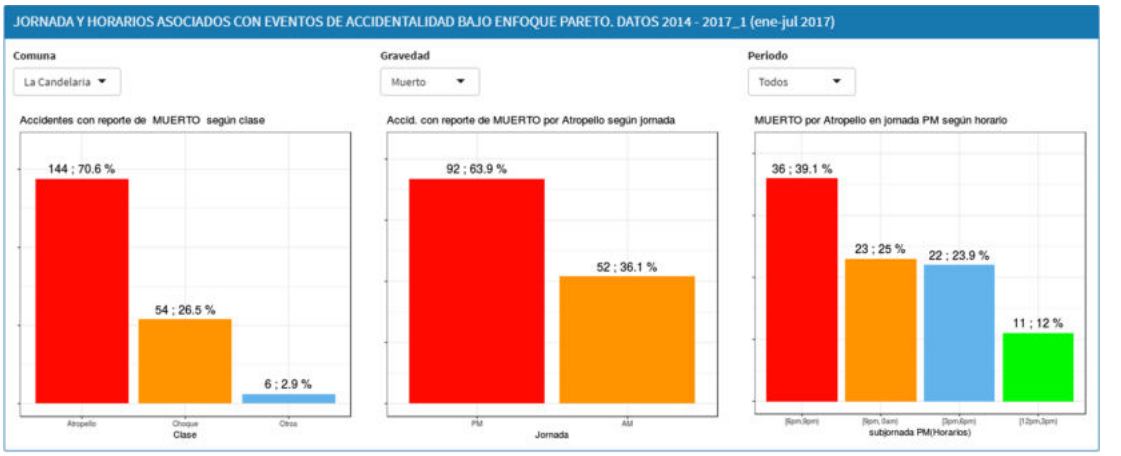

Figura 8. Desglose de eventos bajo enfoque Pareto.

comuna, gravedad, periodo. Primero con base en la clase de evento (atropello, choque, otros). Luego, la categoría de mayor frecuencia se desagrega en jornada y, finalmente, la categoría más frecuente se desagrega en franjas horarias.

Ejemplificando con los eventos de accidentalidad que reportan muerte en la comuna "La Candelaria", y consolidando datos de todos los periodos (2014 – 2017_1), se encuentra que el 70.6% de los casos se debieron a "atropello" (144 casos), seguido de "choque", con una cifra notablemente menor (26.5%; 54 casos). Al considerar solo los 144 casos de eventos que fueron por "atropello", el 63.9% reportaron ocurrencia en la jornada de la tarde (PM; 92 casos). A su vez, de estos 92 casos en la jornada de la tarde, el 39.1% ocurrió entre las 6pm y las 9pm. Nótese, con estas nuevas exploraciones, la posibilidad de buscar diferenciar entre lo "poco –vital" y lo "mucho– trivial". Vale destacar la gran cantidad de combinaciones de comuna –gravedad– periodo que el usuario puede explorar.

CONCLUSIONES

Se ha propuesto una metodología para estudiar (a nivel exploratorio) datos abiertos disponibles sobre accidentalidad vial desde un enfoque de Ciencia de Datos. Esta mirada implica no solo el tratamiento estadístico, sino también la conexión sistemática (bajo un orden lógico) y automática de cada una de las tareas, desde el alistamiento de los datos hasta la visualización de los resultados. Esta metodología consta de cuatro macroprocesos y 15 subetapas en las que se abarca los tres alcances estadísticos (univariado, bivariado y multivariado). Además, aprovecha el potencial de la integración de conjuntos de datos, lo cual aumenta la diversidad de variables y favorece una mejor comprensión del fenómeno y su entorno.

La metodología aporta, en su primer macroproceso, seis preguntas de analítica que guían el resto de los macroprocesos, y ameritan enfoques descriptivos y relacionales para saldarlas. Estas pueden ser empleadas por otros investigadores para aportar nuevas formas de hacerles frente, o para reproducir el estudio, o para crear nuevas preguntas a partir de estas, según las particularidades contextuales, la disponibilidad de datos y los intereses.

La metodología se pone a prueba usando cuatro conjuntos de datos abiertos sobre la ciudad de Medellín. En primera instancia toma como unidad de análisis el evento de accidentalidad y, en segunda instancia, la comuna. De cara al usuario, se cuenta con una aplicación web (macroproceso último de la metodología) usando tecnologías emergentes, en la que de forma automática e interactiva se facilita la transferencia de información oportuna y la exploración de hallazgos por parte del usuario. Los lectores pueden explorar muchos otros hallazgos y creencias directamente en el enlace https://idinnov.shinyapps.io/pasa-segura-medellin/

La importancia de esta metodología se debe a que cada vez el fenómeno del Big Data y de la Ciencia de

Datos y sus múltiples vertientes están impregnando esferas económicas, sociales, ambientales, políticas, etc., y se están comenzando a proliferar cursos y educación formal o continua sobre métodos, técnicas, herramientas, etc. Sin embargo, el uso de estos no garantiza por sí solo la generación de información de valor para la toma de decisiones, sino que depende fundamentalmente de cómo los diferentes recursos y capacidades se articulan de forma lógica y con propósitos específicos (orquestación de recursos [12]). Justamente, esta metodología aporta esos elementos, a nivel exploratorio, usando datos abiertos de accidentalidad vial (caso Medellín), y se prueba en un contexto emergente que demanda nuevas políticas y planes de acción para su prevención. Se espera, entonces, motivar la reflexión y discusión entre los actores organizacionales y gubernamentales, así como en la ciudadanía, sobre los resultados del caso de estudio. Pero también, motivar nuevos desarrollos en Ciencia de Datos, no solo para contextos de accidentalidad vial, sino también para las muchas otras esferas donde esta nueva disciplina tiene mucho por aportar. Esto último también se enmarca en macrometas estatales y educativas con respecto a ciudades debidamente informadas, que aprovechen al máximo la era digital.

A nivel de implicaciones para la toma de decisiones, las visualizaciones arrojadas por la metodología, a través de https://idinnov.shinyapps.io/pasa-segura-medellin/ posibilitan la familiarización descriptiva con el tema, la exploración de relaciones entre variables, la localización concreta de sucesos, la inducción de patrones de agrupación y la identificación de algunos factores latentes y observables que presentan asocio con los eventos de accidentalidad. Todo ello, brindando posibilidad al usuario de realizar combinaciones de campos (Ej: Comuna - gravedad - periodo) para una segmentación más detallada, tal como lo sugieren [14], dada la amplia heterogeneidad que presentan los datos de accidentalidad. En cuanto a los portales de visualización de datos, de la ciudad de Medellín y el país en general, este trabajo brinda insumos para nutrir dichos medios de visualización, mediante la incorporación de análisis bivariados y multivariados que es viable ejecutar con los datos abiertos disponibles. También, con el uso, lectura y discusión de los informes arrojados por la aplicación web que materializa los resultados visuales de la metodología, es viable encontrar insumos importantes para la toma de decisiones o para profundizar en estudios de análisis de causas.

REFERENCIAS

[1] Organización Mundial de la Salud - OMS. “Lesiones causadas por el tránsito”. Centro de Prensa. Mayo de 2017. Leído el 04/01/2018 en: http://www.who.int/mediacentre/factsheets/fs358/es/

[2] Ministerio de Tecnologías de la Información y las Comunicaciones - MINTIC. Datos abiertos para Ciudades Inteligentes. Cada vez es más evidente cómo la revolución digital transforma el día a día de los ciudadanos. Sala de Prensa, Columnas Ministro TIC. 28 de junio de 2016. Leído el 04/01/2018, en: http://www.mintic.gov.co/portal/604/w3-article-15599.html

[3] F. Pop. “High Performance Numerical Computing for High Energy Physics: A New Challenge for Big Data Science. Advances in High Energy Physics”. pp. 1-13. 2014.

[4] J. Davis, J. Burgoon, L. Can. “Data Science Inform Environmental Justice and Community Risk Screening for Type 2 Diabetes”. Plos one. April 14. pp. 1-14. 2015.

[5] T. Schoenherr, Cheri Speier. “Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential”. Journal of Business Logistics. Vol. 36 Nº 1, pp. 120-132. 2015.

[6] B. Hazen, C. Boone, J. Ezell and L. Jones-Farmer. “Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications”. International Journal of Production Economics. Vol. 154, pp. 72-80. 2014.

[7] J. I. Pérez-Rave. “Statihouse®: desarrollo tecnológico basado en ciencia de datos para explorar estadísticamente el sector inmobiliario”. Ingeniare. Revista chilena de ingeniería. Vol. 27, Nº 1. pp. 113-130. 2019.

[8] Departamento Administrativo de Ciencia. Tecnología e Innovación - Colciencias (2017). Términos de referencia, versión consulta. “Convocatoria para la formación de ciudadanos en ciencia de datos”. Código M301PR01F02, versión 12. Leído el 04//01/2018, en: http://

www.colciencias.gov.co/convocatorias/innovacion/convocatoria-para-la-formacion-ciudadanos-en-ciencia-datos

[9] C. Costa, G. Chatzimilioudis, D. Zeinalipour-Yazti and M. Mokbel. 2017. "Towards real-time road traffic analytics using Telco Big Data. In 11th International Workshop on Real-Time Business Intelligence and Analytics, BIRTE 2017". Association for Computing Machinery. Leído el 08/01/2018, en: https://experts.umn.edu/en/publications/towards-real-time-road-traffic-analytics-using-telco-big-data

[10] J. Pérez-Rave. Pasa-Segura-Medellín. Vers.3, 2018-1: Aplicación web sobre accidentalidad vial en Medellín usando Ciencia de Datos. Grupo de investigación IDINNOV, IDINNOV S.A.S. 2017. https://idinnov.shinyapps.io/pasa-segura-medellin/

[11] J. Barney. "Firm resources and sustained competitive advantage". Journal of Management. Vol. 17, pp. 99-120. 1991.

[12] D. Teece, G. Pisano & Shuen and A. "Dynamic. capabilities and strategic management". Strategic Management Journal. pp. 509-533. 1997.

[13] A. Popovič, R. Hackney, R. Tassabehji and MCastelli. "The impact of big data analytics on firms' high value business performance". Information Systems Frontiers. pp. 1-14. 2016.

[14] M. Raihan, M. Hossain and T. Hasan. "Data mining in road crash analysis: the context of developing countries". International journal of injury control and safety promotion. pp. 1-12. 2017.

[15] R Project. "R: A language and environment for statistical computing. R Foundation for Statistical Computing". Austria. 2008.

[16] R Studio Team. "RStudio: Integrated Development for R. RStudio". Inc., Boston. USA. 2015.

[17] Alcaldía de Medellín. GeoMedellín. Movilidad. Datos de accidentalidad 2014-2017 (ene-jul 2017). Leído el 08 de enero de 2018. en: https://geomedellin-m-medellin.opendata.arcgis.com

[18] Alcaldía de Medellín. Inversión por comuna y corregimiento 2016. Leído el 28 de diciembre de 2018, en: https://www.datos.gov.co

[19] Alcaldía de Medellín. Encuesta de Calidad de Vida 2016. Leído el 28 de diciembre de 2018, en: https://www.medellin.gov.co

[20] Alcaldía de Medellín. Proyecciones de población 2014-2017. Leído el 28 de diciembre de 2018, en: https://www.datos.gov.co

[21] H. Varian "Big data: New tricks for econometrics". The Journal of Economic Perspectives. Vol. 28 Nº 2, pp. 3-27, 2014.

[22] W. Chang, J. Cheng, J. Allaire, Y. Xie and J. shiny McPherson. "Web Application Framework for R. R package version 0.12.2". 2015.

Anexo 1. Evidencia de consistencia entre algunas visualizaciones de Pasa-Segura-Medellín y la API GeoMedellín para datos "Accidentalidad 2016" de Medellín.

* https://geomedellin-m-medellin.opendata.arcgis.com/datasets/accidentalidad-2016
** https://idinnov.shinyapps.io/pasa-segura-medellin/