



Datos masivos geoespaciales para identificación de patrones de riesgo en la RNC

Juan Carlos Vázquez Paulino
Elsa María Morales Bautista

Publicación Técnica No. 540
Sanfandila, Qro, 2019

SECRETARÍA DE COMUNICACIONES Y TRANSPORTES
INSTITUTO MEXICANO DEL TRANSPORTE

**Datos masivos geoespaciales para identificación de
patrones de riesgo en la RNC**

Publicación Técnica No. 540
Sanfandila, Qro, 2019

Esta investigación fue realizada en la Unidad de Sistemas de Información Geoespacial del Instituto Mexicano del Transporte, por el Lic. Juan Carlos Vázquez Paulino y la Ing. Elsa María Morales Bautista; es el producto final del proyecto de investigación interna OI-02/17 “Datos masivos geoespaciales para la identificación de patrones de riesgos reportados por usuarios en la RFC” y forma parte del portafolio de proyectos de la Coordinación de Economía de los Transportes y Desarrollo Regional.

Se agradece la colaboración del M. G. Miguel Ángel Backhoff Pohls, Jefe de la Unidad de Sistemas de Información Geoespacial del IMT, por sus acertados comentarios y apoyo para la realización de esta publicación.

Asimismo, se agradecen los comentarios del Dr. Guillermo Torres Vargas, Jefe de la División de Estudios Económicos y Sociales del Transporte.

Contenido

Índice de figuras		iv
Sinopsis		ix
Abstract		xi
Resumen	Ejecutivo	xiii
Capítulo 1.	Introducción	1
Capítulo 2.	Marco teórico	3
Capítulo 3.	Plataforma Waze	27
Conclusiones		89
Bibliografía		91

Índice de figuras

Figura 2.1	Escala del tamaño de los datos	5
Figura 2.2	Procesos de los datos masivos	14
Figura 3.1	Pantalla de mapa en vivo de Waze.	28
Figura 3.2	Pantalla de mapa en vivo de Waze, acercamiento a la Ciudad de México	29
Figura 3.3	Atributos de las alertas emitidas por la plataforma Waze	37
Figura 3.4	Tipos de alerta y subtipos relacionados a Waze	38
Figura 3.5	Tipos de carretera definidos por la plataforma Waze	38
Figura 3.6	Captura de pantalla del <i>feed</i> en formato Json provenientes de Json	41
Figura 3.7	Captura de pantalla del feed en formato XML	44
Figura 3.8	Tabla resumen de las lecturas generadas por la plataforma Waze, para Mayo 2017	46
Figura 3.9	Gráfica de la Tabla resumen de las lecturas generadas por la plataforma Waze, para Mayo 2017	47
Figura 3.10	Tabla resumen de las lecturas generadas por la plataforma Waze, para Junio 2017	48
Figura 3.11	Gráfica de la Tabla resumen de las lecturas generadas por la plataforma Waze, para Junio 2017	49
Figura 3.12	Tabla resumen de las lecturas generadas por la	50

	plataforma Waze, para Julio 2017	
Figura 3.13	Gráfica de la Tabla resumen de las lecturas generadas por la plataforma Waze, para Julio 2017	51
Figura 3.14	Código fuente para sumar los archivos Excel en un solo archivo por día	53
Figura 3.15	Campos resultantes de la unificación de tablas	54
Figura 3.16	Eventos del mes de mayo 2017	56
Figura 3.17	Eventos que representan un riesgo, mayo 2017	57
Figura 3.18	Eventos que representan un riesgo, por horas, mayo 2017	58
Figura 3.19	Eventos que representan un riesgo, acumulado por día, mayo 2017	59
Figura 3.20	Densidad de puntos que representan riesgo, mayo 2017	60
Figura 3.21	Conjunto de mapas que muestran la cantidad de puntos que representan riesgo representados por hora/día, mayo 2017.	62
Figura 3.22	Accidentes menores del mes de mayo 2017	63
Figura 3.23	Accidentes mayores del mes de mayo 2017	64
Figura 3.24	Eventos del mes de junio 2017	65
Figura 3.25	Eventos que representan un riesgo, junio 2017	66
Figura 3.26	Eventos que representan un riesgo, por horas, junio 2017	67
Figura 3.27	Eventos que representan un riesgo, acumulados por día, junio 2017	68

Figura 3.28	Densidad de puntos que representan riesgo, junio 2017	69
Figura 3.29	Conjunto de mapas que muestran la cantidad de puntos que representan riesgo representados por hora/día, junio 2017	71
Figura 3.30	Accidentes mayores del mes de junio 2017	72
Figura 3.31	Accidentes menores del mes de junio 2017	73
Figura 3.32	Eventos del mes de julio 2017	74
Figura 3.33	Eventos que representan un riesgo, julio 2017	75
Figura 3.34	Eventos que representan un riesgo, acumulados por día, julio 2017	76
Figura 3.35	Eventos que representan un riesgo, por horas, julio 2017	77
Figura 3.36	Densidad de puntos que representan riesgo, julio 2017	78
Figura 3.37	Conjunto de mapas que muestran la cantidad de puntos que representan riesgo representados por hora/día, julio 2017	80
Figura 3.38	Accidentes mayores del mes de julio 2017	81
Figura 3.39	Accidentes menores del mes de julio 2017	82
Figura 3.40	Densidad de eventos de riesgo, mayo 2017, acercamiento a la región Querétaro – Ciudad de México	83
Figura 3.41	Densidad de eventos de riesgo, junio 2017, acercamiento a la región Querétaro – Ciudad de México	84
Figura 3.42	Densidad de eventos de riesgo, julio 2017, acercamiento a la región Querétaro – Ciudad de México	85
Figura 3.43	Densidad de eventos de riesgo, mayo 2017,	86

acercamiento a la región Guadalajara

Figura 3.44	Densidad de eventos de riesgo, junio 2017, acercamiento a la región Guadalajara	87
-------------	--	----

Figura 3.45	Densidad de eventos de riesgo, julio 2017, acercamiento a la región Guadalajara	88
-------------	--	----

Sinopsis

En este trabajo, se define lo que se entiende por datos masivos. Se presentan diversas definiciones destacando el hecho de que el tamaño es sólo una dimensión de los datos masivos, otras dimensiones, como la velocidad, la variedad, la veracidad, la viabilidad, la visualización y el valor son igualmente importantes. Se revisan las técnicas de análisis para datos de texto, audio, vídeo y medios sociales.

Se revisan los datos masivos geoespaciales (*geospatial big data*) y se identifican algunas fuentes de datos, también se analiza el modo en que se relacionan con el transporte y su posible aplicación en distintas áreas.

Se identifican las fases principales del proceso de los datos masivos geoespaciales, así como las principales técnicas de representación de éstos y los distintos tipos de algoritmos para procesamiento de datos.

Como parte de los objetivos del proyecto, se realizaron ejercicios con el uso de datos de Waze, los cuales se descargaron durante determinado lapso y al procesarlos y generar mapas digitales se pueden identificar patrones de ocurrencia de eventos capturados por la comunidad.

Abstract

In this paper, we first define what is meant by big data. There are several definitions highlighting the fact that size is only one dimension of big data, other dimensions such as speed, variety, accuracy, viability, visualization and value are equally important. Analysis techniques for text, audio, video and social media data are reviewed.

We review the geospatial data (geospatial big data) and identify some data sources. We also analyze the way they relate to transport and its possible application in different areas.

The main phases of the process of the massive geospatial data are identified, as well as the main techniques of representation of these and the different types of algorithms for data processing.

As part of the objectives of the project, exercises were carried out with the use of Waze data, which were downloaded during a certain period and when processing them and generating digital maps, patterns of occurrence of events captured by the community can be identified.

Resumen ejecutivo

Esta investigación tuvo como objetivo conocer y utilizar una vertiente de los datos masivos geoespaciales para aplicarlos y relacionarlos con un aspecto del transporte en México, en este caso, la red federal de carreteras del país. Los datos utilizados son los que se generan a través de un servicio de generación de datos de tráfico e identificación de diversos eventos a través del *crowdsourcing* (colaboración participativa) de la plataforma Waze.

Se hace una breve revisión del concepto de datos masivos geoespaciales, así como de sus principales aspectos como conocimiento fundamental para entender los datos y la información generada a partir de ellos.

Se revisaron diversas maneras de procesar los datos de Waze y se exploraron métodos de análisis geoespacial para identificar patrones espaciales y temporales de ocurrencia y, de este modo conocer, por ejemplo, a qué hora se generan las mayores concentraciones de eventos relacionados a riesgo en el área de estudio; en qué lugares se dan estas concentraciones de eventos y cómo se relacionan con la red federal de carreteras.

El resultado de estos análisis puede ser utilizado por diversos usuarios de la red federal de carreteras para identificar patrones espacio-temporales de tráfico y comportamiento y generar políticas de traslado, distribución, elaboración de rutas óptimas, uso de horarios escalonados y diferenciados en sus rutas de transporte y distribución.

Conocer el panorama de lo que está sucediendo sobre el terreno le da oportunidad a la analítica geoespacial en el sector de transporte para aprovechar las herramientas de datos y análisis predictivo y de este modo ayudar a la Secretaría de Comunicaciones y Transportes, así como a instituciones públicas y privadas relacionadas al transporte a mejorar las operaciones, reducir los costos y servir mejor a los viajeros.

1 Introducción

El Big Data emerge como una nueva etapa del manejo intensivo de información y a su vez como un nuevo paradigma para la generación de conocimiento. La viabilidad de este proceso está sujeta a los cambios tecnológicos que se dan de manera muy rápida en las comunicaciones de modo que la tendencia es promover la digitalización y la convergencia tecnológica de las distintas fuentes de datos. De este modo, a nivel de usuario se traduce en el acceso remoto y en tiempo real a la información para tomar decisiones; a nivel de planeación permite la identificación y descubrimiento de patrones (espaciales y temporales), en el comportamiento de los conjuntos de datos.

A partir de este análisis es posible diseñar políticas efectivas orientadas a sectores del transporte y logística, áreas de planeación, construcción y mantenimiento.

En este sentido, se ha estado trabajando con los datos generados a través de Waze, a partir de un convenio de colaboración que permite disponer de la información que el universo de usuarios de la plataforma en México levanta y comparte a través de la aplicación.

Para esto se han descargado los datos correspondientes a tres meses del año 2017 (mayo, junio y julio)

Este conjunto de datos se debe usar de manera efectiva y para esto se requiere algo diferente de las estadísticas tradicionales, donde solo se hacen gráficas y análisis sin representación geoespacial. Ahora se hace necesario utilizar análisis de tendencias espaciales, contigüidad, continuidad, distancia y tiempo, concentración de eventos, pronósticos de ubicación y comportamiento, entre otras.

Tal como se menciona en la publicación técnica 502, el *crowdsourcing* - (del inglés *crowd* –multitud– y *outsourcing* –recursos externos–) se podría traducir al español como colaboración abierta distribuida o externalización abierta de tareas y consiste en trasladar tareas que, tradicionalmente, realizaban empleados o contratistas, dejándolas a cargo de un grupo numeroso de personas o una comunidad a través de una convocatoria abierta.

Jeff Howe¹, uno de los primeros autores en emplear el término, estableció que el concepto de "crowdsourcing" depende esencialmente del hecho de que, debido a

¹ <http://www.penguinrandomhouse.com/books/83579/crowdsourcing-by-jeff-howe/9780307396211/>

que es una convocatoria abierta a un grupo indeterminado de personas, reúne a los más aptos para ejercer las tareas, para responder ante problemas complejos y para así contribuir aportando las ideas más frescas y relevantes. Por ejemplo, se podría invitar al público a desarrollar una nueva tecnología o a llevar a cabo una tarea de diseño (diseño basado en la comunidad o diseño participativo distribuido), a mejorar e implementar los pasos de un algoritmo o ayudar a capturar, sistematizar y analizar grandes cantidades de datos (ciencia ciudadana).

El término se ha hecho popular entre las empresas y autores como forma abreviada de la tendencia a impulsar la colaboración en masa, posibilitada por las nuevas tecnologías como la Web 2.0 o las redes sociales para así lograr objetivos de negocios o eventualmente propuestas sociales. Sin embargo, el término y sus modelos de negocio también ha generado controversia y crítica.²

Ventajas

- Compilación de una gran variedad de propuestas de alta calidad
- Disminución de costos
- Retroalimentación interna y permanente
- Generación continúa de ideas innovadoras

Desventajas

- Puede generarse mucha basura, por lo que se deben revisar los datos
- Procesos lentos

²<http://www.empresasenred.es/empresasenred/blog/%C2%BFqu%C3%A9-es-crowdsourcing-y-ventajas-tiene>

2 Marco teórico

2.1 Definición de los datos masivos (Big Data)

De acuerdo con la publicación técnica 502 “Datos masivos geoespaciales para el transporte”, los datos masivos son dos palabras de moda en internet; por lo que se debe buscar su aplicación y los posibles beneficios que podrían generar en las áreas del conocimiento humano.

El creciente flujo de datos, que proviene de los diferentes tipos de sensores, sistemas de mensajería y redes sociales, además de sistemas de medición y observación más tradicionales, ya ha invadido muchos aspectos de la existencia cotidiana. Por un lado, los datos masivos (incluidos los geoespaciales), tienen un gran potencial para beneficiar y ampliar el conocimiento de muchas áreas de estudio como el cambio climático, la vigilancia de enfermedades, la respuesta a desastres, el transporte, el monitoreo de infraestructuras críticas, la logística, la accidentabilidad, etc. Por otra parte, los beneficios de los datos masivos para la sociedad suelen estar limitados por cuestiones necesarias como la privacidad de los datos, la confidencialidad y la seguridad, así como por la dificultad de obtener los conjuntos de datos y el costo asociado a su obtención.

Junto con sus definiciones, los datos masivos se describen a menudo por sus características únicas. Laney (2001) propuso tres dimensiones que caracterizan los desafíos y oportunidades de aprovechar los datos masivos: Volumen, Velocidad y Variedad (3Vs)³. A la par de las 3Vs, la característica Veracidad se ha añadido para describir la integridad y calidad de los datos. También se han sugerido Vs adicionales tales como variabilidad, validez, volatilidad, visibilidad, valor o visualización y, debido a que no necesariamente expresan cualidades de magnitud, solo se mencionan en ciertas ocasiones y bajo ciertas condiciones.

Gartner Inc. define el Big data como “el conjunto de información con alto volumen, alta velocidad y alta variedad que demanda innovadoras y efectivas formas de procesar la información para aumentar su valor para la toma de decisiones”⁴.

³ <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>

⁴ <http://www.gartner.com/it-glossary/big-data>

2.1.1 Volumen

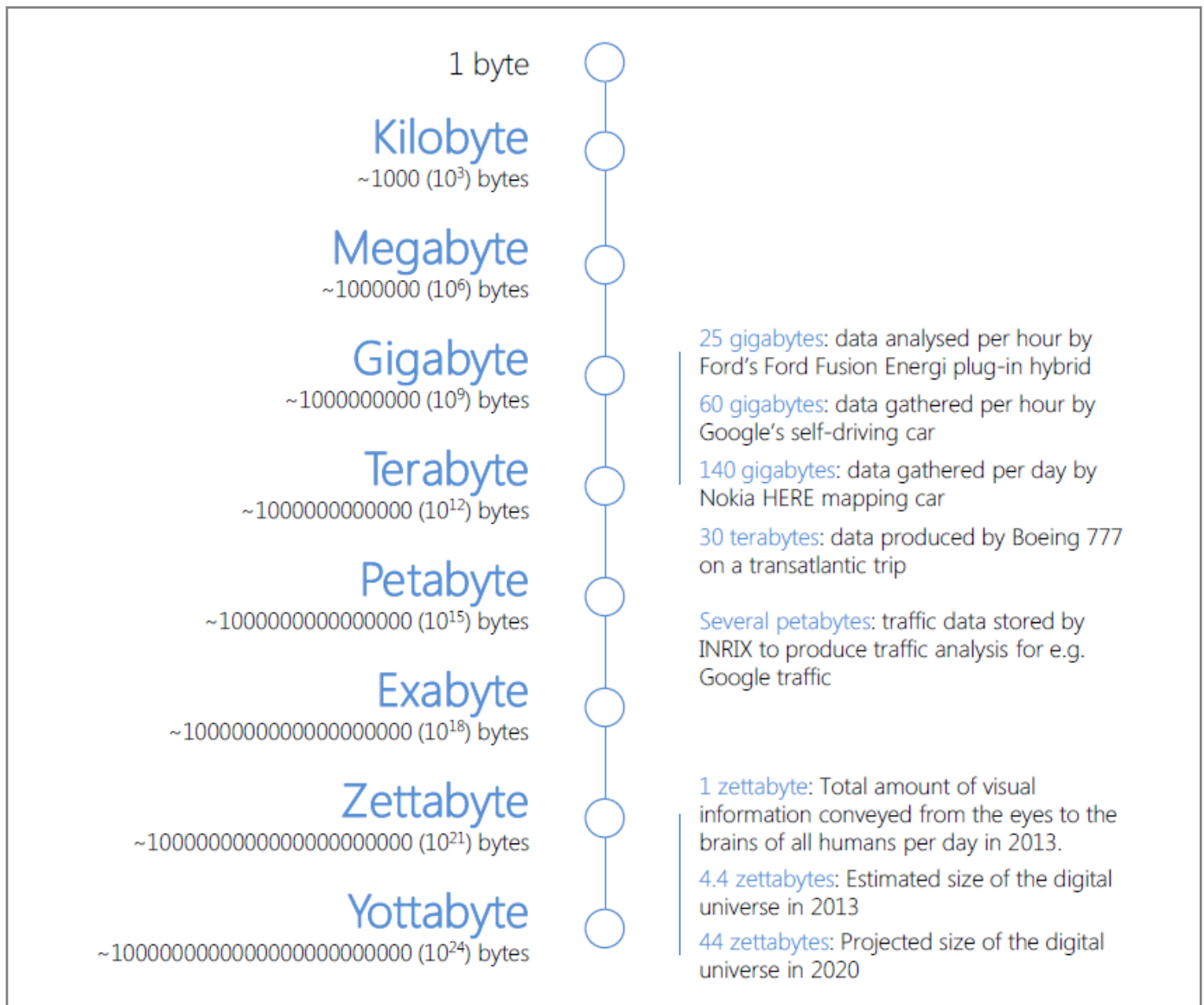
El volumen se refiere a la magnitud de los datos. El tamaño de los datos masivos se reporta en múltiples terabytes o petabytes. Muchos usuarios consideran que los datos masivos deben ser arriba de un terabyte (un terabyte equivale a 1500 discos compactos o 220 dvds. Un petabyte equivale a 1024 terabytes).

De acuerdo con Gandomi (2014), dos conjuntos de datos del mismo tamaño podrían requerir distintas tecnologías de proceso si se toma en cuenta su tipo, por ejemplo, datos tabulares vs datos de video, por lo que las consideraciones acerca del volumen son imprácticas dependiendo de la fuente de los datos.

El reporte *Big Data and Transport*⁵, se refiere a una estimación del año 2013 que dimensiona el tamaño total del "universo digital" (entendiendo por contenido digital las fotografías, películas, videos de vigilancia, datos producidos por sensores y dispositivos asociados, contenido de internet, correo electrónico, mensajes cortos de celular, metadatos de llamadas telefónicas y archivos de audio) en 4.4 zettabytes⁶ y se proyecta que el tamaño de este universo crecerá hasta 44 zettabytes (ver *Figura 2.1 Escala del tamaño de los datos*) en 2020. Estas estimaciones representan una cantidad asombrosa de datos y una porción significativa se refiere a eventos y personas (tarjetas de crédito y transacciones de pago, videos de vigilancia, salidas de sensores de vehículos, señales de acceso Wi-Fi, texto colaborativo e imágenes en redes sociales).

⁵ Big Data and Transport. International Transport Forum. OECD. 2015

⁶ Ibidem.



Fuente: Nokia HERE, Forbes, Idealab, GE, ITF

Figura 2.1 Escala del tamaño de los datos.

2.1.2 Velocidad

La velocidad se refiere a la rapidez con que los datos son generados, almacenados y procesados y de esta manera satisfacer la demanda de datos en tiempo real (Dasgupta, 2017). La proliferación de dispositivos digitales tales como los teléfonos “inteligentes” y otros sensores permiten ahora una rapidez nunca antes vista en la creación de datos y así se acrecienta la necesidad de análisis en tiempo real y planeación basada en evidencias. Los datos generados en estos dispositivos a través de diversas aplicaciones producen grandes cantidades de información que puede ser usada para satisfacer alguna necesidad de los usuarios en tiempo real. Estos datos proveen información acerca de consumidores, como la posición geoespacial, aspectos demográficos y patrones de consumo.

Los sistemas tradicionales de administración y manejo de datos no son capaces de almacenar, procesar y analizar grandes volúmenes de manera instantánea.

2.1.3 Variedad

La variedad se refiere a que los datos pueden provenir de fuentes heterogéneas o diversas. Los avances tecnológicos permiten la generación de datos estructurados, semiestructurados y sin estructura. Los datos estructurados, según Cukier (s/f), representan el 5% de los datos existentes y se refieren a los datos tabulares de hojas de cálculo y bases de datos relacionales. El texto, imágenes, audio y videos son ejemplos de datos sin estructura. Los que no tienen estructura y los semiestructurados no se ajustan a estándares estrictos. El formato XML (*eXtensible Markup Language* por sus siglas en inglés)⁷ es un ejemplo de datos semiestructurados. Esta gran cantidad de datos, al provenir de distintas fuentes y con diferentes temporalidades se pueden insertar en alguno de los siguientes tipos de estructuras⁸:

- Datos estructurados (*Structured Data*): Datos que tienen bien definidos su longitud y su formato, como las fechas, los números o las cadenas de caracteres. Se almacenan en tablas. Un ejemplo son las bases de datos relacionales y las hojas de cálculo.

⁷ XML, siglas en inglés de *eXtensible Markup Language*, traducido como "Lenguaje de Marcado Extensible" o "Lenguaje de Marcas Extensible", es un meta-lenguaje que permite definir lenguajes de marcas desarrollado por el World Wide Web Consortium (W3C) utilizado para almacenar datos en forma legible. Proviene del lenguaje SGML y permite definir la gramática de lenguajes específicos (de la misma manera que HTML es a su vez un lenguaje definido por SGML) para estructurar documentos grandes. A diferencia de otros lenguajes, XML da soporte a bases de datos, siendo útil cuando varias aplicaciones deben comunicarse entre sí o integrar información. <http://definicion.de/xml/>

⁸ <http://consultec-ti.com/blog/big-data-analiza-toda-informacion-que-no-puede-ser-procesada-con-herramientas-tradicionales/>

- Datos no estructurados (*Unstructured Data*): Datos en el formato tal y como fueron recolectados, carecen de un formato específico. No se pueden almacenar dentro de una tabla ya que no se puede separar su información a tipos básicos de datos. Algunos ejemplos son los PDF, documentos multimedia, e-mails o documentos de texto.
- Datos semiestructurados (*Semistructured Data*): Datos que no se limitan a campos determinados, pero que contiene marcadores para separar los diferentes elementos. Es una información poco regular como para ser gestionada de una forma estándar. Estos datos poseen sus propios metadatos semiestructurados que describen los objetos y las relaciones entre ellos y pueden acabar siendo aceptados por convención. Un ejemplo es el HTML, el XML o el JSON.

2.1.4 Veracidad

La veracidad se refiere a la fiabilidad de los datos inherente para ciertas fuentes. En ocasiones los datos tienen una veracidad baja y se debe buscar las fuentes más confiables. Por ejemplo, los sentimientos en las redes sociales es un aspecto difícil de obtener y de comprobar. Según un estudio de IBM, uno de cada tres directores o altos directivos aún no confían en los datos masivos ni en los beneficios que estos les podrían generar (IBM).

2.1.5 Viabilidad

La viabilidad se refiere a la capacidad que tiene el usuario de los datos para manejarlos, administrarlos, procesarlos y transformarlos en información útil para sus propios fines. Estos datos son generados por múltiples fuentes que tienen distintas configuraciones, formatos de salida, velocidades de proceso y se hace necesario conectar, limpiar, transformar y unir todos estos datos en un mismo conjunto.

2.1.6 Visualización

Se refiere a la generación de procedimientos para visualizar los grandes conjuntos de datos. Las visualizaciones permiten identificar patrones, tendencias y relaciones entre puntos y zonas específicas.

2.1.7 Valor

El valor se refiere a que generalmente los datos recibidos tienen poco valor si se comparan con su volumen; sin embargo, al aplicarles procesos de análisis especializados se genera valor agregado al conjunto de datos.

No existen límites definidos en ninguna de las características mencionadas anteriormente para poder decir cuando un conjunto de datos entra en la categoría

de Big Data, ya que estas son dependientes una de otra. Cuando una de las características cambia, es probable que otra también lo haga y hasta una tercera lo hará también.

2.2 Datos masivos geoespaciales (*geospatial big data*)

Cada vez más, los conjuntos de datos con localización son de un tamaño, variedad y tasa de actualización que excede la capacidad de las tecnologías de computación enfocada a los procesos geoespaciales, estos datos se llaman *GeoSpatial Big Data* (GSBD por sus siglas en inglés) o *Datos Masivos Geoespaciales* (DMG), incluyen trayectorias de teléfonos celulares y dispositivos GPS, mediciones de motores de vehículos y mapas de carreteras detallados. La gran diversidad de fuentes de datos masivos geoespaciales aumenta sustancialmente la diversidad de métodos de solución. Los nuevos algoritmos pueden surgir a medida que se encuentren disponibles los conjuntos de datos masivos geoespaciales y de esta manera se crea la necesidad de una arquitectura flexible para integrar rápidamente nuevos conjuntos de datos y algoritmos asociados.

El componente geoespacial de los datos masivos geoespaciales puede venir en diversas formas tales como códigos postales, direcciones IP, localización geográfica (coordenadas x,y,z), nombres de calles o carreteras, recientemente se comienza a usar el análisis del contenido para identificar la ubicación exacta o lo más cercana posible de los datos o eventos relacionados.

De acuerdo con Muñiz (2014), "el 80% de los datos tienen una componente geográfica", esto quiere decir que gran parte de los datos en el mundo pueden ser georreferenciados y también indica la importancia del manejo geoespacial de los datos masivos. Los datos geoespaciales describen objetos con relación al espacio geográfico, con coordenadas de ubicación en un sistema de referencia espacial. Estos datos tradicionalmente se recopilan mediante la topografía terrestre, la fotogrametría, la teledetección y más recientemente, mediante el escaneo láser, la cartografía móvil, los contenidos geo-etiquetados, la información geográfica participativa o colaborativa, los sistemas mundiales de navegación por satélite y sensores geoposicionados. Los datos geoespaciales pueden presentar al menos una de las 3Vs pero las otras Vs mencionadas anteriormente también son relevantes.

El volumen creciente y el formato variable de los datos masivos geoespaciales recolectados plantean retos adicionales en cuanto al almacenamiento, gestión,

procesamiento, análisis, visualización y verificación de la calidad de los datos. Shekhar (2012) afirma que "el tamaño, la variedad y la tasa de actualización de los conjuntos de datos exceden la capacidad de las tecnologías de computación espacial y de bases de datos espaciales comúnmente utilizadas para aprender, gestionar y procesar los datos con un esfuerzo razonable". La comprobación de la calidad de los datos masivos geoespaciales y de los productos entregados a los usuarios finales se observa como uno de los grandes desafíos.

Los datos geoespaciales, que son abstracciones y observaciones de una realidad continua, son por naturaleza inciertos, idealmente estampados en el tiempo y a menudo incompletos. En consecuencia, los datos grandes geoespaciales, con sus características definitorias de ser grande (voluminoso), heterogéneo (variedad), procesado en tiempo real (velocidad), inconsistente (variabilidad) y, por tanto, también de calidad variable (veracidad), deben sufrir incertidumbre, desfase y fragmentación. Sin embargo, si bien ciertos efectos sobre la calidad de los datos se enfatizan para los grandes datos geoespaciales, los fenómenos que se describen siguen siendo los mismos.

2.2.1 Fuentes de datos masivos geoespaciales

Desde Google Maps hasta los dispositivos GPS (*Global Positioning System*), la sociedad se ha beneficiado enormemente de los servicios de posicionamiento y localización. Cada vez más, sin embargo, el tamaño, la variedad y la tasa de actualización de los conjuntos de datos exceden la capacidad de las tecnologías de computación geoespacial y de bases de datos geoespaciales comúnmente utilizadas para aprender, administrar y procesar los datos con un esfuerzo razonable.

Por ejemplo, los conjuntos de datos que proporcionan velocidades cada minuto para cada segmento de carretera, datos de rastreo GPS de teléfonos celulares, mediciones del consumo de combustible de un motor de acuerdo al tipo de terreno y altitud, emisiones de gases, fotos georreferenciadas tomadas con una cámara de un dron, barridos LiDAR, avisos de congestionamientos a través de plataformas colaborativas como Waze, sucesos y "sentimientos" geolocalizados a través de Twitter, Facebook, colecciones de fotos como Instagram, rastreo de mensajes de texto y llamadas, todos demuestran la gran cantidad de fuentes de datos masivos geoespaciales con posibilidad de ser almacenados, procesados y analizados.

Cuando se combinan, estos datos revelan patrones hasta ahora insospechados o no observados en la vida cotidiana que pueden beneficiar tanto a los individuos como a la sociedad. Existe también el riesgo de que los patrones o tendencias puedan abrir nuevas vías para el mal uso de los datos y la manipulación potencial de los individuos y su comportamiento. Es así que los datos masivos pueden ser vistos como una oportunidad y un desafío al mismo tiempo.

Algunas de éstas pueden ser:

- Imágenes de satélite.
- Datos existentes (BD, texto, video, fotos).
- Dispositivos móviles.
- GPS (*Global Positioning System*).
- Sensores (llantas, motores, RFID, tanques, válvulas).
- Internet de las cosas.
- Plataformas específicas (*waze, twitter, openstreetmaps, google maps*).
- Radar.
- LiDAR.
- Foto y video (ambos con geoposición) que proviene de sensores colocados sobre drones.
- *Crowdsourcing*.

2.3 Los datos masivos geoespaciales y el transporte

De acuerdo con la 1ª Ley de Tobler⁹ con respecto a la geografía y que establece “*todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes*”, entonces, el transporte (en su expresión más básica) se trata simplemente de conectar ubicaciones y luego relacionarlas con flujos. Estas ubicaciones pueden ser cercanas, estar bien conectadas y mostrar altos niveles de acceso, como en áreas urbanas. Los flujos entre estos lugares pueden referirse a personas o bienes y pueden referirse al número de vehículos, tipo, frecuencia, capacidad, etc.

⁹ Waldo Tobler. 1930. Geógrafo.

Los dispositivos y redes móviles en los hogares, escuelas, espacios recreativos y lugares de trabajo están conectados a Internet, proporcionando accesibilidad y seguimiento, incluso los viajes están conectados: de hecho, la industria del transporte es un líder en la creación de Internet de las Cosas (*IoT Internet of Things*, por sus siglas en inglés), al generar grandes volúmenes de datos cada día a través de sensores en los sistemas de conteo de pasajeros, localización y conteo de vehículos, sistemas de pago en carreteras, videos de vigilancia y recolección de características de las superficies de rodamiento de carreteras. Con el transcurso del tiempo, sin embargo, esos terabytes de datos recopilados han añadido complejidad a las operaciones de la institución y consumen cantidades inmensas de almacenamiento en los centros de datos.

Los sistemas de transporte que hacen uso de la información de las cámaras y microcontroladores para optimizar el flujo de tránsito público, vigilar el medio ambiente y ejecutar acciones de seguridad son conocidos como sistemas de transporte inteligente (ITS). En general, la información para ITS se extrae a partir de dos tipos de redes de sensores personalizados, sensores fijos y sensores móviles (Calabrese 2013).

Existen aplicaciones para el transporte inteligente, mediante la visualización y análisis del uso en tiempo real de las redes de transporte y seguridad mediante el procesamiento de datos sobre el funcionamiento del vehículo y su entorno para evitar o minimizar conflictos potencialmente peligrosos.

La exactitud de localización derivada de estos métodos varía en precisión, confiabilidad y puntualidad. Algunos de los datos pueden ser imprecisos pero producidos en tiempo real (con un retraso de milisegundos a segundos), algunos pueden ser muy precisos pero sólo disponibles ex-post a través de los registros de la máquina. Gran parte de los datos de localización producidos por dispositivos personales o sistemas embebidos en vehículos; sin embargo, son precisos y entregados casi en tiempo real. Es probable que los avances en las arquitecturas de los sensores aumenten la cantidad de datos con ubicación, al igual que la capacidad de procesar los datos capturados se hace menos costosa.

Los datos masivos geoespaciales se vislumbran como muy prometedores para mejorar la planificación y la gestión de la actividad del transporte aumentando radicalmente la disponibilidad de datos relacionados con la movilidad casi en tiempo real, para esto se requiere contar con los datos de la infraestructura, tales como carreteras, puentes, vías, aeropuertos, puertos y estaciones de autobuses. Asimismo, el acceso a datos más detallados y en relación con el funcionamiento de los vehículos y del medio ambiente en el que operan, mejorará la seguridad del transporte. Estos tres campos - *operaciones, planificación y seguridad* - son áreas en las que las autoridades deben evaluar críticamente dónde y cómo los datos disponibles y las perspectivas relacionadas, pueden mejorar la política de transporte.

Las autoridades de transporte disponen hoy de herramientas de información y de planificación avanzada y cada una tiene un propósito específico; sin embargo, mediante la fusión de estas fuentes de datos se puede crear una imagen mucho más completa de la dinámica y los factores que contribuyen a la eficacia de la red y, en consecuencia, la satisfacción (en todos sus aspectos) de los usuarios finales.

- El gran volumen de los datos masivos geoespaciales se debe a la gran cantidad de agentes (vehículos, personas y bienes) que están en camino en cualquier momento. Además, para poder predecir la demanda o el tráfico del transporte, no sólo se requieren datos en tiempo real sino también datos históricos.
- La gran variedad de los datos masivos geoespaciales en el transporte (como GNSS¹⁰, sensores inerciales, compás, sensores de rueda, radar, escaneado láser, reconocimiento de matrículas, ciclos de inducción, peaje electrónico o boletaje, sensores de estacionamiento) que se combinan en grandes análisis de datos
- La variabilidad de los datos masivos geoespaciales en el transporte proviene tanto de la variedad entre los diferentes canales de datos como de la falta de fiabilidad de cualquiera de estas fuentes. Por ejemplo, la información geográfica voluntaria puede ser inconsistente si los "ciudadanos como sensores" (Goodchild) o "ciudadanos como bases de datos" (Richter & Winter) no están de acuerdo con sus observaciones o están *discapacitados* por la falta de canales de comunicación (por ejemplo, un sensor se está moviendo fuera del alcance de la cobertura WiFi / teléfono celular). En otro ejemplo, el posicionamiento por satélite tiene algunos descriptores de calidad bien conocidos, pero entre las calles de una ciudad estas medidas varían con la ubicación del sensor, el canal de transmisión o el tiempo de transmisión.
- La veracidad de los datos masivos geoespaciales en el transporte indica que los datos son de distintas calidades. Esto se extiende también a tasas de muestreo irregulares (espaciales y temporales), errores de entrada, redundancia, corrupción, falta de sincronización o una variedad de propósitos de recolección (criterios de diseño de las bases de datos, semántica).
- Dado que la gran recopilación de datos requiere conectividad, también se requieren medidas para tratar con contribuciones maliciosas, ataques y robo, así como la privacidad. Esto último es particularmente cierto para los datos geoespaciales recolectados para los movimientos de seguimiento en tiempo real de personas y vehículos.

¹⁰ GNSS(*Global Navigation Satellite System*), es el acrónimo que se refiere al conjunto de tecnologías de sistemas de navegación por satélite que proveen de posicionamiento geoespacial con cobertura global de manera autónoma.

- Típicamente, el valor de los datos masivos geoespaciales (análisis) se ve en la información para el apoyo a la toma de decisiones. Los métodos analíticos tales como la minería de datos y el aprendizaje automático permiten solamente el razonamiento inductivo en datos grandes, es decir, la detección de correlaciones globales o predicciones basadas en estas correlaciones. En el transporte, un ejemplo es el descubrimiento de patrones en la ocurrencia de “embotellamientos” así como de accidentes de tráfico.

Algunas de las áreas de aplicación que se pueden relacionar al transporte son:

- Tendencias sociales (*twitter*, *facebook*) y colaboración participativa: mapas, levantamiento de información, encuestas de origen-destino
- Medición de datos espacio-temporales para identificación de patrones: logística, percepción remota, dirección del viento, precipitación, oleaje
- Seguimiento de datos GPS: movilidad, accesibilidad, cálculo de rutas alternas
- Perfiles históricos y de predicción: contaminación, velocidades, causa de embotellamientos, rendimiento de maquinaria
- Identificación de relaciones topológicas: nubes de puntos masivas, optimización y planeación del transporte

2.4 Proceso de los datos masivos geoespaciales

Los datos masivos - en el transporte - han surgido de la convergencia de los costos decrecientes de recolección, almacenamiento y procesamiento, así como de la difusión de los datos. La disminución de los costos de los sensores ha llevado a una proliferación de plataformas de detección que transforman grandes extensiones del mundo analógico en señales procesadas digitalmente. La reducción de los costos de almacenamiento de datos ha permitido retener los datos que se habían descartado anteriormente. Como lo señaló el historiador de ciencias Dyson (S/F), "Big Data es lo que sucedió cuando el costo de almacenar la información fue menor que el costo de tomar la decisión de tirarlo".

Para que las organizaciones hagan un uso eficiente de los datos requieren procesos que les permitan extraer y generar información correcta a partir de los datos masivos. Labrinidis & Jagadish (2015), han identificado cinco fases principales agrupadas en dos grandes procesos (*Figura 2.2 Procesos de los datos masivos*).



Traducción y adaptación a partir de <http://cra.org/crc/wp-content/uploads/sites/2/2015/05/bigdatawhitepaper.pdf>

Figura 2.2 Procesos de los datos masivos.

Administración de datos

De acuerdo con la *Figura 2.2 Procesos de los datos masivos*, el proceso de administración de datos se refiere a subprocesos y tecnologías de soporte que adquieren y almacenan los datos y los preparan para recuperarlos para su análisis.

Adquisición y almacenamiento

En los últimos años, junto con la disponibilidad de nuevos sensores, han surgido nuevas formas de recolección de datos geospaciales que han dado lugar a fuentes de datos completamente nuevas y tipos de datos de naturaleza geográfica. Los datos adquiridos por el público, llamados *información geográfica*

voluntaria (*Volunteered Geographic Information VGI*, por sus siglas en inglés) y los datos de las redes de sensor con geoposición han llevado a una mayor disponibilidad de información espacial. Mientras, hasta hace poco tiempo, los conjuntos de datos generados por una autoridad eran la mayoría, ahora estos nuevos tipos de datos vienen a complementar, ampliar y enriquecer el conjunto de datos geográficos en términos de variación temática y por el hecho de que están más centrados en el usuario. Esto último es especialmente cierto para los datos recopilados a través de medios sociales (Li 2016).

La recopilación de datos geoespaciales está evolucionando de ausencia de datos a un panorama con gran disponibilidad de éstos. Mientras que hace unos años la captura de datos geoespaciales se basaba en dispositivos técnicamente exigentes, precisos, caros y complicados, ahora la adquisición de datos geoespaciales es un proceso implementado en dispositivos móviles de uso cotidiano para mucha gente. Estos dispositivos son capaces de adquirir información geoespacial a un nivel sin precedentes, con respecto a la resolución geométrica, temporal y temática; son pequeños, fáciles de manejar y capaces de adquirir datos incluso sin que el usuario se percate.

Las personas ayudan a capturar *información geográfica voluntaria* relacionada con el tráfico o la movilidad. La adquisición de datos se debe a un proceso consciente de un usuario, que explícitamente selecciona objetos, observa sus características y aporta esta información, generalmente a través de una plataforma; por ejemplo *Waze* y *OpenStreetMap*. En el caso de *Waze*, la captura de los puntos de congestionamiento y los distintos tipos de eventos son sólo subproductos que generan los usuarios voluntariamente y por otro lado, el principal producto es todo el conjunto de lecturas GPS que se registran sin que el usuario intervenga y se envían del teléfono celular a los servidores de la plataforma, en donde se procesan y al final se generan mapas dinámicos que muestran las calles, avenidas o carreteras a partir de la congestión vial que presenta cada uno de ellos.

En cuanto al almacenamiento se requieren equipos que cumplan con las condiciones solicitadas para efectuar tal función, estos equipos pueden ser locales o también se puede utilizar el almacenamiento en la nube aunque para las dos vertientes se hace necesario implementar mecanismos de organización que permita almacenar cada conjunto de datos en una forma lógica y correcta de tal manera que cuando se requiera utilizarlos sea fácil identificarlos, descargarlos y procesarlos con los métodos establecidos para ello. Es indispensable la creación de los metadatos de cada uno de los conjuntos de datos, en donde se describen las características principales de éstos. Para este proyecto se utilizó un disco duro local con capacidad de 500 GB. El arreglo fue colocarlos en directorios y subdirectorios, por lo que existe un directorio para cada mes y subdirectorios para la hora de levantamiento correspondiente.

Por ejemplo, *mayo/07*, *mayo/09*, *mayo/11*, *mayo/13*, *mayo/15*, *mayo/17*, *mayo/19*, *mayo21*, *mayo/23*

Extracción, limpieza y anotación

Más allá de las cuestiones de disponibilidad y costos de recolección, un factor importante a considerar al seleccionar una fuente de datos es su aptitud para el análisis. Las técnicas de extracción se refieren a todas las formas en que la información se extrae de un conjunto de datos determinado. Una vez validados los campos que dan origen a un identificador único y a la geoposición (por ejemplo, el nombre, el origen y el destino, la longitud y la latitud), se puede realizar una serie de operaciones para limpiar, transformar y modelar los datos en busca de conclusiones significativas.

Se han desarrollado o adaptado una gama de técnicas y herramientas para agrupar, manipular, limpiar y extraer datos masivos. Éstos se basan en la experiencia de distintas áreas del conocimiento como la estadística, informática, matemáticas aplicadas y economía.

La limpieza de datos y su preparación para el uso analítico es una tarea no trivial que puede implicar un costo significativo. Los datos estructurados deben ser analizados y los faltantes o datos potencialmente incorrectos contabilizados. Los datos no estructurados deben ser correctamente interpretados, categorizados y etiquetados consistentemente.

En algunos casos, la "*edición manual de datos*" sigue siendo un componente necesario del flujo de recolección y análisis de datos. Este trabajo requiere una gran inversión en tiempo y recursos que representan entre el 50% y el 80% del tiempo de los "*científicos de datos*" según algunas estimaciones de Lohr (2014). La preparación de los datos puede facilitarse en gran medida mediante el uso de algoritmos compensatorios apropiados, pero la elección del algoritmo puede conducir a errores de imputación o predicción si interpreta incorrectamente los valores faltantes o minimiza los valores atípicos que podrían ser importantes.

En el caso de estudio, los datos descargados fueron categorizados por hora y por día, utilizando la estructura de datos indicada en la parte de Adquisición y Almacenamiento; se modificó el nombre a la extensión de archivo que por defecto tiene y se cambia por .XML.

Enseguida se transforma a formato XLS utilizando el macro en Excel desarrollado para tal fin y que se muestra más adelante; se cambian nombres de campos, esto para asegurar que el macro no presente problemas y los nombres de campo sean compatibles con el diseño de la base de datos geoespacial que se utilizará más adelante.

Por ejemplo, *mayo/07*

02mayo2017_07

02mayo2017_09

02mayo2017_11

.

.

.

Al cambiar la extensión, quedan de la siguiente manera:

02mayo2017_07.xml

02mayo2017_09.xml

02mayo2017_11.xml

.

Integración, agregación, representación y fusión

La integración de datos es el proceso que permite combinar datos heterogéneos de muchas fuentes diferentes en la forma y estructura de una única aplicación. Este proceso de integración de datos facilita que sus diferentes tipos, tales como matrices de datos, documentos y tablas, sean fusionados por usuarios, organizaciones y aplicaciones para un uso personal, de procesos de negocio o de investigación.

La integración de datos soporta el procesamiento analítico de grandes conjuntos de datos alineando, combinando y presentando cada conjunto de datos y fuentes remotas y externas, para cumplir con los objetivos del integrador.

La integración de datos se implementa generalmente en un almacén de datos (*data-warehouse*) mediante software especializado que aloja grandes repositorios de datos de recursos internos y externos. Los datos se extraen, se mezclan y se presentan de forma unificada. Por ejemplo, el conjunto completo de datos de un usuario puede incluir datos extraídos y combinados de marketing, ventas y operaciones, que se combinan para formar un informe completo¹¹.

Un proyecto de integración de datos generalmente implica los siguientes pasos:

¹¹ <http://www.powerdata.es/integracion-de-datos>

- Acceso a los datos desde todas las fuentes y localizaciones, tanto si se trata de locales, en la nube o de una combinación de ambos.
- En la integración de datos existe una técnica en donde los registros de una fuente de datos mapean registros en otra. Se trata de un tipo de preparación de datos esencial para que las analíticas y otras aplicaciones sean capaces de utilizar los datos con éxito.
- Para resolver un problema se necesita conocimiento específico sobre él y también se requiere representar este conocimiento de alguna manera útil y eficaz. La forma de representar dicho conocimiento dependerá del problema a analizar y se deberá encontrar la mejor manera de representarlo.

En el caso de agregación de datos, los conjuntos de datos se emparejan y se combinan con base en atributos y variables compartidos, pero se conserva el conjunto de cada conjunto de datos independiente. Este método es adecuado para incrementar el descubrimiento de conocimiento a través del análisis de datos contextuales.

La fusión de datos es un paso especialmente importante en el uso de entradas de múltiples plataformas de sensores. Por ejemplo, los algoritmos de fusión de datos ayudan a procesar entradas de sensores de movimiento, acelerómetros, magnetómetros, datos de señal celular, cámaras, escáneres láser y chips GPS. Todas estas fuentes de datos contribuyen a crear, por ejemplo, una representación precisa de la ubicación de un auto en una calle, donde dicha fusión de datos es necesaria para el desarrollo de vehículos de conducción autónomos. Aquí, la representación final del vehículo (la ubicación exacta, el tamaño, la dirección y velocidad) es todo lo que se retiene, eliminando así la necesidad de almacenar cada flujo de datos individuales de cada sensor. El aspecto desafiante de la fusión de datos está en la extracción de características sobresalientes de múltiples conjuntos de datos generados para diferentes usos.

En general, las técnicas de fusión de datos buscan integrar precisión y semántica. Los datos recolectados de fuentes múltiples suelen contener incoherencias en términos de resolución. Por ejemplo, hay GPS que registran una posición cada segundo mientras que algunas estaciones base lo hacen a cada 15 segundos. La semántica se refiere al sujeto que está siendo representado. En un caso, los datos rastrean un vehículo y la otra vertiente se refiere al usuario del auto y con quién se comunica.

Para esto, se sugiere la siguiente lista de atributos que debe cumplir un paquete de datos:

- Localización (sistema de coordenadas, latitud y longitud) del dispositivo transmisor o acción reportada.

- Tiempo de transmisión de datos o acción reportada.
- Categoría de datos (datos basados en la ubicación del teléfono móvil, datos GPS, boletín de noticias).
- Formato de datos (valor único, matriz, vector, texto, imagen, etc.)
- Representación de datos (por ejemplo, unidad de medida).
- Semántica de los datos (por ejemplo, vehículo de seguimiento o teléfono móvil).

En el proyecto se verifico que los datos cumplieran con las condiciones anteriores, esto es, cuentan con coordenadas, se refieren a un tiempo determinado (día/hora), son datos que fueron capturados por la plataforma participativa Waze, el formato de los datos corresponde a un punto determinado en cierto lapso y representa un evento de algún tipo (ver diccionario de eventos de Waze) relacionado al tráfico en algún lugar. La unidad de medida, dependiendo si es línea o punto es el día y la hora para los puntos o la velocidad relativa del tramo si es línea.

Enseguida se fusionan los datos en una única tabla, se convierten a formato XLS (Microsoft Excel), se verifican los nombres de los campos, se borran los campos con datos que no sirven o que contienen basura. Se normaliza la base datos y se coloca en un subdirectorío de respaldo.

El siguiente proceso es la representación de los datos en un Sistema de Información Geográfica, en este caso, ArcGIS; para esto se transforman los datos y se generan *shapes* de puntos que los contienen o representan.

Se genera un *shape* por cada día de levantamiento y éste contiene las lecturas waze de este día, agrupadas por hora de captura.

Análisis

De acuerdo con la *Figura 2.2 Procesos de los datos masivos*, el proceso de análisis se refiere a las técnicas usadas para analizar y generar valor agregado e identificar patrones derivados de los datos masivos.

Modelado y análisis

Construir y ejecutar modelos ayuda a probar hipótesis sobre el impacto y la importancia de diferentes variables en los sistemas del mundo real. Al simplificar la simulación de fenómenos del mundo real, los modelos ayudan a caracterizar, comprender, cuantificar y visualizar relaciones que son difíciles de comprender en sistemas complejos. Los modelos de construcción requieren datos sobre las condiciones de referencia y la comprensión de la naturaleza de las relaciones, correlativas o causales, entre múltiples fenómenos. La llegada de los datos masivos ha aumentado drásticamente la escala, el alcance y la accesibilidad de los ejercicios de modelado, aunque debe tenerse en cuenta que la capacidad de

éstos para seguir con precisión el mundo real está vinculada no sólo, o principalmente, a la cantidad y la calidad de los datos de referencia. La construcción del modelo y la garantía de que las preguntas correctas se formulan siguen siendo esenciales para proporcionar resultados de alto valor. Los modelos bien contruidos basados en datos escasos pueden ser tan o más eficaces que los modelos mal diseñados que trabajan en conjuntos masivos de datos en tiempo real. Con esta advertencia en mente, las fuentes y las técnicas de datos masivos han permitido construir nuevos modelos que proporcionen nuevas preguntas y nuevas perspectivas.

Estructuración espacial de datos

Todos los modelos de datos espaciales, incluyendo el modelo de datos vectoriales tipo espagueti¹², el modelo de datos de red, el modelo de datos topológicos y el modelo regular (ráster) e irregular (Voronoi¹³, árbol kd¹⁴, árbol de partición binario¹⁵), se pueden utilizar para manejar datos masivos geoespaciales. Sin embargo, existen algunos modelos que son más adecuados para manejar conjuntos de datos muy grandes y otros que son menos adecuados para datos masivos geoespaciales. Como la red y los modelos de datos topológicos necesitan almacenar la conectividad y la adyacencia, no son adecuados para manejar grandes flujos de datos geoespaciales a menos que un algoritmo muy eficiente realice la indexación de datos, así como la actualización de la conectividad y/o la topología en tiempo real. En tales casos, la única manera posible de satisfacer las

¹² Almacena información de los puntos como pares de coordenadas, las líneas como una sucesión de pares de coordenadas y los polígonos cadena de pares de coordenadas con repetición del primer par de coordenadas que indica que es un elemento cerrado.

¹³ El diagrama de Voronoi de un conjunto de puntos en el plano es la división de dicho plano en regiones, de tal forma, que a cada punto le asigna una región del plano formada por los puntos que son más cercanos a él que a ninguno de los otros objetos. http://www.abc.es/ciencia/abci-diagrama-voronoi-forma-matematica-dividir-mundo-201704241101_noticia.html

¹⁴ En ciencias de la computación, un Árbol kd (abreviatura de árbol k-dimensional) es una estructura de datos de particionado del espacio que organiza los puntos en un Espacio euclídeo de k dimensiones. http://es.unionpedia.org/%C3%81rbol_kd

¹⁵ Es una estructura de datos usados para organizar objetos dentro de un espacio. Dentro del campo de gráfica de computadores, tiene aplicaciones en la remoción de áreas ocultas y en el trazado de rayos. <http://www.symbolcraft.com/products/bsptrees/spanish/>

necesidades de proceso en tiempo real es utilizar un método de indexación de datos espaciales que pueda mantener sus prestaciones con un gran flujo de datos.

Los actuales métodos de indexación de datos espaciales no pueden manejar grandes flujos de datos geoespaciales porque su eficiencia se reduce a medida que los nuevos flujos de datos espaciales superan la capacidad de extensión del índice de datos espaciales. La geoestadística es adecuada para manejar grandes datos. Ofrece oportunidades para resumir los datos y expresar medidas de variación e incertidumbre. La gran preocupación, sin embargo, es que muchos de los procesos y procedimientos se desarrollan para los conjuntos de datos más pequeños. En particular, gran parte del análisis estadístico espacial se realiza en conjuntos de datos que se recogen en una escala puntual (como datos de campo, datos meteorológicos o datos administrativos) y de un contenido relativamente pequeño, o se centra en los conjuntos de datos de imagen relativamente grandes que tienen un carácter muy específico. La estadística espacial depende de la noción de dependencia espacial (y espaciotemporal) y tal dependencia depende a su vez, de la noción de distancia entre puntos. Las estructuras de datos actuales como tales suelen ser capaces de manejar los grandes datos también, pero lo más probable es que se desarrollen procedimientos específicos que sean capaces de resolver problemas relativamente novedosos (como combinar datos en el dominio espacio-tiempo) o que tienen que abordar preguntas y problemas específicos, es decir, seleccionar datos de un conjunto de datos grande para una aplicación de modelo particular.

Algunos autores como Lee (S/F) y Shekhar (2012), ya han señalado la necesidad de una programación paralela y distribuida para manejar los grandes conjuntos de datos en el contexto general o incluso en el contexto geoespacial, así como la aplicación de los métodos de minería de datos, a pesar de que difieren de los métodos tradicionales de análisis de bases de datos que no presuponen un modelo que describe las relaciones en los datos ni requieren consultas específicas sobre las que basar el análisis, más bien estos enfoques permiten que los datos hablen por sí mismos, basándose en algoritmos para descubrir patrones que no son evidentes en los conjuntos de datos únicos y unidos.

Los algoritmos de minería de datos realizan diferentes tipos de operaciones¹⁶:

- Clasificación, donde los objetos o eventos se clasifican según categorías conocidas (por ejemplo, las compañías de seguros emplean algoritmos de

¹⁶ Un algoritmo en minería de datos (o aprendizaje automático) es un conjunto de heurísticas y cálculos que permiten crear un modelo a partir de datos. Para crear un modelo, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias. El algoritmo usa los resultados de este análisis en un gran número de iteraciones para determinar los parámetros óptimos para crear el modelo de minería de datos. <https://msdn.microsoft.com/es-es/library/ms175595.aspx>

clasificación para asignar categorías de riesgo de accidente a los conductores que comparten ciertas características).

- Agrupamiento (*Clustering*), donde se buscan patrones de similitud en los datos brutos (aplicable a *Waze*).
- Regresión o predicción numérica, donde las cantidades numéricas son predichas según el análisis de regresión.
- Asociación, donde se identifican las relaciones entre elementos de conjuntos de datos únicos o unidos.
- Detección de anomalías, donde se identifican los valores atípicos o pausas de patrones en los conjuntos de datos.
- Resumen, tabulación y presentación de características destacadas dentro de conjuntos de datos.
- Los enfoques de minería de datos pueden basarse en ejemplos de relaciones que proporcionan los operadores humanos y se utilizan para guiar el proceso o mediante una operación no supervisada donde los patrones se descubren algorítmicamente.

Para el modelado se utilizó la técnica de *Densidad de Puntos*, en donde se calcula la densidad de las entidades de punto alrededor de cada celda ráster de salida. Conceptualmente, la vecindad se define alrededor de cada centro de celdas ráster y la cantidad de puntos que están dentro de la vecindad se totaliza y divide por el área de la vecindad.

Si se selecciona una unidad de área, la densidad que se calculó para la celda se multiplica por el factor apropiado antes de que se escriba en el ráster de salida. Por ejemplo, si las unidades de terreno de entrada están en metros, la comparación de un factor de escala de unidades de metros a kilómetros ocasionará que los valores sean diferentes por un multiplicador de 1.000.000 (1.000×1.000).

Los posibles usos incluyen la búsqueda de densidad de viviendas, observaciones de vida silvestre, informes de delitos o para el estudio de caso del proyecto, la densidad de eventos relacionados a riesgos identificados por usuarios de *Waze*. El campo de población se puede utilizar para ponderar algunos puntos más que otros, según su significado, o para permitir que un punto represente varias observaciones. Por ejemplo, algunos riesgos se pueden ponderar más que otros al determinar los niveles de riesgo en general.

Aumentar el radio no cambiará en gran medida los valores de densidad que se calcularon, aunque caerán más puntos dentro de la vecindad más grande, este

número se dividirá por un área más grande cuando calcule la densidad. El efecto principal de un radio más grande es que la densidad se calcula considerando una mayor cantidad de puntos, que pueden estar más lejos de la celda ráster. Esto tiene como resultado un ráster de salida más generalizado¹⁷.

Interpretación

Desde hace tiempo, los seres humanos han confiado en la representación gráfica o pictórica de los datos para que los registros de información sean accesibles, comprensibles y, lo que es más importante, atractivos para la mente humana. En consecuencia, la reciente explosión de datos también ha visto un aumento en las herramientas que se centran en la visualización efectiva de los datos. Muchas herramientas sobresalen en representar la información usando métodos tradicionales como tablas, histogramas, gráficos circulares y gráficos de barras. Sin embargo, los gráficos de barras presentados en documentos extensos o presentaciones de diapositivas no suelen adaptarse a una audiencia más allá de los profesionales. A medida que los conjuntos de datos se hacen más grandes y los esfuerzos para explotar estos datos buscan llegar a más personas, el lenguaje de visualización de datos debe adaptarse y mejorar.

Los ciudadanos viven en un mundo cada vez más visual, observan pantallas de diferentes tamaños con resoluciones que se incrementan cada que cambian de dispositivo. Las visualizaciones sirven no sólo para la entrega de información sino también para generar interés e impactar; son presentaciones de información enmarcadas en la convergencia del arte, los medios digitales y la tecnología de la información.

Visualización de datos

Son las técnicas utilizadas para generar imágenes, diagramas o animaciones para comunicar los resultados del análisis de datos, tales como mapas de tráfico. Las técnicas de visualización se utilizan durante y después del análisis de datos para dar sentido a la información.

Por lo tanto, las visualizaciones son ampliamente reconocidas como parte del proceso de análisis (es decir, no sólo la comunicación), en las que se puede explorar los datos y construir hipótesis durante este proceso (Chen & Zhang, 2014). Es importante señalar que las visualizaciones se han conceptualizado comúnmente como herramientas de comunicación. Si bien esto es cierto y las

¹⁷ <http://desktop.arcgis.com/es/arcmap/10.3/tools/spatial-analyst-toolbox/how-point-density-works.htm>

visualizaciones son muy importantes en la comunicación de hipótesis, resultados e ideas, en el caso de los datos masivos, su papel en la exploración juega un papel muy importante. Los datos masivos tienen muchas cosas para mostrar y a menudo se muestran pantallas muy "ocupadas" o sobresaturadas de información, por lo que se hace necesario un pre-análisis visual de las aplicaciones de datos masivos.

Un SIG es una caja de herramientas muy completa y de amplia utilización en la ciencia de datos, esto por su capacidad para procesar datos espaciales y no espaciales (atributos), incluso cuando no están perfectamente estructurados, a través de medios computacionales y visuales. La ciencia de la información geográfica y sus aplicaciones asociadas, como la teledetección y la geoinformática, han estado tratando con grandes conjuntos de datos durante un tiempo relativamente largo, incluso antes de que el término datos masivos tomara impulso en la ciencia, la cultura popular y los negocios (Cöltekin & Reichenbacher, 2011). A escala urbana, regional y nacional, las plataformas de suministro de información capaces de combinar capas de información de manera comprensible pueden aumentar la eficiencia general y la sostenibilidad de la planificación y regulación de la infraestructura del transporte en sus diversos ámbitos y facilitar las labores de las instituciones encargadas de su administración.

Para promover el acceso universal y la compatibilidad, las visualizaciones deben poder ser exportadas y distribuidas en varios formatos, desde imágenes a vídeos o páginas web. Las grandes pantallas interactivas son más propensas a atraer a los usuarios, sobre todo, los principiantes. La compatibilidad de la plataforma de visualización es importante ya que permite la integración inteligente con otras plataformas como los portales web públicos o privados. También permite agregar nuevos módulos de terceros en un entorno de código abierto.

La difusión de los productos de análisis de datos, así como la distribución de algunas formas de datos, se realiza con papel o documentos digitales. Sin embargo, la utilidad de estos canales está disminuyendo ya que estos medios son generalmente estáticos. El acceso en línea a datos tabulares o geográficos que son directamente utilizables por diversas plataformas de software puede ser valioso para algunas formas de análisis de datos. Cada vez más, la difusión de datos se producirá a través de Interfaces de Programación de Aplicaciones¹⁸ o formatos de datos que permiten la integración de estos datos directamente en diferentes aplicaciones móviles o de otros tipos. El acceso a datos dirigido a aplicaciones en teléfonos móviles basado en API ha contribuido a muchos nuevos servicios que facilitan enormemente la navegación, los viajes, la logística y otros

¹⁸ Una API es un conjunto de funciones y procedimientos que cumplen una o muchas funciones con el fin de ser utilizadas por otro software. Las siglas API vienen del inglés *Application Programming Interface*. Una API permite implementar las funciones y procedimientos que engloba determinado proyecto sin la necesidad de programarlas de nuevo. En términos de programación, es una capa de abstracción. <https://hipertextual.com/archivo/2014/05/que-es-api/>

servicios relacionados con el transporte para individuos y empresas. La difusión de datos a través de las API desempeñará un papel cada vez más importante en el acceso de los ciudadanos y el uso de datos relacionados con la movilidad, pero los beneficios finales de este canal de difusión de datos dependerán de los términos de uso asociados con esos datos.

Ahora bien, estos términos de uso de los datos abarcan desde términos completamente "abiertos" sin restricciones de uso y redistribución, hasta términos altamente restringidos que permiten el acceso comercial a los datos sólo en un conjunto limitado de condiciones. Los defensores de los datos abiertos proponen que la mayor cantidad posible de datos sea de uso abierto. Esto incluye (casi) todos los datos recopilados por el gobierno, ya que sostienen que los ciudadanos ya los pagan con sus impuestos y por lo tanto se debería tener acceso libre a ellos. Por supuesto, hay problemas con el modelo de datos abiertos relacionados con la seguridad y la privacidad. Sin embargo, muchas autoridades están tratando de abrir el acceso a gran parte de sus datos, especialmente en el transporte. Por otro lado, las limitaciones que se imponen al acceso a los datos recogidos por empresas comerciales son comprensibles y necesarias, ya que las empresas que recopilan datos y prestan servicios basados en el uso y análisis del mismo, deben generar valor y ganancias económicas a sus propietarios. Para ello, se debe generalizar cuidadosamente, hacer hincapié en lo importante, eliminar lo poco importante, agrupar la información tanto de forma temática como perceptual y prestar atención a la jerarquía visual cuando se diseñan las pantallas.

3 Plataforma Waze

3.1 Cómo funciona

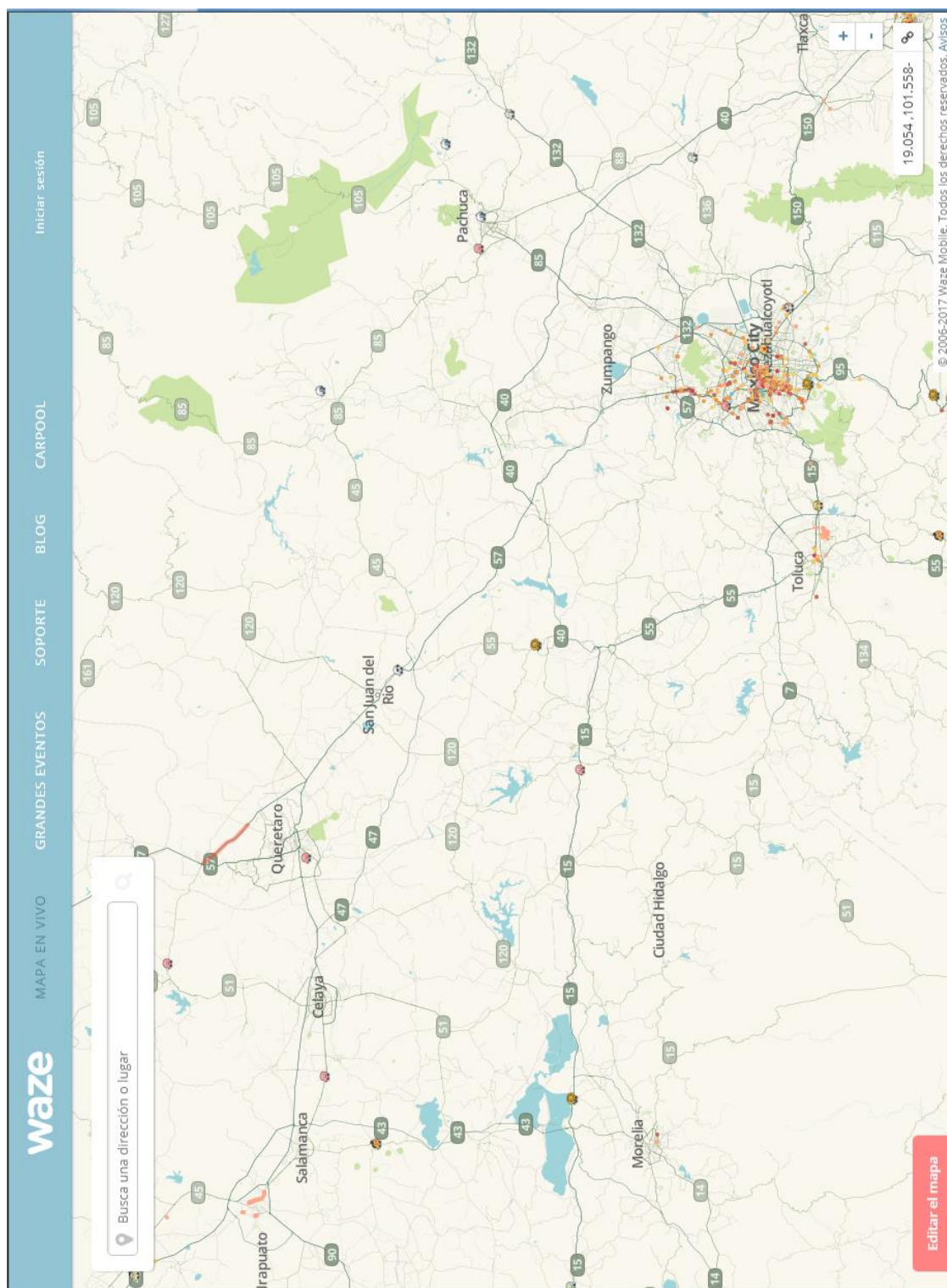
El *crowdsourcing* (Introducción, pag. 12), es la base fundamental de la aplicación Waze, la cual fue creada en 2008 en Israel, utiliza navegación GPS, proporciona información del tránsito, además de varias funciones sociales. Los Wazers (usuarios de waze) se pueden informar unos a otros sobre el tránsito, controles policiales, obras, radares, accidentes y eventos meteorológicos. Al ser información generada por los propios usuarios, cuantos más usen la plataforma, será mejor y habrá mas datos. En 2013 Waze fue adquirida por parte de Google por 1.3 miles de millones de dólares, fundamentalmente para tener acceso a la información social de la plataforma y para recolectar información en tiempo real en una forma más barata¹⁹. Se estima que en todo el mundo la aplicación cuenta con 70 millones de usuarios y para la Ciudad de México se cree que hay 1.5 millones de usuarios²⁰.

Existen dos maneras de participar en la plataforma, la primera es cuando los usuarios conducen con la aplicación abierta para contribuir pasivamente con información del tránsito; también pueden tomar un rol más activo al compartir alertas de incidentes en el tránsito como accidentes, controles policiales o cualquier peligro que encuentren en la vía. Esto ayuda a otros usuarios que están en la zona porque reciben un aviso de lo que sucede más adelante en su ruta. Existe una comunidad de editores de mapas que aseguran que los datos estén lo más actualizados que sea posible²¹, ver Figura 3.1 y Figura 3.2

¹⁹ <https://www.entrepreneur.com/article/266030>

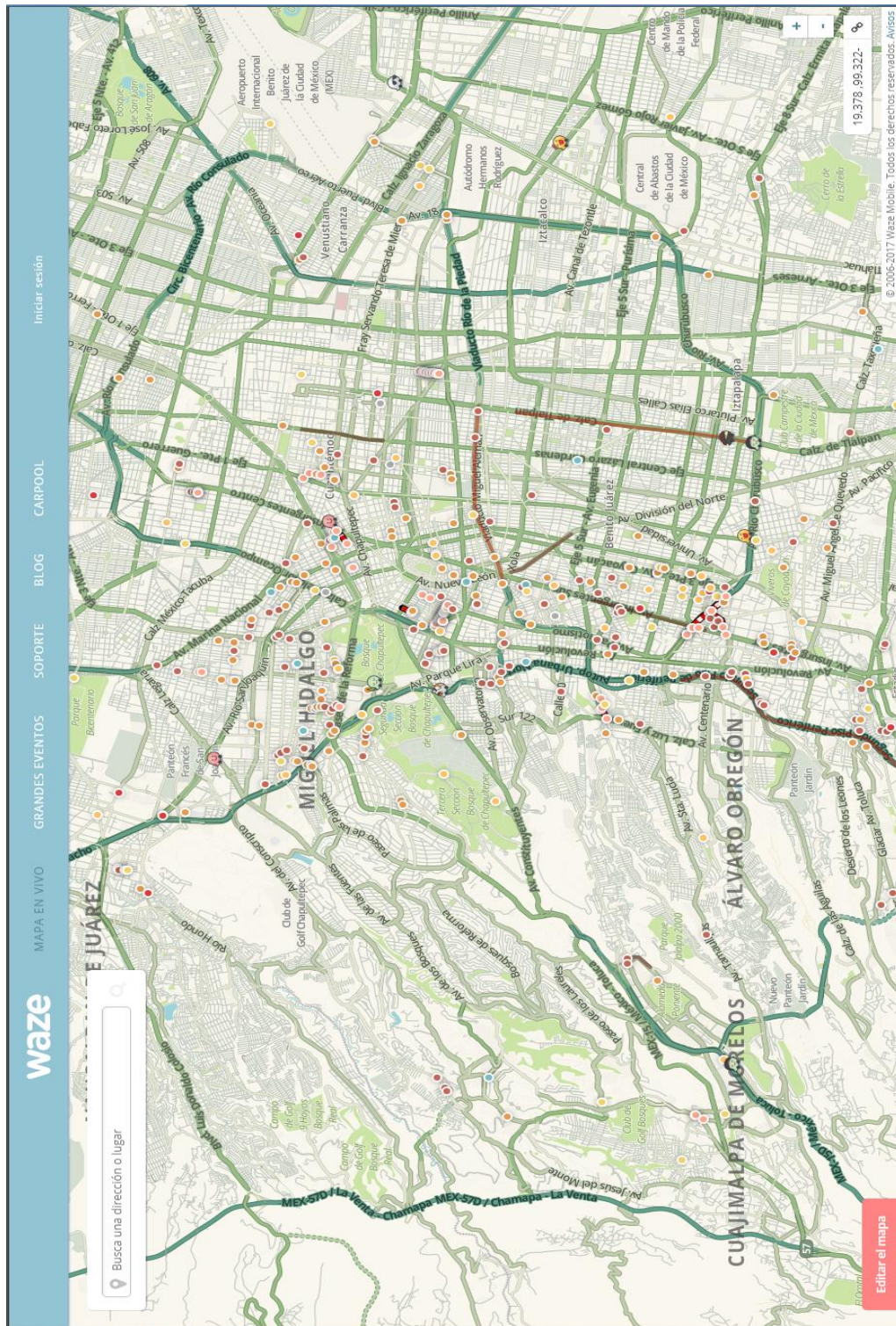
²⁰ Estos son números estimados ya que la plataforma, por políticas internas, no publica el número de usuarios totales a nivel mundial y tampoco por país.

²¹ Tomado de ¿Cómo funciona Waze? <https://support.google.com/waze/answer/6078702?hl=es>



Fuente: waze.com

Figura 3.1 Pantalla de mapa en vivo de Waze.



Fuente: waze.com

Figura 3.2 Pantalla de mapa en vivo de Waze, acercamiento a la Ciudad de México.

3.2 Descripción del programa *Citizen Connected Partner*

El Programa de Ciudadanos Conectados (*Citizen Connected Partner - CCP*) es una asociación continua entre Waze y varias agencias gubernamentales internacionales para compartir datos de cierre de caminos e incidentes que se distribuyen públicamente. Los participantes en este programa (a Octubre 2016), divididos por regiones son²²:

North America

1. Alabama – Department of Transportation
2. California – Caltrans
3. California – City of Cupertino City Hall
4. California – City of Los Angeles
5. California – City of Sacramento
6. California – City of San Francisco
7. California – Los Angeles Metropolitan Transportation Authority
8. California – Paramedics Plus (Genesis Pulse)
9. California – Town of Los Gatos
10. Canada – Ville de Montreal
11. Colorado – City of Colorado Springs
12. Colorado – Douglas County
13. District of Columbia – D.C. Department of Transportation
14. Florida – City of Miami Beach
15. Florida – City of Tampa
16. Florida – Florida Department of Transportation
17. Florida – Paramedics Plus (Genesis Pulse)
18. Florida – Miami-Dade County
19. Florida – Sunstar EMS (Genesis Pulse)
20. Georgia – Bartow County
21. Georgia – City of Atlanta

²² https://wiki.waze.com/wiki/Connected_Citizens_Program

22. Georgia – City of Johns Creek
23. Georgia – City of West Jackson
24. Georgia – Georgia Department of Transportation
25. Georgia – Georgia Emergency Management & Homeland Security Agency
26. Illinois – City of Evanston
27. Illinois – City of Naperville
28. Indiana – City of Bloomington
29. Indiana – Paramedics Plus (Genesis Pulse)
30. Indiana – Three Rivers Ambulance Authority (Genesis Pulse)
31. Iowa – Iowa Department of Transportation
32. Kentucky – City of Louisville
33. Kentucky – Kentucky Transportation Cabinet
34. Louisiana – City-Parish of Baton Rouge
35. Louisiana – Louisiana Department of Transportation and Development
36. Maine – Maine Department of Transportation
37. Maryland – University of Maryland
38. Maryland – St. Mary's Emergency Services and Technology
39. Massachusetts – City of Boston
40. Massachusetts – City of Cambridge (Kleinfelder East)
41. Massachusetts – Capital Strategic Solutions
42. Massachusetts – Massachusetts Department of Transportation
43. Missouri – CoxHealth (Genesis Pulse)
44. Missouri – Mercy EMS (Genesis Pulse)
45. Missouri – Taney County Ambulance Directory (Genesis Pulse)
46. National – SeeClickFix (nonprofit partner)
47. National – United States Department of Transportation
48. Nebraska – Nebraska Department of Roads
49. Nevada - Regional Transportation Commission of Southern Nevada (RTCSVN)
50. New Hampshire – New Hampshire Department of Transportation
51. New Jersey – City of Jersey City

52. New Jersey – Jersey City EMS (Genesis Pulse)
53. North Carolina – City of Charlotte
54. North Carolina – City of Greensboro
55. North Carolina – City of Raleigh
56. Ohio – Town of Dublin
57. Oregon – Oregon Department of Transportation
58. Pennsylvania – Pennsylvania Department of Transportation
59. Pennsylvania – Pennsylvania Turnpike Authority
60. Pennsylvania – Wilkes-Barre Township Police Department
61. Rhode Island – City of Providence
62. Rhode Island – Rhode Island Turnpike & Bridge Authority
63. South Dakota – Paramedics Plus (Genesis Pulse)
64. Tennessee – Tennessee Department of Transportation
65. Texas – CareFlite (Genesis Pulse)
66. Texas – Champion EMS (Genesis Pulse)
67. Texas – ETMC EMS (Genesis Pulse)
68. Texas - City of Fort Worth
69. Texas – LifeNet EMS (Genesis Pulse)
70. Utah – Utah Department of Transportation
71. Vermont – Vermont Department of Transportation
72. Virginia – City of Arlington
73. Virginia – Portsmouth Police Department

Latin America

1. Brazil – City of Petropolis
2. Brazil – City of Vitoria
3. Brazil – Juiz de Fora Secretary of Transport and Transit, Secretaria de Transporte e Transito
4. Brazil – Rio de Janeiro Center for Traffic Operations (COR)
5. Colombia – Bogotá Instituto de Desarrollo Urbano (IDU)
6. Colombia – Medellín Alcaldía de Medellín
7. Costa Rica – Ministry of Transport

8. México – City of Puebla
9. México – Delegación Miguel Hidalgo (México City)
10. México – La Sultana de Norte (Monterrey)
11. Perú – Municipalidad de Miraflores

Europe

1. Belgium – City of Ghent
2. Estonia – Tarktee (Smart Roads)
3. France – Department of Var
4. France – Northern France, Tollway Authority
5. Hungary – BKK Center for Budapest Transport
6. Latvia – City of Riga
7. Latvia – Latvia State Roads
8. Lithuania – Lithuania Road Administration
9. Netherlands – National Data Warehouse for Traffic Information
10. Portugal – Brisa/Via Verde (Portugal Tollway Authority)
11. Rome – Rome Center for Mobility
12. Spain – City of Barcelona
13. Spain – Government of Catalonia
14. United Kingdom – Transport for London

Middle East

1. Israel - City of Tel-Aviv
2. Israel - Holon Municipality

Asia-Pacific

1. Indonesia - City of Jakarta
2. Australia - Transportation Management Centre of New South Wales

Los objetivos del programa son reducir la congestión, aumentar la eficiencia de la respuesta a incidentes y tomar decisiones basadas en datos de lo que está sucediendo en la infraestructura vial, ya sean calles o carreteras.

El Programa de Ciudadanos Conectados de Waze es un intercambio de datos bidireccional que permite transferir información en ambas direcciones para ayudar

a la toma de decisiones y de este modo lograr un impacto concreto en la comunidad en cuanto a identificación de rutas alternativas al tráfico, localización de accidentes y eventos que afectan el transporte en las zonas urbanas y carreteras que las conectan.

Lanzado en octubre de 2014 con 10 socios de distintas ciudades, el programa se ha expandido a más de 100 socios, incluyendo agencias gubernamentales de ciudades, estados y países, así como organizaciones sin fines de lucro.

En este modo, *Waze* proporciona en tiempo real información sobre cierres de calles o carreteras, tráfico, rangos de velocidad y diversos eventos directamente desde el origen, en este caso, los propios conductores. El CCP al proporcionar más datos, da a los *Wazers* una mayor capacidad para evitar los cierres de carretera y los lugares con tráfico.

Los miembros de CCP reciben la información de los incidentes en tiempo real más rápido que otros métodos de transferencia de informes. *Waze* identifica de una manera precisa y verifica dónde ocurren los incidentes, creando tiempos de respuesta más rápidos. Los socios pueden utilizar estándares de datos diseñados por *Waze* para el cierre y la notificación de incidentes para reducir la fragmentación de datos y, de este modo, promover la agregación de datos de transporte así como de distintas instancias de gobierno.

En el programa CCP se encuentran disponibles las herramientas que permiten realizar algunas de las siguientes actividades:

- Especificación de incidentes.
- Herramienta de visualización de tráfico.
- Mensajes para aviso de tráfico inusual.
- Mapa en vivo de *Waze*.
- Herramienta de cierre de carretera.
- Foro de socios en línea.
- Herramienta para eventos de crisis o situaciones de emergencia, por ejemplo, tormentas o lluvias torrenciales.

Waze pone sus datos a disposición de los socios CCP a través de un *feed*²³ en formato JSON o XML que se actualiza cada 2 minutos.

3.3 Especificaciones de los datos de Waze

²³ Una fuente web o canal web (en inglés *web feed*) es un medio de redifusión de contenido web. <http://www.wordreference.com/es/translation.asp?tranword=feed>

En el documento *Waze Traffic Dataspec*²⁴ se detalla la estructura y el contenido del *feed*²⁵ de datos de tráfico en tiempo real de Waze. Los mapas de Waze proporcionan información sobre rutas específicas para ayudar a los automovilistas a evitar zonas de tráfico. El tráfico se actualiza dinámicamente mediante una red de asociados que cada día crece más. La representación más exacta es a través de polilíneas y, cuando el tráfico tiene una base geográfica para representarlos, es independiente de los atributos específicos del mapa.

Se proporciona información sobre tráfico y eventos que afectan las condiciones del camino, ya sea a través de los conductores que usan Waze (wazares) o mediante fuentes externas. Los *wazers*²⁶ proporcionan informes de ruta, o UGI (incidentes generados por el usuario), que incluyen información detallada sobre tráfico específico y condiciones del camino.

Se pueden emitir informes desde la ubicación en la que se encuentran actualmente o, si ya no está en la ubicación, dentro de los 30 minutos posteriores al evento. También se permite proporcionar alertas automáticas para lo que llamamos incidentes de tráfico inusual que afectan a una gran cantidad de usuarios y quedan fuera de los patrones de tráfico normales para un día y hora determinados.

El período de tiempo que afecta los resultados puede ser ajustado internamente por el servidor de Waze para proporcionar mejores resultados, en función de las condiciones locales. Esto incluyen las condiciones del clima, autos detenido en el sobre la vialidad, etc.

Los datos de tráfico de Waze constan de la siguiente información:

1. Información general: marca de tiempo del archivo, área geográfica desde la que se recuperaron los datos, etc.
2. Alertas de tráfico: incidentes de tráfico informados por los usuarios.
3. Atascos de tráfico: información de ralentización del tráfico generada por el servicio en función de la ubicación y la velocidad del usuario.
4. Tráfico inusual (irregularidades): alertas y atascos que afectan a un número excepcionalmente grande de usuarios.

Nota: de vez en cuando, se pueden agregar nuevos campos al *feed*. Una vez que esos campos estén permanentemente disponibles, este documento será modificado y redistribuido a los socios relevantes.

²⁴ Waze-traffic-dataspecv2.7.1.pdf

²⁵ Definición de *feed*

Alertas de tráfico

La sección 'Alertas' incluye todos los datos de tráfico informados por los usuarios a través de la aplicación móvil Waze.

Confiabilidad

Cada alerta obtiene un puntaje de confiabilidad basado en las reacciones de otros usuarios ('Pulgar hacia arriba', 'No hay', etc.) y el nivel de confianza del usuario (Los Wazers ganan niveles contribuyendo al mapa, comenzando en el nivel 1 y alcanzando hasta nivel 6. Cuanto mayor sea el nivel, más experimentado y confiable será el Wazer.) El puntaje (010) indica qué tan confiable es el informe.

Confianza

Cada alerta obtiene un puntaje de confianza basado en las reacciones de otros usuarios ('Pulgares arriba', 'No hay'). El puntaje oscila entre 0 y 10. Una puntuación más alta indica más comentarios positivos de los usuarios de Waze.

Los datos incluyen una matriz de alertas con los siguientes atributos:

Elemento	Valor	Descripción
pubDate	Hora	Fecha de publicación
Georss:point	Coordenadas	Localización por reporte (latitud – longitud)
Linqmap:uuid	Texto	ID único
Linqmap:magvar	Entero (0-359)	Dirección del evento (dirección del conductor al momento del reporte, 0 grados al norte de acuerdo con el dispositivo del conductor).
Linqmap:type	Ver la tabla de tipos de alerta	Tipo de evento
Linqmap:subtype	Ver la tabla de subtipos de alertas	Subtipo de evento
Linqmap:reportDescription	Texto	Descripción del reporte (cuando se encuentra disponible)
Linqmap:Street	Texto	Nombre de la calle (puede ser nulo)
Linqmap:city	Texto	Ciudad y nombre del estado
Linqmap:country	Texto	De acuerdo con los códigos de la

		norma ISO 3166-1
Linqmap:roadType	Entero	Tipo Carretera (ver la tabla de tipos de carretera)

Fuente: waze.com

Figura 3.3 Atributos de las alertas emitidas por la plataforma Waze.

Tipo Alerta	ScoreTipo	Subtipo Alerta
ACCIDENT	1	• ACCIDENT_MINOR • ACCIDENT_MAJOR • NO_SUBTYPE
JAM	2	• JAM_MODERATE_TRAFFIC • JAM_HEAVY_TRAFFIC • JAM_STAND_STILL_TRAFFIC • JAM_LIGHT_TRAFFIC • NO_SUBTYPE
WEATHERHAZARD / HAZARD	4/5	• HAZARD_ON_ROAD • HAZARD_ON_SHOULDER • HAZARD_WEATHER • HAZARD_ON_ROAD_OBJECT • HAZARD_ON_ROAD_POT_HOLE • HAZARD_ON_ROAD_ROAD_KILL • HAZARD_ON_SHOULDER_CAR_STOPPED • HAZARD_ON_SHOULDER_ANIMALS • HAZARD_ON_SHOULDER_MISSING_SIGN • HAZARD_WEATHER_FOG • HAZARD_WEATHER_HAIL • HAZARD_WEATHER_HEAVY_RAIN • HAZARD_WEATHER_HEAVY_SNOW • HAZARD_WEATHER_FLOOD • HAZARD_WEATHER_MONSOON • HAZARD_WEATHER_TORNADO • HAZARD_WEATHER_HEAT_WAVE • HAZARD_WEATHER_HURRICANE • HAZARD_WEATHER_FREEZING_RAIN • HAZARD_ON_ROAD_LANE_CLOSED

		• HAZARD_ON_ROAD_OIL • HAZARD_ON_ROAD_ICE • HAZARD_ON_ROAD_CONSTRUCTION • HAZARD_ON_ROAD_CAR_STOPPED • NO_SUBTYPE
ROAD_CLOSED	3	• ROAD_CLOSED_HAZARD • ROAD_CLOSED_CONSTRUCTION • ROAD_CLOSED_EVENT • NO_SUBTYPE

Fuente: waze.com

Figura 3.4 Tipos de alerta y subtipos relacionados en Waze.

Valor	Tipo
1	Calle
2	Calle principal
3	Autopistas
4	Rampas de acceso
5	Caminos
6	Primarias
7	Secundarias

Fuente: waze.com

Figura 3.5 Tipos de carretera definidos en la plataforma Waze.

3.3.1 Formas de distribución de datos

3.3.1.1 JSON (JavaScript Object Notation)

Es un formato para el intercambio de datos, básicamente JSON describe los datos con una sintaxis dedicada que se usa para identificar y gestionar los datos. JSON nació como una alternativa a XML. Una de las mayores ventajas que tiene el uso de JSON es que puede ser leído por cualquier lenguaje de programación. Por lo tanto, puede ser usado para el intercambio de información entre distintas tecnologías.

JSON (*JavaScript Object Notation* - Notación de Objetos de JavaScript) es un formato ligero de intercambio de datos. Está basado en un subconjunto del

Lenguaje de Programación JavaScript, Standard ECMA-262 3rd Edition de diciembre 1999²⁷. JSON es un formato de texto que es completamente independiente del lenguaje pero utiliza convenciones que son ampliamente conocidas por los programadores de la familia de lenguajes C, incluyendo C, C++, C#, Java, JavaScript, Perl, Python, y muchos otros. Estas propiedades hacen que JSON sea un lenguaje ideal para el intercambio de datos.

JSON está constituido por dos estructuras:

Una colección de pares de nombre/valor. En varios lenguajes esto es conocido como un objeto, registro, estructura, diccionario, tabla hash, lista de claves o un arreglo asociativo.

Una lista ordenada de valores. En la mayoría de los lenguajes, esto se implementa como arreglos, vectores, listas o secuencias.

Estas son estructuras universales; virtualmente todos los lenguajes de programación las soportan de una forma u otra. Es razonable que un formato de intercambio de datos, que es independiente del lenguaje de programación, se base en estas estructuras.

Un archivo JSON, se define de esta forma:

Un objeto o conjunto desordenado de pares nombre/valor. Un objeto comienza con { (llave de apertura) y termine con } (llave de cierre). Cada nombre es seguido por : (dos puntos) y los pares nombre/valor están separados por , (coma).²⁸

3.3.1.2 Muestra de datos JSON generados por Waze

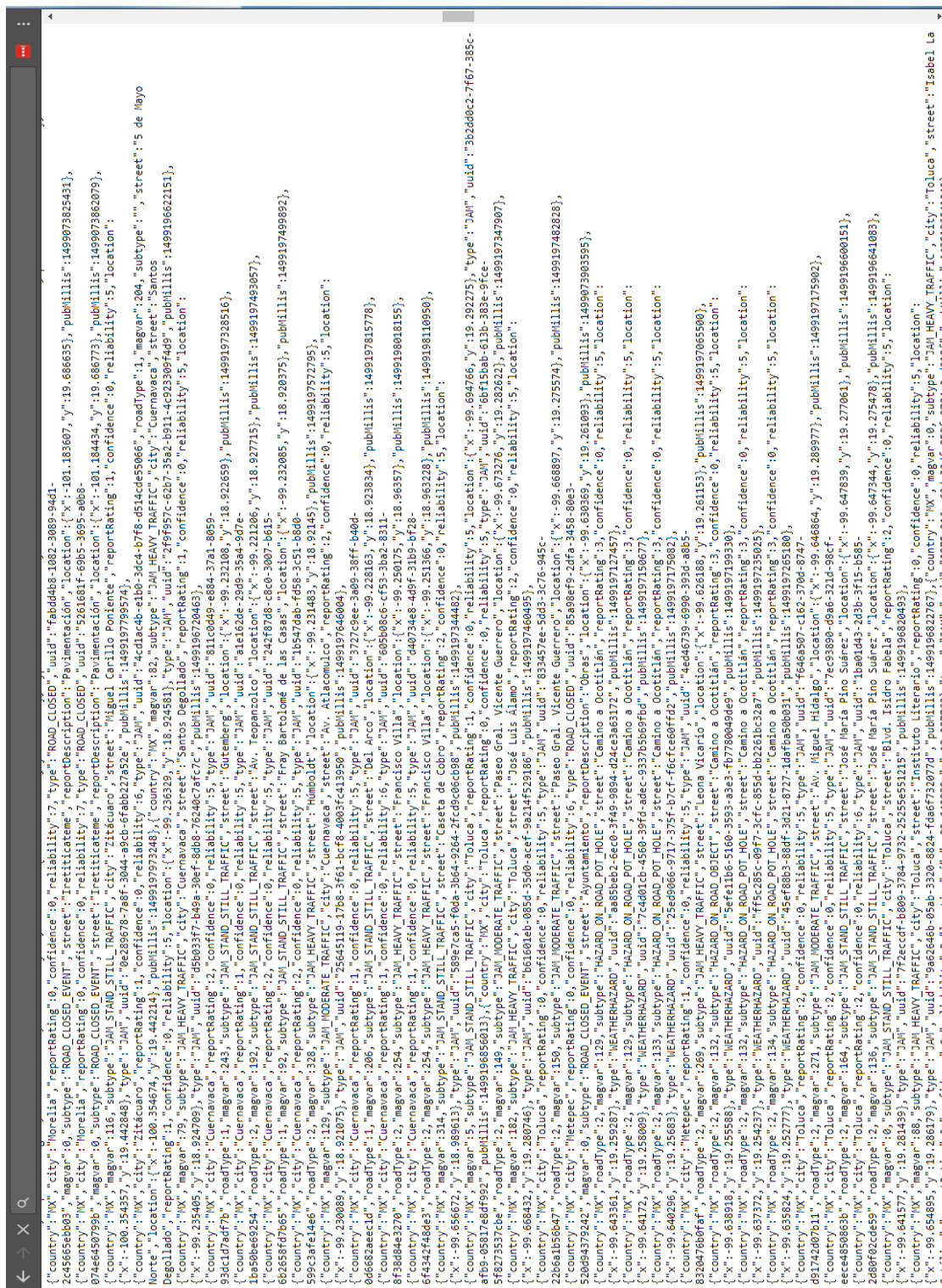
Ejemplo:

```
{ "country": "MX", "roadType": 2, "magvar": 44, "subtype": "JAM_STAND_STILL_TRAFFIC", "reportRating": 0, "confidence": 0, "reliability": 5, "location": { "x": -99.22162, "y": 19.32204 }, "type": "JAM", "uuid": "191035d1-e70b-3cbc-9847-76d85652cee1", "pubMillis": 1499197477095 }, { "country": "MX", "city": "Magdalena Contreras", "reportRating": 0, "confidence": 0, "reliability": 5, "type": "JAM", "uuid": "414f68bb-8c82-3ee7-8f81-d182b6a406b6", "roadType": 2, "magvar": 37, "subtype": "JAM_HEAVY_TRAFFIC", "street": "Av . México", "location": { "x": -99.222258, "y": 19.321257 }, "pubMillis": 1499197658949 }, { "country": "MX", "city": "Magdalena
```

²⁷ <http://www.json.org/json-es.html>

²⁸ Ibidem.

```
Contreras", "reportRating":1, "confidence":0, "reliability":5, "type":"JAM", "uuid":"8
80586d1-fc2b-391f-b518-
feela8f67233", "roadType":2, "magvar":42, "subtype":"JAM_STAND_STILL_TRAFFIC", "stree
t":"Av. Contreras", "location":{"x":-
99.21805, "y":19.324541}, "pubMillis":1499197717544}, {"country":"MX", "city":"Magdal
ena
Contreras", "reportRating":2, "confidence":0, "reliability":5, "type":"JAM", "uuid":"1
23bc295-e781-310b-8396-
b79414aca889", "roadType":1, "magvar":149, "subtype":"JAM_STAND_STILL_TRAFFIC", "stre
et":"Veracruz", "location":{"x":-
99.227671, "y":19.319699}, "pubMillis":1499197720769}
```



Fuente: waze.com

Figura 3.6 Captura de pantalla del *feed* en formato Json provenientes de Waze.

3.3.2 XML (*eXtensible Markup Language*)

XML, (siglas en inglés de *eXtensible Markup Language*), traducido como "Lenguaje de Marcado Extensible" o "Lenguaje de Marcas Extensible", es un meta-lenguaje que permite definir lenguajes de marcas desarrollado por el World Wide Web Consortium (W3C) utilizado para almacenar datos en forma legible. Proviene del lenguaje SGML y permite definir la gramática de lenguajes específicos (de la misma manera que HTML es a su vez un lenguaje definido por SGML) para estructurar documentos grandes²⁹. A diferencia de otros lenguajes, XML da soporte a bases de datos, siendo útil cuando varias aplicaciones deben comunicarse entre sí o integrar información.

XML es un estándar para el intercambio de información estructurada entre diferentes plataformas. Se puede usar en bases de datos, editores de texto, hojas de cálculo. XML es una tecnología sencilla que tiene a su alrededor otras que la complementan y la hacen mucho más grande, con unas posibilidades mucho mayores. Tiene un papel muy importante en la actualidad ya que permite la compatibilidad entre sistemas para compartir la información de una manera segura, fiable y fácil.

Un archivo XML se representa de la siguiente forma:

- Elementos: Pieza lógica del marcado, se representa con una cadena de texto (dato) encerrada entre etiquetas. Pueden existir elementos vacíos (
). Los elementos pueden contener atributos.
- Instrucciones: órdenes especiales para ser utilizadas por la aplicación que procesa <?xml-stylesheet type="text/css" href="estilo.css">
- Las instrucciones XML. Comienzan por <? Y terminan por ?>.
- Comentarios: Información que no forma parte del documento. Comienzan por <!-- y terminan por -->.
- Declaraciones de tipo: Especifican información acerca del documento. <!DOCTYPE persona SYSTEM "persona.dtd">
- Secciones CDATA: Se trata de un conjunto de caracteres que no deben ser interpretados por el procesador. <![CDATA[Aquí se puede meter cualquier carácter, como <, &, >, ... Sin que sean interpretados como marcación]]³⁰>

²⁹ <https://www.w3.org/XML/>

³⁰ Ibidem.

3.3.2.1 Muestra de datos XML generados por Waze

```

<rss xmlns:georss="http://www.georss.org/georss"
xmlns:linqmap="http://www.linqmap.com" version="2.0">
<channel>
<title>GeoRSS</title>
<description>GeoRSS</description>
<georss:box>-85.000000 -179.000000 85.000000 179.000000</georss:box>
<linqmap:time>
Fri May 26 17:58:00 +0000 2017,Fri May 26 17:59:00 +0000 2017
</linqmap:time>
<item>
<title>alert</title>
<pubDate>Tue Jul 4 20:35:54 +0000 2017</pubDate>
<georss:point>19.393525 -99.135527</georss:point>
<linqmap:uuid>c400896b-3692-3510-823e-9ed0b07f0a7c</linqmap:uuid>
<linqmap:magvar>97</linqmap:magvar>
<linqmap:type>WEATHERHAZARD</linqmap:type>
<linqmap:subtype>HAZARD_ON_ROAD_CAR_STOPPED</linqmap:subtype>
<linqmap:street>Eje 4 Sur - Napoleón (carril derecho)</linqmap:street>
<linqmap:city>Benito Juárez</linqmap:city>
<linqmap:country>MX</linqmap:country>
<linqmap:roadType>6</linqmap:roadType>
<linqmap:reportRating>0</linqmap:reportRating>
<confidence>0</confidence>
<linqmap:reliability>5</linqmap:reliability>
</item>
<item>
<title>alert</title>
<pubDate>Tue Jul 4 20:36:06 +0000 2017</pubDate>
<linqmap:uuid>227e18c9-bd22-3408-9a4b-69845c94d0de</linqmap:uuid>
<linqmap:magvar>126</linqmap:magvar>
<linqmap:type>JAM</linqmap:type>
<linqmap:subtype>JAM_STAND_STILL_TRAFFIC</linqmap:subtype>
<linqmap:street>Eje 1 Nte. - Fuerza Aérea Mexicana</linqmap:street>
<linqmap:city>Venustiano Carranza</linqmap:city>
<linqmap:country>MX</linqmap:country>
<linqmap:roadType>6</linqmap:roadType>
<linqmap:reportRating>3</linqmap:reportRating>
<confidence>0</confidence>
<linqmap:reliability>5</linqmap:reliability>
</item>
<item>
<title>alert</title>
<pubDate>Tue Jul 4 20:37:18 +0000 2017</pubDate>
<georss:point>19.418267 -99.079619</georss:point>
<linqmap:uuid>d4512759-ce60-3819-b12a-10244f3ddfbe</linqmap:uuid>
<linqmap:magvar>126</linqmap:magvar>
<linqmap:type>JAM</linqmap:type>
<linqmap:subtype>JAM_STAND_STILL_TRAFFIC</linqmap:subtype>
<linqmap:street>Eje 1 Nte. - Fuerza Aérea Mexicana</linqmap:street>
<linqmap:city>Venustiano Carranza</linqmap:city>
<linqmap:country>MX</linqmap:country>
<linqmap:roadType>6</linqmap:roadType>

```

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml version="1.0" encoding="UTF-8" ?>
<rss xmlns:georss="http://www.georss.org/georss" xmlns:lingmap="http://www.lingmap.com" version="2.0">
  <channel>
    <title>Georss</title>
    <description>Georss</description>
    <georss:box>12.243391 -119.619140 34.334364 -85.869140</georss:box>
    <lingmap:time>
      Tue Jul 4 20:36:00 +0000 2017, Tue Jul 4 20:37:00 +0000 2017
    </lingmap:time>
  </channel>
  <item>
    <title>Alert</title>
    <pubDate>Mon Jul 3 09:24:09 +0000 2017</pubDate>
    <georss:point>19.513276 -101.611448</georss:point>
    <lingmap:uid>fa70d76a-c862-3d47-8e97-c7085ade683e</lingmap:uid>
    <lingmap:magvar>0</lingmap:magvar>
    <lingmap:type>ROAD_CLOSED</lingmap:type>
    <lingmap:subtype>ROAD_CLOSED_EVENT</lingmap:subtype>
    <lingmap:reportDescription>Mantenimiento vial</lingmap:reportDescription>
    <lingmap:street>Ponce de León</lingmap:street>
    <lingmap:city>Pátzcuaro</lingmap:city>
    <lingmap:country>MX</lingmap:country>
    <lingmap:reportRating>0</lingmap:reportRating>
    <confidence>0</confidence>
    <lingmap:reliability>6</lingmap:reliability>
  </item>
  <item>
    <title>Alert</title>
    <pubDate>Mon Jul 3 09:23:35 +0000 2017</pubDate>
    <georss:point>19.513342 -101.612019</georss:point>
    <lingmap:uid>4053450e-474f-3806-b582-3ffb5e873d1</lingmap:uid>
    <lingmap:magvar>0</lingmap:magvar>
    <lingmap:type>ROAD_CLOSED</lingmap:type>
    <lingmap:subtype>ROAD_CLOSED_EVENT</lingmap:subtype>
    <lingmap:reportDescription>Mantenimiento vial</lingmap:reportDescription>
    <lingmap:street>Ponce de León</lingmap:street>
    <lingmap:city>Pátzcuaro</lingmap:city>
    <lingmap:country>MX</lingmap:country>
    <lingmap:reportRating>0</lingmap:reportRating>
    <confidence>0</confidence>
    <lingmap:reliability>6</lingmap:reliability>
  </item>
  <item>
    <title>Alert</title>
    <pubDate>Mon Jul 3 09:23:37 +0000 2017</pubDate>
    <georss:point>19.513276 -101.611448</georss:point>
    <lingmap:uid>9a5dea09-6d03-3d00-9014-b567db7a08d6</lingmap:uid>
    <lingmap:magvar>0</lingmap:magvar>
    <lingmap:type>ROAD_CLOSED</lingmap:type>
    <lingmap:subtype>ROAD_CLOSED_EVENT</lingmap:subtype>
    <lingmap:reportDescription>Mantenimiento vial</lingmap:reportDescription>
    <lingmap:street>Ponce de León</lingmap:street>
    <lingmap:city>Pátzcuaro</lingmap:city>
    <lingmap:country>MX</lingmap:country>
    <lingmap:reportRating>0</lingmap:reportRating>
    <confidence>0</confidence>
    <lingmap:reliability>6</lingmap:reliability>
  </item>
  <item>
    <title>Alert</title>
    <pubDate>Mon Jul 3 09:26:59 +0000 2017</pubDate>
    <georss:point>19.51313 -101.610185</georss:point>
```

Fuente: waze.com

Figura 3.7 Captura de pantalla del feed en formato XML.

3.4 Proceso y transformación de datos

El proceso de los datos se refiere a los pasos que se deben realizar para generar capas de información geoespacial, ya que la cantidad de datos que se genera es muy grande y estos se actualizan de manera continua (cada media hora), para el propósito del presente proyecto se seleccionó un periodo de tiempo para procesar solamente los datos de ese intervalo.

El periodo seleccionado fue del 4 de mayo de 2017 al 31 julio del mismo año. Este conjunto de datos se descargó, transformó, revisó, editó, se procesó y convirtió en información geoespacial compatible con el software ArcGIS.

Finalmente, se aplicó análisis espacial a través de un Sistema de Información Geográfica para obtener mapas de densidad, identificación de principales eventos relacionados a riesgos que podrían afectar las carreteras; así mismo se pudieron observar las tendencias de saturación de tráfico y eventos diversos de acuerdo con la hora y día de la semana.

3.4.1 Transformación de XML a XLS

Los datos crudos se descargan y recaban el siguiente *feed*, el cual fue proporcionado por Waze al IMT de acuerdo al convenio celebrado, https://world-geoss.waze.com/rtservers/web/TGeoSStk=ccp_partner&ccp_partner_name=SantiagoQueretaro&format=XML&types=traffic,alerts,irregularities&polygon=-117.961,33.588;-102.126,33.036;-87.01,33.235;-85.712,12.582;-116.866,13.644;-117.961,33.588

Los archivos se renombran de acuerdo al día, año y hora de descarga, las cuales para esta aplicación son cada 2 horas a partir de las 0500 hrs y hasta las 2300 del mismo día, todos los días de la semana, durante el periodo indicado anteriormente.

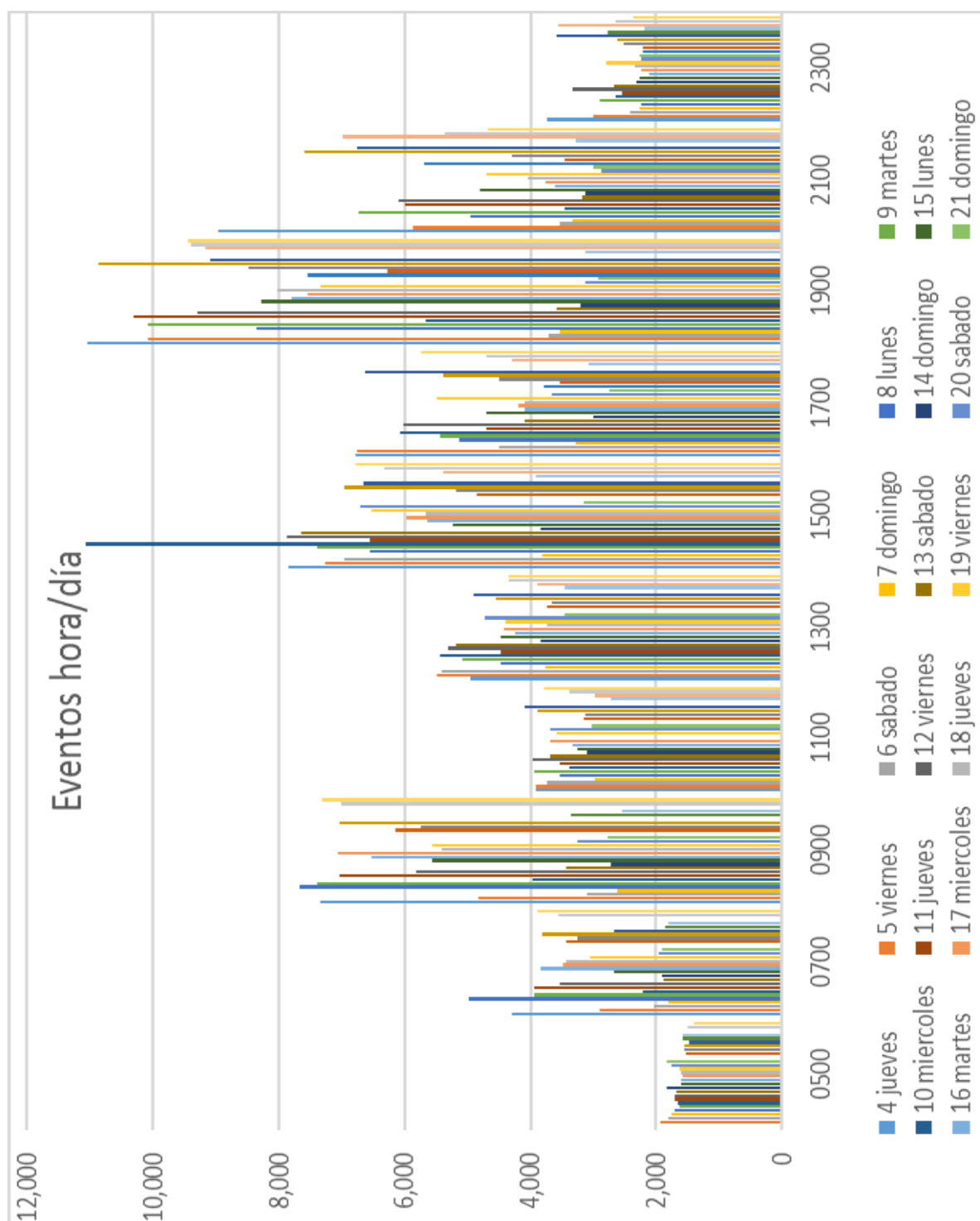
El siguiente paso es transformar los archivos resultantes del formato XML a formato XLS.

Lecturas de usuarios de Waze, incluye puntos y líneas			Hora										total
Mes	Día		1	2	3	4	5	6	7	8	9	10	
			0500	0700	0900	1100	1300	1500	1700	1900	2100	2300	
Mayo	4	jueves	0	4,286	7,344	3,907	4,943	7,859	6,781	11,046	8,955	3,741	58,862
	5	viernes	1,921	2,891	4,825	3,915	5,484	7,268	6,760	10,079	5,854	3,000	51,997
	6	sabado	1,806	2,027	3,105	3,734	5,405	6,955	4,490	3,718	3,520	2,419	37,179
	7	domingo	1,766	1,795	2,622	2,976	3,757	3,799	3,265	3,533	3,333	2,259	29,105
	8	lunes	1,710	4,983	7,670	3,525	4,479	6,545	5,118	8,344	4,952	2,240	49,566
	9	martes	1,619	3,926	7,402	3,948	5,078	7,396	5,432	10,088	6,732	2,894	54,515
	10	miercoles	1,641	2,219	3,974	3,370	5,430	11,076	6,058	5,674	3,453	2,646	45,541
	11	jueves	1,712	3,942	7,043	3,537	4,463	6,556	4,691	10,307	5,985	2,538	50,774
	12	viernes	1,694	3,537	5,819	3,959	5,312	7,878	6,027	9,304	6,106	3,324	52,960
	13	sabado	1,680	1,873	3,438	3,694	5,191	7,646	4,086	3,589	3,175	2,660	37,032
	14	domingo	1,827	1,894	2,714	3,107	3,827	3,846	2,994	3,200	3,133	2,312	28,854
	15	lunes	1,610	2,676	5,563	3,241	4,465	5,237	4,706	8,289	4,805	2,253	42,845
	16	martes	1,608	3,843	6,517	3,321	4,251	5,633	4,098	7,785	3,618	2,117	42,791
	17	miercoles	1,585	3,490	7,056	3,691	4,413	5,975	4,193	7,536	3,758	2,224	43,921
	18	jueves	1,593	3,432	5,421	0	3,742	5,654	4,080	8,025	4,036	2,335	38,318
	19	viernes	1,615	3,043	5,565	3,570	4,406	6,530	5,496	7,346	4,689	2,793	45,053
	20	sabado	1,750	1,963	3,244	3,675	4,726	6,706	3,661	3,124	2,864	2,224	33,937
	21	domingo	1,827	1,907	2,758	3,019	3,450	3,157	2,751	2,926	3,010	2,252	27,057
	22	lunes	0	0	0	0	0	0	3,795	7,530	5,693	2,213	19,231
	23	martes	1,524	3,421	6,140	3,160	3,726	4,852	3,541	6,267	3,466	2,222	38,319
	24	miercoles	1,548	3,241	5,744	3,124	3,670	5,174	4,496	8,477	4,294	2,511	42,279
	25	jueves	1,546	3,799	7,045	3,894	4,553	6,954	5,390	10,857	7,605	2,617	54,260
	26	viernes	1,472	2,663	0	4,082	4,904	6,655	6,637	9,085	6,745	3,574	45,817
	27	sabado	1,588	1,854	3,352	0	0	0	0	0	0	2,766	9,560
	28	domingo	1,565	1,809	2,541	2,712	3,462	3,908	3,082	3,133	3,289	2,174	27,675
	29	lunes	0	0	0	2,979	3,877	5,376	4,298	9,178	6,972	3,568	36,248
	30	martes	1,507	3,549	7,019	3,391	4,354	6,334	4,704	9,407	5,353	2,638	48,256
	31	miercoles	1,409	3,884	7,306	3,792	4,354	6,780	5,731	9,450	4,684	2,369	49,759
			41,123	77,947	131,227	87,323	115,722	161,749	126,361	197,297	130,079	72,883	1,141,711

Fuente: Elaboración propia

Figura 3.8 Tabla resumen de las lecturas generadas por la plataforma Waze, para Mayo 2017.

Se observa que la mayor lectura acumulada para el mes de mayo fue el día 4, mientras que la hora con el mayor número de lecturas fue las 1900 horas del mismo día. Los horarios que presentan el mayor número de lecturas son las 0900 hrs, las 1500 horas y las 1900. Los días con el mayor número son los jueves y viernes. Los horarios con menor número de lecturas son las 0500 horas, 1100 horas y 23 horas. Los días con el menor número de lecturas son los sábados y domingos. La suma acumulada para los horarios y días de mayo con datos fue de 1,141,711 lecturas. De este conjunto se seleccionan los eventos que representan puntos y posteriormente los que indican algún riesgo, para mayo el subconjunto resultante es de 16,661 registros.



Fuente: Elaboración propia

Figura 3.9 Gráfica de la Tabla resumen de las lecturas generadas por la plataforma Waze, para Mayo 2017.

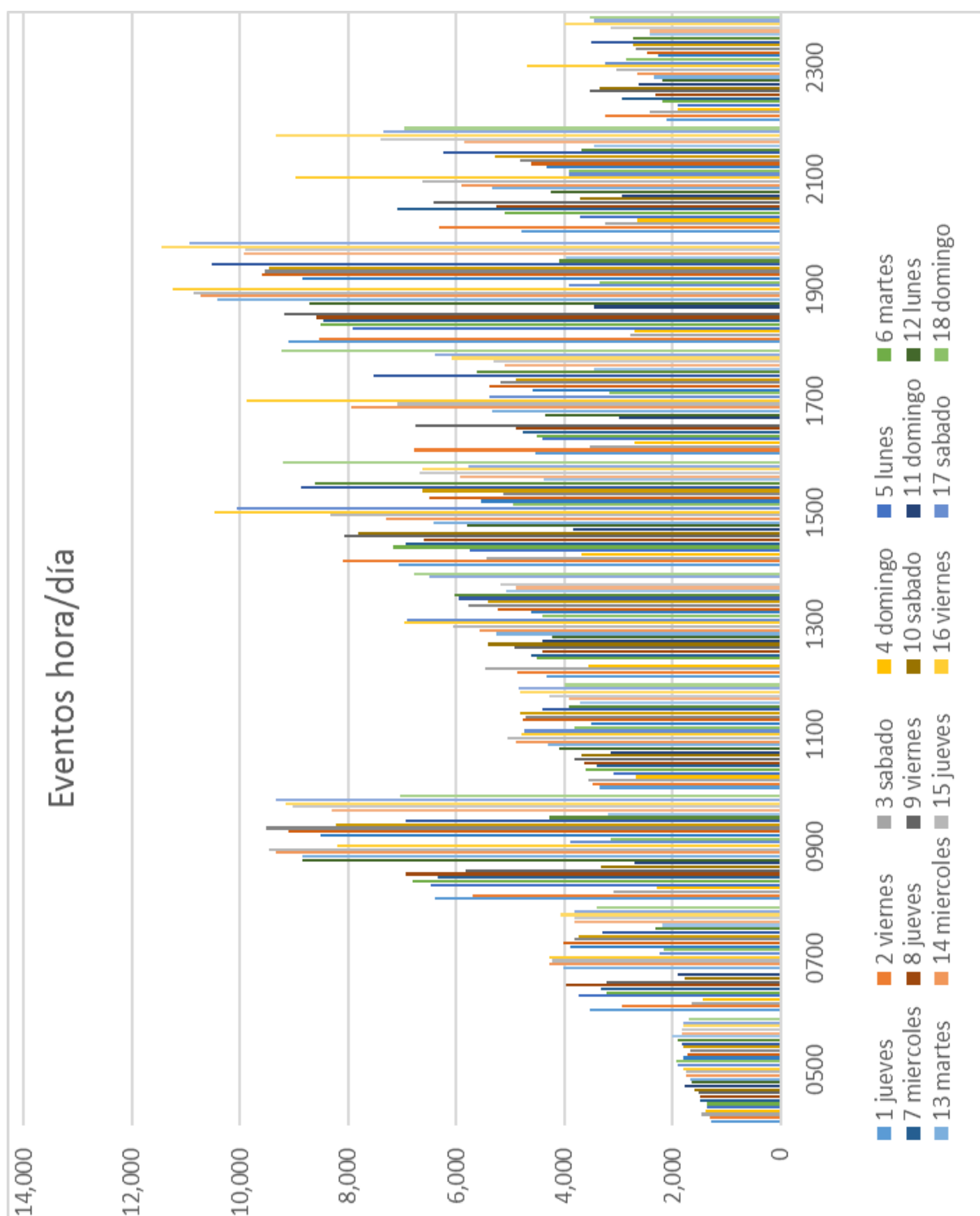
Lecturas de usuarios de Waze, incluye puntos y líneas			Hora										total
			1	2	3	4	5	6	7	8	9	10	
Mes	Día		0500	0700	0900	1100	1300	1500	1700	1900	2100	2300	
Junio	1	jueves	1,293	3,527	6,392	3,350	4,322	7,071	4,540	9,118	4,786	2,105	46,504
	2	viernes	1,308	2,948	5,700	3,484	4,862	8,092	6,788	8,535	6,331	3,240	51,288
	3	sabado	1,456	1,638	3,086	3,550	5,478	5,443	3,534	2,794	3,242	2,425	32,646
	4	domingo	1,382	1,442	2,301	2,689	3,558	3,696	2,696	2,711	2,664	1,916	25,055
	5	lunes	1,353	3,733	6,484	3,084	0	5,744	4,416	7,925	3,712	1,909	38,360
	6	martes	1,367	3,228	6,812	3,600	4,518	7,166	4,516	8,520	5,116	2,191	47,034
	7	miercoles	1,479	3,331	6,342	3,396	4,620	6,942	4,759	8,455	7,084	2,941	49,349
	8	jueves	1,490	3,960	6,927	3,634	4,418	6,608	4,902	8,585	5,272	2,325	48,121
	9	viernes	1,521	3,227	5,826	3,827	4,935	8,076	6,770	9,174	6,417	3,537	53,310
	10	sabado	1,592	1,786	3,330	3,694	5,412	7,808	0	0	3,723	3,362	30,707
	11	domingo	1,769	1,897	2,716	3,146	4,398	3,847	2,983	3,441	2,934	2,634	29,765
	12	lunes	1,643	0	8,841	4,101	4,235	5,794	4,354	8,725	4,247	2,186	44,126
	13	martes	1,667	4,028	8,844	4,298	5,257	6,427	5,346	10,413	5,337	2,338	53,955
	14	miercoles	1,738	4,286	9,341	4,901	5,563	7,287	7,943	10,720	5,916	2,646	60,341
	15	jueves	1,761	4,233	9,460	5,059	6,054	8,335	7,101	10,866	6,630	3,030	62,529
	16	viernes	1,792	4,275	8,214	4,807	6,962	10,487	9,878	11,240	8,975	4,683	71,313
	17	sabado	1,911	2,234	3,892	4,734	6,916	10,069	5,388	3,930	3,914	3,250	46,238
	18	domingo	1,924	2,151	3,131	3,826	4,398	4,951	3,168	3,349	3,928	2,871	33,697
	19	lunes	1,792	3,887	8,520	3,493	4,624	5,531	4,590	8,853	4,328	2,259	47,877
	20	martes	1,724	4,018	9,105	4,767	5,229	6,491	5,394	9,604	4,606	2,481	53,419
	21	miercoles	1,683	3,815	9,509	4,710	5,775	5,122	5,177	9,548	4,831	2,678	52,848
	22	jueves	1,805	3,731	8,239	4,809	5,424	6,621	4,887	9,465	5,286	2,737	53,004
	23	viernes	1,820	3,307	6,927	4,405	5,966	8,882	7,536	10,530	6,231	3,501	59,105
	24	sabado	1,904	2,307	4,267	3,929	6,031	8,623	5,628	4,091	3,684	2,729	43,193
	25	domingo	1,997	2,194	3,205	3,717	5,078	4,371	3,459	3,981	3,454	2,428	33,884
	26	lunes	1,816	3,815	8,304	3,925	4,898	5,928	5,101	9,941	5,847	2,429	52,004
	27	martes	1,815	3,803	9,023	4,284	5,191	6,681	5,299	9,910	7,391	3,144	56,541
	28	miercoles	1,794	4,079	9,146	4,821	0	6,632	6,087	11,443	9,351	3,987	57,340
	29	jueves	1,811	3,826	9,349	4,847	6,510	5,783	6,385	10,936	7,363	3,465	60,275
	30	viernes	1,689	3,391	7,051	3,983	6,774	9,214	9,228	0	6,966	3,535	51,831
			50,096	94,097	200,284	120,870	147,406	203,722	157,853	226,803	159,566	84,962	1,445,659

Fuente: Elaboración propia

Figura 3.10 Tabla resumen de las lecturas generadas por la plataforma Waze, para Junio 2017.

Se observa que la mayor lectura acumulada para el mes de junio fue el día 16, mientras que la hora con el mayor número de lecturas fue las 1900 horas del mismo día.

Los horarios que presentan el mayor número de lecturas son las 0900 hrs, las 1500 horas y las 1900. Los días con el mayor número son los jueves y viernes. Los horarios con menor número de lecturas son las 0500 horas, 1100 horas y 23 horas. Los días con el menor número de lecturas son los sábados y domingos. La suma acumulada para los horarios y días de junio con datos fue de 1,445,659 lecturas.



Fuente: Elaboración propia

Figura 3.11 Gráfica de la Tabla resumen de las lecturas generadas por la plataforma Waze, para Junio 2017.

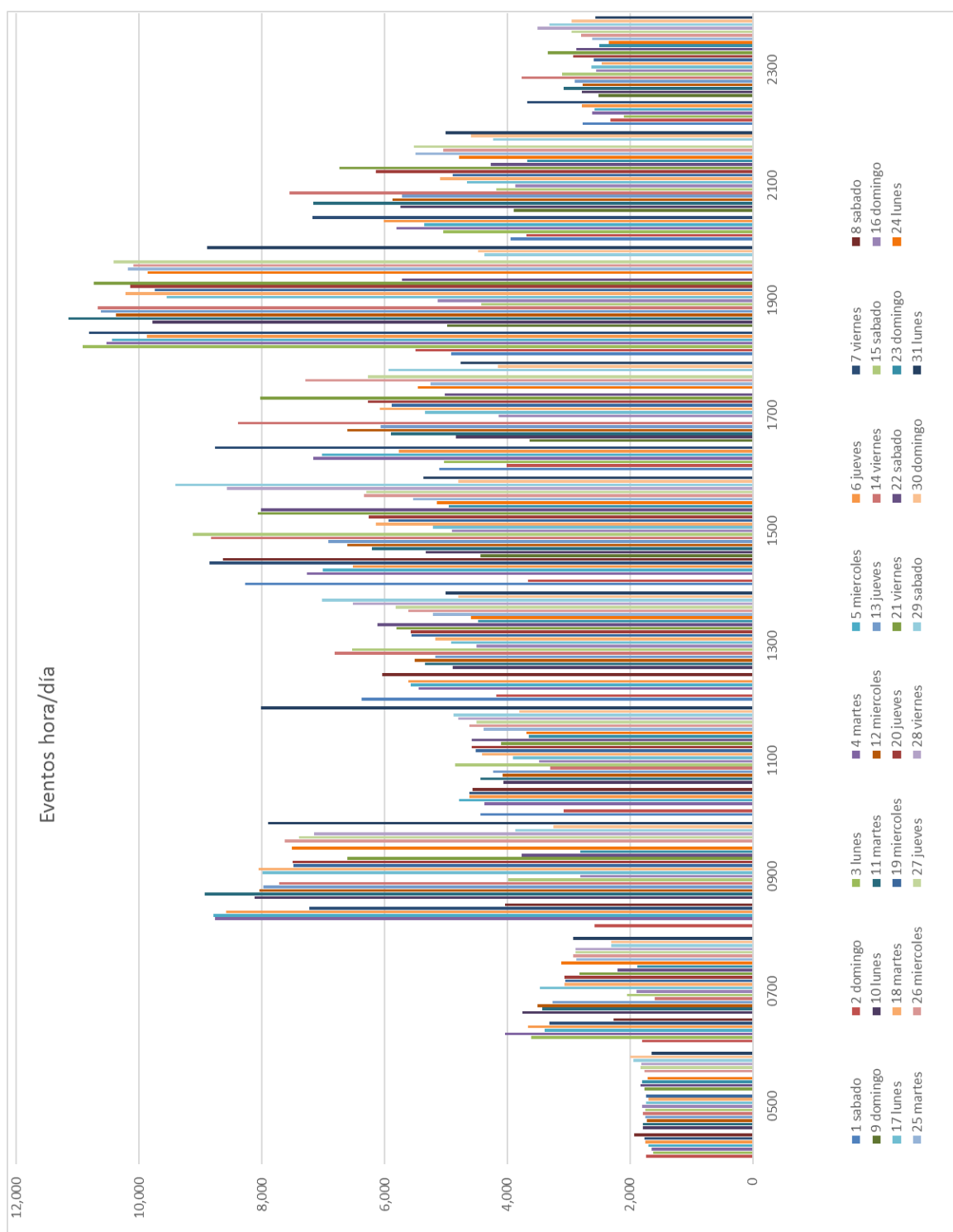
Lecturas de usuarios de Waze, incluye puntos y líneas			Hora										total
			1	2	3	4	5	6	7	8	9	10	
Mes			0500	0700	0900	1100	1300	1500	1700	1900	2100	2300	
Julio	1	sabado	0	0	0	4,434	6,374	8,271	5,109	4,909	3,945	2,774	35,816
	2	domingo	1,739	1,799	2,574	3,080	4,176	3,661	4,008	5,495	3,687	2,324	32,543
	3	lunes	1,617	3,610	0	0	0	0	5,035	10,923	5,040	2,097	28,322
	4	martes	1,653	4,034	8,764	4,366	5,438	7,267	7,157	10,527	5,801	2,622	57,629
	5	miercoles	1,699	3,395	8,786	4,780	5,568	7,001	7,022	10,447	5,355	2,571	56,624
	6	jueves	1,745	3,657	8,579	4,624	5,609	6,513	5,769	9,878	6,006	2,788	55,168
	7	viernes	1,759	3,307	7,225	4,623	0	8,859	8,769	10,821	7,178	3,671	56,212
	8	sabado	1,930	2,271	4,033	4,566	6,044	8,629	0	0	0	0	27,473
	9	domingo	0	0	0	0	0	4,441	3,640	4,985	3,892	2,519	19,477
	10	lunes	1,788	3,757	8,115	4,066	4,895	5,326	4,831	9,777	5,735	2,784	51,074
	11	martes	1,795	3,427	8,930	4,432	5,336	6,209	5,898	11,152	7,165	3,075	57,419
	12	miercoles	1,730	3,508	8,041	4,071	5,507	6,603	6,608	10,377	5,869	2,772	55,086
	13	jueves	1,755	3,260	7,969	4,226	5,172	6,917	6,065	10,617	5,717	2,906	54,604
	14	viernes	1,796	1,590	7,717	3,306	6,807	8,826	8,395	10,669	7,545	3,768	60,419
	15	sabado	1,745	2,043	3,991	4,854	6,533	9,119	0	4,427	4,174	3,112	39,998
	16	domingo	1,799	1,899	2,814	3,478	4,507	4,900	4,146	5,130	3,863	2,554	35,090
	17	lunes	1,737	3,468	7,994	3,908	4,915	5,216	5,337	9,545	4,654	2,629	49,403
	18	martes	1,696	3,062	8,058	4,406	5,170	6,142	6,073	10,228	5,090	2,462	52,387
	19	miercoles	1,737	3,052	7,487	4,520	5,557	5,937	5,884	9,743	4,883	2,587	51,387
	20	jueves	0	3,072	7,497	4,579	5,576	6,260	6,265	10,146	6,136	2,928	52,459
	21	viernes	1,759	2,819	6,604	4,106	5,805	8,065	8,030	10,738	6,731	3,339	57,996
	22	sabado	1,826	2,203	3,759	4,574	6,115	8,013	5,023	5,718	4,263	2,874	44,368
	23	domingo	1,798	1,882	2,814	3,644	4,482	4,954	0	0	3,680	2,500	25,754
	24	lunes	1,713	3,116	7,508	3,686	4,593	5,147	5,462	9,856	4,783	2,339	48,203
	25	martes	0	2,869	0	4,388	5,209	5,539	5,248	10,189	5,492	2,612	41,546
	26	miercoles	1,758	2,920	7,624	4,615	5,606	6,339	7,295	10,096	5,040	2,800	54,093
	27	jueves	1,823	2,881	7,400	4,496	5,816	6,293	6,276	10,418	5,519	2,951	53,873
	28	viernes	1,819	2,890	7,155	4,795	6,518	8,566	0	0	0	3,503	35,246
	29	sabado	1,951	2,306	3,871	4,876	7,014	9,406	5,938	4,371	4,236	3,318	47,287
	30	domingo	1,994	2,301	3,250	3,804	4,802	4,793	4,151	4,481	4,587	2,949	37,112
	31	lunes	1,649	2,928	7,904	8,011	5,003	5,365	4,760	8,897	4,999	2,570	52,086
			47,810	83,326	176,463	127,314	154,147	198,577	158,194	244,560	151,065	84,698	1,426,154

Fuente: Elaboración propia

Figura 3.12 Tabla resumen de las lecturas generadas por la plataforma Waze, para Julio 2017.

Se observa que la mayor lectura acumulada para el mes de julio fue el día 14, mientras que la hora con el mayor número de lecturas fue las 1900 horas del día 11.

Los horarios que presentan el mayor número de lecturas son las 0900 hrs, las 1500 horas y las 1900. Los días con el mayor número son los martes, jueves y viernes. Los horarios con menor número de lecturas son las 0500 horas, 1100 horas y 23 horas. Los días con el menor número de lecturas son los sábados y domingos. La suma acumulada para los horarios y días de junio con datos fue de 1,426,154 lecturas.



Fuente: Elaboración propia

Figura 3.13 Gráfica de la Tabla resumen de las lecturas generadas por la plataforma Waze, para Julio 2017.

Dentro de Excel, se ejecuta el siguiente código, el cual suma las tablas de cada una de las horas en una sola tabla por día.

```
*****Aplicacion Unificar archivos de excel(*.xlsx) en un solo libro-----  
*****Desarrollado por: Freddy Maihuire*****  
*****Sitio web: http://www.cursoslab.com*****
```

Sub Open_Files()

Dim Hoja As Object

Application.ScreenUpdating = False

'Definir la variable como tipo Variante

Dim X As Variant

'Abrir cuadro de dialogo

X = Application.GetOpenFilename _

("Excel Files (*.xlsx), *.xlsx", 2, "Abrir archivos", , True)

'Validar si se seleccionaron archivos

If IsArray(X) Then ' Si se seleccionan

'Crea Libro nuevo

Workbooks.Add

'Captura nombre de archivo destino donde se grabaran los archivos seleccionados

A = ActiveWorkbook.Name

*/*****

For y = LBound(X) To RBound(X)

Application.StatusBar = "Importando Archivos: " & X(y)

Workbooks.Open X(y)

b = ActiveWorkbook.Name

For Each Hoja In ActiveWorkbook.Sheets

Hoja.Copy after:=Workbooks(A).Sheets(Workbooks(A).Sheets.Count)

Next

Workbooks(b).Close False

Next

Application.StatusBar = "Listo"

```
Call Unir_Hojas
End If
Application.ScreenUpdating = False
End Sub

Sub Unir_Hojas()
Dim Sig As Byte, Eliminar As Boolean
For Sig = 2 To Worksheets.Count
    Worksheets(Sig).UseRange.Copy _
    Worksheets(1).Range("a1000000").End(xlUp).Offset(1)
Next
Application.DisplayAlerts = False
For Sig = 2 To Worksheets.Count
    Worksheets(2).Delete
Next
Application.DisplayAlerts = True

End Sub
```

Fuente: www.cursoslab.com

Figura 3.14 Código fuente para sumar los archivos de Excel en un solo archivo por día.

Una vez que se han sumado los archivos de cada hora en un archivo por día, se pueden agregar por semana o por mes, aunque esto incrementará el volumen de datos a manejar y los procesos posteriores se ralentizarán y consumirán más tiempo.

Para esta aplicación se sumaron en un archivo por mes, por lo que el resultado fue un archivo xls para mayo, junio y julio.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	version	title	descriptio	ns2:box	ns1:time	title2	p	ns2:point	ns1:uuid	ns1:										
7086	31mayo	0900	GeoRSS	12.243391 -1	Wed May 31	alert	Wed	19.407393 -99.268318	e3584c76-45											
7087	31mayo	0900	GeoRSS	12.243391 -1	Wed May 31	alert	Wed	19.409112 -99.253546	a424b508-f14											
7088	31mayo	0900	GeoRSS	12.243391 -1	Wed May 31	alert	Wed	19.413693 -99.2295	483625ad-e2											
7089	31mayo	0900	GeoRSS	12.243391 -1	Wed May 31	alert	Wed	19.390527 -99.291397	911e9a75-fd											
7090	31mayo	0900	GeoRSS	12.243391 -1	Wed May 31	alert	Wed	19.408195 -99.267753	a9f4b985-90f											
7091	31mayo	0900	GeoRSS	12.243391 -1	Wed May 31	alert	Wed	19.408601 -99.254048	a049214d-25											
7092	31mayo	0900	GeoRSS	12.243391 -1	Wed May 31	alert	Wed	19.407332 -99.233576	8600f32a-ce8											
7093	31mayo	0900	GeoRSS	12.243391 -1	Wed May 31	alert	Wed	19.393633 -99.259457	8547fa06-51t											
7094	31mayo	0900	GeoRSS	12.243391 -1	Wed May 31	alert	Wed	19.382899 -99.278218	f3ed83e9-71											
7095	31mayo	0900	GeoRSS	12.243391 -1	Wed May 31	alert	Wed	19.405498 -99.278553	468e83f3-05											
7096	31mayo	0900	GeoRSS	12.243391 -1	Wed May 31	alert	Wed	19.408947 -99.26683	0845d967-82											
7097	31mayo	0900	GeoRSS																	
7098	31mayo	0900	GeoRSS	magv	ns1:type	ns1:subty	ns1:repor	ns1:street	ns1:city	ns1:count	ns1:repor	confidenc	ns1:reliab	ns1:roadT	ns					
7099	31mayo	0900	GeoRSS	214 JAM	JAM_STAND_STILL_TRAFF	Paseo de la t	Huixquilucar MX			2	0	5	7							
7100	31mayo	0900	GeoRSS	58 WEATHERHA	HAZARD_ON_ROAD_CAR	Av. de los Bc	Huixquilucar MX			1	0	8	7							
7101	31mayo	0900	GeoRSS	10 JAM	JAM_STAND_STILL_TRAFF	Paseo de las	Miguel Hidal MX			2	0	5	7							
7102	31mayo	0900	GeoRSS	32 JAM	JAM_MODERATE_TRAFFIC	Av. Jesús del	Huixquilucar MX			2	0	5	7							
7103	31mayo	0900	GeoRSS	214 JAM	JAM_HEAVY_TRAFFIC	Paseo de la t	Huixquilucar MX			2	0	5	7							
7104	31mayo	0900	GeoRSS	35 JAM	JAM_HEAVY_TRAFFIC	Av. de los Bc	Huixquilucar MX			4	0	6	7							
7105	31mayo	0900	GeoRSS	218 WEATHERHA	HAZARD_ON_ROAD_CAR	STOPPED	Miguel Hidal MX			2	0	6	7							
7106	31mayo	0900	GeoRSS	344 JAM	JAM_HEAVY_TRAFFIC	Paseo de los Ahuehuetes	MX			0	0	5	7							
7107	31mayo	0900	GeoRSS	43 JAM	JAM_STAND_STILL_TRAFF	Loma de la P	Cuajimalpa c MX			3	0	5	7							
7108	31mayo	0900	GeoRSS	342 JAM	JAM_MODERATE_TRAFFIC	Palma Criolla	Huixquilucar MX			1	0	5	7							
7109	31mayo	0900	GeoRSS	239 JAM	JAM_STAND_STILL_TRAFF	Paseo de la t	Huixquilucar MX			1	0	5	7							
7110	31mayo	0900	GeoRSS	41 JAM	JAM_HEAVY_TRAFFIC	Av. de los Bc	Huixquilucar MX			2	0	5	7							
7111	31mayo	0900	GeoRSS	73 JAM	JAM_STAND_STILL_TRAFF	Glaciar	Álvaro Obreg MX			2	1	7	7							
7112	31mayo	0900	GeoRSS	178 JAM	JAM_STAND_STILL_TRAFF	Don Manueli	Álvaro Obreg MX			2	0	5	7							
7113	31mayo	0900	GeoRSS	87 JAM	JAM_STAND_STILL_TRAFF	Rómulo O'Fa	Álvaro Obreg MX			2	0	5	7							
7114	31mayo	0900	GeoRSS	107 JAM	JAM_HEAVY_TRAFFIC	Calz. Sta. Cat	Álvaro Obreg MX			4	0	5	7							
7115	31mayo	0900	GeoRSS	97 JAM	JAM_STAND_STILL_TRAFF	Calz. Sta. Cat	Álvaro Obreg MX			3	0	5	7							
7116	31mayo	0900	GeoRSS	266 JAM	JAM_STAND_STILL_TRAFF	Jalapa	Álvaro Obreg MX			2	0	5	7							
				38 JAM	JAM_STAND_STILL_TRAFF	Rómulo O'Fa	Álvaro Obreg MX			3	0	5	7							
				86 WEATHERHA	HAZARD_ON_SHOULDER	Rómulo O'Fa	Álvaro Obreg MX			2	1	7	7							
				76 JAM	JAM_STAND_STILL_TRAFF	Glaciar	Álvaro Obreg MX			3	0	5	7							
				265 JAM	JAM_HEAVY_TRAFFIC	Av. Toluca	Álvaro Obreg MX			2	0	5	7							
				264 JAM		Av. Universi	Coyoacán MX			0	0	6	7							
				264 JAM	JAM_HEAVY_TRAFFIC	Av. Universi	Coyoacán MX			2	0	6	7							
				274 JAM	JAM_STAND_STILL_TRAFF	Molinos	Benito Juárez MX			3	1	7	7							
				90 JAM	JAM_STAND_STILL_TRAFF	Molinos	Álvaro Obreg MX			2	0	5	7							
				122 JAM	JAM_HEAVY_TRAFFIC	Barranca del	Álvaro Obreg MX			0	0	6	7							
				273 JAM	JAM_STAND_STILL_TRAFF	Molinos	Benito Juárez MX			2	0	6	7							
				72 JAM	JAM_HEAVY_TRAFFIC	Av. Sta. Lucía	Álvaro Obreg MX			1	0	5	7							
				99 JAM	JAM_STAND_STILL_TRAFF	Barranca del	Álvaro Obreg MX			0	1	8	7							
				78 ACCIDENT	ACCIDENT MINOR	Barranca del	Álvaro Obreg MX			1	0	5	7							

Fuente: Elaboración propia

Figura 3.15 Campos resultantes de la unificación de tablas.

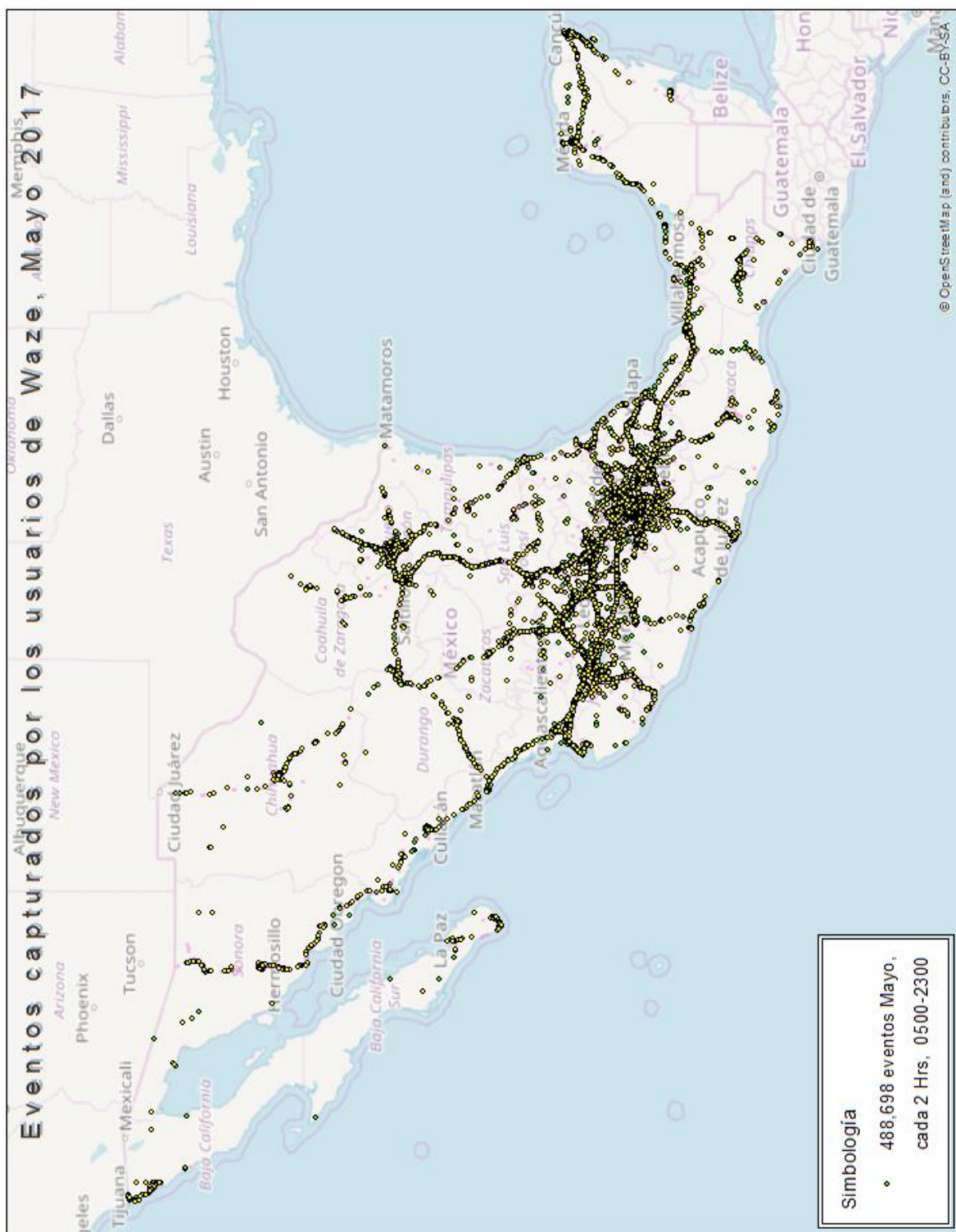
En la Figura 3.15 se pueden observar los campos que provienen directamente de Waze, entre estos se incluyen la versión, descripción, área geográfica donde fueron levantados los datos, coordenadas del punto (x,y), identificador único del evento, tipo y subtipo de evento, nombre de la calle o carretera donde sucedió, nombre del estado, país (en este caso, México), otros datos, etc.

Los eventos se dividen en dos tipos:

- Puntos
- Líneas

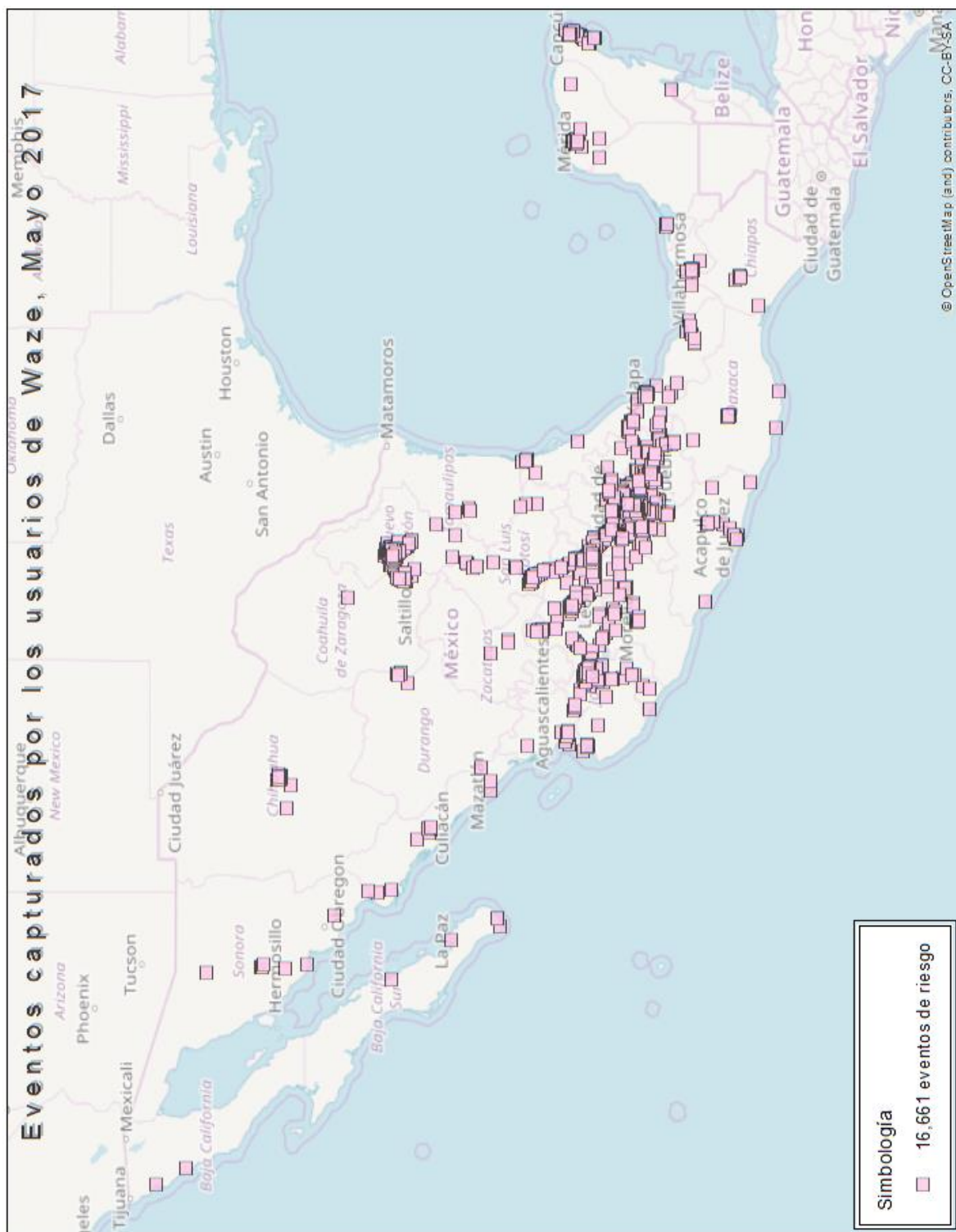
Cuando son puntos el formato de las coordenadas es xxx.xxx yyy.yyy en una misma celda, por lo que se hace necesario separar los valores en distintas celdas para poder reconstruir en ArcGIS los eventos y representarlos mediante puntos.

Cuando son líneas el formato de las coordenadas es xxx.xxx yyy.yyy x1x1x1.y1y1y1 x2x2x2.y2y2y2n en una misma celda y, al igual que los puntos se deben separar los valores en distintas celdas para poder reconstruir los eventos y representarlos mediante líneas cuando se requiera. Esto requiere un proceso diferente ya que una línea se compone de varios o muchos vértices. Para los fines del proyecto solamente se utilizaron los puntos, aunque más adelante se podrían manejar también las líneas y modelar redes de carreteras y vialidades a partir de sus datos.



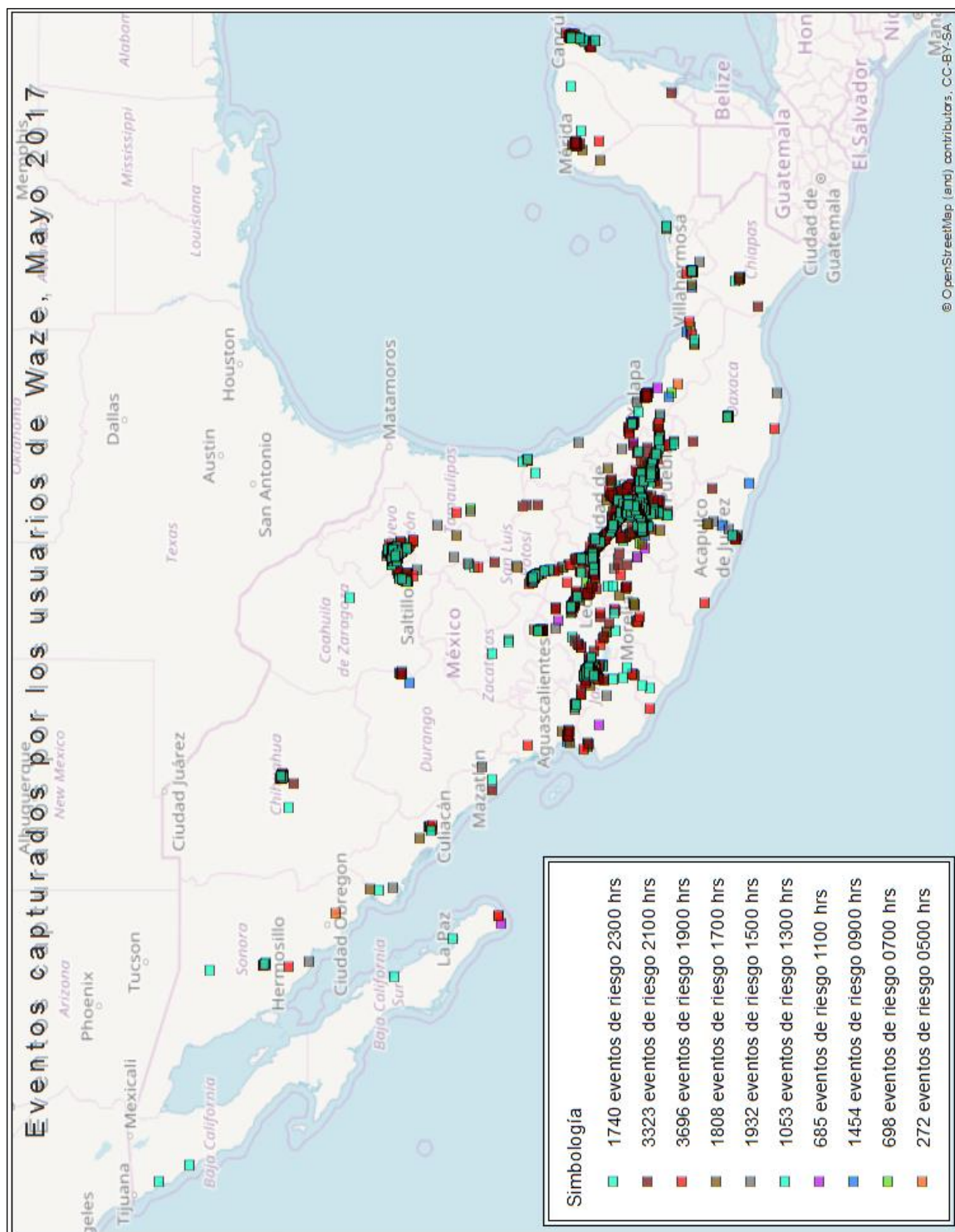
Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.16 Eventos del mes de mayo 2017.



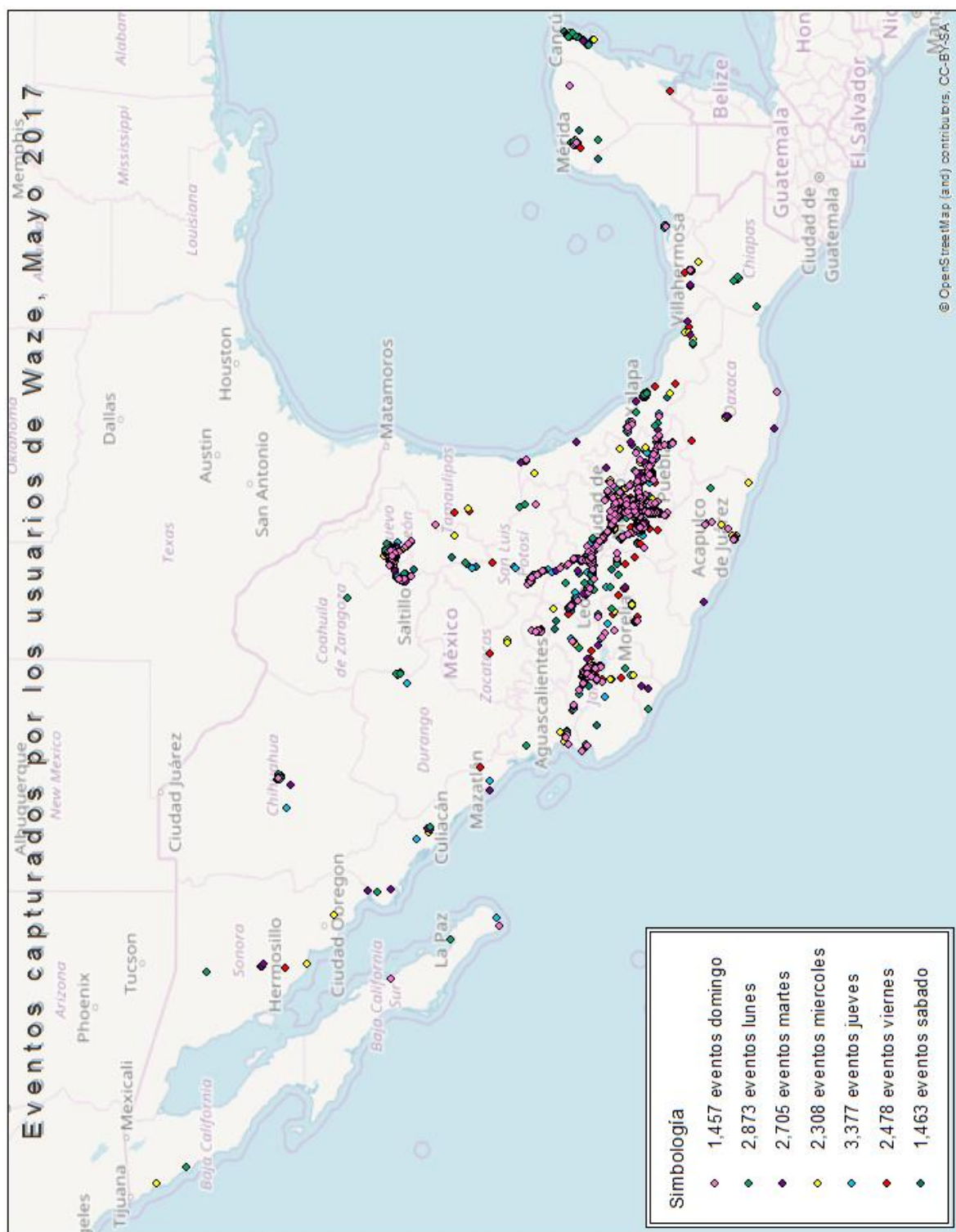
Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.17 Eventos que representan un riesgo, mayo 2017.



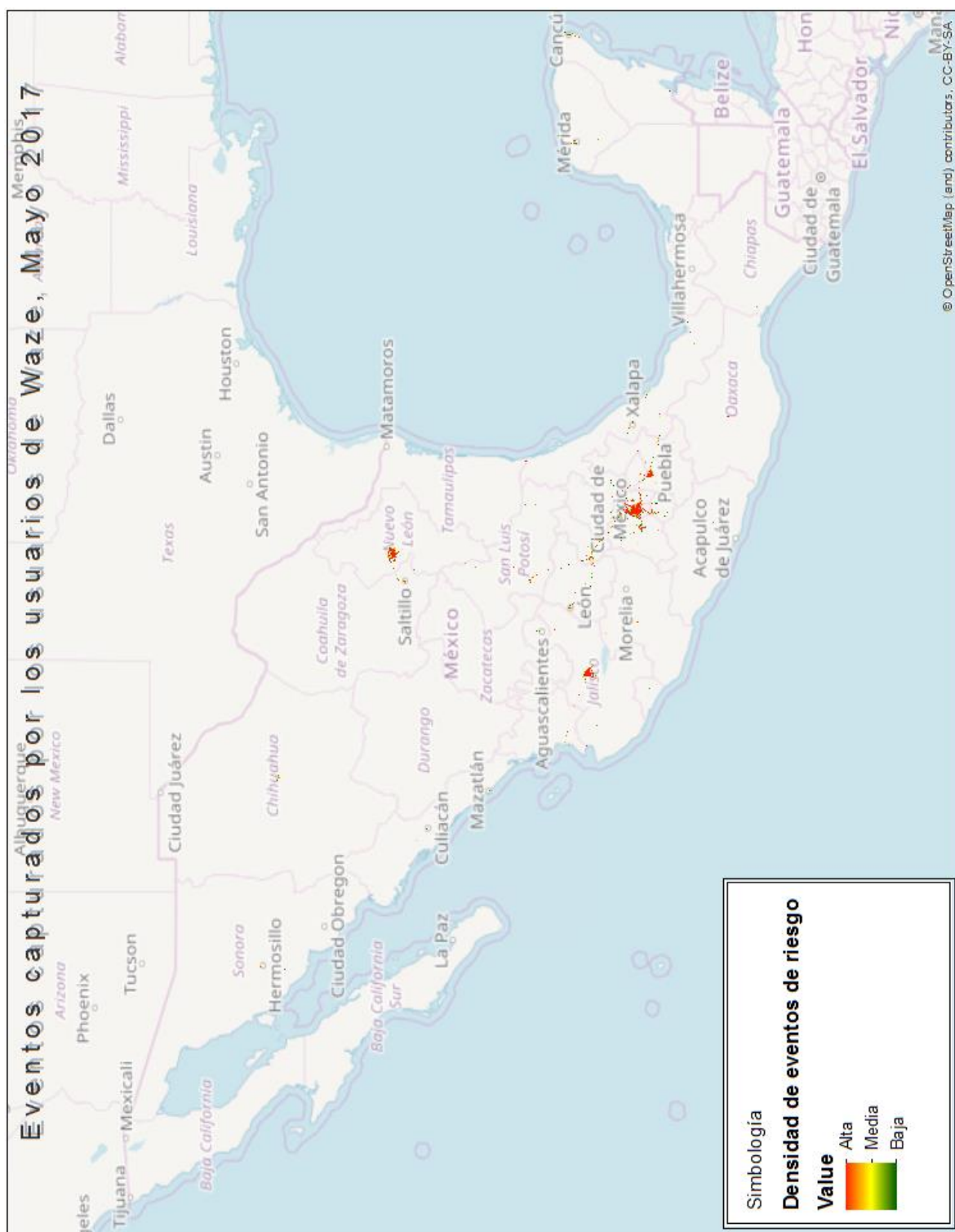
Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.18 Eventos que representan un riesgo, por horas, mayo 2017.



Fuente: Elaboración propia a partir de datos de Waze.

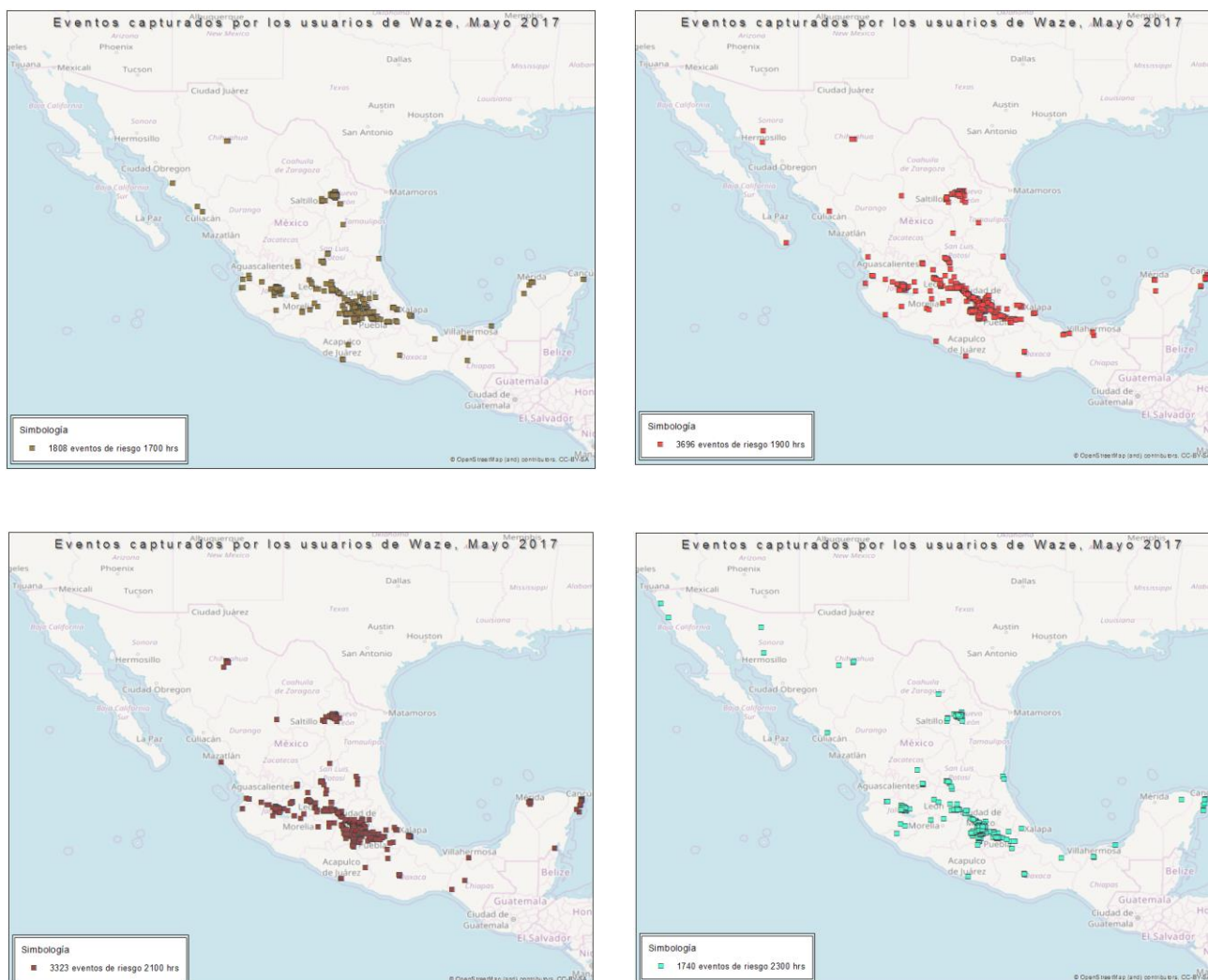
Figura 3.19 Eventos que representan un riesgo, acumulados por día, mayo 2017.



Fuente: Elaboración propia a partir de datos de Waze.

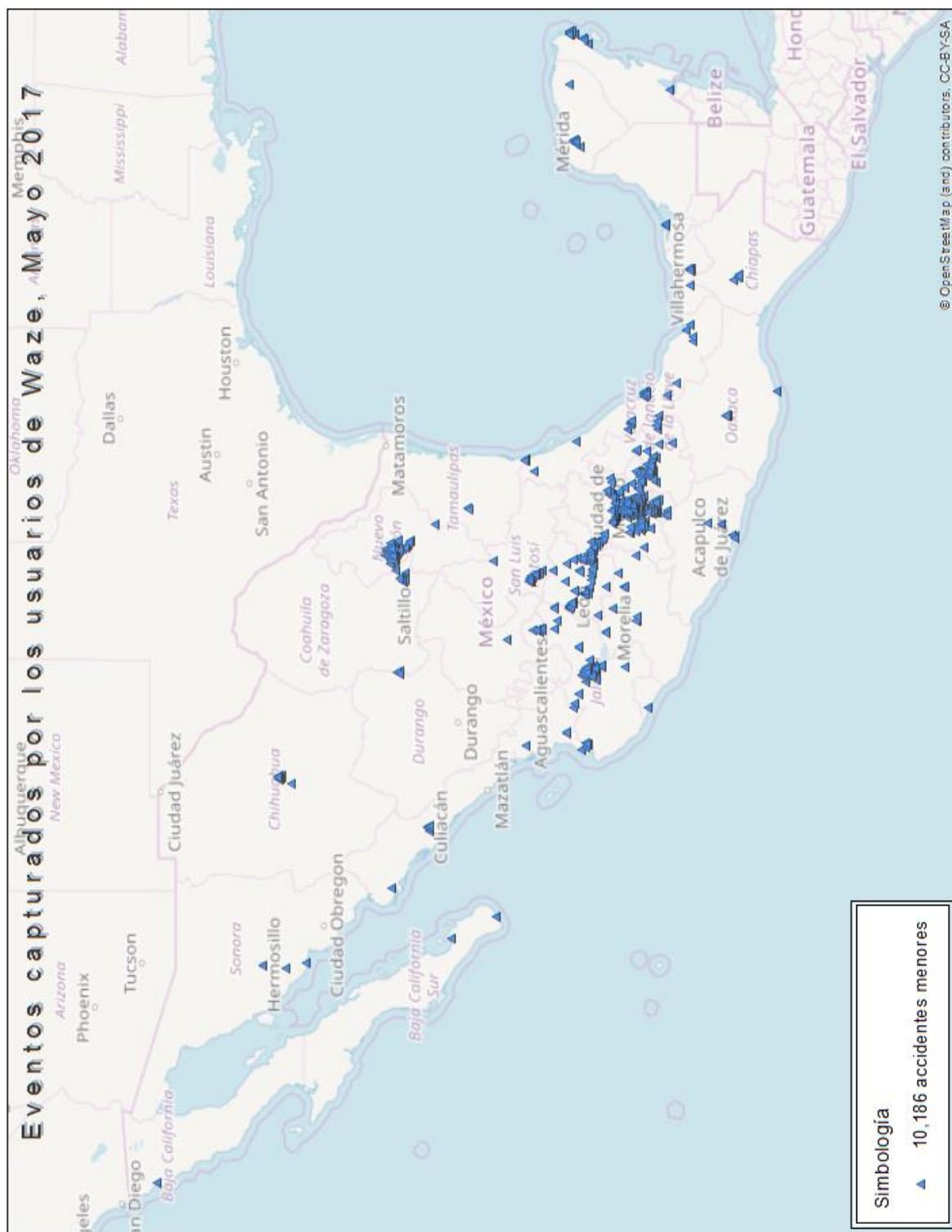
Figura 3.20 Densidad de puntos que representan riesgo, mayo 2017.





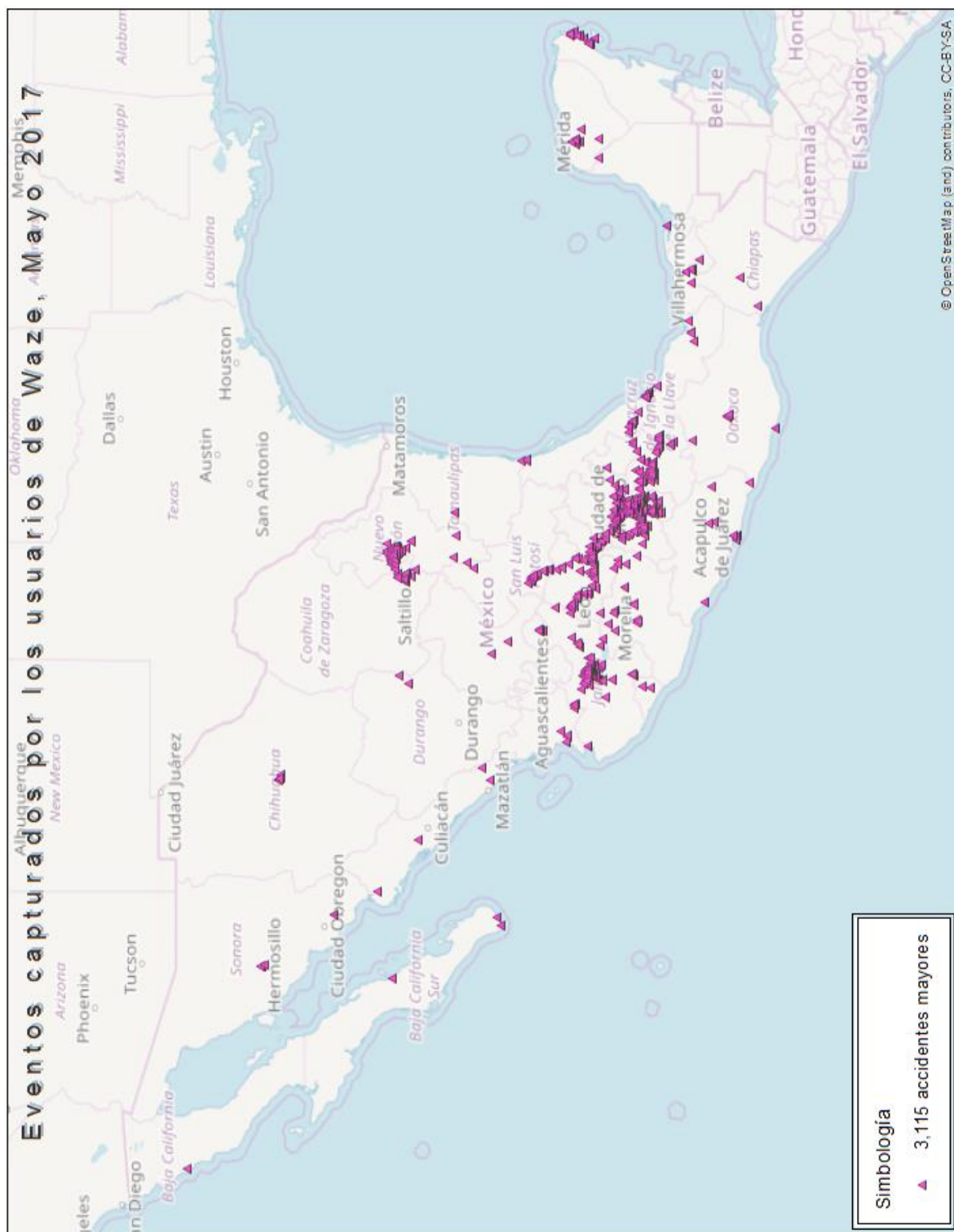
Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.21 Conjunto de mapas que muestran la cantidad de puntos que representan riesgo representados por hora/día, mayo 2017. Los horarios van desde las 0500, 0700, 0900, 1100, 1300, 1500, 1700, 1900, 2100 y 2300 horas.



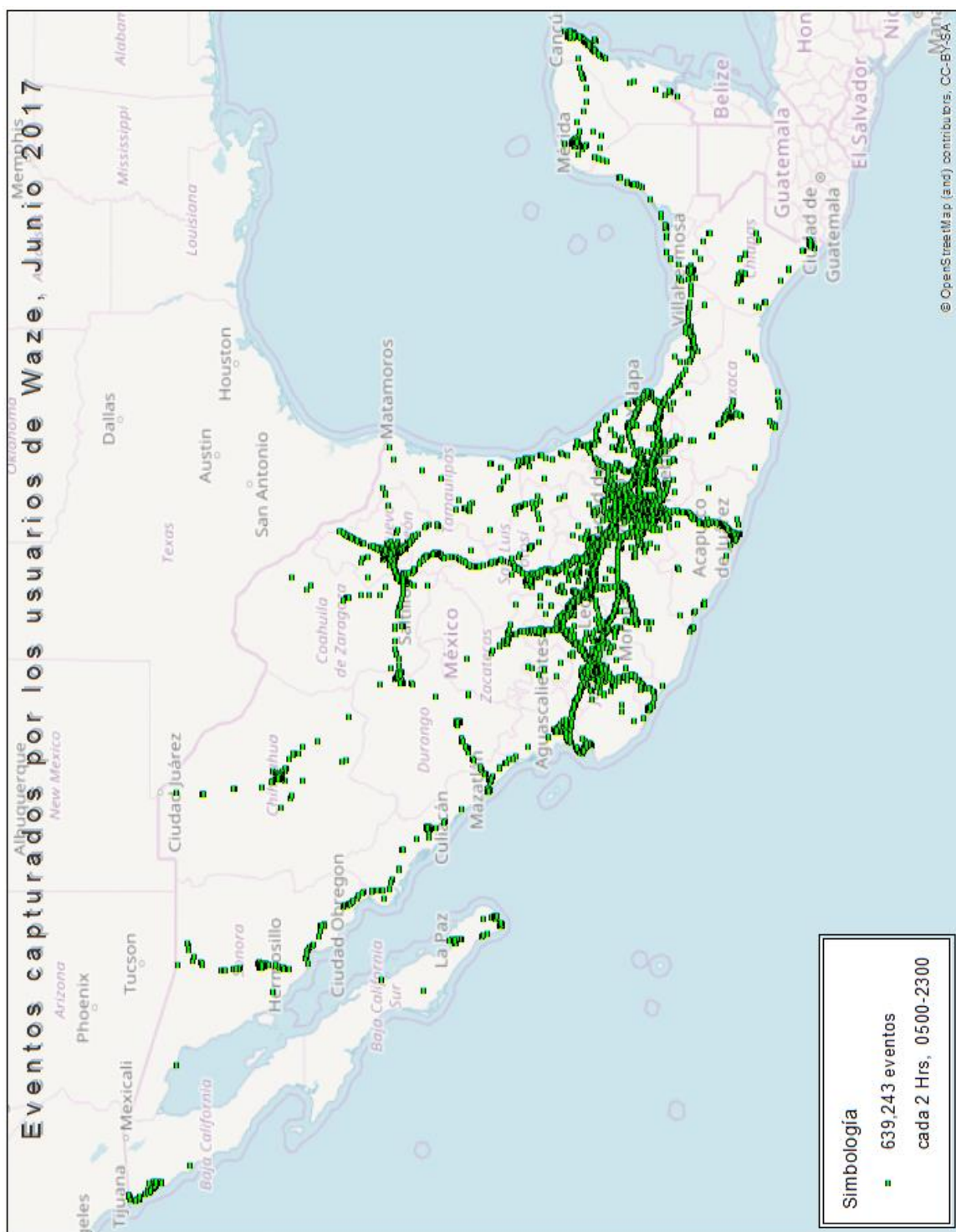
Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.22 Accidentes menores del mes de mayo 2017.



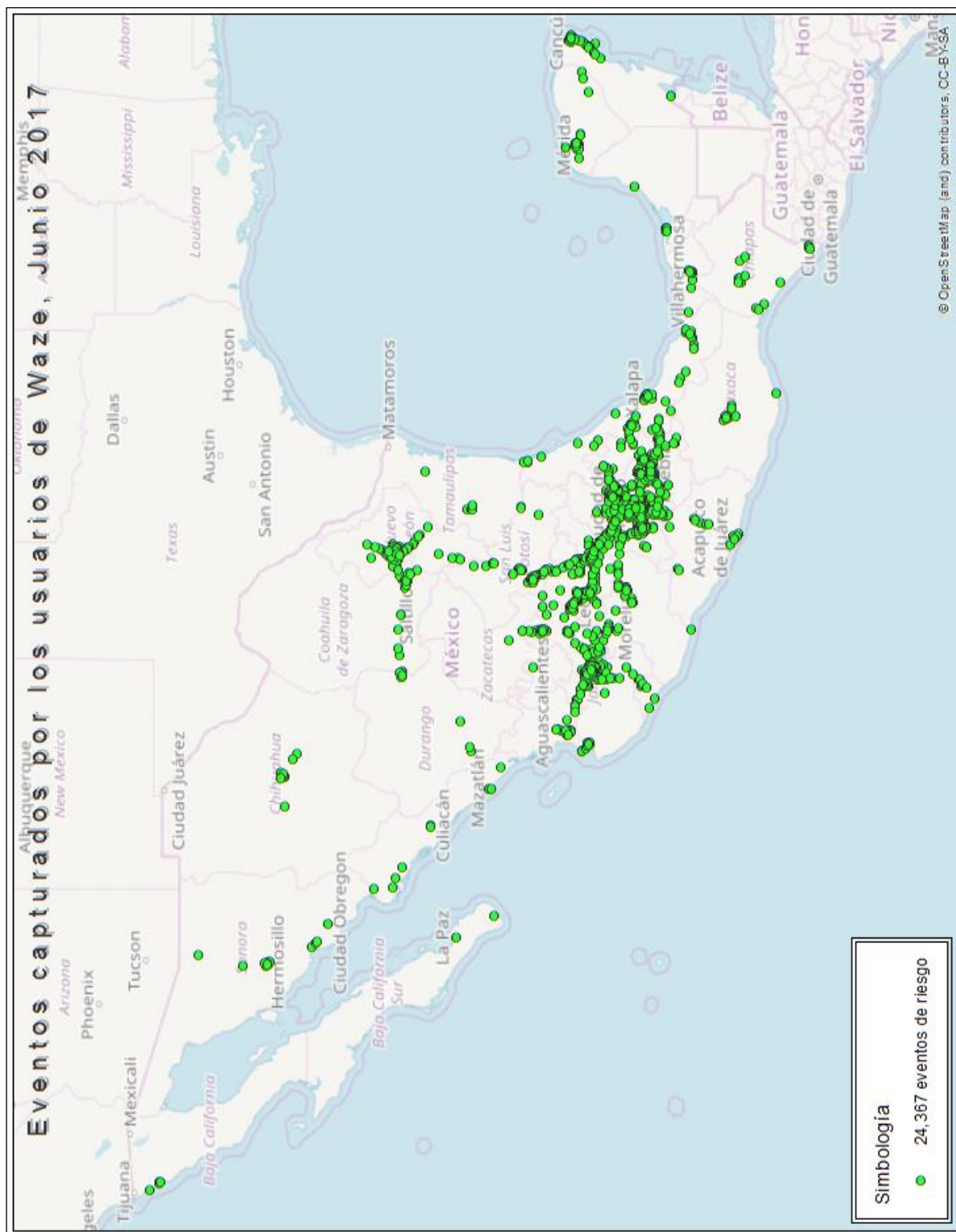
Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.23 Accidentes mayores del mes de mayo 2017.



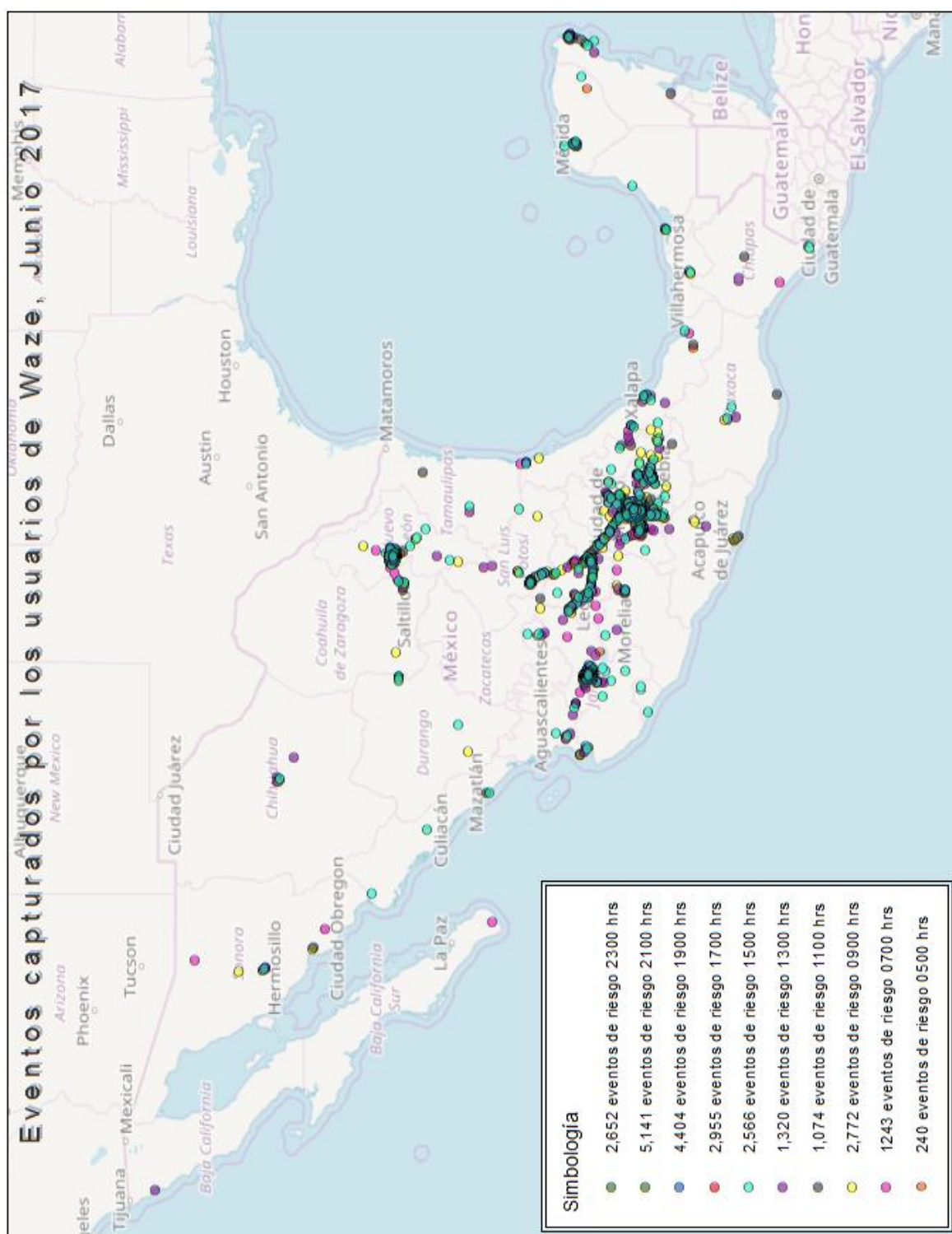
Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.24 Eventos del mes de junio 2017.



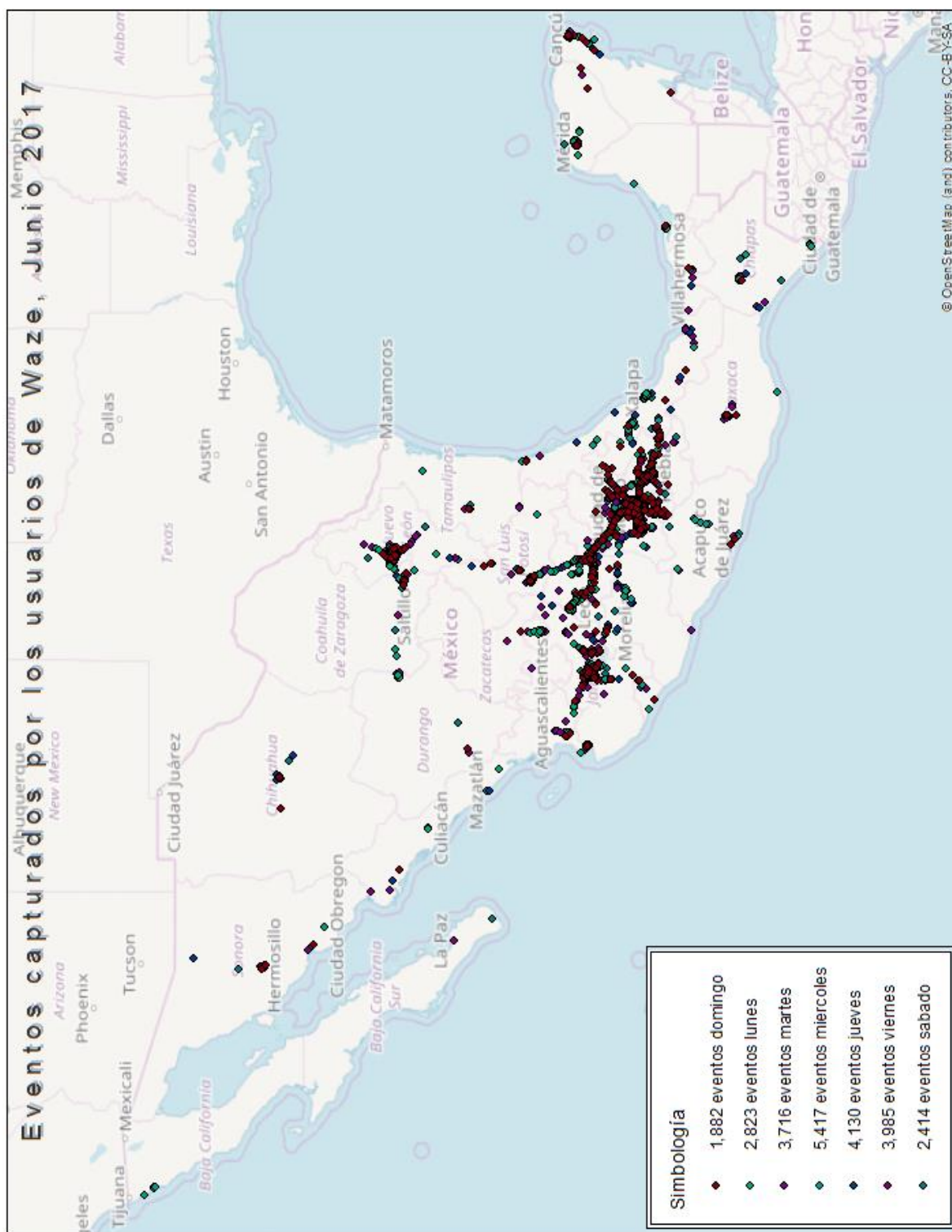
Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.25 Eventos que representan un riesgo, junio 2017.



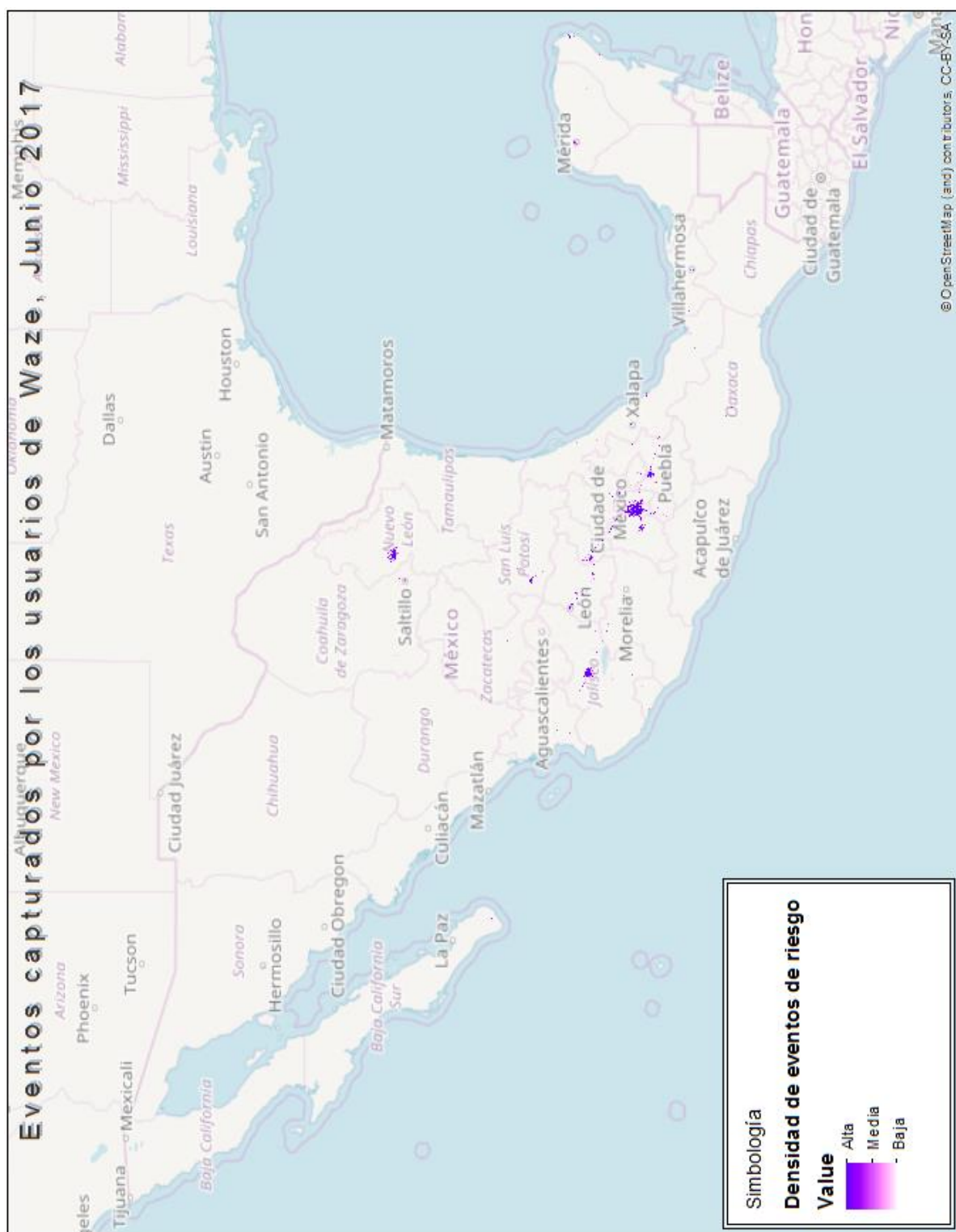
Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.26 Eventos que representan un riesgo, por horas, junio 2017.



Fuente: Elaboración propia a partir de datos de Waze.

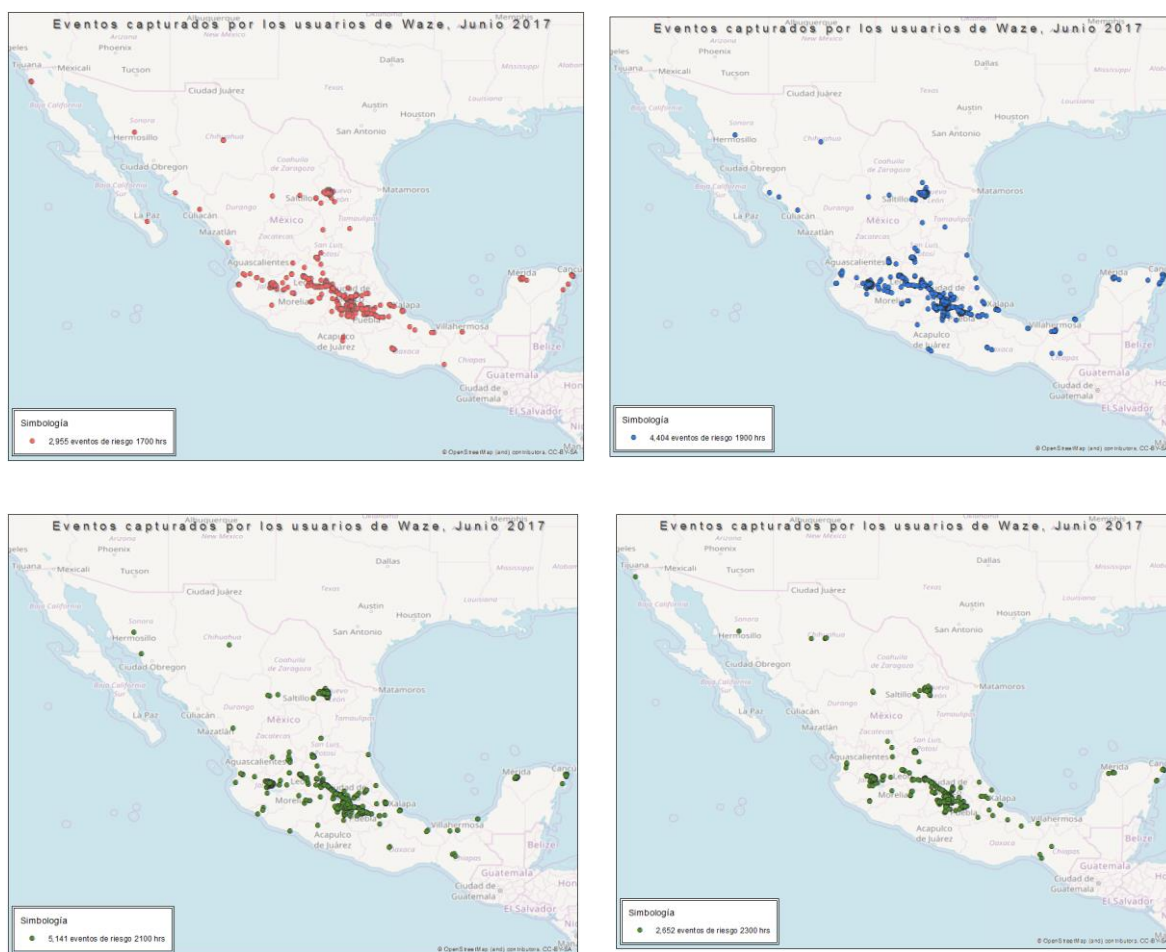
Figura 3.27 Eventos que representan un riesgo, acumulados por día, junio 2017.



Fuente: Elaboración propia a partir de datos de Waze.

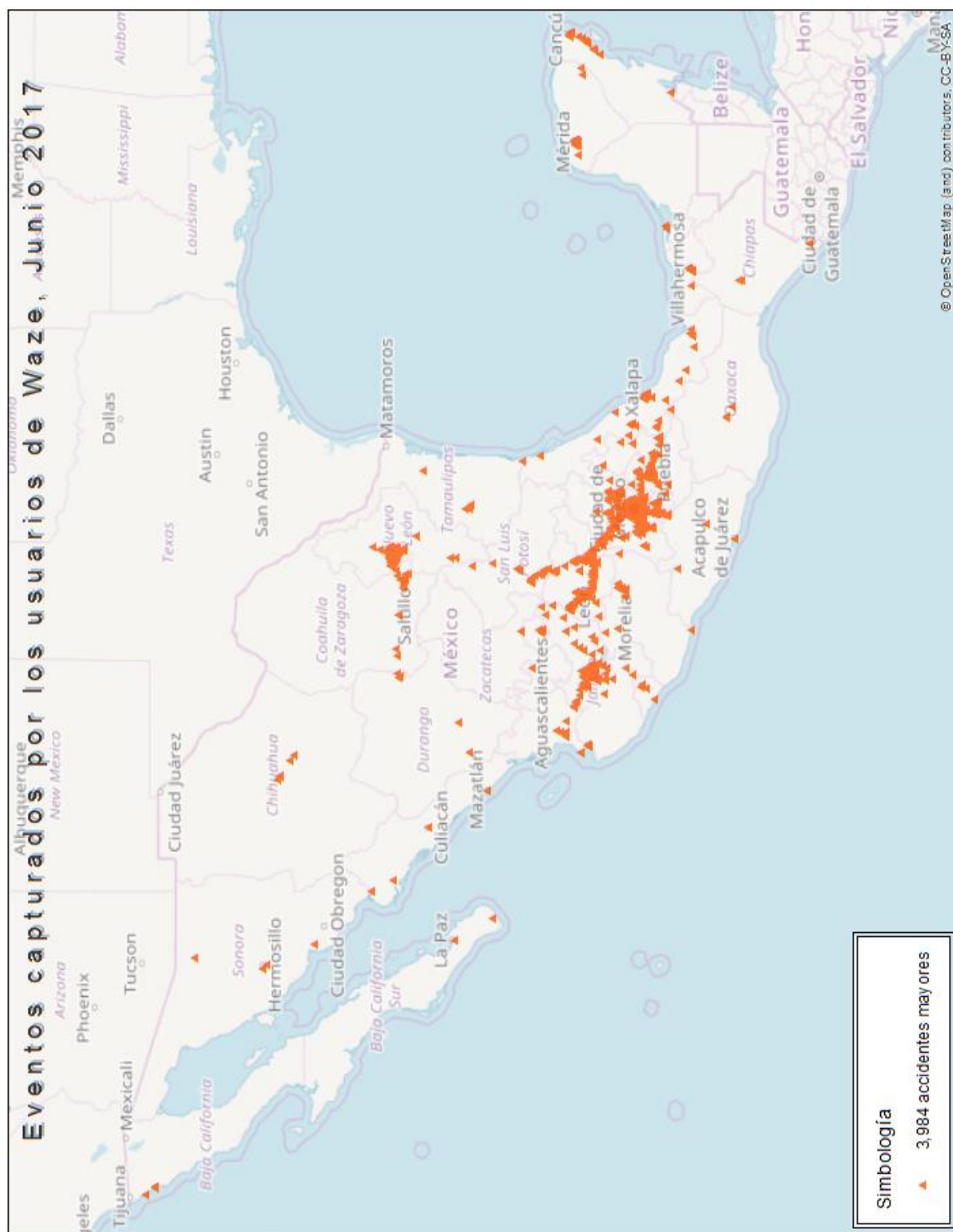
Figura 3.28 Densidad de puntos que representan riesgo, junio 2017.





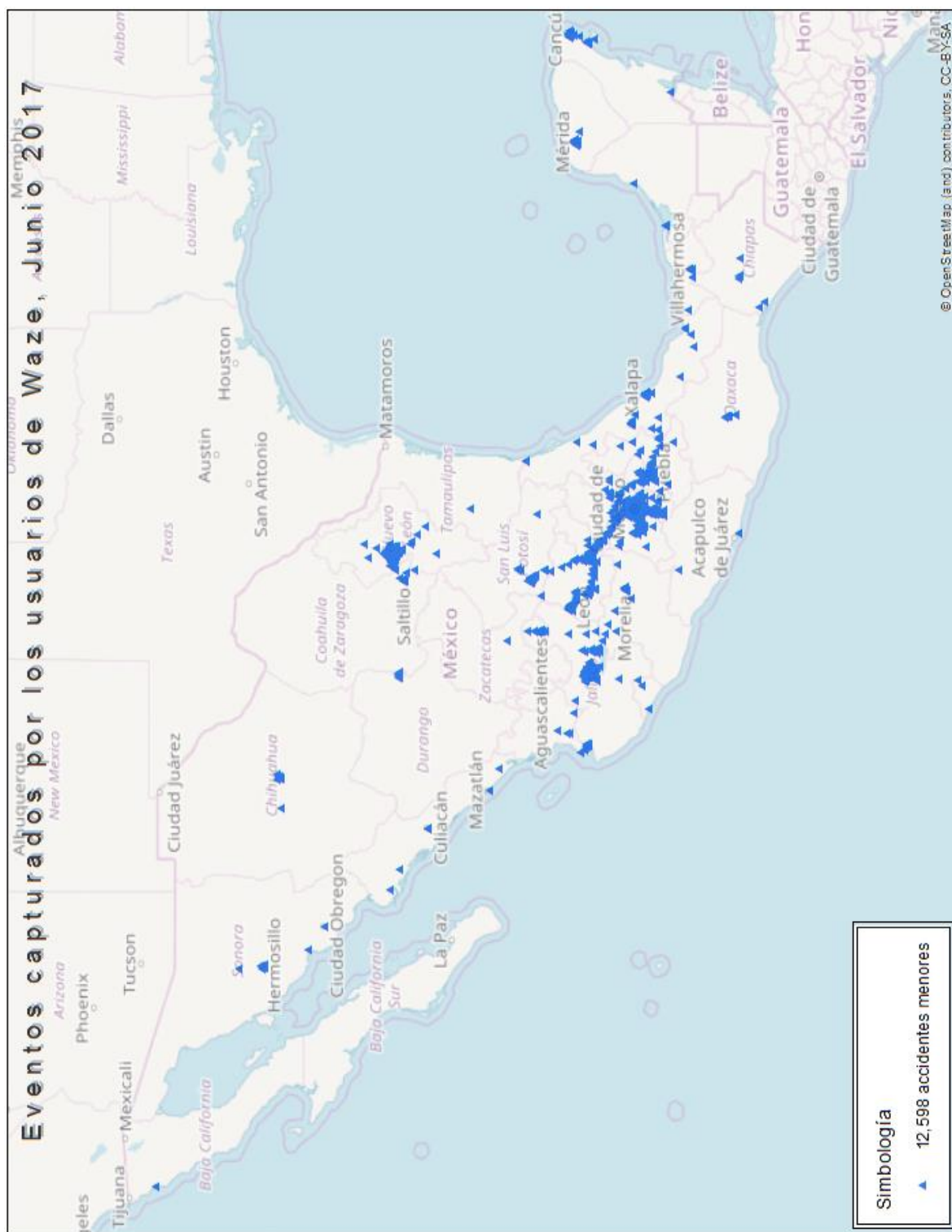
Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.29 Conjunto de mapas que muestran la cantidad de puntos que representan riesgo representados por hora/día, junio 2017. Los horarios van desde las 0500, 0700, 0900, 1100, 1300, 1500, 1700, 1900, 2100 y 2300 horas.



Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.30 Accidentes mayores del mes de junio 2017.

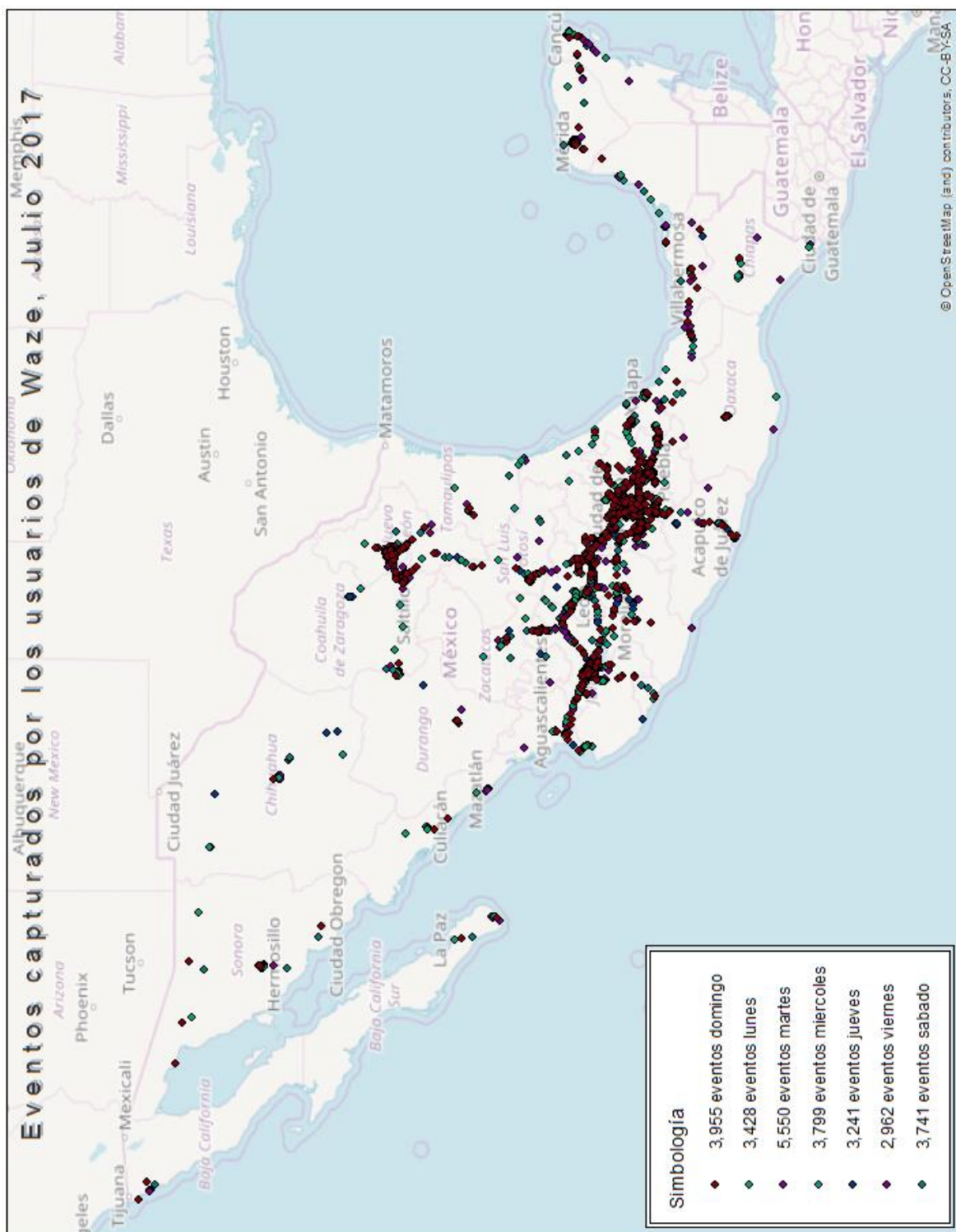


Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.31 Accidentes menores del mes de junio 2017.

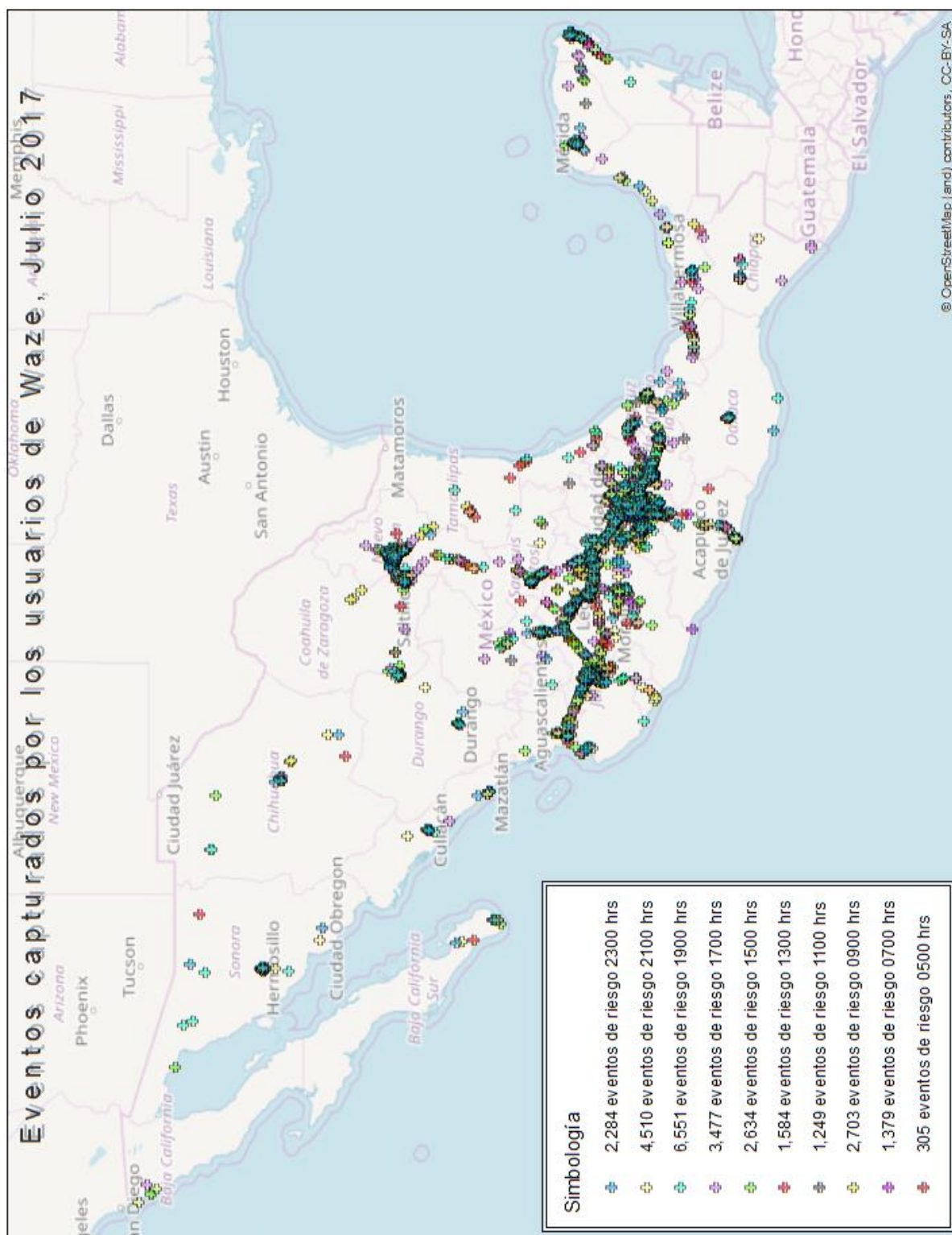
Figura 3.32 Eventos del mes de julio 2017.

Figura 3.33 Eventos que representan un riesgo, julio 2017.



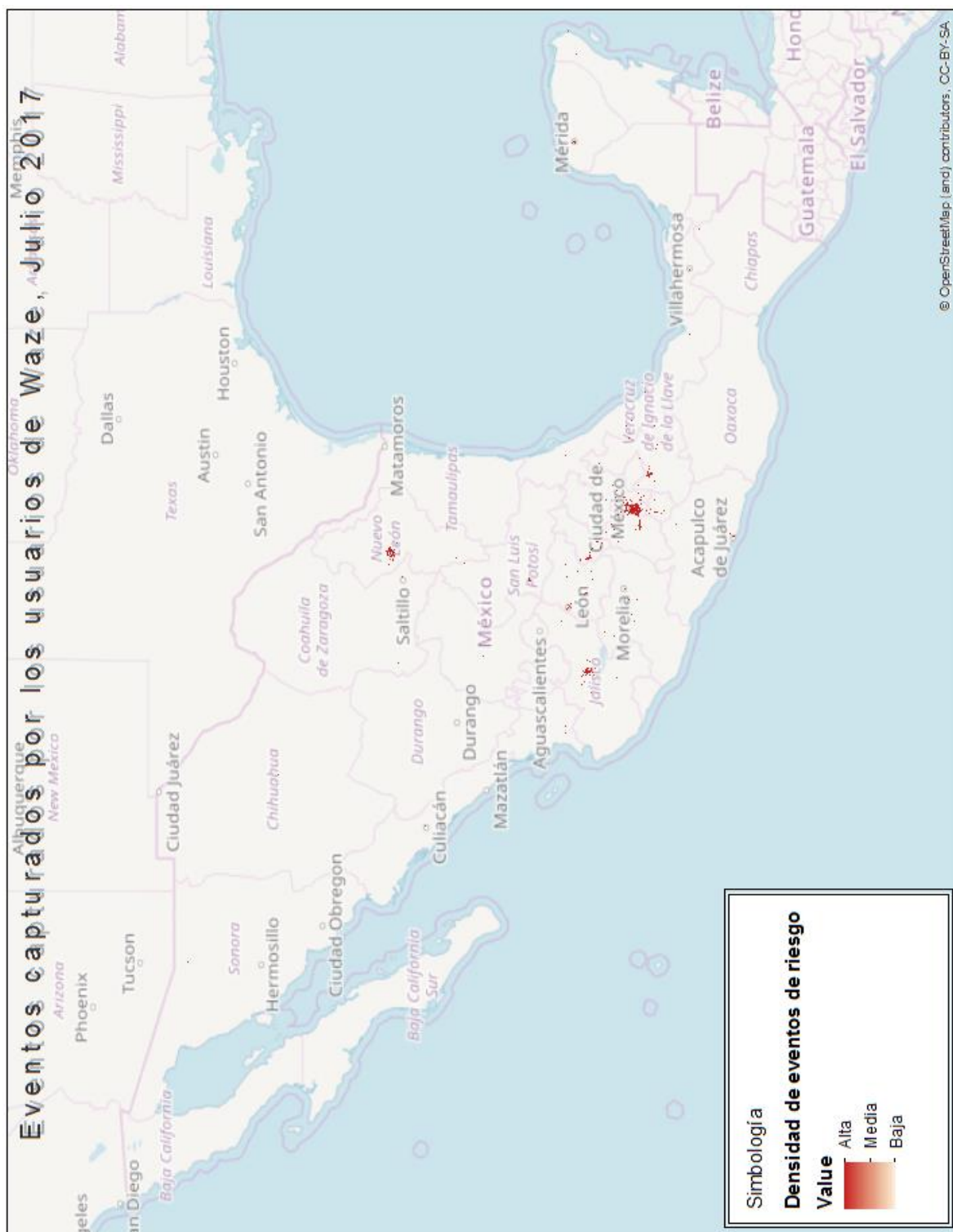
Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.34 Eventos que representan un riesgo, acumulados por día, julio 2017.



Fuente: Elaboración propia a partir de datos de Waze.

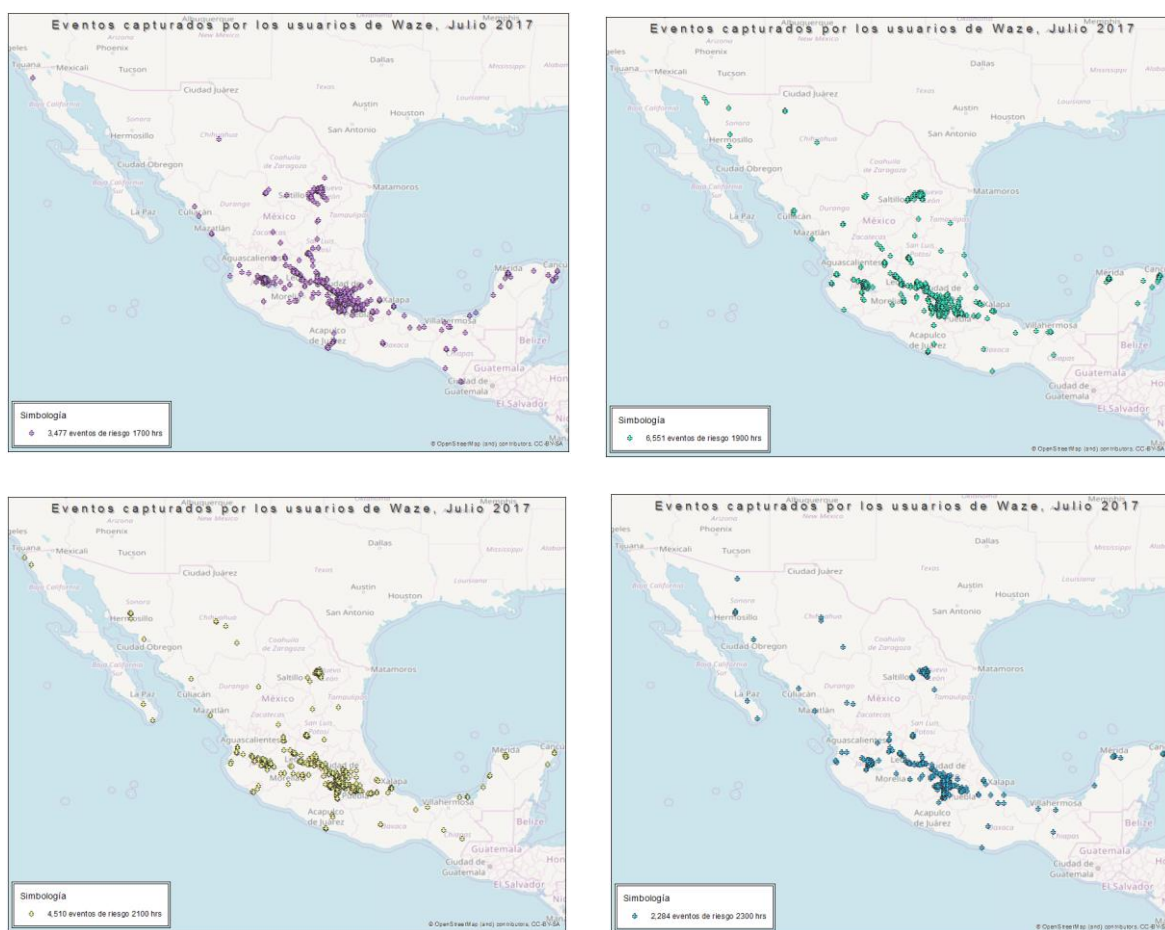
Figura 3.35 Eventos que representan un riesgo, por horas, julio 2017.



Fuente: Elaboración propia a partir de datos de Waze.

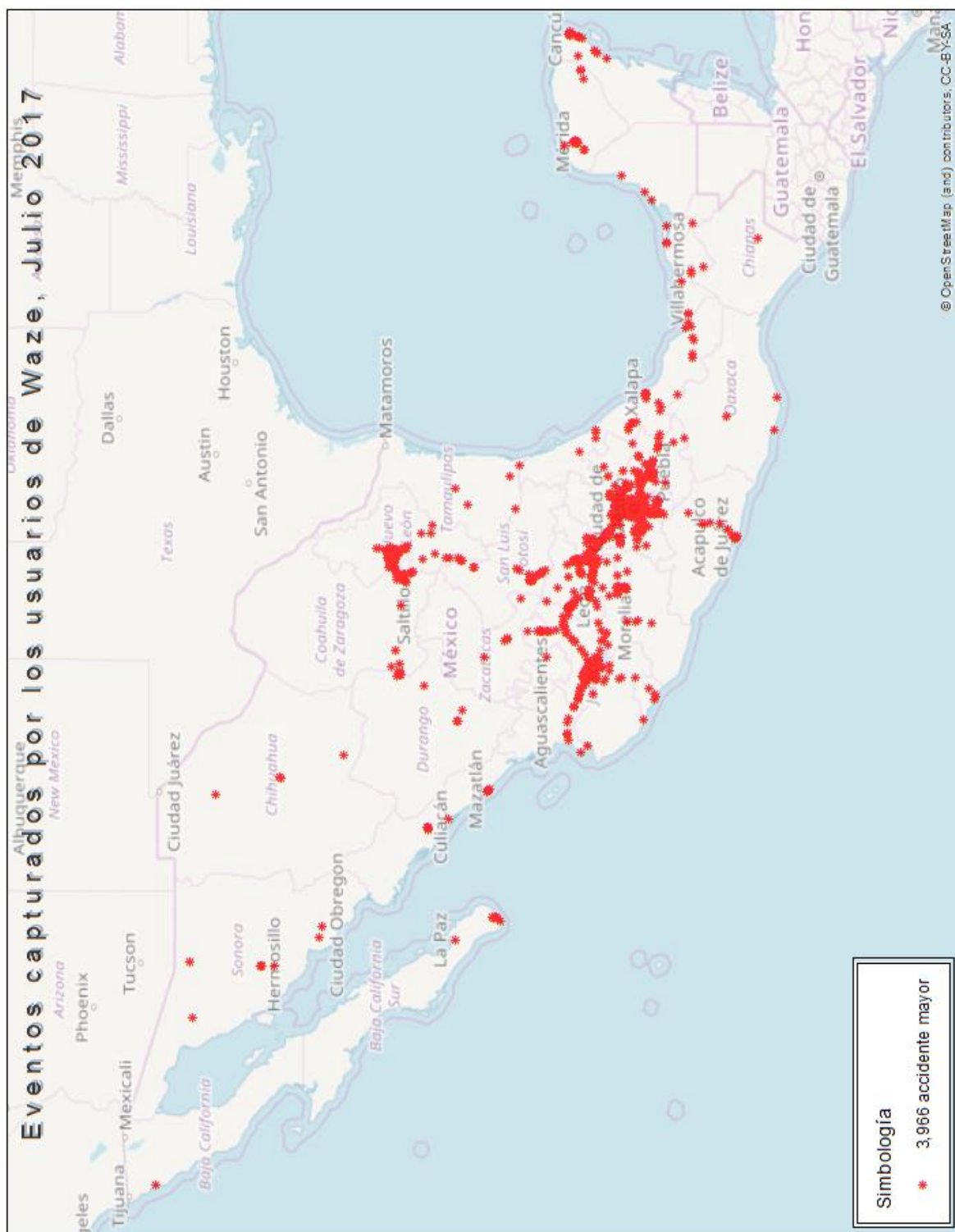
Figura 3.36 Densidad de puntos que representan riesgo, julio 2017.





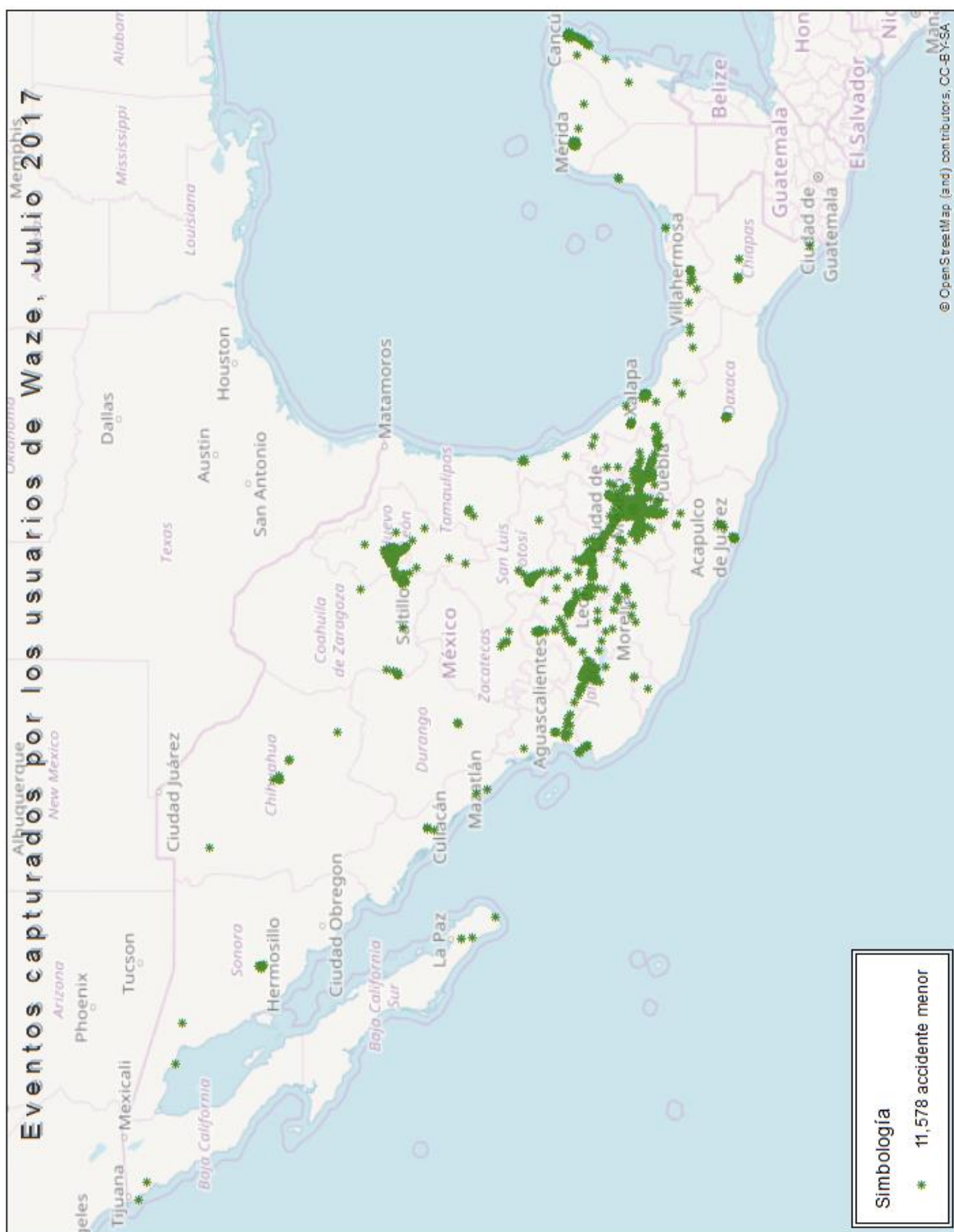
Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.37 Conjunto de mapas que muestran la cantidad de puntos que representan riesgo representados por hora/día, julio 2017. Los horarios van desde las 0500, 0700, 0900, 1100, 1300, 1500, 1700, 1900, 2100 y 2300 horas.



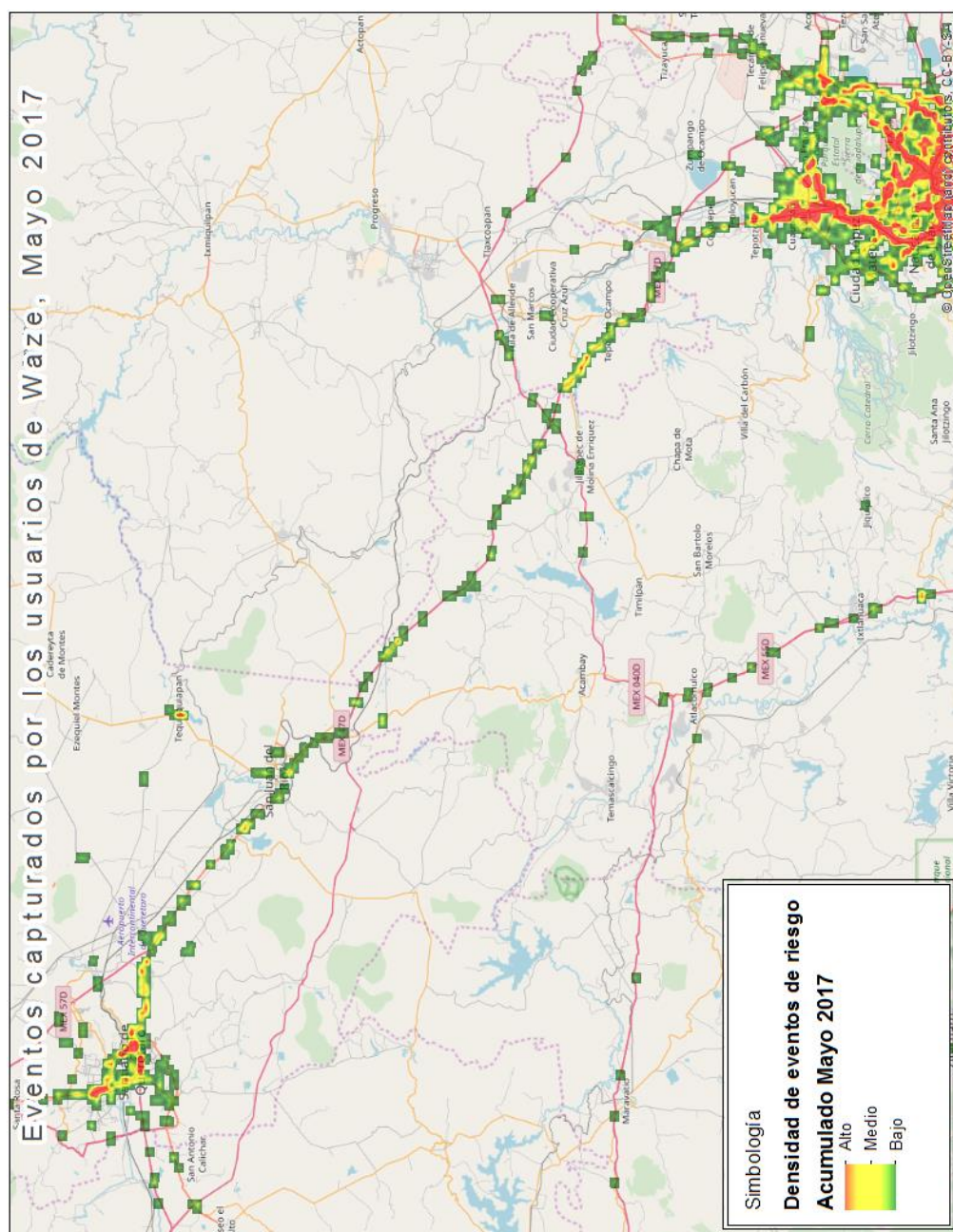
Fuente: Elaboración propia a partir de datos de Waze.

Figura 3.38 Accidentes mayores del mes de julio 2017.



Fuente: Elaboración propia a partir de datos de Waze.

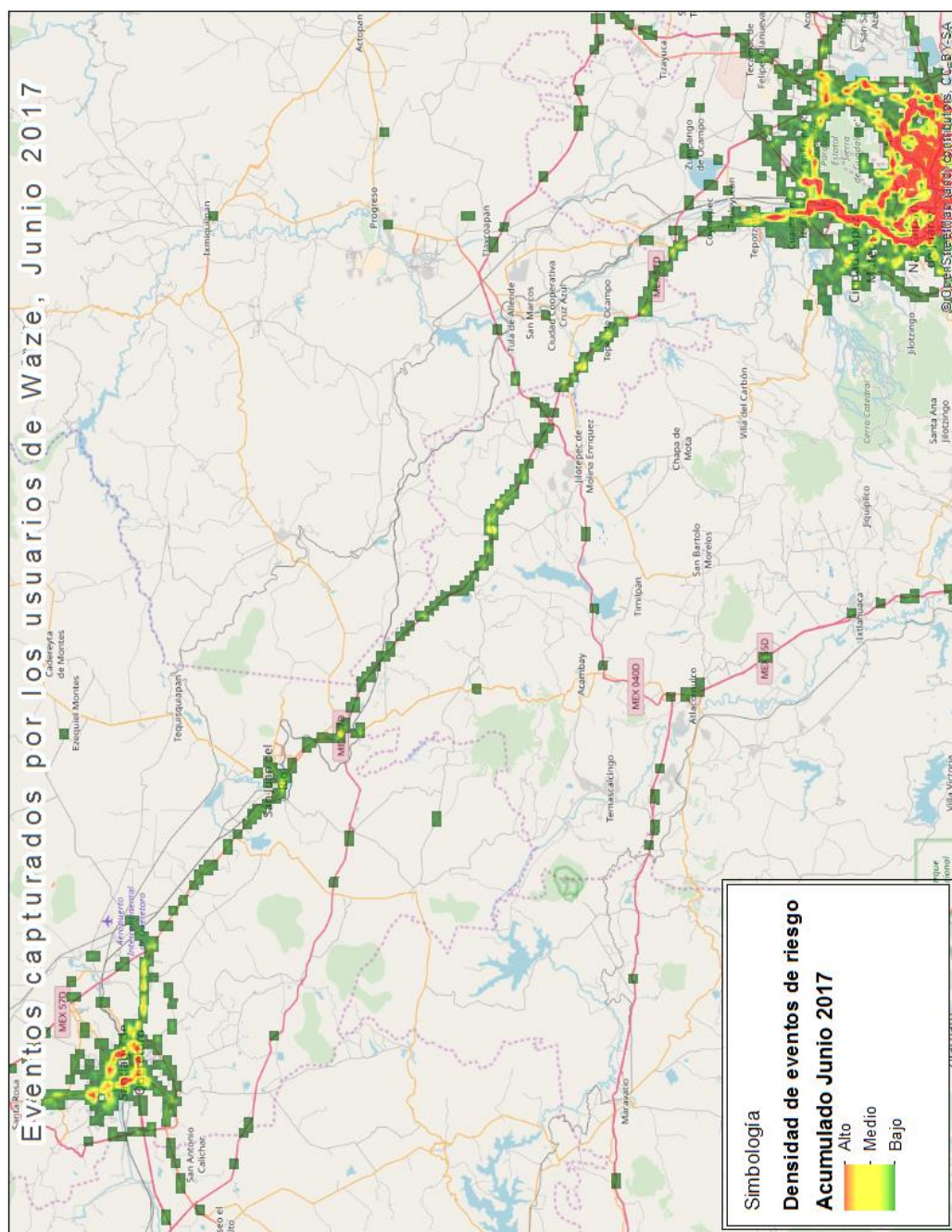
Figura 3.39 Accidentes menores del mes de julio 2017.



Fuente: Elaboración propia a partir de los datos de Waze

Figura 3.40 Densidad de eventos de riesgo, mayo 2017, acercamiento a la región Querétaro – Ciudad de México.

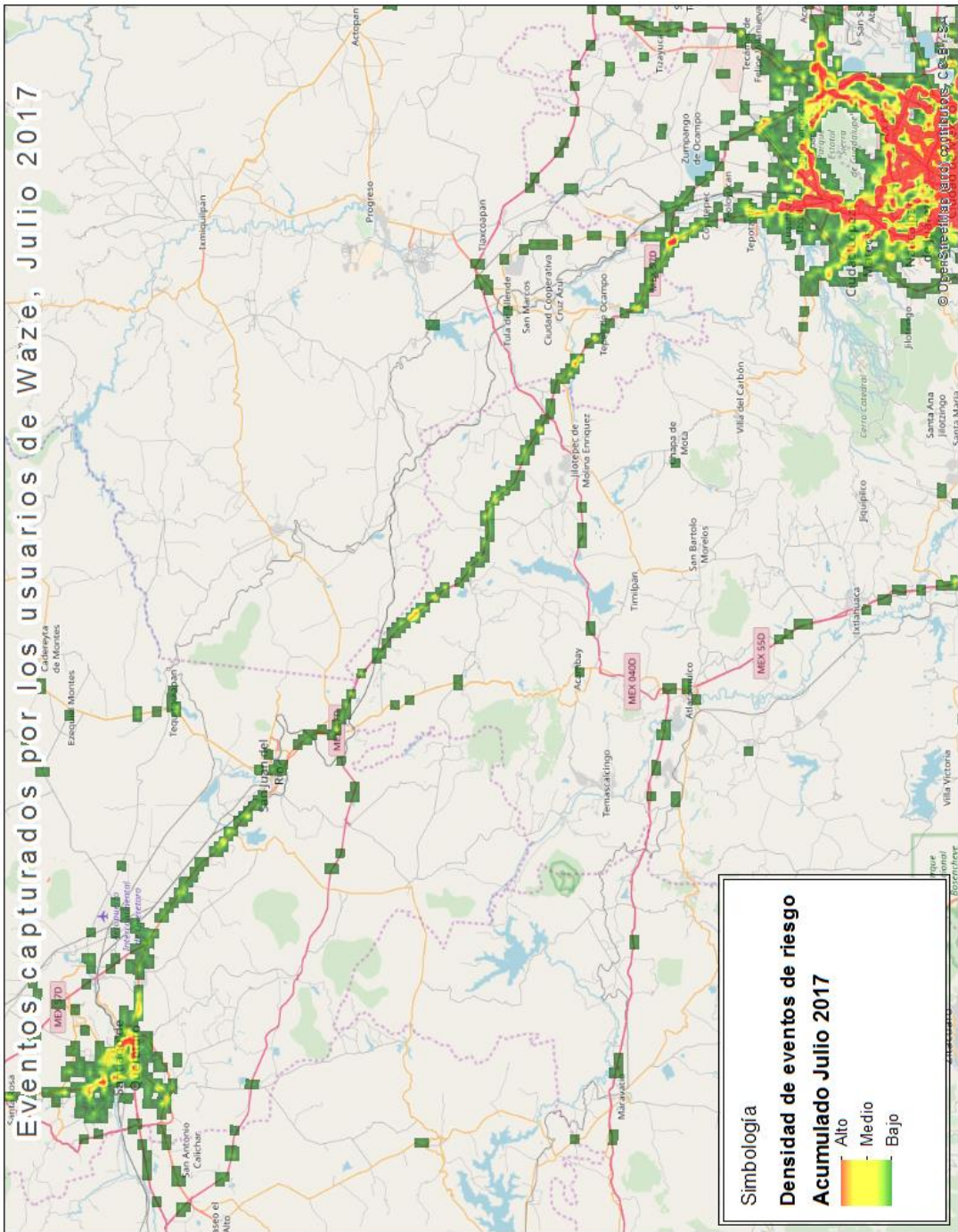
Se observa que los eventos asociados a riesgo se concentran en las zonas urbanas y en menor medida en las carreteras, en este ejemplo, la carretera 57 México – Querétaro. En este caso se muestra el acumulado mensual para mayo 2017 (16,661 eventos).



Fuente: Elaboración propia a partir de los datos de Waze

Figura 3.41 Densidad de eventos de riesgo, junio 2017, acercamiento a la región Querétaro – Ciudad de México.

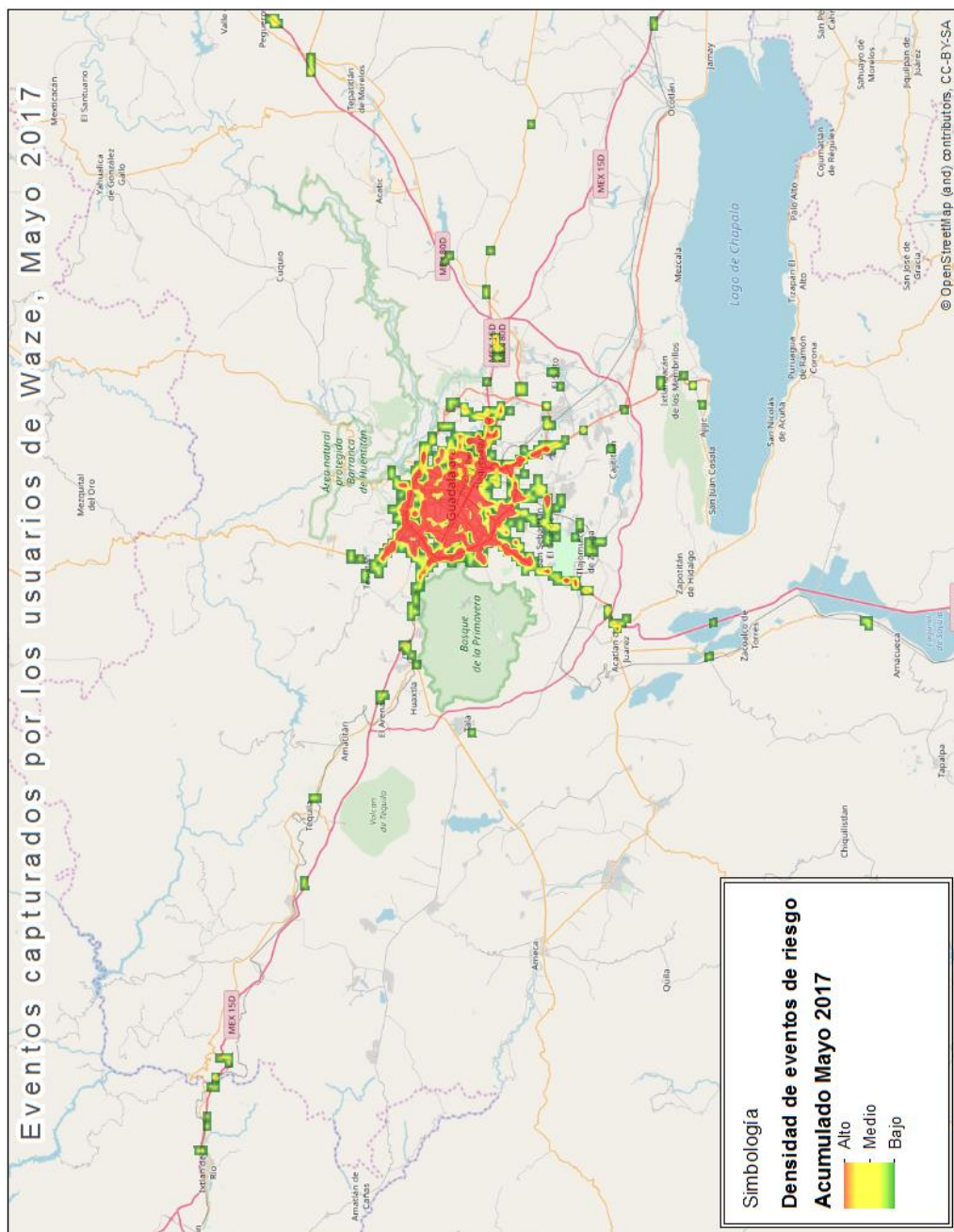
Para el mes de junio de 2017 se registran 24,367 eventos asociados a riesgo, se observa que hay un incremento de eventos en las carreteras, esto se debe a que la temporada de lluvias ha iniciado.



Fuente: Elaboración propia a partir de los datos de Waze

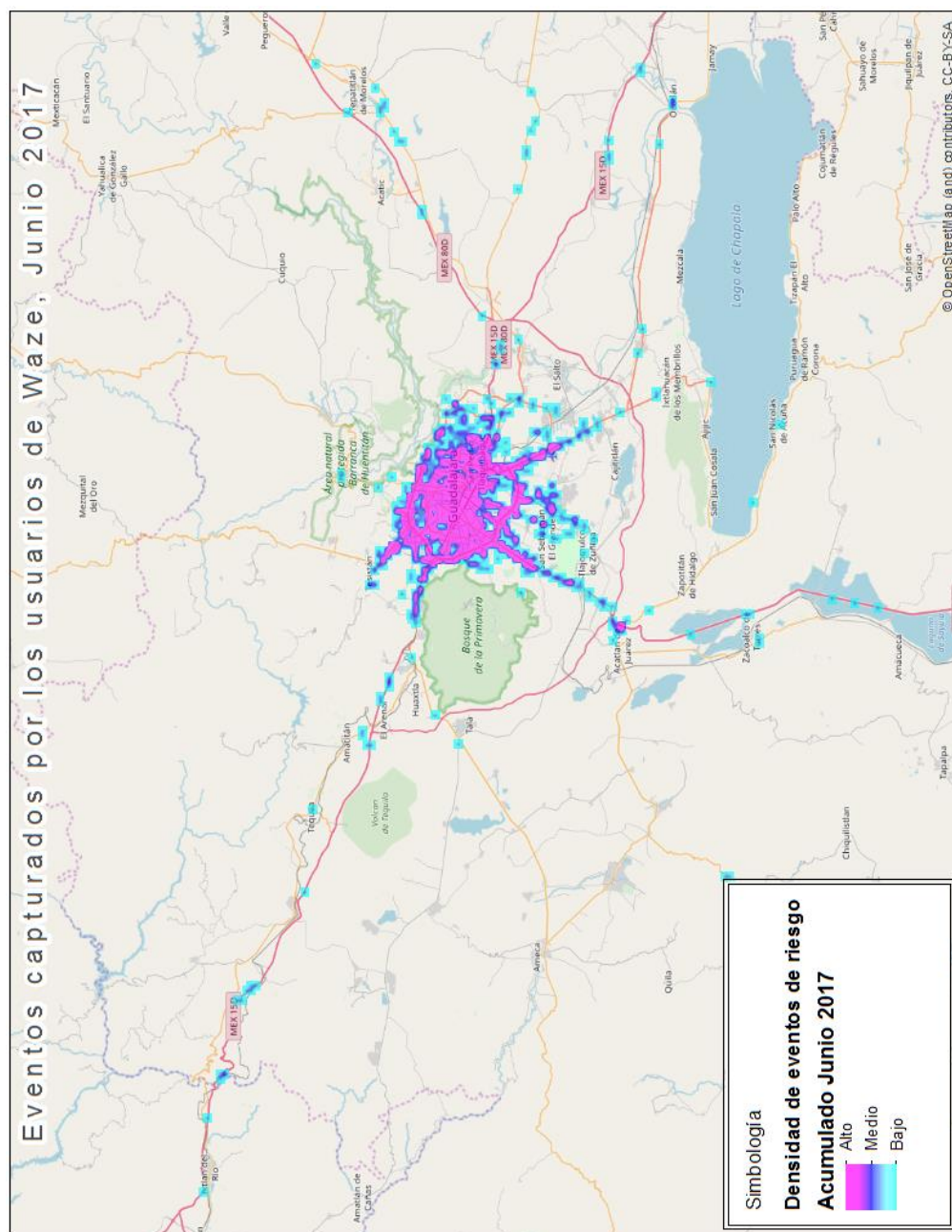
Figura 3.42 Densidad de eventos de riesgo, julio 2017, acercamiento a la región Querétaro – Ciudad de México.

En julio de 2017 se registran 26,676 eventos asociados a riesgo, se pueden identificar zonas en donde se concentran dichos eventos.



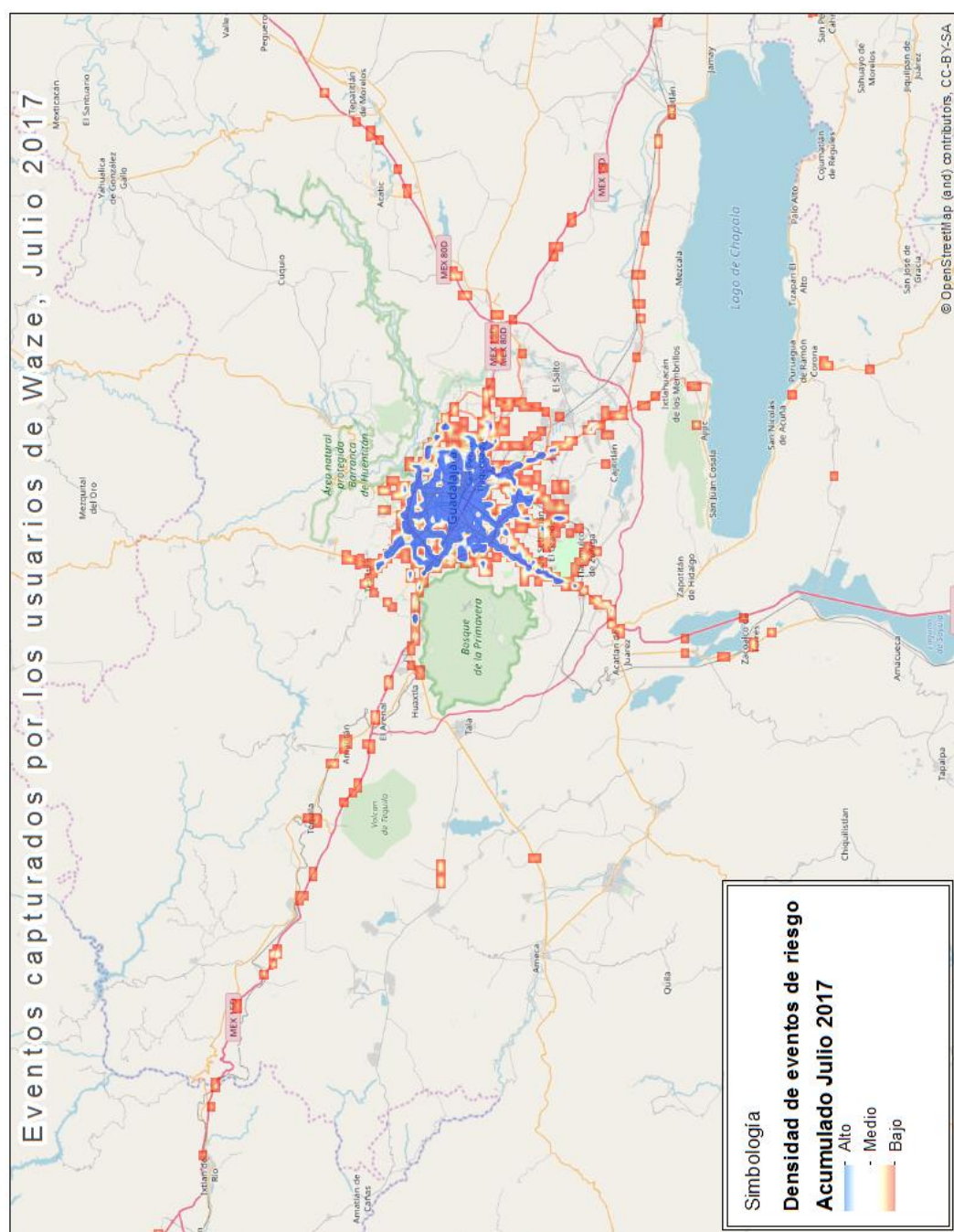
Fuente: Elaboración propia a partir de los datos de Waze

Figura 3.43 Densidad de eventos de riesgo, mayo 2017, acercamiento a la región Guadalajara.



Fuente: Elaboración propia a partir de los datos de Waze

Figura 3.44 Densidad de eventos de riesgo, junio 2017, acercamiento a la región Guadalajara.



Fuente: Elaboración propia a partir de los datos de Waze

Figura 3.45 Densidad de eventos de riesgo, julio 2017, acercamiento a la región Guadalajara.

Conclusiones

Los datos de *Waze* procesados en este estudio podrán servir para predecir el impacto en carreteras, autopistas y redes de transporte público causado por cierres, trabajos de mantenimiento en carreteras, eventos relacionados a riesgos y recomendar cambios en los cronogramas de tránsito, políticas de prevención y atención. Se podrá detectar y predecir la probable ocurrencia de incidentes no planeados como un accidente de tráfico o eventos meteorológicos y con esta información sugerir respuestas óptimas de acción para disminuir los daños y el tiempo de espera.

Es importante tener en cuenta que las aplicaciones de Big Data no se darán de forma automática. Su implementación no sólo está sujeta a la disponibilidad de expertos en su manejo o la previa inversión en TIC, sino también a las barreras que puedan imponer el contexto institucional, político y cultural particular de los sectores, como así también los intereses económicos de los actores involucrados. Asimismo, un tema central en la aplicación de políticas exitosas será el tratamiento de la privacidad de los datos de las personas para que no haya un daño en la integridad de las mismas y un descrédito de la herramienta que obstaculice un mejor aprovechamiento de sus beneficios.

Como se mencionó anteriormente, ya existen conjuntos de datos masivos en las organizaciones de transporte, por lo que estas entidades y dependencias deben decidir cuál de ellas utilizar y en dónde buscar datos complementarios. Este primer obstáculo del proceso de recolección de datos incluye la extracción y recopilación de datos en cualquier formato desde cualquier lugar en el que se encuentren disponibles. Luego deben ser colocados en un repositorio de datos, preferiblemente uno que no esté restringido por las restricciones tradicionales de gestión de datos, antes de ser limpiado y organizado en modelos estandarizados adecuados para el análisis.

Enseguida, se debe superar el siguiente obstáculo del proceso, que es comprender la historia contada por los datos, esto ocurre mediante la fusión de datos específicos que fueron limpiados y organizados en el primer paso, con otros datos modelados y aplicando técnicas de análisis predictivo para llegar a un cuadro más completo. El análisis estadístico, la simulación y la optimización se pueden aplicar para explotar las relaciones identificadas en los datos, para planificar y predecir lo que es probable que suceda en diferentes escenarios. Estas ideas deben ser presentadas en visualizaciones de datos que sean convincentes e informes intuitivos que revelen y comuniquen la información que importa.

Los datos masivos geoespaciales presentan desafíos y oportunidades, en particular para la gestión integral del transporte. A partir de herramientas geoinformáticas se pueden realizar modelos de administración del tránsito, transporte público y planificación urbana. También se pueden realizar aplicaciones dinámicas continuas de información del transporte tales como la detección de tráfico, identificación de patrones de comportamiento y uso, visualización en tiempo real.

En cuanto al ejemplo de utilización de datos de Waze, en donde se observan los incidentes de tráfico, peligros y eventos meteorológicos reportados por sus 50 millones de usuarios, es posible vincular estos eventos con herramientas de visualización predictivas. Así, serán posibles las visualizaciones en intervalos de tiempo donde se podrá observar cómo se propaga la congestión desde un punto o zona y, posteriormente implementar medidas de mitigación.

En la Unidad de Sistemas de Información Geoespacial se ha estado trabajando para identificar el potencial de utilización de estas herramientas, así como para obtener fuentes de información actualizada y verificada que puedan utilizarse para las líneas de investigación aplicada y desarrollo que están en marcha. Todo esto con la finalidad de proponer soluciones y metodologías que pueden apoyar a la labor sustantiva de la Secretaría de Comunicaciones y Transportes en cuanto a la operación, seguridad, conservación, planeación, seguimiento y administración en general de la infraestructura carretera del país.

Para tal fin se sugieren las siguientes acciones:

- Generar y proponer la creación de bases de datos masivos relacionadas al transporte
- Difundir en diversos medios y con otras instituciones tanto públicas como privadas, la generación, uso y aprovechamiento de los datos masivos geoespaciales
- Continuar con la investigación de los métodos, software (propietario y abierto) acerca del tema
- Homologar la generación de los datos masivos geoespaciales para el transporte a través de lineamientos de recopilación y compartición
- Identificar nuevas fuentes de datos masivos geoespaciales y probar su utilidad para los fines y objetivos planteados en la SCT
- Intensificar el uso de TICs con la finalidad de mejorar la movilidad, la sustentabilidad y el aprovechamiento de los datos

Bibliografía

- Calabrese F. *Understanding individual mobility patterns from urban sensing data: A mobile phone trace example*. Department of Real Estate, National University of Singapore, Singapore. 2013.
- Chen, P. Zhang, C. *Data-intensive applications, challenges, techniques and technologies: A survey on Big data*. University of Macau, China. 2014.
- Cöltekin, Arzu. Reichenbacher, Tumasch. *High Quality Geographic Services and Bandwidth Limitations*. GIScience-GIVA, Department of Geography, University of Zurich, Switzerland. 2011
- Cukier K. *Big Data. La revolución de los datos masivos*. Turner.
- Dasgupta Arup. *The Continuum: Big Data, Cloud & Internet of Things*. IBM Blog. EUA. 2017. <https://www.ibm.com/blogs/internet-of-things/big-data-cloud-iot/>
- Dyson, George. *The Big Data Problem Will Also Be A Problem For Science 2.0*. edge.org
- Gandomi A. Haider M. *Beyond the hype: big data concepts, methods and analytics*. Ryerson University. Ontario, Canada. 2014
- IBM. *The four V's of Big Data*. 2014
- Labrinidis A. Jagadish H. *Challenges and Opportunities with Big Data*. Universidad de Pittsburgh. Universidad de Michigan. 2015.
- Laney Doug. *3D Data Management: Controlling Data Volume, Velocity and Variety*. MetaGroup. 2001.
- Lee, R. *Big Data & Analytics*. Executive Consulting Services LLC.
- Li, Songnian; Dragicevic, Suzana; Anton, François; Sester, Monika; Winter, Stephan; Coltekin, Arzu; Pettit, Chris; Jiang, Bin; Haworth, James; Stein, Alfred; Cheng, Tao. *Geospatial Big Data Handling Theory and Methods: A Review and Research Challenges*. ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 115, 2016.
- Lohr Steve. *For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights*. The New York Times. 2014.

<https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html>

- Muñiz G. Rafael. *Marketing en el siglo XXI*. 5ª Edición. Centro de Estudios Financieros. España. 2014.
- Richter Kai-Florian, Winter Stephan. *Citizens as database: Conscious ubiquity in data collection*. International Symposium on Spatial and Temporal Databases. 2011. Lecture Notes in Computer Science, vol 6849. Springer, Berlin, Heidelberg
- Shekhar S. *Spatial Big Data*. Department of Computer Science and Engineering, University of Minnesota. 2012. <http://www-users.cs.umn.edu/~shekhar/talk/2013/13.5.sbd.asu.pdf>
- Vázquez P. *Datos masivos geoespaciales aplicados al transporte*. Publicación técnica No. 502. Instituto Mexicano del Transporte. México. 2017



Km 12+000 Carretera Estatal 431 "El colorado-Galindo"
Parque Tecnológico San Fandila
Mpio. Pedro Escobedo, Querétaro, México
CP 76703
Tel +52 (442) 216 9777 ext. 2610
Fax +52 (442) 216 9671

publicaciones@imt.mx

<http://www.imt.mx/>

Esta publicación fue desarrollada en el marco de un sistema de gestión de
calidad certificada bajo la norma ISO 9001:2015