INSTITUTO MEXICANO DEL TRANSPORTE

Análisis geoespacial de tráfico en la RNC, utilizando datos colaborativos

Juan Carlos Vázquez Paulino
Elsa María Morales Bautista
Jonatan Omar González Moreno
Miguel Ángel Backhoff Pohls

Publicación Técnica No. 614
Sanfandila, Qro.
2020

ISSN 0188-7297

Esta investigación fue realizada en la Unidad de Sistemas de Información Geoespacial del Instituto Mexicano del Transporte, por el Lic. Juan Carlos Vázquez Paulino, la Ing. Elsa María Morales Bautista, el Geog. Jonatán Omar González Moreno y el M.G. Miguel Ángel Backhoff Pohls; es el producto final del proyecto de investigación interna OI-04/19 “Análisis geoespacial de tráfico en la RNC, utilizando datos colaborativos” y forma parte del portafolio de proyectos de la Coordinación de Economía de los Transportes y Desarrollo Regional.

Asimismo, se agradecen los atinados comentarios del Dr. Guillermo Torres Vargas, Coordinador de Estudios Económicos y Sociales del Transporte.

Contenido

Página

Índice de figuras iv

Sinopsis vii

Abstract ix

Resumen Ejecutivo xi

Introducción 1

1 Antecedentes 3

2 Proceso de los datos 8

3 Plataforma Waze 19

4 Análisis 27

5 Conclusiones 50

6 Bibliografía 52

Índice de figuras

Página

Figura 3.1 22
Figura 3.2 23
Figura 3.3 24
Figura 4.1 30
Figura 4.2 32
Figura 4.3 33
Figura 4.4 34
Figura 4.5 34
Figura 4.6 35
Figura 4.7 35
Figura 4.8 36
Figura 4.9 36
Figura 4.10 37
Figura 4.11 37
Figura 4.12 38
Figura 4.13 38
Figura 4.14 39
Figura 4.15 39
Figura 4.16 40
Figura 4.17 40
Figura 4.18 41

Figura 4.19 42
Figura 4.20 43
Figura 4.21 44
Figura 4.22 44
Figura 4.23 45
Figura 4.24 46
Figura 4.25 47
Figura 4.26 48
Figura 4.27 49

Sinopsis

En este trabajo, se da una breve definición acerca de los datos masivos. Se presentan algunas de sus características tales como el tamaño, la velocidad, la variedad, la veracidad, la viabilidad, la visualización y el valor, todos estos de suma importancia.

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. Es así que se generaron dos pruebas piloto utilizando diferentes plataformas y métodos para la visualización y análisis del tráfico en diversos lugares de México.

Abstract

In this paper, a brief definition about big data is given. Some of its characteristics are presented such as size, speed, variety, veracity, feasibility, visualization and value, all of which are of utmost importance.

The main phases of the geospatial massive data process are identified, as well as the main representation techniques 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 of time and by processing them and generating digital maps, patterns of occurrence of events captured by the community can be identified. Thus, two pilot tests were generated using different platforms and methods for the visualization and analysis of traffic in various places in Mexico.

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.

Introducción

Este proyecto forma parte de la línea de investigación de Sistemas de Información Geoestadística para el Transporte y es útil para que los responsables de la toma de decisiones de la SCT, tengan información, procesos y métodos que les permitan administrar y operar eficientemente la infraestructura carretera. Así mismo, esta investigación contribuirá al cumplimiento del objetivo 6 “Desarrollar integralmente y a largo plazo al sector con la creación y adaptación de tecnología y la generación de capacidades nacionales”, del Programa Sectorial de Comunicaciones y Transportes , y también se encuentra alineado a los objetivos 3.5 “Hacer del desarrollo científico, tecnológico y la innovación pilares para el progreso económico y social sostenible” y, 4.9 “Contar con una infraestructura de transporte que se refleje en menores costos para realizar la actividad económica” del Plan Nacional de Desarrollo 2013-2018 . Asimismo, se contribuirá con el cumplimiento de algunos de los Objetivos de Desarrollo Sostenible de la ONU (Sustainable Developmente Goals), por ejemplo, el Objetivo 8 Trabajo Decente y Crecimiento Económico, Objetivo 9 Industria, Innovación e Infraestructura, Objetivo 11 Ciudades y Comunidades Sostenibles, Objetivo 12 Producción y Consumo Responsables

En publicaciones anteriores se han mostrado los resultados acerca de la investigación y aplicación de los *Datos Masivos Geoespaciales al Transporte* (Publicación Técnica 502) y en la Publicación Técnica 540, *Datos masivos geoespaciales para identificación de patrones de riesgo en la RNC*, se menciona el proceso de los datos de tráfico obtenidos de la plataforma Waze, en donde se muestran los resultados de aplicar análisis geoespacial para identificar patrones de riesgo de acuerdo a días, horas y lugares en la Red Nacional de Caminos.

El Instituto Mexicano del Transporte, pertenece al Programa de Ciudadanos Conectados (Connected Citizens CCP) de Waze en donde se obtienen los datos de tráfico, velocidades, obras y embotellamientos, y es posible descargarlos para toda la República Mexicana. Estos datos se procesan, transforman, se les aplica técnicas de análisis geoespacial y con los resultados obtenidos se pueden generar e identificar patrones de ubicación y ocurrencia temporal sobre accidentes, eventos identificados por usuarios de la plataforma, riesgos o peligros que afectan las carreteras y vialidades urbanas.

Este conjunto de datos masivos geoespaciales y su aplicación para identificar concentraciones de tráfico en tiempo real en la Red Nacional de Caminos, así como en vialidades urbanas donde existan datos, supone una importante oportunidad tecnológica para responder a la creciente demanda de este tipo de servicios. Esta oportunidad es estratégica y responde a demandas concretas de usuarios oficiales,

particulares y corporativos para el aprovechamiento de la gestión de estos grandes volúmenes de datos.

Más allá del impacto tecnológico, se ofrecerá un gran avance para todos aquellos usuarios e instituciones que deseen utilizar y explotar datos geográficos que tienen también el aspecto temporal dentro de sus soluciones. Se trata de añadir de forma factible y a bajo costo una dimensión de análisis y calidad en las soluciones que puede ser aplicada en múltiples ámbitos: todo tipo de aplicaciones basadas en mapas/cartografía, sistemas de localización, logística, control de tráfico, investigaciones de mercados, estudio del impacto ambiental, prevención y atención de riesgos y desastres, etc.

Actualmente en las carreteras del país, así como en las vialidades urbanas el nivel de congestión vehicular que se presentan en distintos intervalos de tiempo, hace que las personas, instituciones y empresas usuarias de la red viaria, experimenten perdidas en su productividad a causa de un incremento en el costo generalizado del transporte, que repercute en muchas ocasiones en la pérdida de bienes o productos de alta densidad económica, mismos que tendrán que ser repuestos, reconstruidos o ensamblados nuevamente, con un consecuente incremento en el costo del ciclo de producción con una disminución en los márgenes de utilidad esperados.

Algunas causas de la congestión en carreteras como en vialidades urbanas se deben entre otras cosas, al mal diseño de las obras civiles, que tienen impacto en la seguridad de los usuarios así como a eventos ligados a situaciones consideradas como pico o tope en el comportamiento temporal del tránsito (horarios de entradas y salidas escolares, laborales y de acceso y salida a parques industriales, eventos comunitarios sociales o religiosos).

Los datos masivos geoespaciales generados a partir del convenio establecido con la plataforma Waze, cuentan con variables que después de ser procesadas y al aplicárseles análisis espacial se pueden utilizar para identificar patrones de ocurrencia en determinada ubicación, así como patrones de temporalidad, ya sea por horas, días, semanas o meses.

Al utilizar estos datos será posible identificar, por ejemplo, la concentración de zonas donde se presenta tráfico denso, clasificación de zonas con paso libre, restringido o completamente cerradas de acuerdo a rangos de velocidad en carreteras o vialidades.

Así se podrán definir o calcular rutas alternas para desviar el flujo vehicular y de este modo los usuarios ahorrarían tiempo, recursos, combustible y aumentaría la seguridad, disminuiría el desgaste de los vehículos, etc. Se podrán generar predicciones sobre frecuencias de ocurrencia de tráfico, diseño de rutas más confortables y se contará con información para la gestión del tráfico.

1. Antecedentes

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 al 2019 y lo que va del año 2020.

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 Howe[1], uno de los primeros autores en emplear el término, estableció que el concepto de "crowdsourcing" depende esencialmente del hecho de que, debido a 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).

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

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 negocios también han 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

Este concepto es sumamente importante ya que los datos requeridos para este proyecto se generan a partir de la plataforma Waze y quienes recaban o capturan los datos son los usuarios de dicha plataforma, por lo que sí en algún momento la plataforma o los usuarios dejarán de hacerlo no se dispondría de ellos.

1.1 Definición de los datos masivos (Big Data)

De acuerdo con la publicación técnica 502 "*Datos masivos geoespaciales para el transporte*"[3], 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,

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

[3] https://imt.mx/archivos/Publicaciones/PublicacionTecnica/pt502.pdf

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)[4]. 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”[5].

1.2 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

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

[5] http://www.gartner.com/it-glossary/big-data

(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.

1.3 Los datos masivos geoespaciales y el transporte

De acuerdo con la 1ª Ley de Tobler[6] 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.

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.

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.

En cualquier momento una gran cantidad de vehículos, personas y bienes se encuentran en camino de un lugar a otro, ya sea cercano, a una distancia intermedia o muy lejanos, es por ello que para poder predecir para poder predecir la demanda o el tráfico del transporte, no sólo se requieren datos en tiempo real sino también datos históricos.

Los métodos de análisis tales como la minería de datos y el aprendizaje automático permiten aplicar el razonamiento inductivo en datos grandes, es decir, la detección

[6] Waldo Tobler. 1930. Geógrafo.

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.

2. Proceso de los datos

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".

2.1 Administración de datos

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 geoespaciales 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. 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.

Entonces, la adquisición de los datos colaborativos de Waze se hace a partir de un *feed* (fuente) de datos compartido por la plataforma y en donde se puede observar y descargarlos en formato XML.

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.

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. 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.

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 completo[7].

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

- 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

[7] http://www.powerdata.es/integracion-de-datos

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 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), al ser datos que fueron capturados por la plataforma participativa Waze, el formato 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 subdirectorio 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

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 construidos 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 espagueti[8], el modelo de datos de red, el modelo de datos topológicos y el modelo regular (ráster) e irregular (Voronoi[9], árbol kd[10], árbol de partición binario[11]),

[8] 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.

[9] 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

[10] 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

[11] 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/

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 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 operaciones[12]:

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

[12] 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

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 x 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 generalizado[13].

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.

[13] http://desktop.arcgis.com/es/arcmap/10.3/tools/spatial-analyst-toolbox/how-point-density-works.htm

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 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 compartibilidad, 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 compartibilidad 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 Aplicaciones[14] 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 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.

[14] 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/

3. Plataforma Waze

3.1 Cómo funciona

El *crowdsourcing* (*Antecedentes, pag. 17*), 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 meteorólogicos. 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 barata[15]. 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 usuarios[16].

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 posible[17], ver Figura 3.1, Figura 3.2 y Figura 3.3

[15] https://www.entrepreneur.com/article/266030

[16] 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.

[17] Tomado de ¿Cómo funciona Waze? https://support.google.com/waze/answer/6078702?hl=es

3.2 Descripción del programa *Citizen Connected Partner (actualización noviembre 2020)*

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.

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*[18] en formato JSON o XML que se actualiza cada 2 minutos.

3.3 Especificaciones de los datos de Waze

En el documento *Waze Traffic Dataspec*[19] se detalla la estructura y el contenido del *feed*[20] 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.

[18] 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

[19] Waze-traffic-dataspecv2.7.1.pdf

[20] Definición de *feed*

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*[21] 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.
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 3.1 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_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 3.2 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 3.3 Tipos de carretera definidos en la plataforma Waze.

3.3.1 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 grandes[22]. 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.

[22] https://www.w3.org/XML/

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 (<br/>). 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]][23]>

3.3.1.1 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>
```

[23] Ibidem.

```
<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>
```

4. Análisis

4.1 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 (la plataforma actualiza cada media hora la fuente o "*feed*"), para el propósito del presente proyecto se realizaron dos pruebas piloto.

En primer lugar se seleccionó un periodo de tiempo para procesar solamente los datos de ese intervalo.

El periodo seleccionado fue del 13 de marzo de 2020 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.

Enseguida, se aplicó análisis espacial 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 el día de la semana.

Los resultados obtenidos se cargaron a la plataforma CARTO[24] en su versión gratuita, en donde se pueden desplegar los datos, navegar a través de todo el territorio, apagar y encender temas así como hacer búsquedas en las capas de información geoespacial. Ahí mismo se podrán observar y consultar capas de información procesadas a partir de datos de Waze y que corresponden a cada viernes a partir del día 13 de marzo 2020, que se utiliza como línea base para el análisis y comparación.

En segundo lugar, de la suite ofrecida por Google, se utilizó Google Data Studio[25], herramienta gratuita que permite convertir datos en informes y paneles o tableros

[24] https://carto.com/

[25] https://datastudio.google.com/navigation/

de control, los cuales son personalizables y permiten crear consultas y compartirlos, genera gráficas, mapas, tablas de datos, tablas dinámicas, etc.

4.1.1 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/rtserver/web/TGeoSStk=ccp_partner&ccp_partner_name=SantiagodeQueretaro&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.

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.

Primer prueba piloto

Derivado de la contingencia sanitaria que se presenta en México a raíz de la epidemia mundial de COVID19[26], se restringieron muchas actividades en el país, entre ellas:

- Escuelas
- Oficinas
- Comercios
- Servicios de transporte

Y al mismo tiempo se implementó el trabajo en casa o “home office” para las personas y empresas que fuera posible y lo pudieran aplicar. Esto como una estrategia para evitar el contagio de la enfermedad a través del contacto

[26] https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q-a-coronaviruses

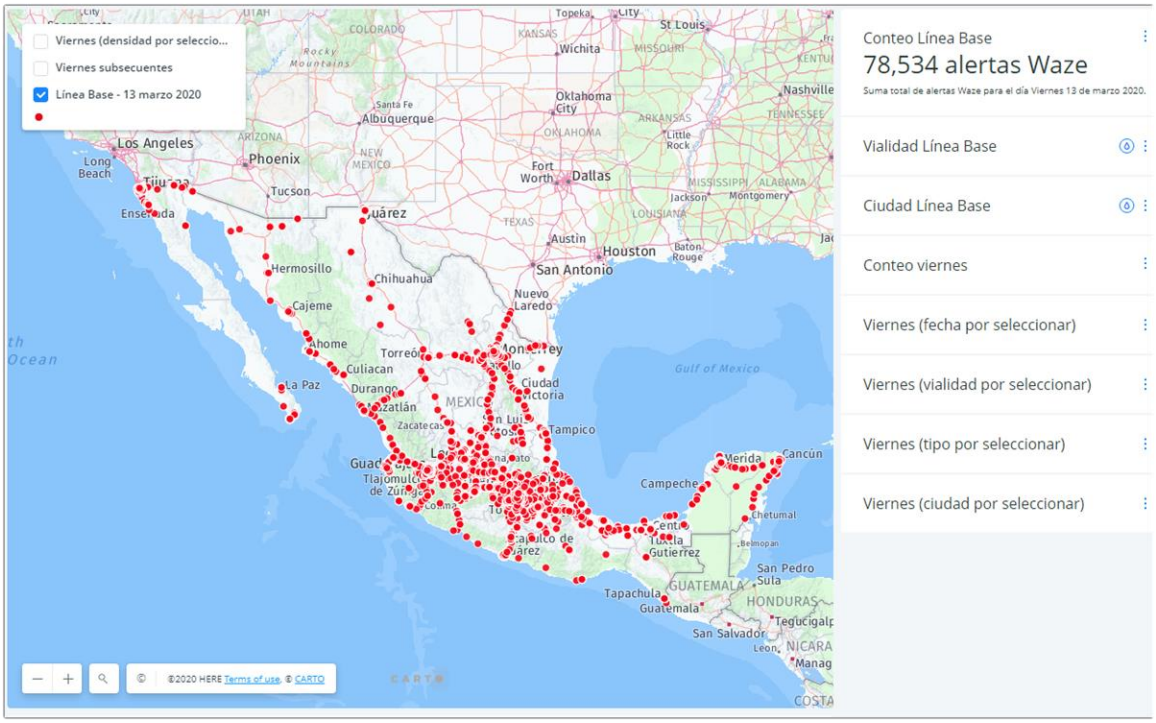
Para la prueba piloto se tomó la línea base temporal el día 13 de marzo de 2020, esto debido a que fue el último viernes con “normalidad” antes de que fuera declarada la contingencia sanitaria.

Posteriormente se procesaron los datos de Waze para los viernes siguientes, para esto se descargaron, transformaron de formato XML a XLS, enseguida se convirtieron a capa geoespacial y se subieron a la plataforma CARTO.

Una vez integrados los datos en la plataforma, se integró también la capa de la Red Nacional de Caminos 2019. De este modo se pudieron hacer comparaciones entre el día definido como “línea base” y los días viernes posteriores.

Se pudo observar que el tráfico disminuyo drásticamente conforme fueron pasando los días de confinamiento.

Un ejemplo del proceso de los datos provenientes de la plataforma Waze y aplicados a identificar concentración, temporalidad y ubicación de eventos relacionados al tráfico vehicular es la utilización de la plataforma CARTO[27].



Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.1 Atributos de las alertas emitidas por la plataforma Waze

[27] https://carto.com/

Esta plataforma permite generar mapas digitales a través de una suscripción con costo aunque también dispone de una versión gratis que ofrece menos opciones aunque es 100% funcional.

De acuerdo a su página web, CARTO es una plataforma para crear aplicaciones de Location Intelligence utilizando los flujos de datos disponibles. Permite crear una aplicación que da la posibilidad a los usuarios de visualizar, analizar y obtener información utilizando sus propios datos, así como para enriquecer su comprensión con flujos de datos disponibles a través de la página de CARTO. Los datos se almacenan como una tabla de base de datos PostgreSQL. Los conjuntos de datos CARTO ocupan espacio de almacenamiento en la cuenta del usuario, lo que significa que la cantidad de conjuntos de datos que puede tener en su cuenta depende de la cantidad de datos ocupados por los diferentes conjuntos de datos.

Un conjunto de datos se crea cuando el usuario importa un archivo a su cuenta usando la interfaz de usuario de Builder o la API de importación CARTO o cuando crea una tabla usando una consulta SQL. Los conjuntos de datos también se crean cuando se importa una tabla de base de datos en CARTO utilizando los conectores de base de datos CARTO.

Los mapas en CARTO se pueden construir con una apariencia variada, usando diferentes bibliotecas (CARTO.js, CARTO VL) o herramientas (Builder), pensando en diferentes públicos objetivos. Sin embargo, se requieren dos elementos principales para construir un mapa en CARTO: style properties y data.

Cuando se abre un mapa CARTO accede a los datos del conjunto y luego se define el estilo para visualizar los datos en el mapa. Esta mezcla de datos y estilo crea la capa del mapa que se representa cuando se carga.

CARTO proporciona diferentes herramientas y bibliotecas que permiten crear aplicaciones de Location Intelligence. En este momento las diferentes opciones para desarrollar aplicaciones son:

CARTO.js: una biblioteca de JavaScript que interactúa con diferentes API de CARTO para crear aplicaciones personalizadas con renderizado ráster.

CARTO VL: una biblioteca de JavaScript que interactúa con diferentes API de CARTO para crear aplicaciones personalizadas que aprovechan la representación vectorial.

Airship: un conjunto de herramientas diseñadas para facilitar el desarrollo de aplicaciones de Location Intelligence al ofrecer diseños, patrones básicos, plantillas, clases CSS, componentes, widgets y mucho más.

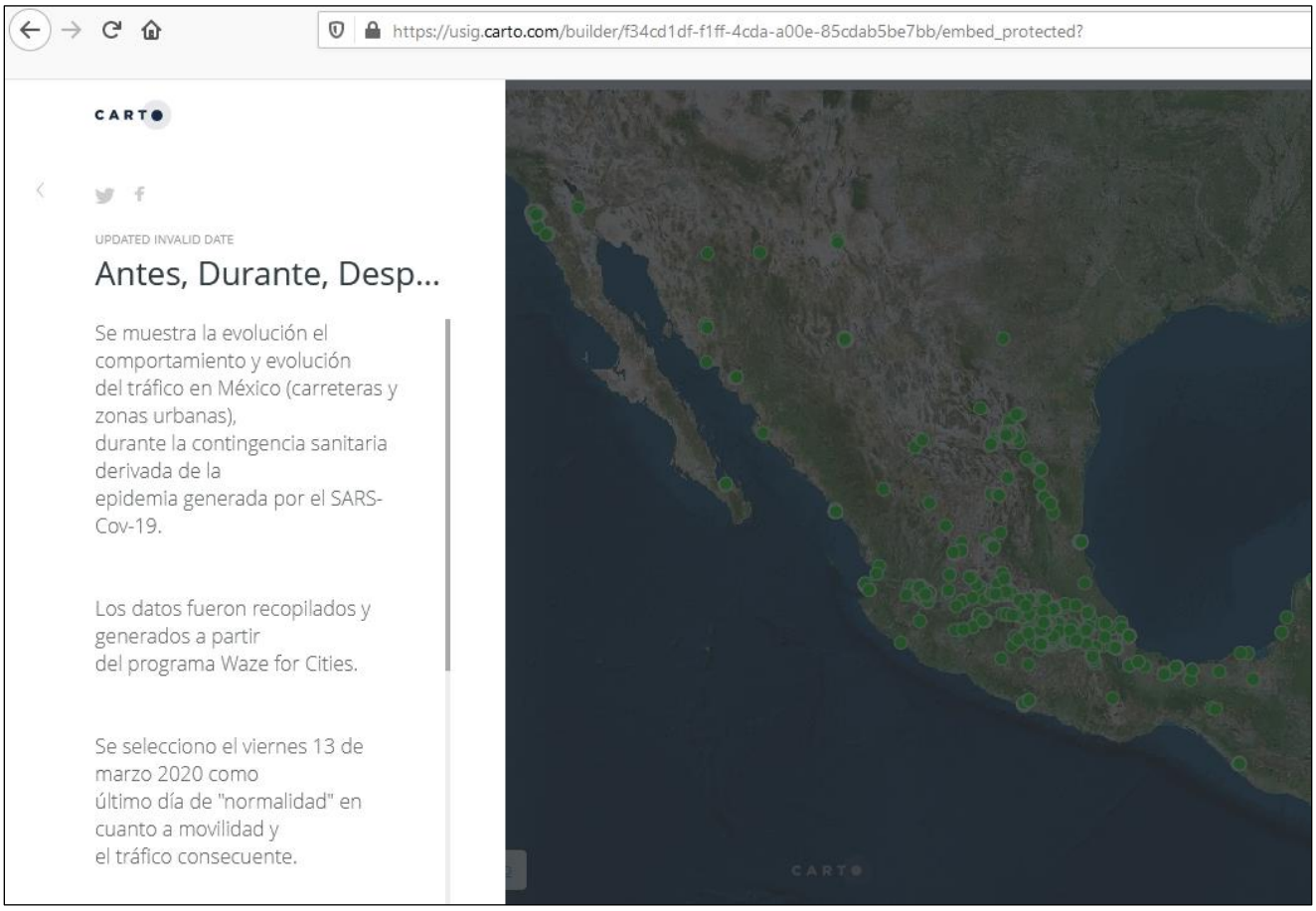
CARTO Frames: un paquete de Python para integrar mapas, análisis y servicios de datos CARTO en flujos de trabajo de ciencia de datos.

Python SDK: un contenedor de las API de CARTO diseñado para preparar, cargar, analizar y enriquecer sus datos.

API de CARTO: la puerta de entrada a todos los servicios que ofrece CARTO.

CARTO está construido sobre una base de datos llamada PostgreSQL. El SQL significa Lenguaje de Consulta Estructurado y se basa en las matemáticas llamadas álgebra relacional.

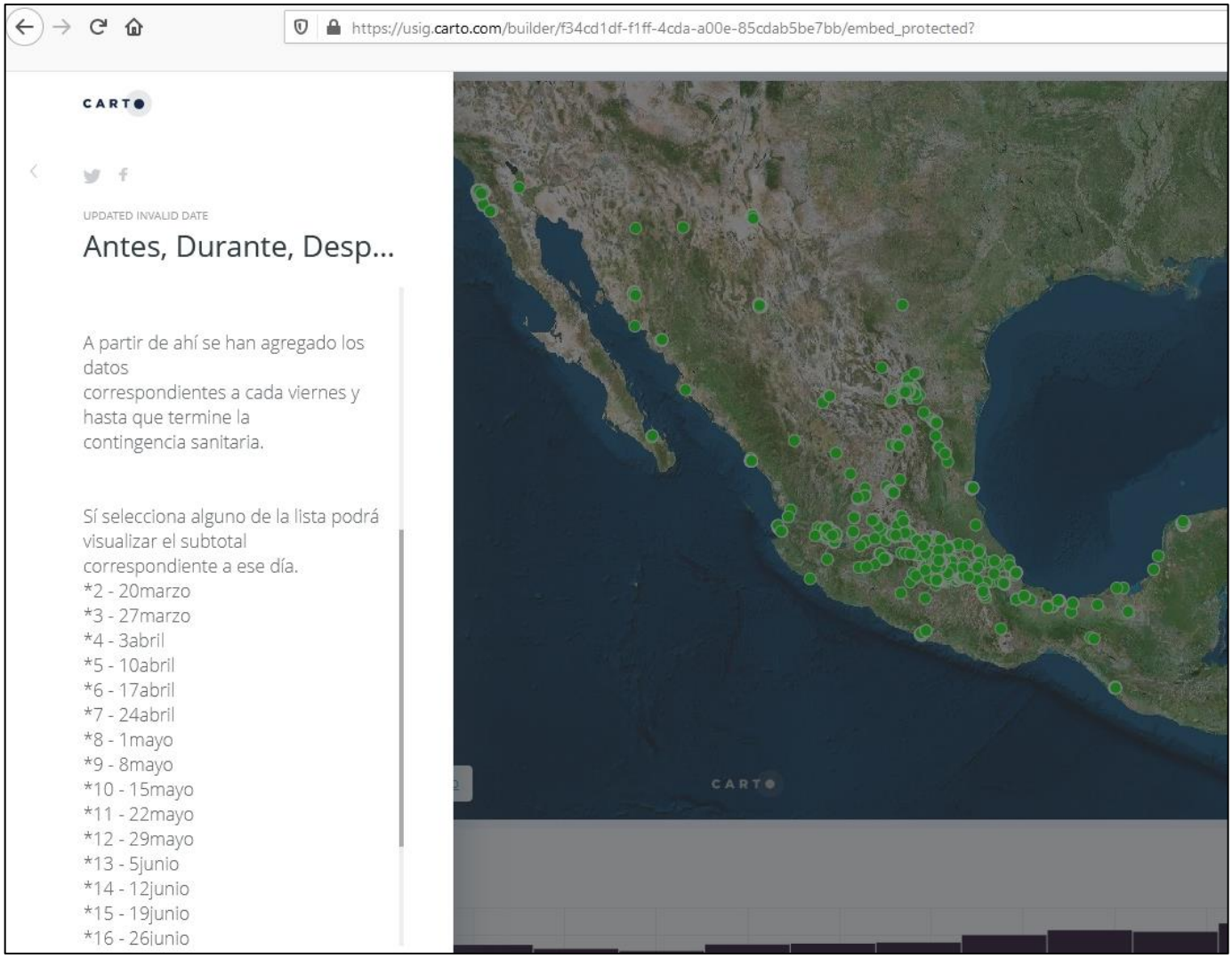
Las consultas SQL funcionan con datos organizados en tablas que son visualmente similares a una hoja de cálculo de Excel, pero tienen mecanismos subyacentes muy diferentes. Los nombres de las tablas están en minúsculas, no contienen espacios y son únicos dentro de una cuenta CARTO. Las tablas tienen columnas que tienen encabezados en minúsculas sin espacios. Los nombres de las columnas no pueden ser los mismos que algunos de las palabras clave en SQL[28].



Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.2 Caratula de inicio de prueba piloto en plataforma CARTO

[28] Información tomada de la página Web de CARTO.



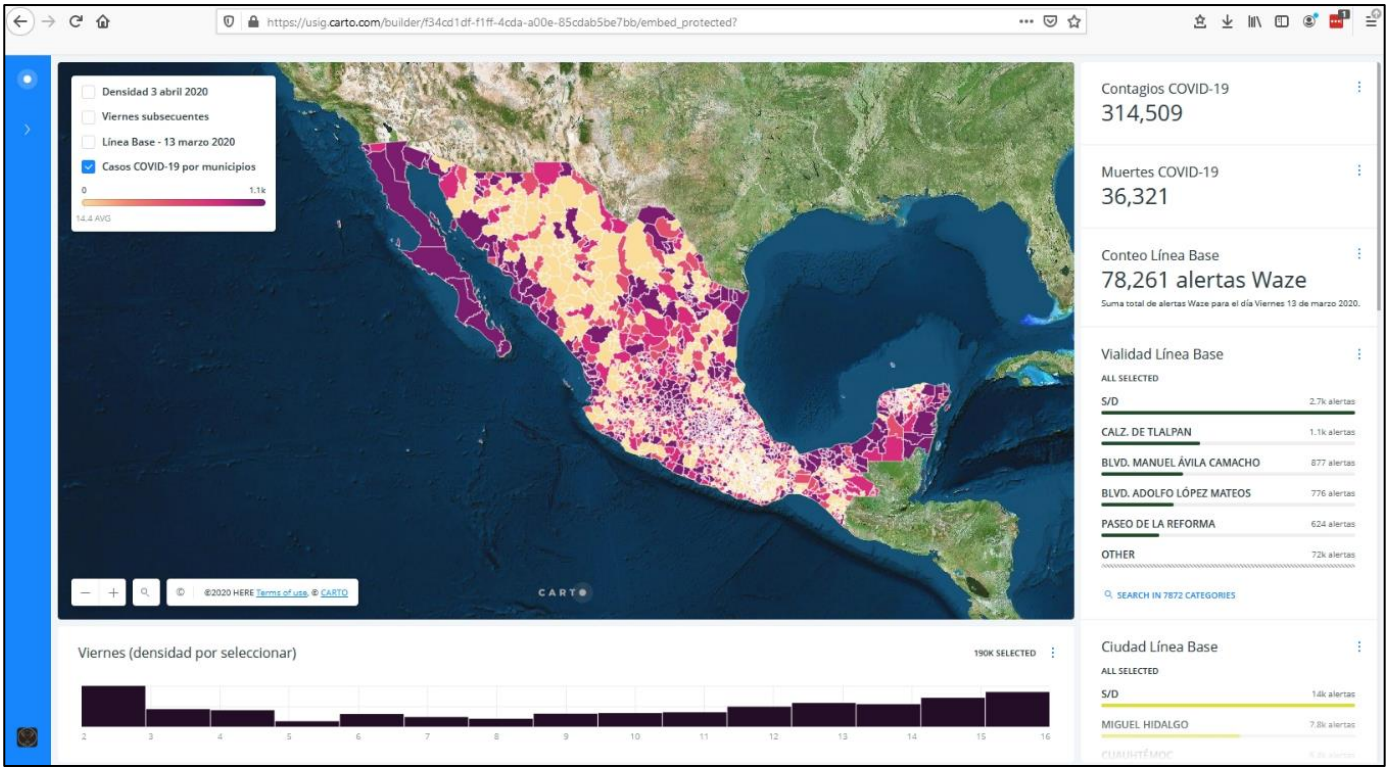
Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.3 Periodo de tiempo de estudio

En la Fig. 4.3 se observa el periodo de tiempo en donde se compara el tráfico para el día (13 marzo 2020) que representa la línea base y los días similares subsecuentes (cada viernes).

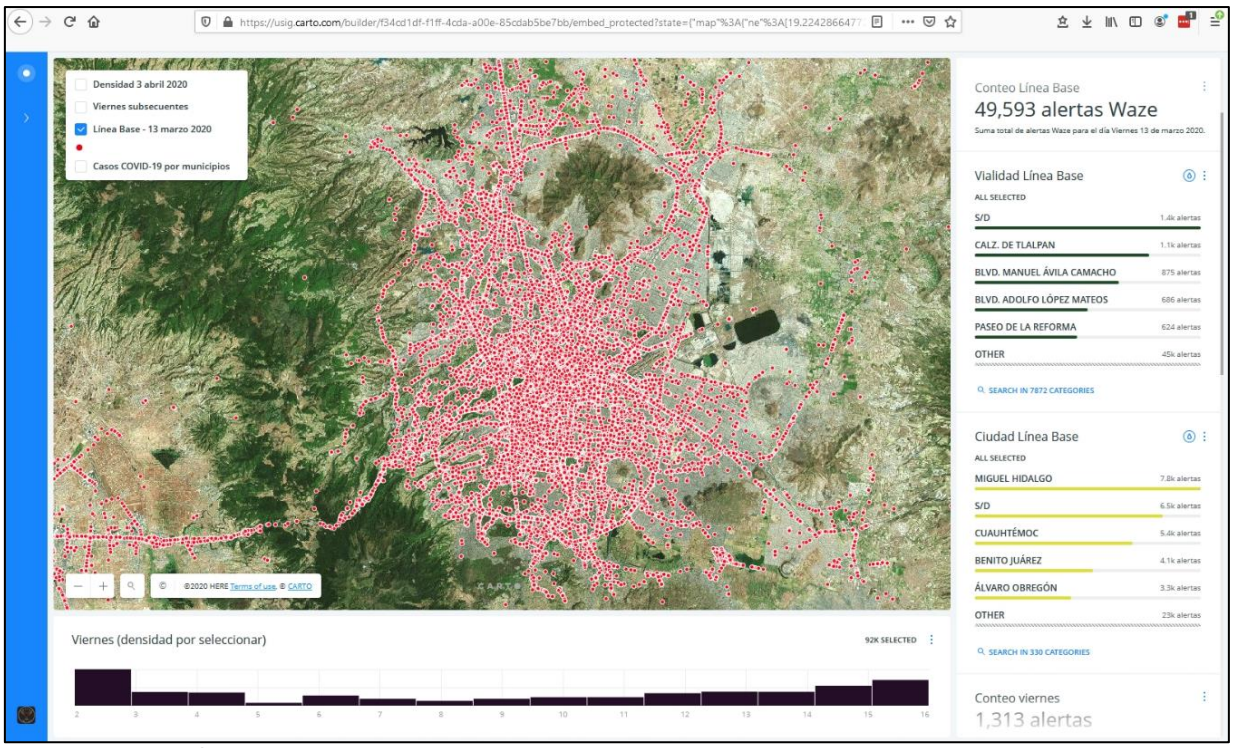
Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.4 Línea base – 13 marzo 2020



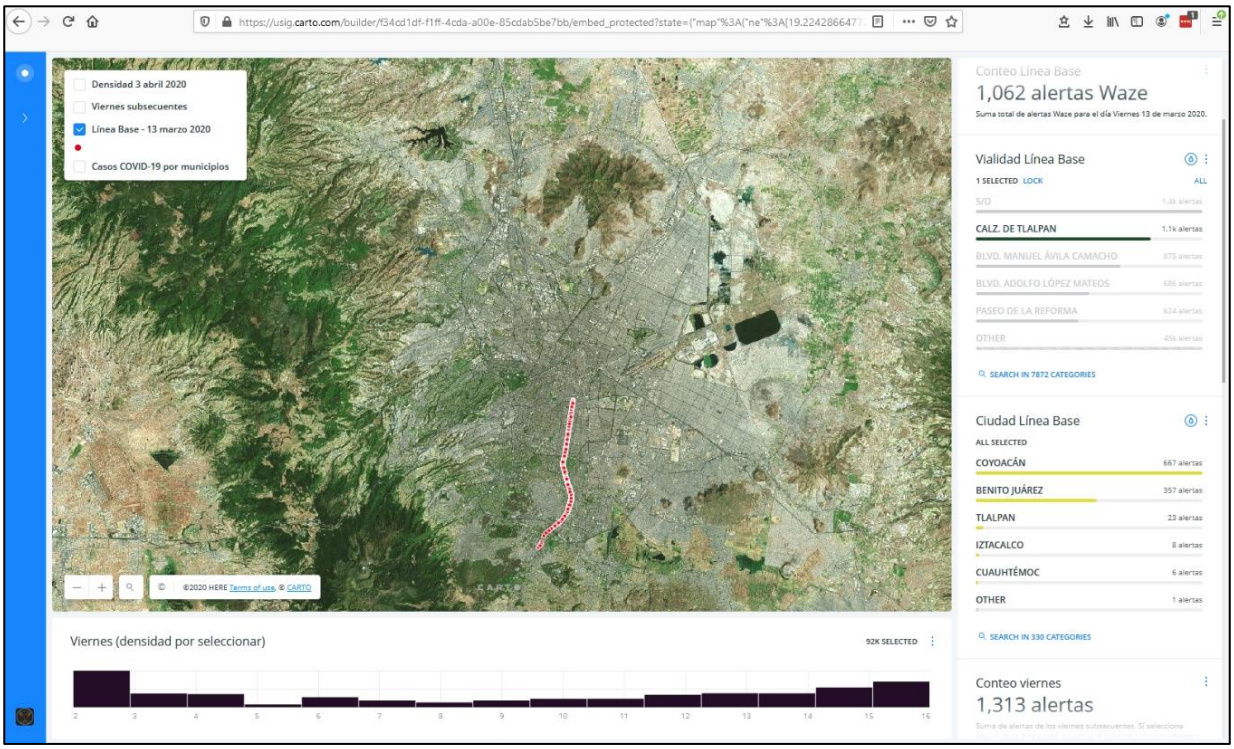
Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.5 Periodo de tiempo de estudio



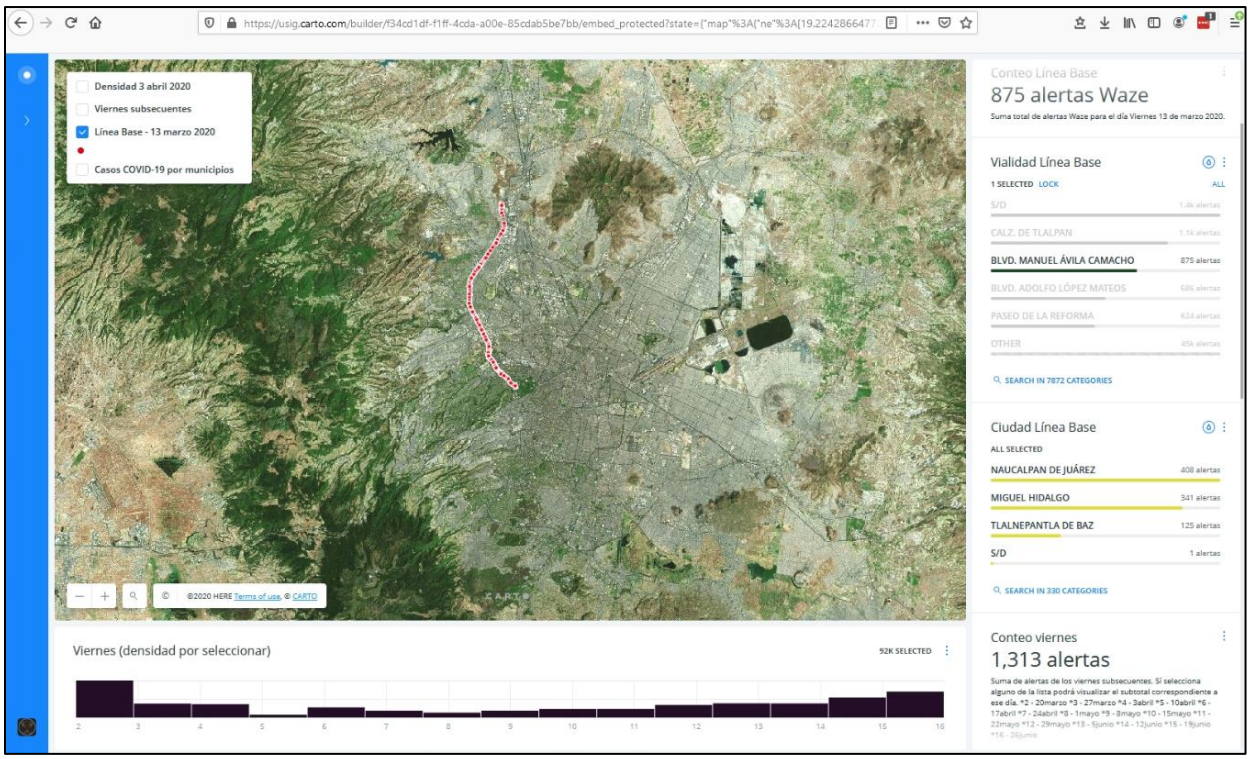
Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.6 Periodo de tiempo de estudio y acercamiento a la zona de la CdMX



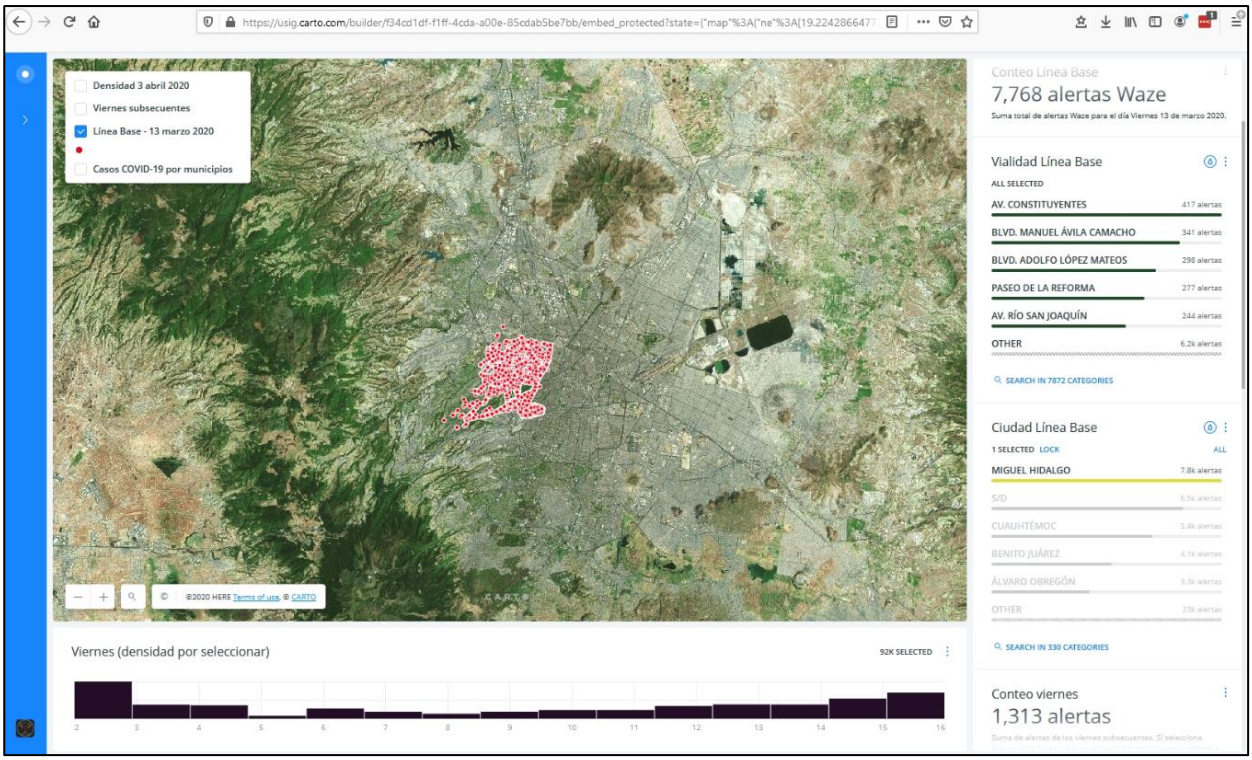
Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.7 Vialidad con más trafico el día 13 de marzo 2020



Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.8 Segunda vialidad con más trafico el día 13 de marzo 2020

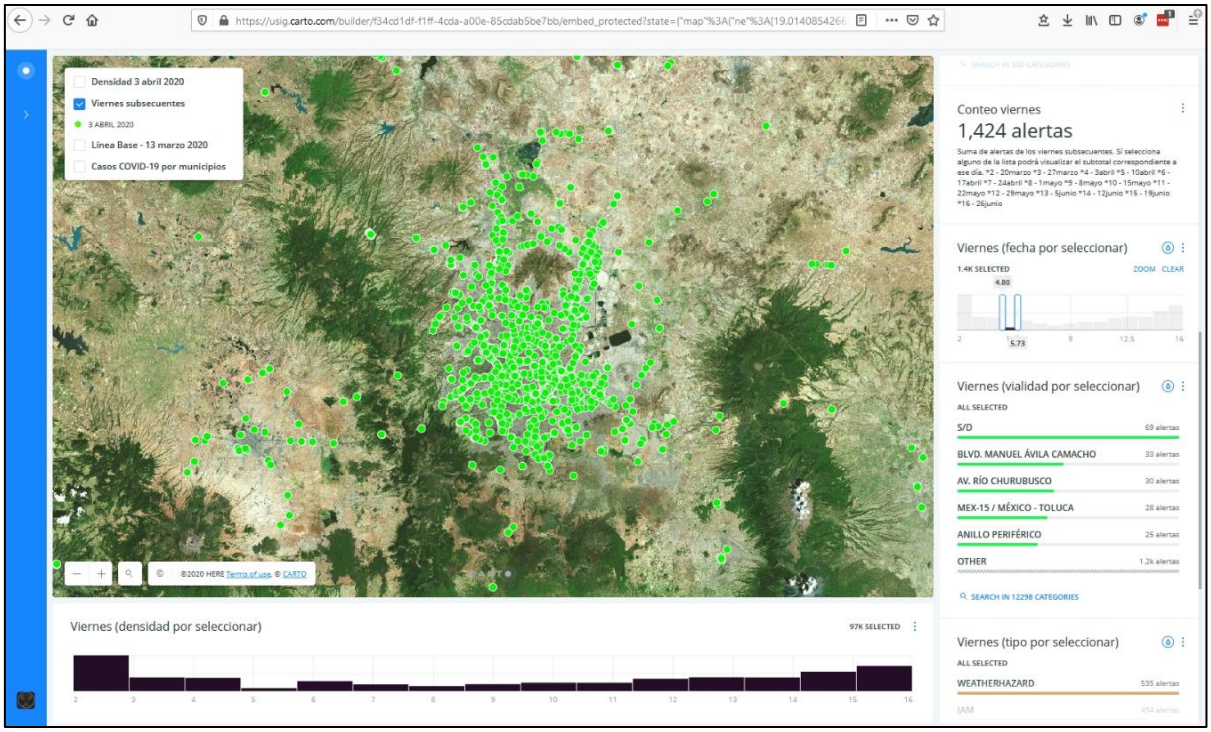


Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.9 Alcaldía en la CdMx con más trafico el día 13 de marzo 2020

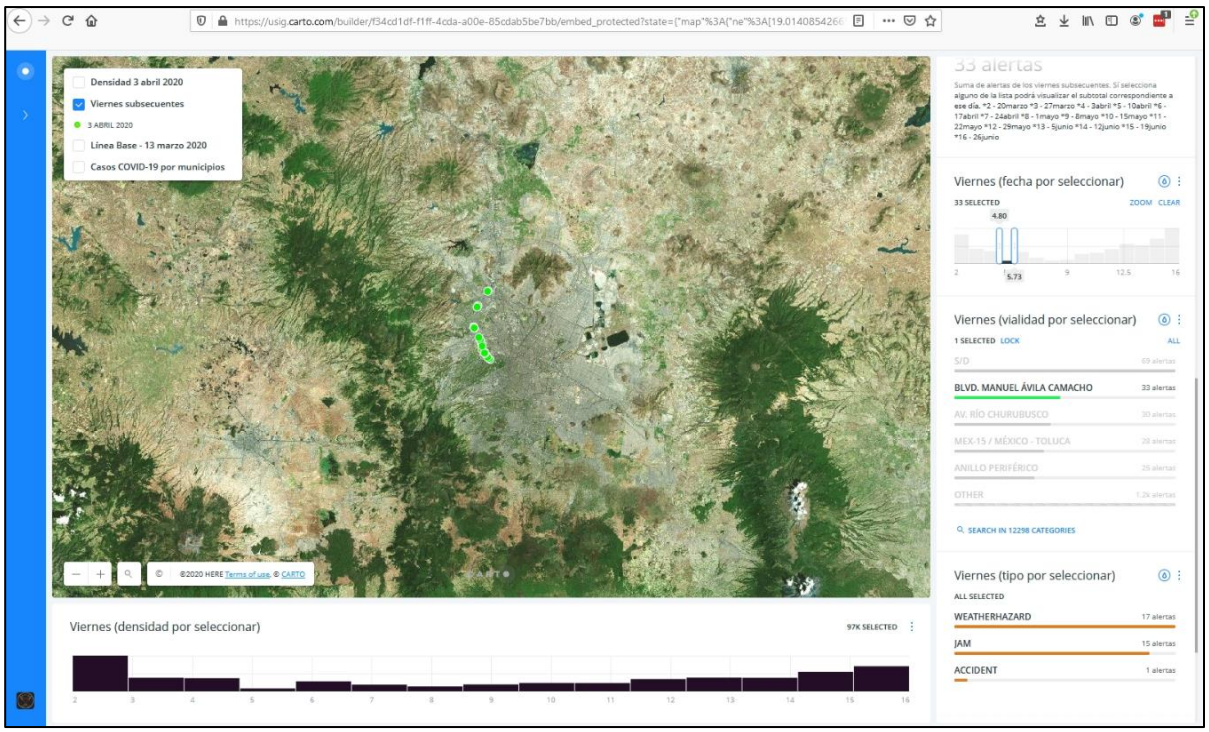
Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.10 Segunda Alcaldía en la CdMx con más trafico el día 13 de marzo 2020



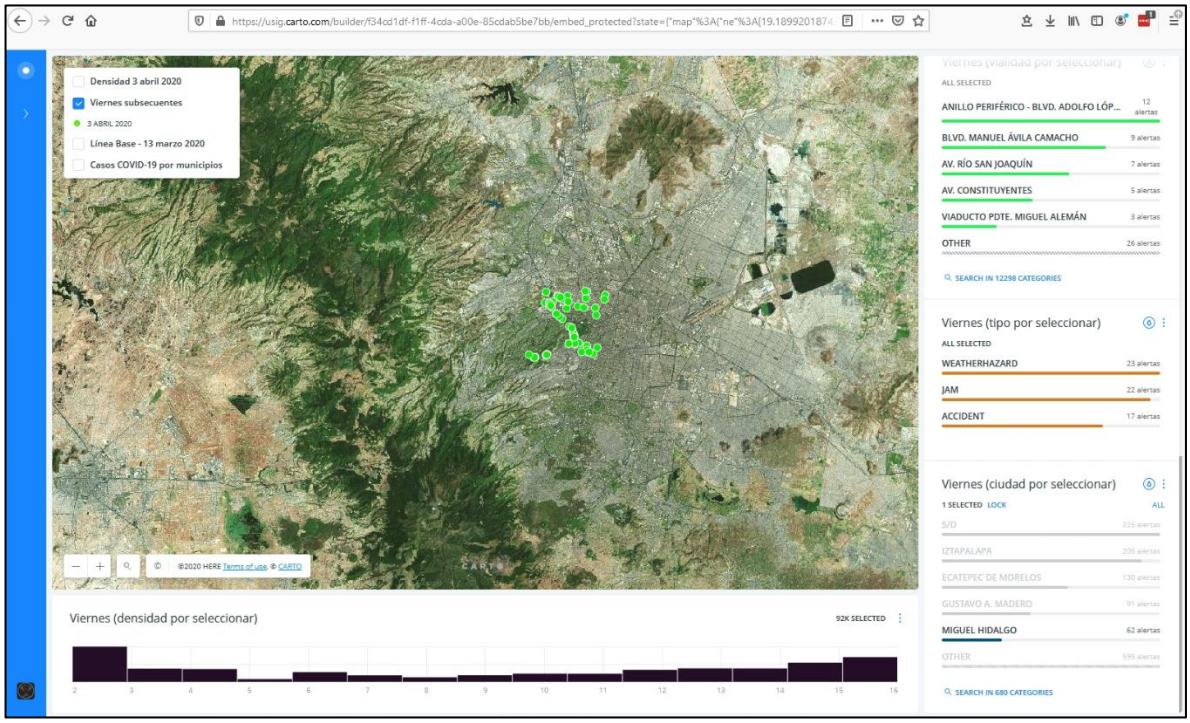
Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.11El día 3 de abril 2020 fue cuando hubo el tráfico de toda la pandemia



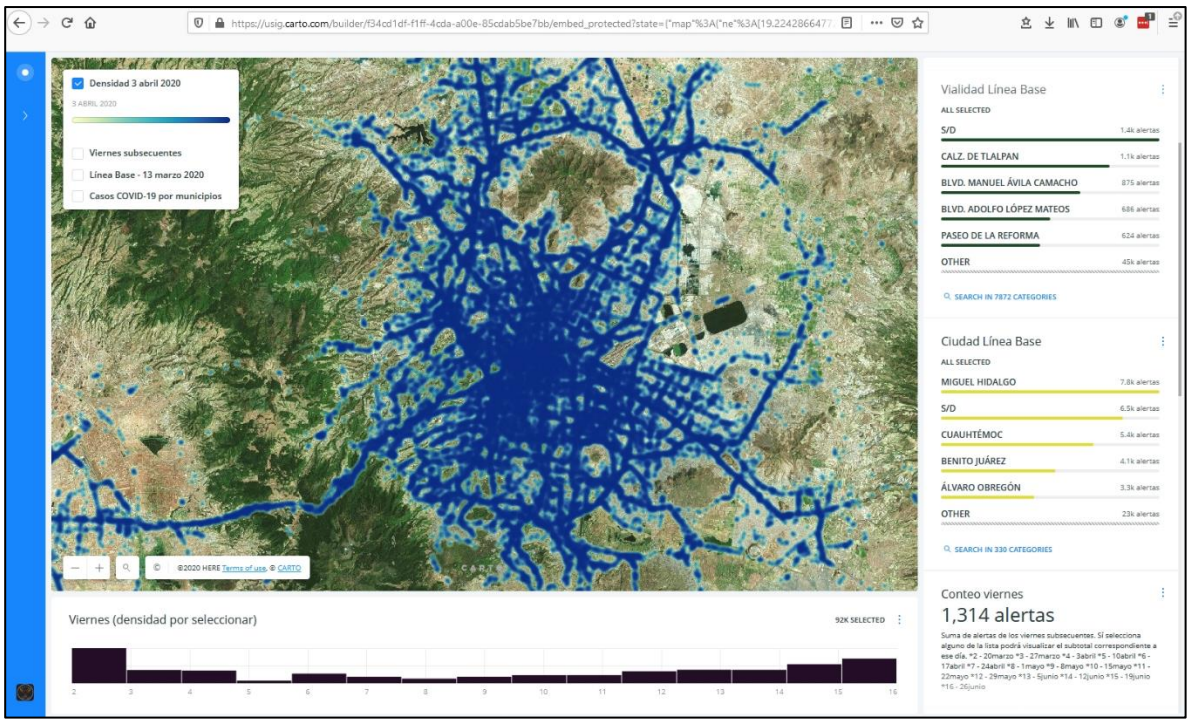
Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.12 El 3 de abril 2020 la vialidad con más trafico – 33 alertas



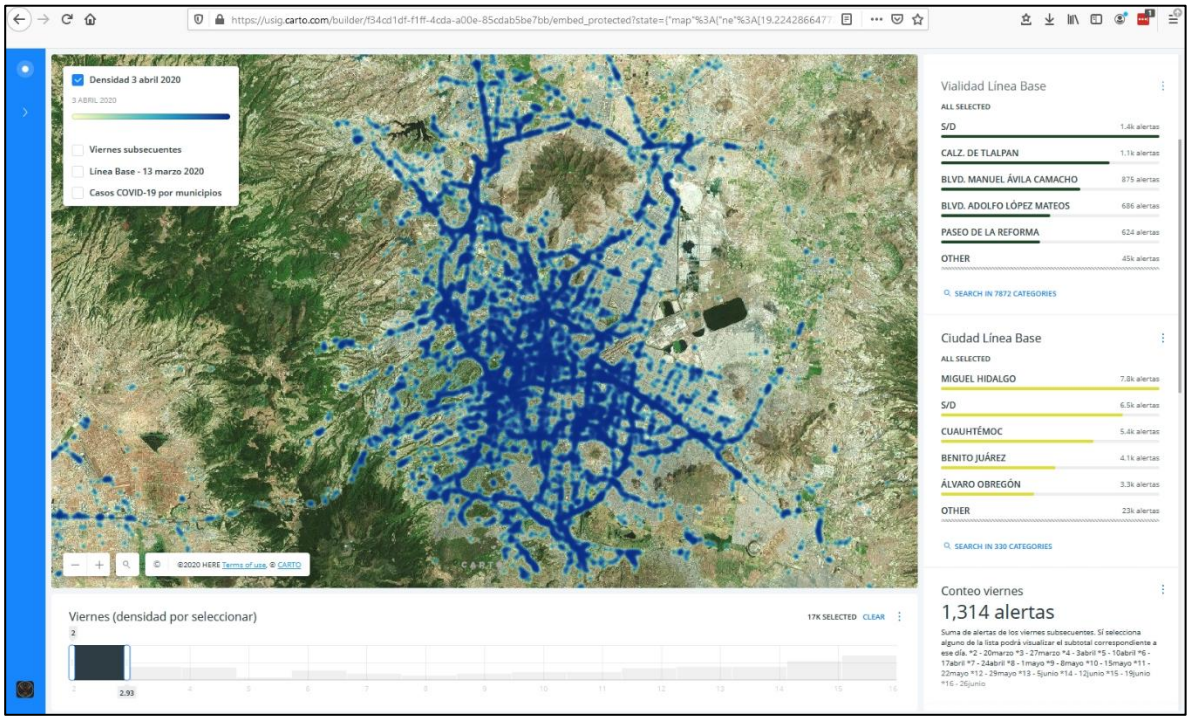
Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.13 El 3 de abril 2020 la alcaldía Miguel Hidalgo tuvo 62 alertas



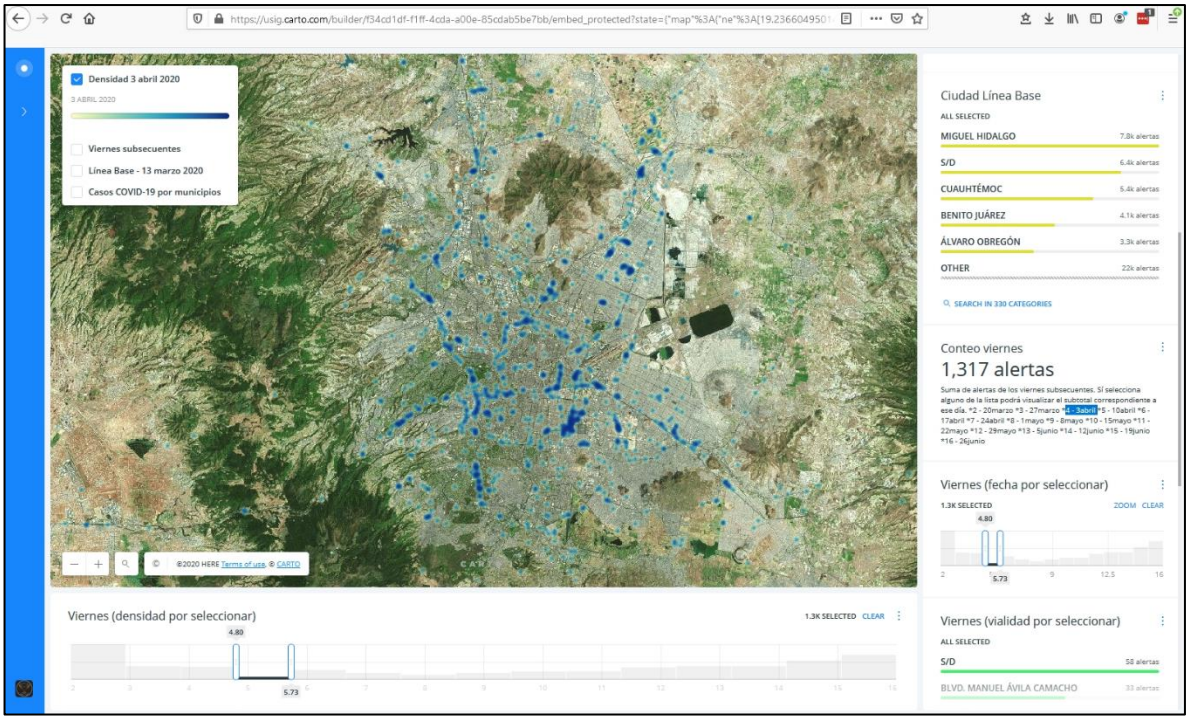
Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.14 Densidad de tráfico para todo el periodo de estudio



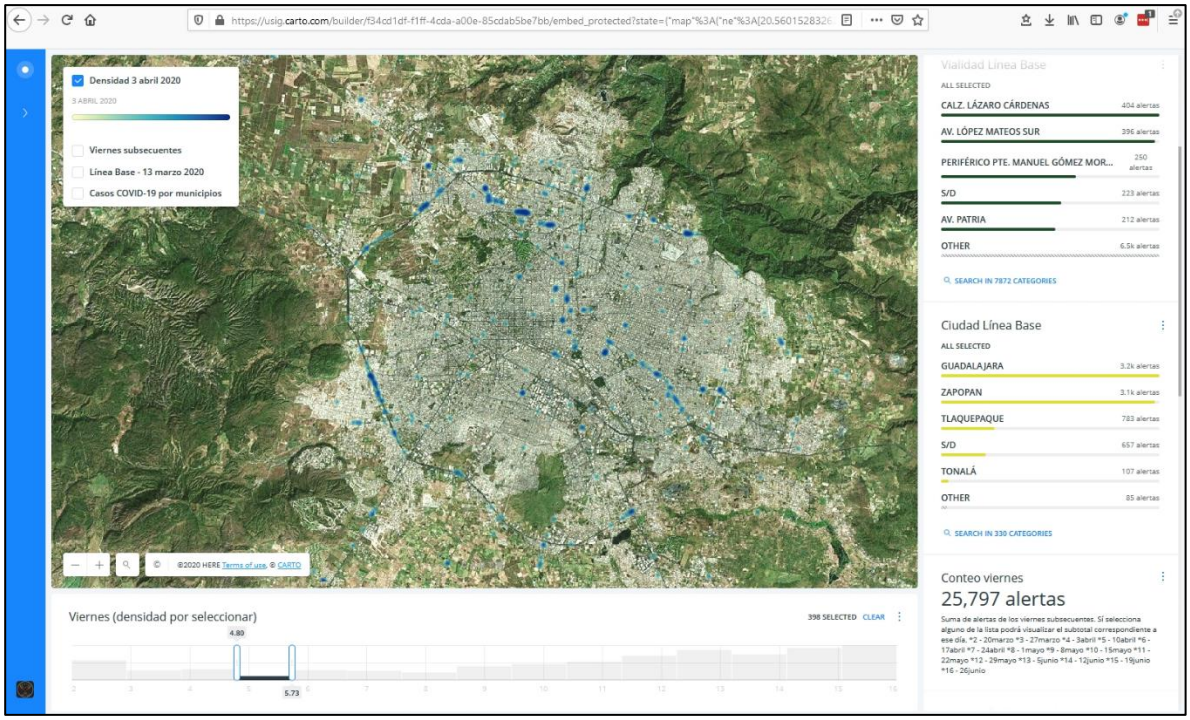
Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.15 Densidad de tráfico para el 20 de marzo 2020



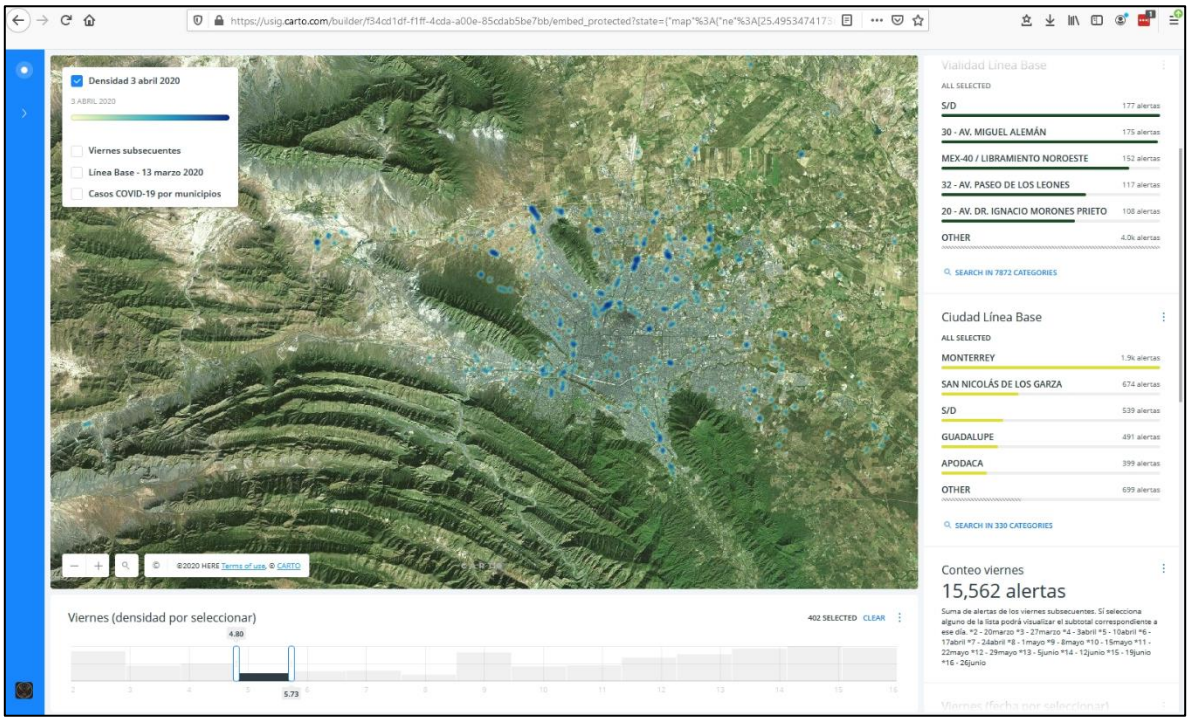
Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.16 Densidad detráfico del 3 de abril 2020 para la CdMx



Fuente: Elaboración propia a partir de la plataforma CARTO

Figura 4.17 Densidad de tráfico del 3 de abril 2020 para Guadalajara, Jal.



Fuente: Elaboración propia a partir de la plataforma CARTO

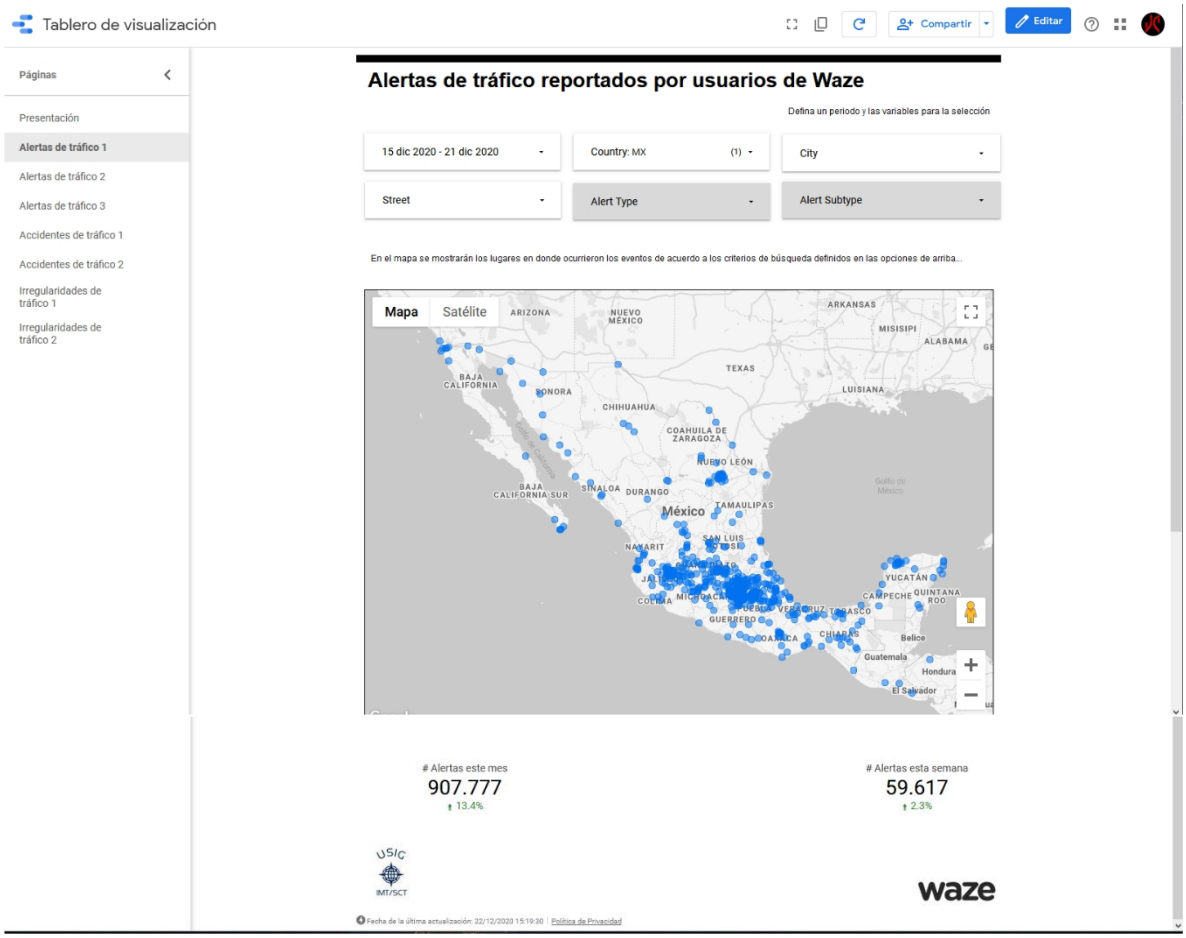
Figura 4.18 Densidad de tráfico del 3 de abril 2020 para Monterrey, N.L.

Segunda prueba piloto

Como parte de los trabajos realizados para este trabajo también se generó un método de consulta de los datos colaborativos de la plataforma Waze, en este sentido se utilizó la aplicación Data Studio de Google y la API que proporciona Waze, el desarrollo se basó en el prototipo desarrollado por Michael Wallman, integrante del equipo de Google.

Fuente: Elaboración propia a partir de la plataforma Google Data Studio

Figura 4.19 Portada de inicio de la aplicación.



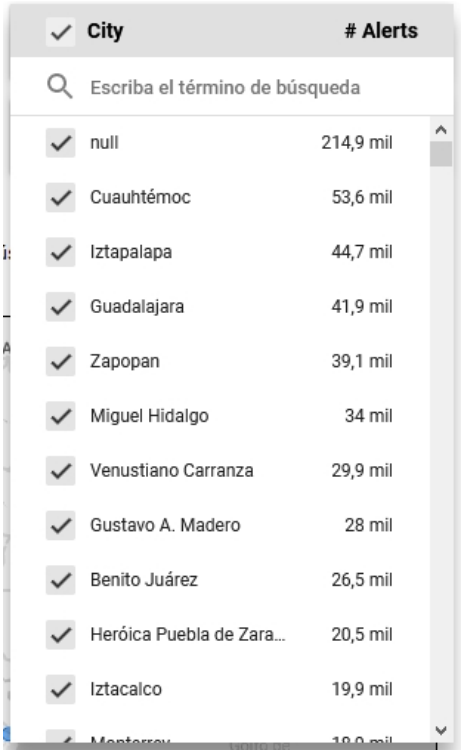
Fuente: Elaboración propia a partir de la plataforma Google Data Studio

Figura 4.20 Alertas de tráfico 1.

En la aplicación es posible consultar diversos tipos de consultas, tal como se observa en la Figura 4.19, por ejemplo, al seleccionar la primera opción *Alertas de tráfico 1*, se podrá visualizar realizar una consulta, las opciones disponibles son: por fecha, por país, por ciudad, por calle, por tipo de alerta y por subtipo.

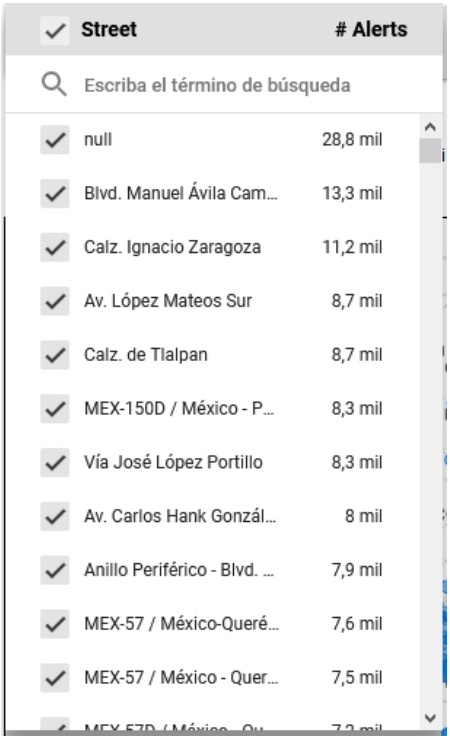
Fuente: Elaboración propia a partir de la plataforma Google Data Studio

Figura 4.21 Opciones de fecha disponibles.



Fuente: Elaboración propia a partir de la plataforma Google Data Studio

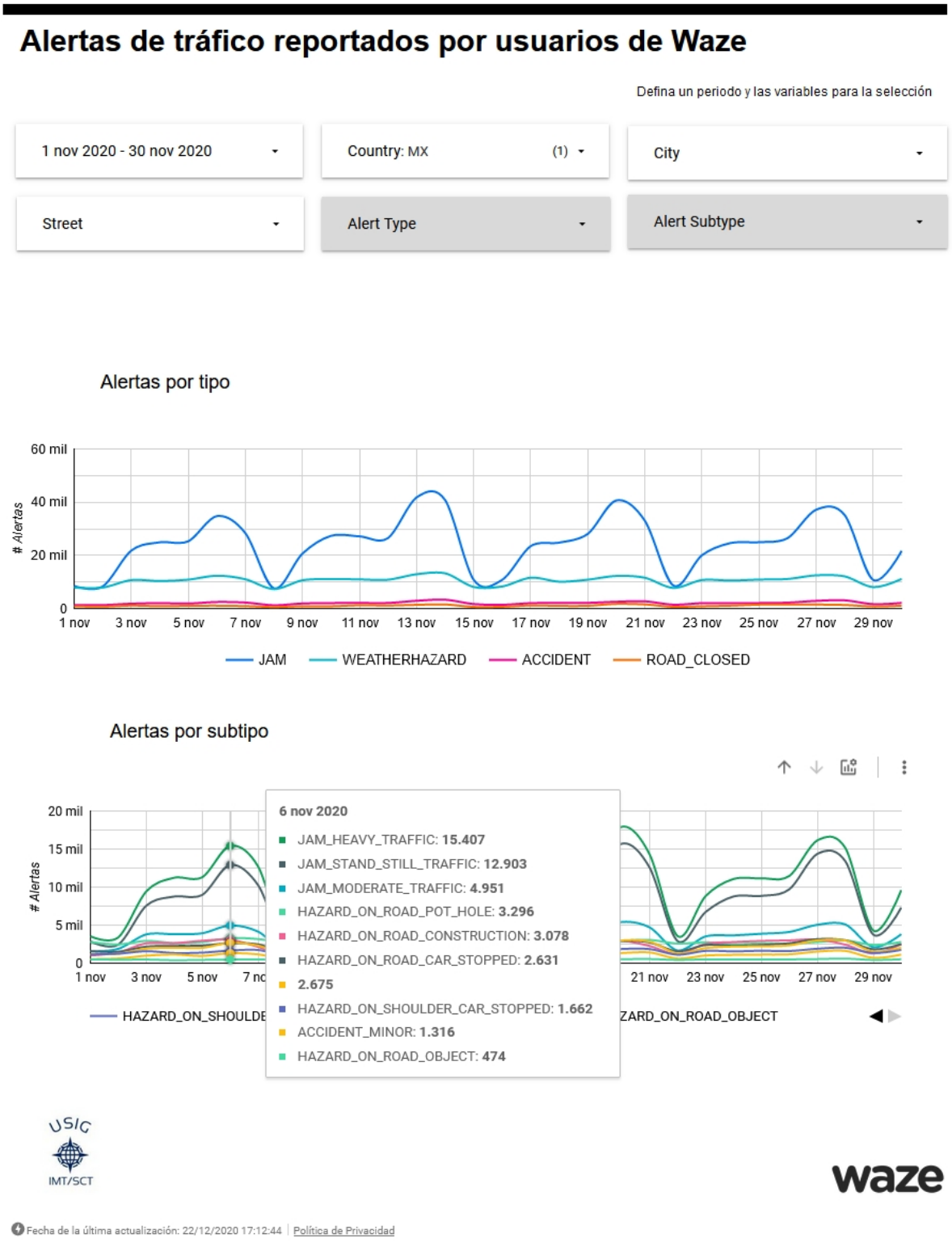
Figura 4.22 Busqueda de alertas a partir de ciudades disponibles.



Fuente: Elaboración propia a partir de la plataforma Google Data Studio

Figura 4.23 Busqueda de alertas a partir de calles disponibles.

Cuando se selecciona *Alertas de tráfico 2* las opciones disponibles son:



Fuente: Elaboración propia a partir de la plataforma Google Data Studio

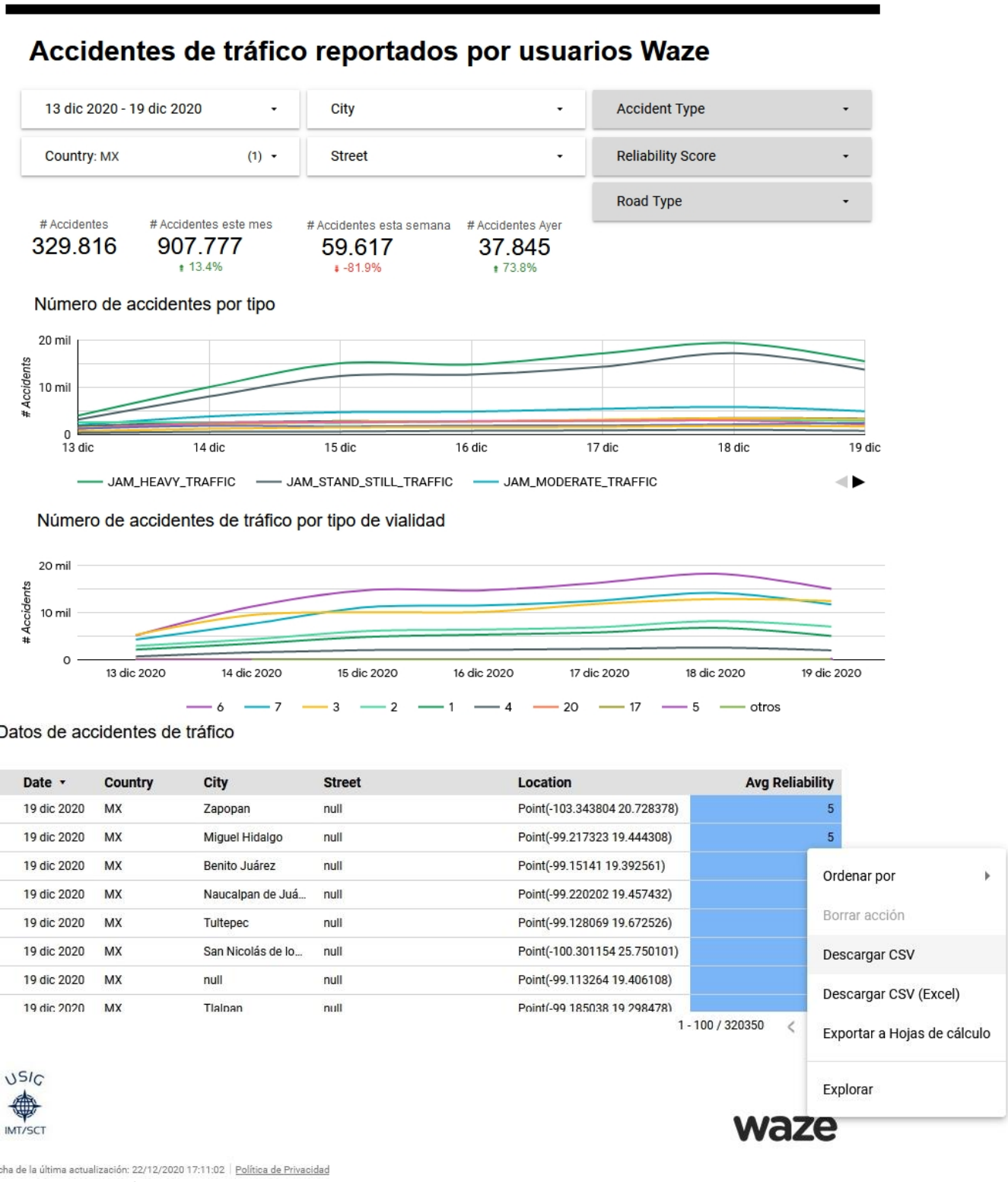
Figura 4.24 Alertas de tráfico 2.

Al seleccionar *Alertas de tráfico 3*, las opciones son:

Fuente: Elaboración propia a partir de la plataforma Google Data Studio

Figura 4.25 Alertas de tráfico 3.

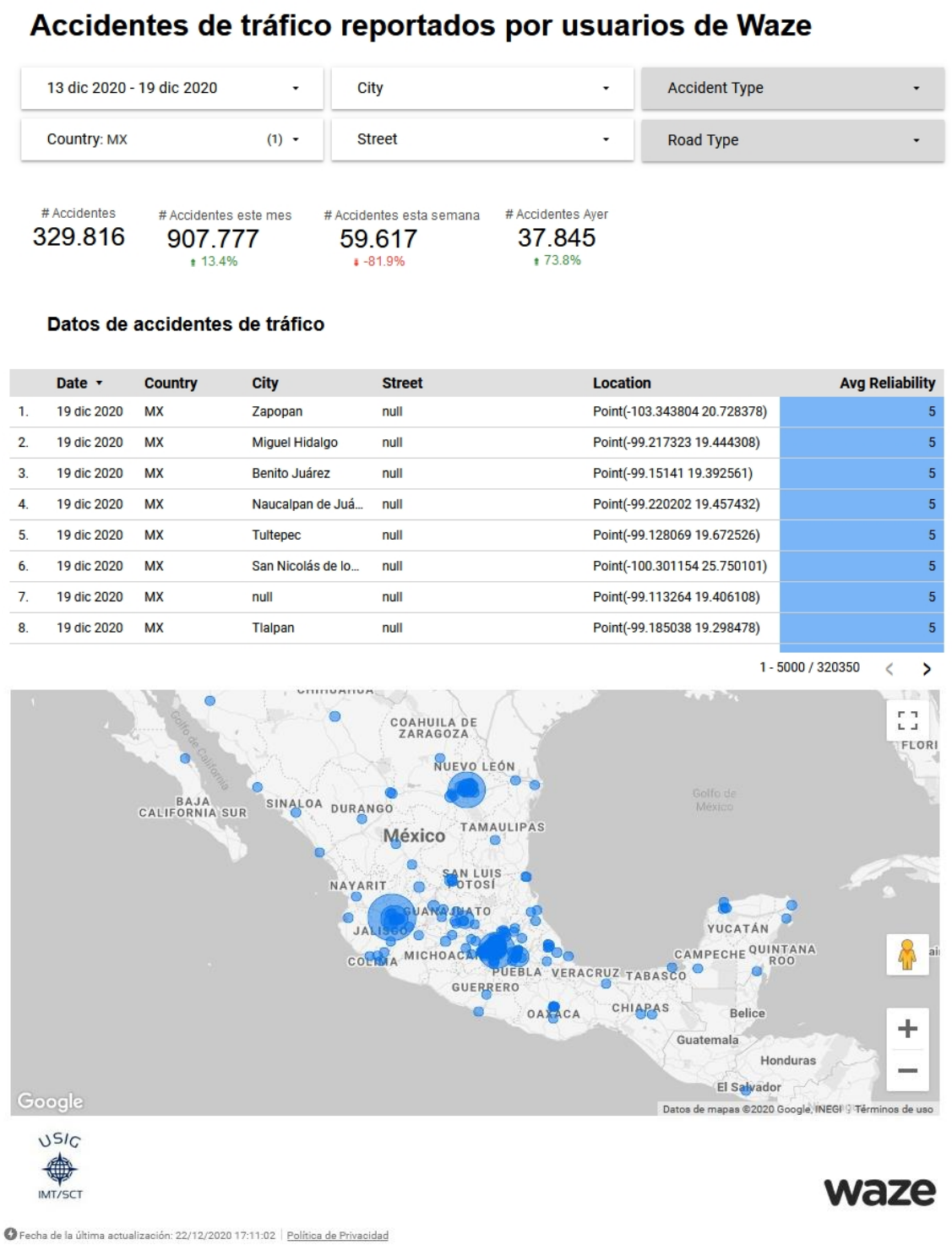
Cuando se selecciona *Accidentes de tráfico 1*, las opciones son:



Fuente: Elaboración propia a partir de la plataforma Google Data Studio

Figura 4.26 Accidentes de tráfico 1.

Opciones de *Accidentes de tráfico 2*



Fuente: Elaboración propia a partir de la plataforma Google Data Studio

Figura 4.27 Accidentes de tráfico 2.

5. Conclusiones

Este proyecto ha recopilado y utilizado las técnicas, métodos y conocimiento de publicaciones anteriores con la finalidad de afinar el uso y aprovechamiento de los datos provenientes de la plataforma Waze a los que tiene acceso la Secretaría de Comunicaciones y Transportes a través del Instituto Mexicano del Transporte.

Estos datos, cuando son procesados y analizados podrían servir para, entre cosas, predecir ocurrencias de eventos relacionados a tráfico, riesgos meteorológicos y situaciones extraordinarias que pudieran afectar tanto la Red Nacional de Caminos como las vialidades urbanas existentes en México así como a las redes de servicio público.

La utilización y aplicación ordenada podría incidir, por ejemplo, en la definición de políticas y criterios de trabajo y atención para casos como sucedió ahora en el año 2020 con la epidemia de SARS-COV2, permitiendo así identificar ocurrencias, relaciones, eventos, lugares o puntos en donde el tráfico sigue existiendo a pesar de la disminución de actividades no esenciales, también podría ayudar a identificar las rutas más rápidas y seguras para la distribución de medicinas o vacunas, el transporte de insumos en general o de personal médico y de apoyo.

Es necesario que la SCT (a través del IMT) se integre al manejo de los datos masivos geoespaciales como parte fundamental de las labores que tiene asignada. Es cierto que se requiere más conocimiento especializado, expertos calificados, equipos e infraestructura de almacenamiento, computo, análisis y despliegue.

Como se mencionó, 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. También es necesario identificar con precisión los objetivos del uso de este tipo de datos y tecnología, estos deberán ajustarse al plan de desarrollo nacional e internacionales actuales.

Los datos masivos geoespaciales presentan desafíos y oportunidades, en particular para la gestión integral del transporte. A partir de herramientas informáticas que manejen el aspecto geográfico, 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 a las pruebas piloto utilizando *CARTO* y *Google Data Studio* con la 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. De este modo es posible analizarlos en diferentes intervalos de tiempo donde se podrá identificar cómo se propaga la congestión desde un punto o zona y, posteriormente implementar medidas de mitigación.

Las pruebas piloto permiten identificar, analizar y comparar tendencias de tráfico, ciudades, municipios y vialidades donde más se presentan incidentes de transito así como ocurrencias de tráfico, riesgos meteorológicos.

En la Unidad de Sistemas de Información Geoespacial se continúa con el trabajo y la investigación enfocada en 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:

- Continuar con la recopilación de la bases de datos masivos geoespaciales 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, challengues, 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 Scince and Engineering, University of Minessota. 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

COMUNICACIONES
SECRETARÍA DE COMUNICACIONES Y TRANSPORTES

INSTITUTO MEXICANO DEL TRANSPORTE

Km 12+000 Carretera Estatal 431 “El Colorado Galindo”
Parque Tecnológico San Fandila, Mpio. Pedro Escobedo,
Querétaro, México. C.P. 76703
Tel: +52 (442) 216 97 77 ext. 2610
Fax: +52 (442) 216 9671

publicaciones@imt.mx

http://www.imt.mx/