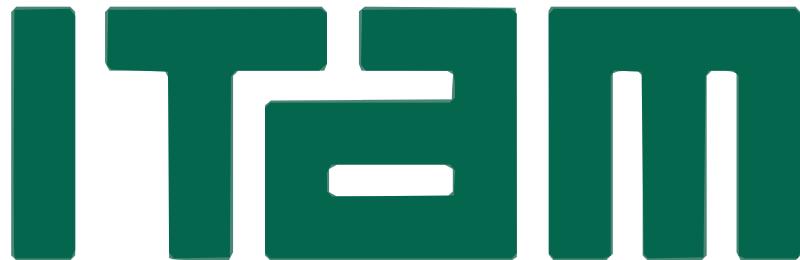


INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO



**USO DE TWITTER COMO RED DE SENORES PARA
DESCRIBIR EL MOVIMIENTO DE PERSONAS EN LA
CIUDAD DE MÉXICO**

TESINA

QUE PARA OBTENER EL TÍTULO DE
INGENIERO EN COMPUTACIÓN

P R E S E N T A

LUIS EDUARDO PÉREZ ESTRADA

ASESOR: DR. CARLOS FERNANDO ESPONDA DARLINGTON

México, D.F.

2014

Autorización para difusión

Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada “USO DE TWITTER COMO RED DE SENSORES PARA DESCRIBIR EL MOVIMIENTO DE PERSONAS EN LA CIUDAD DE MÉXICO”, otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Baillères Jr., autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación.

Luis Eduardo Pérez Estrada

Fecha

Firma

USO DE TWITTER COMO RED DE SENSORES PARA DESCRIBIR EL MOVIMIENTO DE PERSONAS EN LA CIUDAD DE MÉXICO

Luis Eduardo Pérez Estrada

Resumen

La Ciudad de México, notable por su mala planeación urbana y alta densidad de población, sufre de tránsito excesivo y transporte público insuficiente. Para poder dar solución a estos problemas, es necesario tener información acerca de cómo se mueven las personas dentro del área metropolitana. El propósito de este trabajo es describir estos movimientos utilizando únicamente información pública obtenida del servicio de microblogging Twitter. Se explica cómo se obtienen los datos, cómo se analizan, y cómo se visualizan para poder apoyar la toma de decisiones en este tema. Con estos datos se logra identificar zonas de la ciudad, y cómo se divide geográficamente con respecto a las actividades de sus habitantes. Métodos como éste tendrán cada vez mayor versatilidad, precisión y confiabilidad conforme crezca la adopción de tecnologías celulares y de servicios de redes sociales.

Palabras clave: Twitter, microblogging, geolocalización, visualización.

TABLA DE CONTENIDO

1. INTRODUCCIÓN	1
1.1 Motivación	1
1.2 Objetivo	1
1.3 Alcance	2
1.4 Justificación	3
2. MARCO TEÓRICO	4
2.1 Microblogging	4
2.2 Antecedentes en Análisis de Twitter	5
2.3 Grafos	6
2.4 Graph Clustering	7
2.5 Markov Clustering Algorithm	9
3. ANÁLISIS DE LOS DATOS	11
3.1 Recopilación de <i>tweets</i>	11
3.2 Caracterización de los Datos	11
3.2.1 Distribución en el Espacio	12
3.2.2 Distribución en el Tiempo	14
3.2.3 Distribución entre los Usuarios	14
4. DISEÑO E IMPLEMENTACIÓN	18
4.1 Zonas y Rutas	18
4.2 Algoritmo para Construcción de Grafos de Zonas y Rutas	20
4.3 Implementación del Algoritmo 4-1	21
4.4 Métodos de Visualización	22
4.4.1 Mapa de Calor	22
4.4.2 Mapa de Rutas	24
4.4.3 Mapa de Clusters	26
4.5 Video	31
5. CONCLUSIONES	34
BIBLIOGRAFÍA	36

Í N D I C E D E F I G U R A S

FIGURA

3-1 Campos disponibles en la base de datos	12
3-2 Diagrama de burbujas sobre el mapa	13
3-3 Gráfica de cantidad de <i>tweets</i> por día de la semana	14
3-4 Gráfica de cantidad de <i>tweets</i> por hora del día	15
3-5 Gráfica de frecuencia del número de <i>tweets</i> por usuario	15
3-6 Distribución de número de usuarios sobre la longitud de radio por día de la semana y hora del día	17
4-1 – Mapa de calor del área metropolitana comprendiendo el periodo de 18 de junio a 4 de diciembre de 2013	23
4-2 – Mapa de rutas del área metropolitana ilustrando las 500 rutas más comunes en el periodo de 18 de junio a 4 de diciembre de 2013	25
4-3 – Visualización de los diez clusters más importantes con valor de inflación 2.0 comprendiendo el periodo de 18 de junio a 4 de diciembre de 2013.	29
4-4 – Mapas de los diez clusters más importantes, con valor de inflación 3.5 y 4.0 en el periodo de 18 de junio a 4 de diciembre de 2013	30
4-5 – Secuencia de 16 cuadros consecutivos de mapas de calor, cada uno incluyendo <i>tweets</i> de un periodo de 4 horas.	33

1. INTRODUCCIÓN

1.1 Motivación

La Ciudad de México, notable por su mala planeación urbana, sufre problemas de tránsito excesivo en sus vialidades y de transporte público insuficiente. Conocer cuánto y cómo se mueven sus habitantes es vital para generar soluciones correctas cuando se proponen nuevas rutas de transporte colectivo o nuevas obras para automóviles. Una vez conseguida la caracterización del movimiento de personas dentro de la zona metropolitana, se propicia la búsqueda de una manera de detectar eventos anómalos como accidentes, manifestaciones e inundaciones para posibilitar una respuesta oportuna y disminuir el impacto en los habitantes.

Estas características del área, sumadas a la alta densidad de población, a la rápida adopción de los teléfonos celulares con conexión a Internet y al crecimiento de servicios como Twitter, hacen de la Ciudad un objetivo ideal para utilizar la información que generan sus habitantes como sensores de movimiento, de temas de conversación comunes y de sus sentimientos para observar rutas comunes, zonas problemáticas o eventos disruptivos.

1.2 Objetivo

El propósito de este trabajo es describir los movimientos importantes de personas dentro de la zona metropolitana de la Ciudad de México utilizando los datos de geolocalización de las tweets de cuentas públicas locales de Twitter, y generar visualizaciones de estos datos que permitan identificar problemas para apoyar el diseño de soluciones de planeación vial.

Adicionalmente, se busca demostrar la viabilidad de utilizar las cuentas de Twitter como una amplia red de sensores para obtener información de un ambiente, en este caso el transporte en la Ciudad de México, sin necesitar el uso de mediciones directas o incurrir en el costo de desplegar sensores convencionales.

1.3 Alcance

Se utilizaron *tweets* de usuarios localizados en la zona metropolitana de la Ciudad de México obtenidas durante el verano y otoño de 2013 con la API pública de *streaming* proporcionada por Twitter. Específicamente, se utilizaron los datos de localización que contienen las *tweets*: las coordenadas de latitud y longitud compartidas voluntariamente por los usuarios, extraídas de los sistemas de GPS de los dispositivos móviles en las que fueron publicadas. Incluso si no se tiene una ubicación exacta, muchos usuarios eligen compartir el nombre del establecimiento o punto de interés en el que se encuentran, lo que también arroja información sobre su ubicación, aunque con una precisión menor.

Se presenta el análisis descriptivo de los datos obtenidos en este periodo con el propósito de identificar características problemáticas del movimiento de las personas en la ciudad que puedan ser reconocidas algorítmicamente después. Se describen detalladamente los algoritmos utilizados para extraer y representar la estructura de los datos contenidos en las *tweets*, así como las herramientas utilizadas en sus implementaciones. Se proponen tres algoritmos de visualización de dicha estructura para ayudar el reconocimiento de patrones y posibles problemas en el movimiento de personas, se enumeran las herramientas utilizadas para generarlas y se presentan los resultados en imágenes y video.

1.4 Justificación

El trabajo busca proveer una imagen acertada de cómo, cuándo y cuánto se mueven las personas dentro de la zona metropolitana a lo largo de su vida diaria, e identificar patrones y flujos importantes que **puedan ser útiles para la toma de decisiones en materia de planeación urbana,** de transporte colectivo y de vialidad. Además, se busca demostrar el uso de Twitter como una red de sensores pública y de uso gratuito, útil para una gran cantidad de aplicaciones. Considerando la acelerada penetración de la tecnología celular en la población mexicana y la popularidad de los sitios de redes sociales, este tipo de aplicaciones será cada vez más precisa y versátil.

2. MARCO TEÓRICO

2.1 Microblogging

El microblogging es un medio de difusión que consiste en el envío de mensajes breves, imágenes individuales o hipervínculos. Los servicios de microblogging surgen como evolución de la idea del blog tradicional: un sitio en la web compuesto de publicaciones discretas cubriendo una variedad de temas y generalmente desplegados en orden cronológico inverso, favoreciendo el contenido reciente sobre el anterior. Sin embargo, se distingue por la extensión de las publicaciones – la mayoría de los servicios de microblogging establecen un límite de caracteres – así como por su mayor frecuencia de publicación. Adicionalmente, los servicios de microblogging deben su éxito en parte a los múltiples métodos de publicación que ofrecen además de un sitio web tradicional, como mensajes SMS, correo electrónico y aplicaciones móviles. Esto permitió que la popularidad de los microblogs creciera junto con el uso de los teléfonos celulares, y que se convirtieran en una parte central de la experiencia de la web en dispositivos móviles. Además del contenido mismo de las *tweets*, muchos servicios permiten incluir la ubicación en la que el usuario se encontraba en el momento de la publicación, datos con son cada vez más precisos conforme crece el uso de *smartphones* con GPS u otras técnicas de geoposicionamiento integradas.

A partir del año 2005, surgen muchos servicios de microblogging como Tumblr, FriendFeed, Jaiku, e identi.ca. Además de éstos, sitios populares de redes sociales incorporan frecuentemente esta funcionalidad en su producto en la forma de “actualizaciones de estado”. Este es el caso para servicios como Facebook, LinkedIn y, con

un menor grado de éxito, Google Buzz. Sin embargo, entre toda la diversidad de servicios y productos, Twitter claramente emerge como el más popular, tanto en número de usuarios registrados, como número de usuarios activos, número de publicaciones creadas y visitas a su sitio.

El amplio uso de Twitter, así como su presencia en todas las plataformas móviles, el carácter público de sus datos y su impacto rápido y profundo en la cultura y en la sociedad propicia que mucha de la investigación sobre el fenómeno del microblogging se centre en él o cuando menos le incluya. Según cifras de Alexa.com, hoy Twitter es el 9º sitio más popular de la web a nivel mundial y ocupa el 8º lugar en México.

2.2 Antecedentes en Análisis de Twitter

Por su volumen y naturaleza, la información que se hace pública por medio de las redes sociales como Twitter tiene un gran potencial para ser explotada para una amplia variedad de aplicaciones. Utilizando la información de geolocalización y aprovechando el hecho de que las *tweets* son publicadas en tiempo real, se han utilizado técnicas de Data Mining para detectar cuando ocurren terremotos y sus desplazamientos [Sakaki, et. al., 2010], así como para visualizar en un mapa cómo la gente es afectada por la rinitis alérgica (“fiebre del heno”) en temporadas de polen [Takahashi, et. al., 2011]. Además de detectar y describir eventos fuera de lo común, se puede extraer información constantemente de sucesos continuos como el estado del tiempo [Dermibas, et. al., 2010], el enojo y alegría de las personas y como se transmiten estas emociones [Fan, et. al., 2013], como viajan las noticias por Twitter y como este movimiento es afectado por el tono y contenido del mensaje

[Naveed, et. al., 2011]. Incluso se ha conseguido predecir el movimiento de bolsas de valores para conseguir una mayor utilidad en intercambios [Bollena, et. al., 2011] [Chen y Lazer, 2011].

Analizando aspectos de la estructura de la red social se han logrado detectar tempranamente brotes de enfermedades altamente contagiosas [Christakis y Fowler, 2010]. Aún sin contar con la información precisa de localización, se ha demostrado que se puede inferir la localización del usuario utilizando únicamente el contenido de sus *tweets* [Cheng, et. al., 2010] o las relaciones *follower-follower*, es decir, la forma misma del grafo de relaciones entre los usuarios [Davis, et. al., 2011].

Examinando la frecuencia con la que el usuario publica *tweets*, es posible detectar si la cuenta está manejada únicamente por un único dueño, por una agencia de relaciones públicas de una empresa, o por un robot [Tavares y Faisal, 2013]. Esto permite discriminar las *tweets* de usuarios “auténticos”, elevando la probabilidad de que el contenido de la publicación esté reflejando eventos que ocurren alrededor del usuario.

2.3 Grafos

Para construir una representación del movimiento de las personas en la ciudad en un periodo determinado de tiempo, se utilizan grafos que conectan sectores de la ciudad con otros como se describe en el capítulo 4. Formalmente, un grafo G se define como un par ordenado $G = (V, E)$, donde V es un conjunto de nodos o vértices y E es un conjunto de aristas o arcos, cada arista un par de elementos de V . Si las aristas no tienen orientación o

dirección, es decir, los elementos de E son conjuntos no ordenados, se dice que es un grafo no dirigido. En este caso la arista entre dos nodos cualesquiera $A \rightarrow B$ es idéntica a la arista $B \rightarrow A$. De otra manera, si los elementos de E son pares ordenados de vértices, entonces las aristas tienen una dirección de un nodo hacia otro, y se dice que el grafo es dirigido. En este caso la arista $A \rightarrow B$ es un elemento distinto al $B \rightarrow A$, y la existencia de uno no implica la existencia del otro.

Adicionalmente, en la representación elegida para el movimiento de personas se utiliza el concepto de grafo ponderado. Éste es un grafo donde cada una de las aristas de E tiene asociada una etiqueta llamada peso o costo. Para los grafos que se utilizan en este trabajo, los pesos asignados a cada arista son números enteros no negativos, y representan el número de rutas observadas que conectan dos nodos, es decir, una aproximación de qué tan común es que un habitante de la ciudad se mueva de un lugar en particular a otro.

2.4 Graph Clustering

Uno de los patrones que se busca encontrar con este trabajo es, por encima de divisiones políticas, cómo se divide realmente la geografía de la ciudad con respecto a las actividades diarias de sus habitantes. Esto requiere poder identificar, de puntos discretos en el tiempo y en el espacio provistos por *tweets* de los usuarios, grupos de estos puntos similares entre sí que se distingan del resto y que correspondan a sectores de la ciudad que se consideren “parecidas” por la naturaleza del movimiento de las personas que viven, trabajan o transitan ahí.

Una vez representado el movimiento de los habitantes en un grafo, esta tarea se vuelve idéntica al *graph clustering*, o agrupamiento de grafos. Consiste en agrupar (o discriminar) elementos de un conjunto en elementos de menor tamaño llamados clusters, de suerte que los elementos dentro de un cluster sean más similares entre ellos y menos similares a elementos que se encuentran en otro cluster.

Para resolver esta tarea, existe una gran cantidad de algoritmos distintos, cada uno con distintas características como su modelo de cluster (lo que entiende el algoritmo por “grupo”), si permite agrupamiento difuso (la pertenencia a un cluster se da como una probabilidad), si es de particionamiento estricto (todos los elementos pertenecen exactamente a un cluster), si permite que los clusters se traslapen (que un elemento pueda pertenecer a más de un cluster), etc. En general, no existe un “mejor” algoritmo de agrupamiento, sino que la elección del algoritmo se da por la naturaleza de los datos que se desea analizar, o por la coincidencia del modelo de cluster del algoritmo con los resultados que se desea obtener.

Por ejemplo, entre los modelos de cluster que un algoritmo podría utilizar se encuentran los de conectividad, donde la distancia entre los elementos definen el grupo; de centroide, donde cada cluster se caracteriza con un punto; de distribución, donde los clusters se definen por similitud a una distribución estadística; de densidad, donde los clusters son regiones del espacio con mayor concentración de elementos; o simplemente de grupo, donde el algoritmo no ofrece una explicación precisa de su modelo.

Para realizar el agrupamiento de sectores de la ciudad se utiliza el Markov Clustering Algorithm (MCL) [van Dongen, 2000], seleccionado por la manera en que aborda el clustering. Este algoritmo extrae la información de agrupamiento de los datos simulando recorridos aleatorios sobre el grafo, tomando el peso de las aristas para construir las probabilidades en una matriz estocástica.

2.5 Markov Clustering Algorithm

El algoritmo elegido, MCL, extrae los clusters que se encuentran naturalmente en el grafo. Es decir, se encuentran y agrupan nodos que tengan muchas conexiones entre ellos y pocas conexiones al exterior del grupo, además de tener mayores pesos en las conexiones internas al grupo que a las externas.

Si se ve el grafo como un sistema que transita entre los diferentes posibles estados – los vértices – y los pesos de las aristas se interpretan como las probabilidades de moverse a un estado determinado dado el estado actual, se puede definir una matriz de transiciones. Ésta es una matriz estocástica, donde las sumas de los elementos de cada columna es exactamente 1, y donde la i -ésima columna contiene las probabilidades de transición del i -ésimo estado a todos los demás.

El MCL opera sobre esta matriz, alternando la aplicación de dos distintas funciones: la primera, llamada expansión, que corresponde a calcular el producto de la matriz por sí misma utilizando el producto matricial habitual; y la segunda, llamada inflación, que de nuevo corresponde a tomar una potencia de la matriz pero utilizando el producto de

Hadamard, es decir, elevando cada elemento de la matriz a un exponente. La potencia a la que se elevan los elementos en el paso de inflación puede variarse, y de esta manera puede controlarse la granularidad del clustering. Así, pueden obtenerse clusters más grandes y menos fuertemente conectados, o clusters más pequeños pero más fuertemente conectados.

Estas operaciones sobre la matriz de transiciones calculan, determinísticamente, las características de recorridos aleatorios sobre el grafo. El paso de expansión corresponde a calcular las probabilidades de llegar de un nodo a otro con recorridos largos, por lo que si dos nodos pertenecen naturalmente a un cluster, esta probabilidad será relativamente alta por haber muchas maneras de llegar del primero al segundo. Si dos nodos no pertenecen a un cluster, entonces la probabilidad de llegar del primero al segundo será baja. El paso de inflación aplicado posteriormente hará más evidentes las diferencias entre recorridos de un nodo a otro: en caso de que pertenezcan naturalmente a un cluster la probabilidad será mucho mayor que en el caso contrario. El parámetro de inflación determina qué tanto se aumentarán las probabilidades de recorridos dentro de un cluster sobre las de recorridos entre clusters distintos.

El algoritmo continúa con iteraciones de estos dos pasos hasta que el conjunto de todos los vértices queda separado en distintos grupos y estos grupos no se modificaron de una iteración a la siguiente. Al terminar, se cuenta con una matriz de transiciones distinta a la original, donde únicamente sobreviven las aristas entre nodos que pertenecen a un mismo cluster.

3. ANÁLISIS DE LOS DATOS

3.1 Recopilación de *tweets*

En este trabajo se utilizan *tweets* publicadas en México, obtenidas directamente de Twitter utilizando su *Streaming API*, disponible a través de HTTP. En particular, se utiliza el *endpoint* “POST statuses/filter” para poder discriminar las *tweets* por su ubicación.

La *Streaming API* de Twitter permite obtener publicaciones en tiempo real, a diferencia de su API REST que se utiliza para obtener datos de una *tweet*, de un usuario o de otros objetos en particular. Por esta razón, no basta con enviar una solicitud al servicio y recibir una respuesta, sino que para continuar recibiendo datos es necesario tener una conexión persistente, y se debe tener una máquina encendida y con conexión a Internet durante todo el tiempo que se desee capturar.

Se utilizó una instancia de Amazon Web Services para llevar a cabo la recopilación, corriendo un programa de Python utilizando la librería Tweepy para abstraer la comunicación con el servicio de Twitter. Una vez obtenida una *tweet* en formato JSON, la información almacenada en una base de datos PostgreSQL. Los datos se dividen en dos tablas: *tweet*, que contiene los datos de cada una de las *tweets* y *twitter_user*, que contiene la información de sus autores. La base de datos se describe en la figura 3-1.

3.2 Caracterización de los Datos

Para saber si en los datos se pueden detectar problemas de tránsito de la ciudad, es necesario ver como se distribuyen las *tweets* en el tiempo, en el espacio y entre los usuarios.

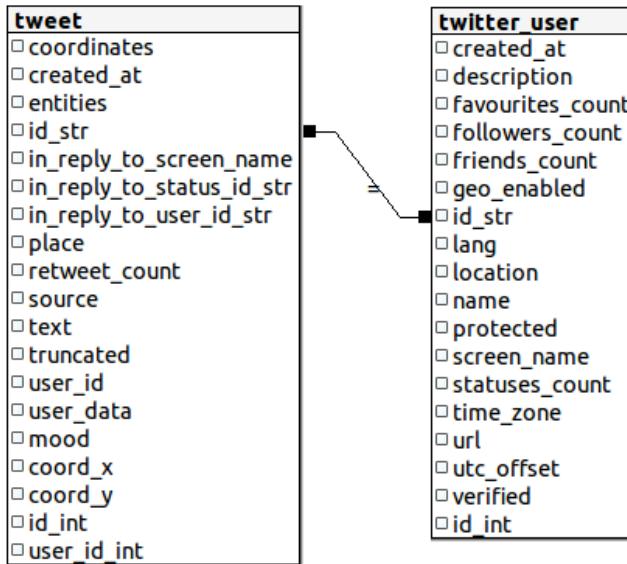


Figura 3-1 – Campos disponibles en la base de datos

3.2.1 Distribución en el Espacio

Aún haciendo la petición a la *Streaming API* de Twitter solamente por *tweets* publicadas en un área geográfica, no puede garantizarse que un *tweet* descargado tenga datos obtenidos del *Global Positioning System* (GPS) o de otras técnicas que proveen la ubicación del dispositivo en el que se publicó con la precisión que se requiere. Esto se debe a los usuarios pueden reportar en su perfil un campo de ubicación, como “Ciudad de México” o “México DF”, y que este campo se utiliza por la API para realizar el filtro. Adicionalmente, incluso si una *tweet* no contiene información de GPS, puede ser incluida por el filtro de la API por tener etiquetado el lugar de publicación, por ejemplo, el nombre de restaurantes, cafeterías, tiendas y otros lugares públicos. En este caso puede obtenerse un rectángulo dentro del cual se sabe que se encuentra el dispositivo, pero por la alta variabilidad en el tamaño de estas cotas, estos datos no son de utilidad para los propósitos de este trabajo.

Del total de 6553288 *tweets* obtenidas utilizando el servicio, 5497244 (83%) de ellas sí contienen información de GPS. De este subconjunto, 5133894 (93%) se ubican a menos de un grado en longitud o latitud del centro de la ciudad de México, por lo que se cuenta con una gran cantidad de información de ubicación utilizable. Para ubicar las *tweets* visualmente de manera muy general, se utilizó un diagrama de burbujas, donde se dibuja un círculo centrado en una ubicación del mapa de la ciudad con radio proporcional al número de *tweets* observadas en ese lugar. Esta visualización simple se presenta en la figura 3-2.

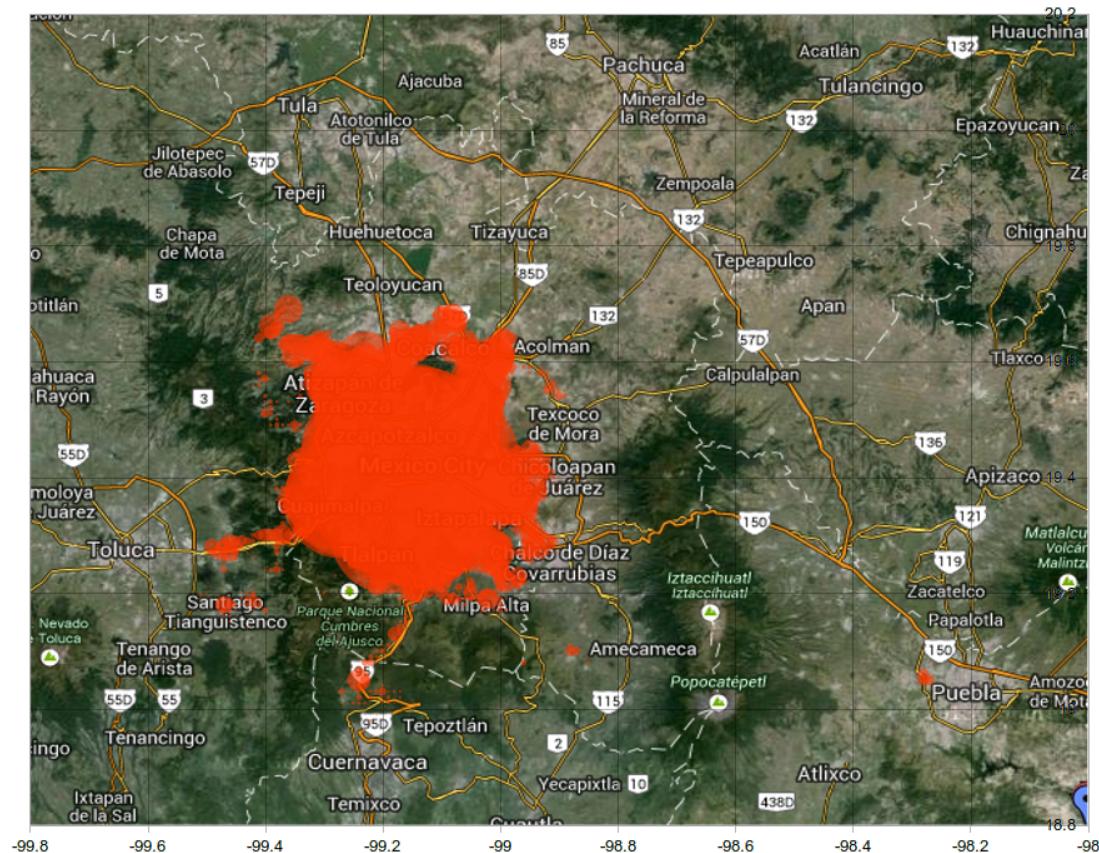


Figura 3-2 – Diagrama de burbujas sobre el mapa

3.2.2 Distribución en el Tiempo

Para obtener la distribución en el tiempo de las *tweets*, se toma el conjunto completo de datos y se agrupan por rangos de tiempo, tomando en cuenta la semana, el día o la hora según la resolución deseada. Por semana, en promedio, se publicaron 312,000 *tweets*. Por día de la semana, se observa un pico en los días jueves y miércoles, como se muestra en la figura 3-3.

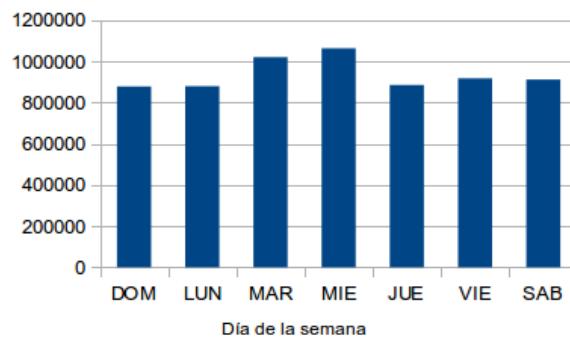


Figura 3-3 – Gráfica de cantidad de *tweets* por día de la semana

Agrupando los datos por la hora del día, se observa un pico de actividad alrededor de las 9:00 de la noche, y pueden verse claramente los ciclos de día y noche, con una muy baja actividad en la madrugada y frecuencias más estables a partir de las 8:00 a.m. y hasta las 6:00 de la tarde. En la figura 3-4 se muestra la gráfica de *tweets* por hora del día.

3.2.3 Distribución entre los Usuarios

En promedio, cada usuario incluido en el conjunto de datos publicó 25 *tweets* en el periodo de tiempo cubierto. En la figura 3-5 se muestra la gráfica con cuántos usuarios coinciden en tener el mismo número de *tweets* publicadas. Puede verse que hay una gran cantidad de

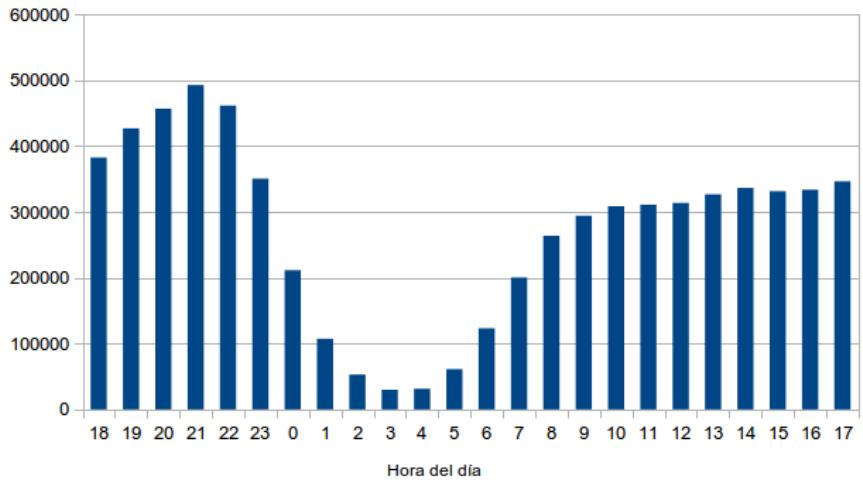


Figura 3-4 – Gráfica de cantidad de *tweets* por hora del día

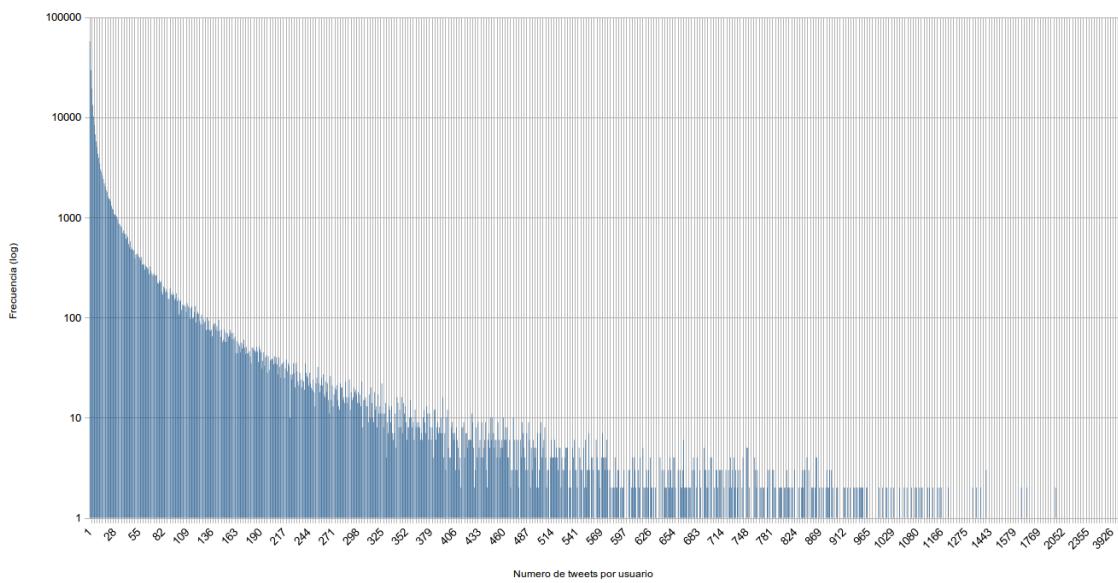


Figura 3-5 – Gráfica de frecuencia del número de *tweets* por usuario

cuentas con pocas *tweets*, y como decrece rápidamente la cantidad de usuarios conforme crece el número de *tweets*. Sin embargo, existen usuarios con decenas o cientos de *tweets* publicadas, por lo que podría ser factible hacer análisis de las publicaciones de un solo usuario.

Si se desea poder confirmar la existencia de un problema con la planeación de la ciudad o de sus sistemas de transporte, es necesario observar el movimiento de las personas individualmente, y tomar en cuenta las distancias que recorren de manera rutinaria. Para esto, se calculó una medida de desplazamiento por persona basada en el “radio” o distancia por la que se aleja del centro de sus actividades.

En primer lugar, se calcula el centroide de todas las *tweets* de cada usuario bajo la suposición de que se observarán más *tweets* suyas en lugares en los que el usuario frecuenta más. Este punto, que puede coincidir o no con la ubicación de alguna de sus *tweets*, se obtiene promediando todas las coordenadas de todas sus publicaciones, tendiendo a acercarse a lugares donde publica frecuentemente y le resta importancia a viajes posiblemente más largos pero poco comunes. Una vez calculado el centroide de las ubicaciones observadas de una persona, se computa el promedio de la distancia entre el centro y cada una de sus *tweets*, dando información sobre el tamaño del área dentro de la cual un usuario tiende a moverse en su vida diaria. La mayoría de los usuarios se mueven, en promedio, dentro de un radio de 10 km. En la figura 3-6 se muestra el número de usuarios distintos por longitud de radio dada.

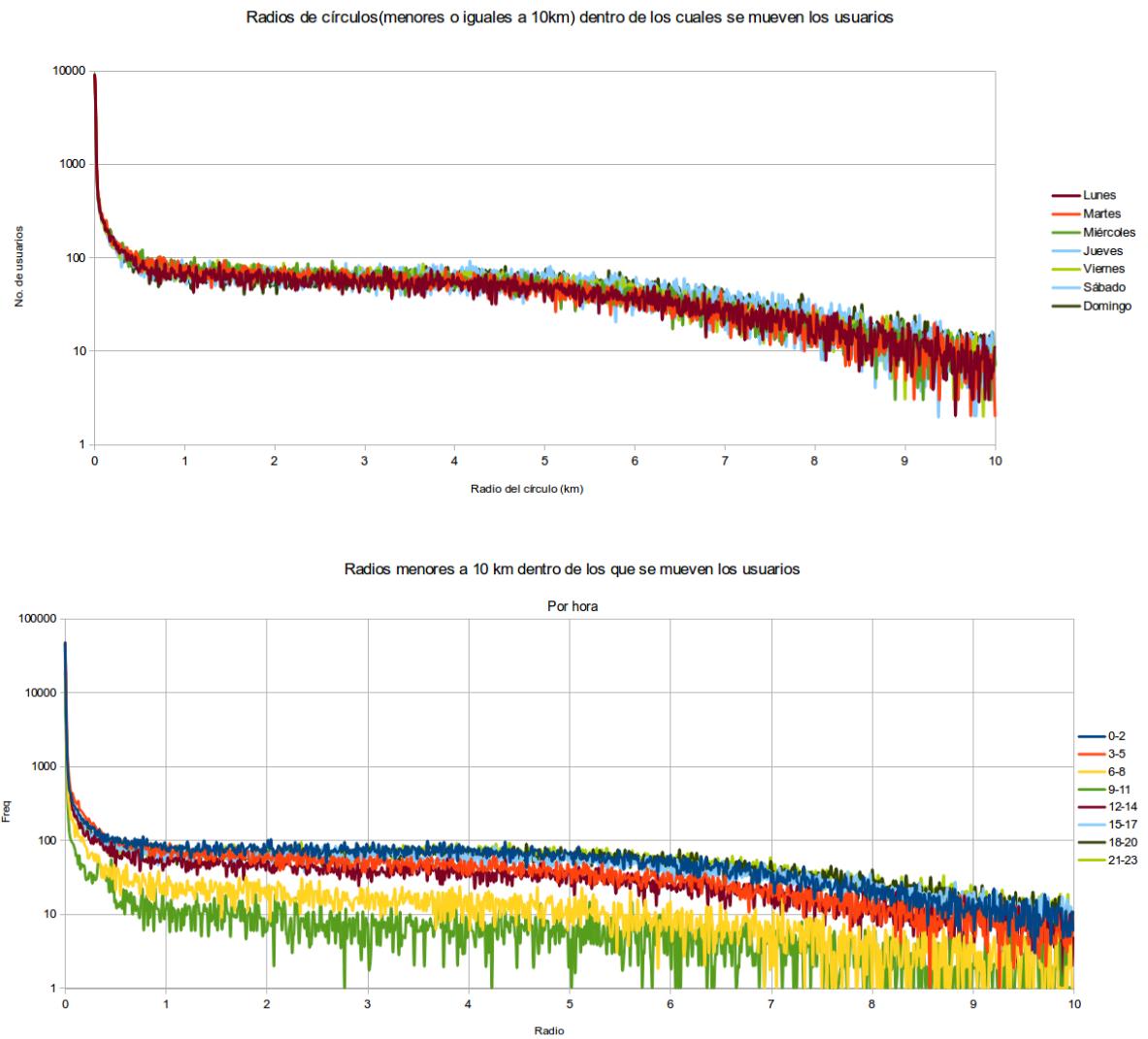


Figura 3-6 – Distribución de número de usuarios sobre la longitud de radio por día de la semana (arriba) y hora del día (abajo).

4. DISEÑO E IMPLEMENTACIÓN

4.1 Zonas y Rutas

Hablar de movimiento de los habitantes de la ciudad a lo largo del día implica poder observar su desplazamiento en el espacio y a lo largo del tiempo. Sin embargo, como no se cuenta con monitoreo continuo del movimiento de un usuario, sino con eventos discretos en puntos arbitrarios de sus trayectos, se debe inferir información acerca del trayecto final a partir de datos incompletos. Para esto se definen los conceptos de “zona” de la ciudad para localizar usuarios en el espacio, y el de “ruta” para conectar zonas de un momento en el tiempo a otro.

Para dividir el espacio del área metropolitana en regiones se utiliza una cuadrícula superpuesta sobre el mapa de proyección Mercator. El mapa se divide en 10,000 celdas rectangulares, cada una de 0.1° de longitud o latitud por lado (aproximadamente 1.1 km por lado). La cuadrícula se centra en el Zócalo de la ciudad, en las coordenadas $19^\circ 26'N$ $99^\circ 8'W$. A cada celda se le asigna un número del 0 al 9999, iniciando con 0 en el noroeste – la esquina superior izquierda en el mapa – y continuando con renglones hacia el este y columnas de norte a sur.

Se dice que un usuario tiene una ruta de una zona de la ciudad a otra si se mueve de una celda origen hacia otra celda destino, y que existe una conexión de la primera a la segunda de un instante en el tiempo a otro. Si bien no se puede saber exactamente la ruta que toma un usuario por la naturaleza de los datos disponibles, una ruta puede ser aproximada por una sucesión de *tweets* del mismo usuario en celdas diferentes, y esta aproximación puede

ser mejorada si se utilizan criterios adicionales que consideren la diferencia en el tiempo entre la publicación de éstas. La dificultad en identificar rutas con exactitud surge de la arbitrariedad del momento en que un usuario elija publicar una *tweet*. Esto impide saber con certeza cuál es el verdadero origen de su trayecto y cual es su destino final, ya que es posible que haga publicaciones a lo largo del viaje y que omita publicar una vez que ha llegado. Por eso, los puntos a lo largo de su movimiento que se observan en los datos puede o no que contengan puntos intermedios entre el origen y el destino. También es posible que se omita el origen o el destino del viaje, lo que implica tener trayectos incompletos dentro de los datos que pueden llevar a conclusiones erróneas acerca de una ruta.

Para mitigar este problema, se considera también el tiempo transcurrido entre una *tweet* y la siguiente, considerando las celdas correspondientes como conectadas siempre y cuando la diferencia en el tiempo se encuentre dentro de un rango predefinido. Al establecer un tiempo mínimo entre *tweets* para considerarlas como etapas distintas del trayecto se reduce la probabilidad de que se creen rutas cortas artificialmente cuando el usuario simplemente publique una o más *tweets* mientras se encuentra en movimiento. Al establecer un tiempo máximo entre *tweets* para considerarlas como parte del mismo viaje, se reduce la probabilidad de que se conecte erróneamente cualquier etapa de un trayecto con el inicio de otro que no esté relacionado, en caso de que el usuario pase mucho tiempo sin publicar nada. En este caso al mínimo y al máximo se asignaron 5 minutos y 12 horas, respectivamente.

4.2 Algoritmo para Construcción de Grafos de Zonas y Rutas

Para analizar y visualizar la estructura de las zonas y sus relaciones con las rutas se construye un grafo donde los vértices son las celdas de la ciudad y son conectados por aristas que corresponden a rutas. Se compone de exactamente 10,000 vértices y a lo más 99,990,000 aristas. Éste es un grafo dirigido y ponderado, ya que contiene información sobre la dirección del movimiento del usuario, y cada arista tiene asignado un peso que corresponde al número de trayectos distintos que conectan un par de celdas. Es posible construir un grafo como éste para cualquier rango de tiempo y para cualquier número de usuarios, utilizando solamente las *tweets* de un periodo específico o publicadas por un conjunto de usuarios en particular.

La construcción del grafo inicia con la obtención de los datos que se considerarán, es decir, extraer del almacenamiento las *tweets* correspondientes al periodo de tiempo que se va a analizar y a los usuarios que se desean observar. Los campos de las *tweets* que se necesitan son ambas coordenadas de geolocalización (latitud y longitud) y el identificador de usuario. Estos datos deben ordenarse por usuario y por orden cronológico.

Luego, para cada *tweet* obtenida de esta manera, recorriéndolas en orden, se debe determinar a qué celda de la cuadrícula corresponden las coordenadas de latitud y longitud. Esta *tweet* se compara con la última observada del mismo usuario, y si ambas pertenecen a celdas distintas y el tiempo transcurrido entre sus publicaciones se encuentra dentro del rango permitido, se crea una arista de la celda anterior a la celda de la *tweet* actual. El pseudocódigo del algoritmo de construcción de grafos se presenta como el algoritmo 4-1:

```

vértices := conjunto vacío

lista_usuarios := los usuarios que se deseen considerar

fecha_inicio, fecha_final := el periodo de tiempo que se desee considerar

resultado_query := seleccionar latitud, longitud, autor y fecha de la tabla de tweets

        donde autor esté en lista_usuarios y fecha_de_creación esté
        entre fecha_inicio y fecha_final

        ordenadas por usuario, fecha_de_creación

para cada tweet en resultado_query:

    celda := determinar_celda(tweet.longitud, tweet.latitud)
    si tweet.autor == ultimo_autor y celda != ultima_celda y
        5 minutos < tweet.fecha - ultima_fecha < 12 horas:
            agregar "ultima_celda → celda" a vértices
    ultima_celda := celda
    ultimo_autor := autor
    ultima_fecha := fecha
regresar vértices

```

Algoritmo 4-1 – Construcción del grafo a partir de los datos.

4.3 Implementación del Algoritmo 4-1

El algoritmo se implementó en un programa de Python, donde se realiza la consulta a la base de datos para obtener todas las *tweets* ordenadas por tiempo. La partición en periodos de tiempo más cortos se realiza dentro del programa, con el propósito de generar grafos no sólo para un día en particular o un periodo de dos horas, por ejemplo, sino para obtener grafos para cada uno de los días para los que se tienen datos, o todos los periodos de dos horas posibles. Esto permite observar y comparar cada periodo con otros de la misma

longitud. Estos grafos se construyen, se serializan y se escriben al disco, donde pueden ser utilizados posteriormente por los programas que generan las visualizaciones.

4.4 Métodos de Visualización

Para mejorar la utilidad de la construcción de estructuras de datos y permitir la interpretación de los datos por personas, a partir de estas estructuras se generan imágenes superpuestas a mapas que ilustran las características de las regiones de la ciudad y sus relaciones de distintas maneras. Estos tres algoritmos toman como entrada un grafo previamente generado para un periodo de tiempo cualquiera, y generan una imagen con una visualización del movimiento en toda el área metropolitana en ese momento

4.4.1 Mapa de Calor

Un mapa de calor es una imagen en dos dimensiones donde el valor de una función para un punto en el plano se representa con colores. Para cada valor que la función arroje para cada uno de los puntos del plano se obtiene otro punto dentro de un continuo de colores cuyos máximo y mínimo han sido mapeados respectivamente a los valores máximo y mínimo de la función.

Se generan mapas de calor para visualizar la cantidad de rutas que tiene como destino una celda particular de la ciudad, es decir, la cantidad de *tweets* que se observan llegando a una celda desde cualquier otra. Esto ilustra la cantidad de personas que entran a una zona de la ciudad, y por lo tanto, el congestionamiento que puede esperarse en ese lugar.

En la figura 4-1 puede observarse como el mayor congestionamiento observado se

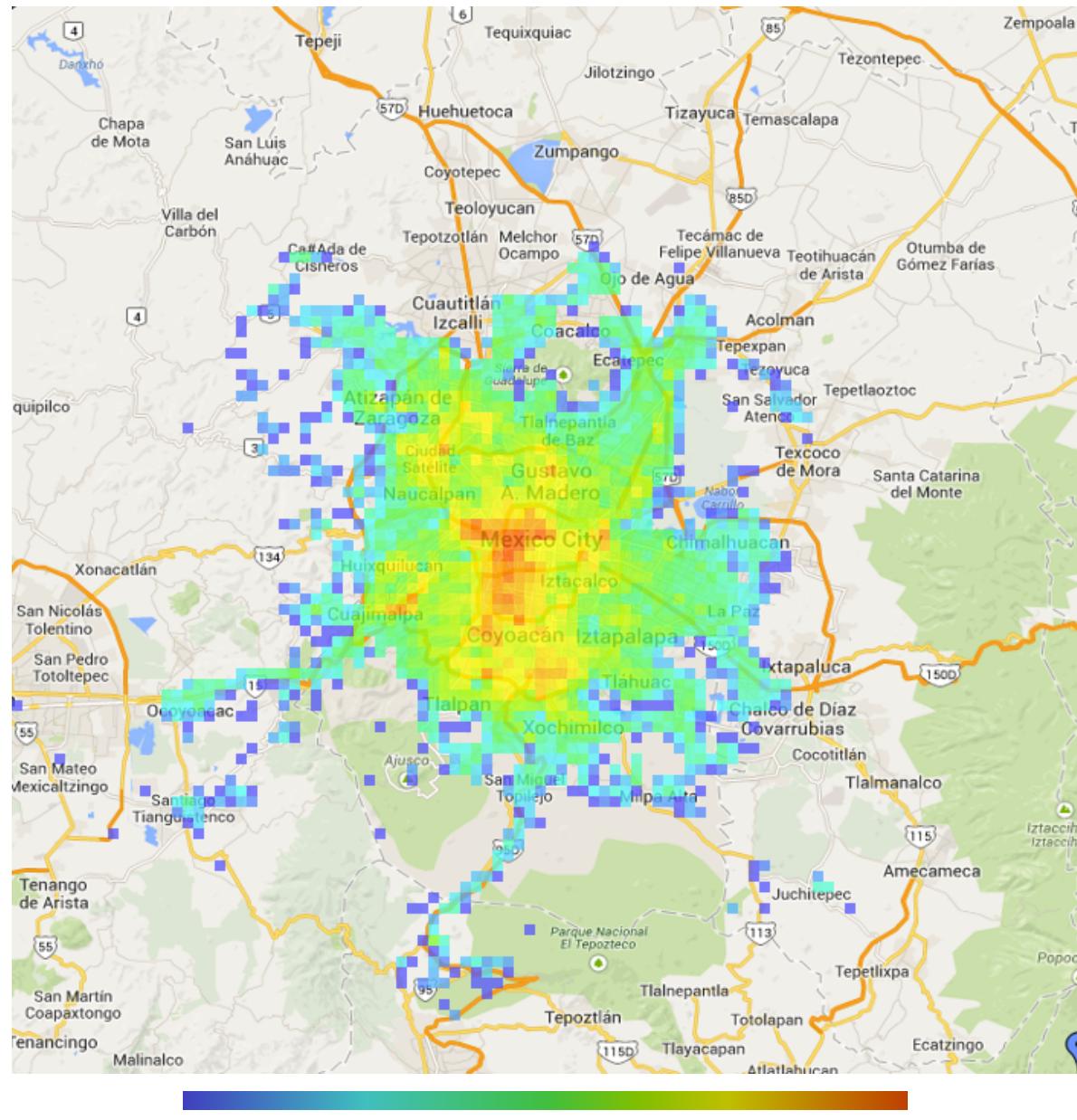


Figura 4-1 – Mapa de calor del área metropolitana comprendiendo el periodo de 18 de junio a 4 de diciembre de 2013

encuentra cerca de la zona centro de la ciudad, en las delegaciones Cuauhtémoc, Benito Juárez y Miguel Hidalgo, con algunos picos de menor intensidad en Coyoacán, Cuajimalpa, Venustiano Carranza, Gustavo A. Madero y – en el Estado de México – en Ciudad Satélite.

Es interesante también como este método de visualización dibuja claramente la trayectoria de carreteras, como la autopista México-Cuernavaca, la México-Toluca, la Marquesa-Santiago Tianguistenco y el tramo de la México-Puebla desde Iztapalapa hasta Chalco. Esto sugiere que los usuarios efectivamente están compartiendo sus ubicaciones a la red mientras se encuentran en movimiento.

4.4.2 Mapa de Rutas

Esta visualización consiste en dibujar sobre el mapa las n rutas más observadas como líneas entre las dos celdas que conecta. Para ordenar las rutas visualmente de menos comunes a más comunes, se utiliza nuevamente una escala de colores.

Para contar cuántas veces se conecta una celda A a otra celda B, se obtiene la suma de los pesos de las aristas $A \rightarrow B$ y $B \rightarrow A$. Esto es equivalente a contar cuántos pares de *tweets* se observaron con una de ellas originándose en la celda A y otra en la celda B. Por ello, esta visualización no da información de la dirección del movimiento y considera el grafo como no dirigido. Una vez obtenido un peso para cada par de celdas, se ordena la lista de pares de mayor a menor y se dibujan a lo más n de ellas sobre el mapa como rectas del color que corresponde según su lugar en la lista.

Si los desplazamientos de las personas dentro del área metropolitana fueran eficientes, es decir, que sus actividades ocurrieran dentro de un espacio reducido sin recorrer grandes distancias, se observaría un mayor peso a rutas cortas. Sin embargo, en la figura 4-2 puede verse un peso importante a rutas que atraviesan el centro de la ciudad de un extremo a otro,

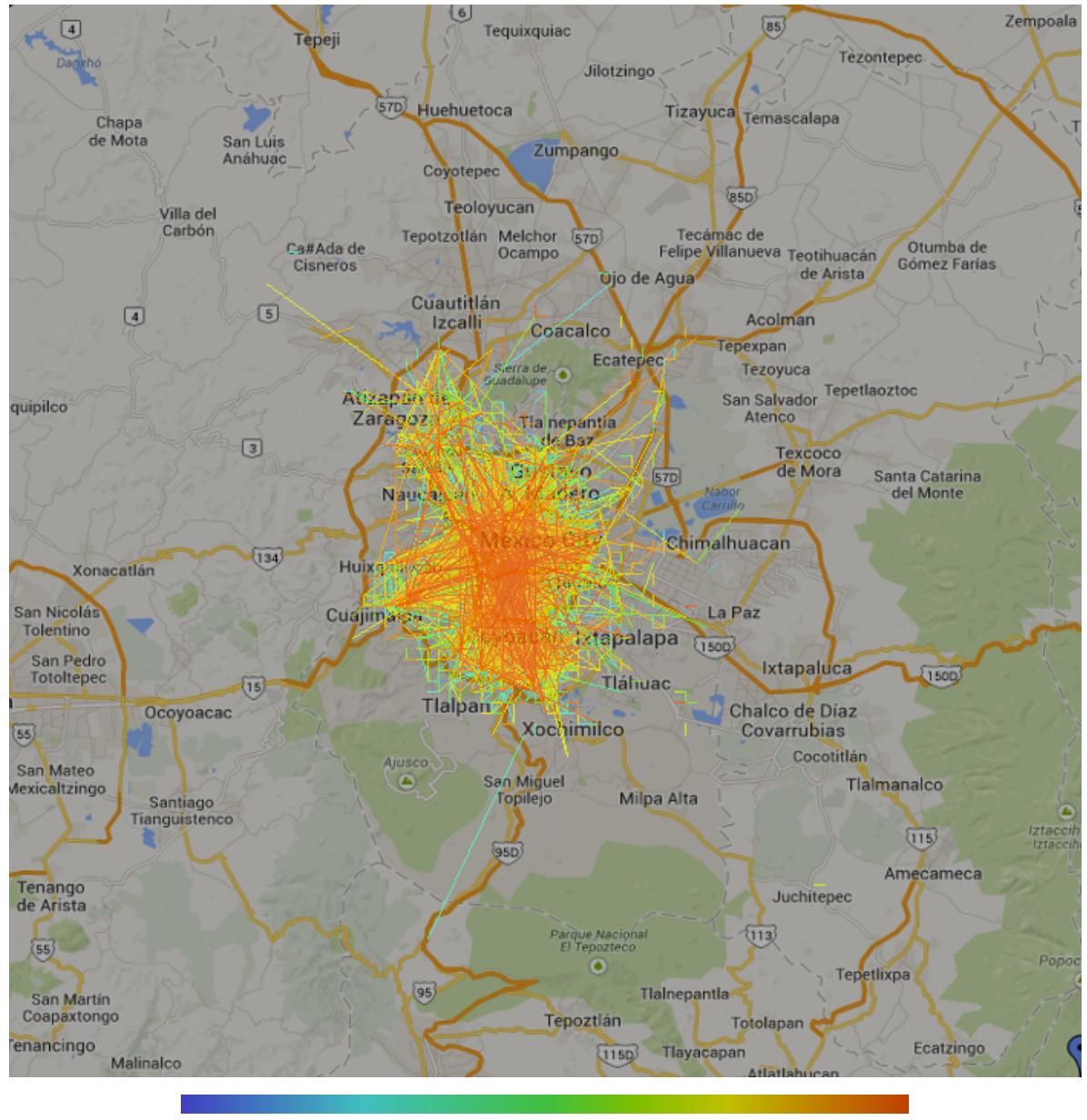


Figura 4-2 – Mapa de rutas del área metropolitana ilustrando las 500 rutas más comunes en el periodo de 18 de junio a 4 de diciembre de 2013

lo que contribuye al congestionamiento. Se observan puntos de donde salen rutas con altos pesos en las zonas de Cuajimalpa, Xochimilco, Gustavo A. Madero, Venustiano Carranza y Tlalnepantla, zonas separadas entre sí con grandes distancias y zonas densamente pobladas.

4.4.3 Mapa de Clusters

Más allá de divisiones políticas en delegaciones y colonias, es interesante saber cómo se divide la ciudad realmente según el movimiento de sus habitantes. Se desea obtener algorítmicamente un resultado que comunique una estructura de los datos similar a la noción intuitiva de “vecindario”, es decir, un sector de la ciudad de cualquier tamaño que esté fuertemente comunicada consigo misma y menos intensamente con otras. La existencia de estos grupos de celdas con un gran número de rutas compartidas entre ellas y comunicación limitada hacia el exterior implicaría que existen sectores en el área metropolitana autocontenido y autosuficientes, y, nuevamente indicarían una organización de la ciudad y por lo tanto un movimiento eficiente.

La búsqueda de estos vecindarios en el grafo de zonas y rutas equivale al problema de clustering: dividir los vértices del grafo en grupos tales que los elementos de un grupo sean más similares entre sí que a vértices de otros grupos. En este caso, se utiliza el número de rutas que comparten a dos celdas como medida de similaridad entre ellas.

Dado que el grafo tiene un número finito de vértices – “estados” en donde puede encontrarse una persona dentro del área metropolitana – y aristas con pesos distintos que salen de estos vértices hacia otros, para un estado dado en el que se encuentre un usuario se puede aproximar la probabilidad de que desplace a cualquier otro estado. Para cada vértice, si se dividen los pesos de cada arista saliente entre el total de los pesos de todas las aristas salientes, se obtendrán pesos con valores entre 0 y 1 cuya suma siempre será 1. De esta manera se obtiene una matriz de transiciones que contiene las probabilidades de que un

usuario se mueva a otro estado dado su estado actual. Esta matriz, junto con el espacio de estados, definen una cadena de Markov que caracteriza el movimiento de los habitantes de la ciudad, y que se puede utilizar para simularlo.

Este principio de simulación de recorridos aleatorios sobre el grafo utilizando las probabilidades que se encuentran en los nodos es lo que utiliza el algoritmo de clustering elegido, llamado Markov Clustering Algorithm (MCL), cuyo funcionamiento se explica en el capítulo 2. El algoritmo aprovecha el hecho de que es más probable que se recorra un camino entre dos celdas del mismo grupo que un camino entre dos grupos distintos para poder discriminarlos.

Ejecutando el MCL sobre el grafo de zonas y rutas se obtiene una lista de clusters, cada uno con la lista de celdas que pertenecen a él. El algoritmo arroja en muchos casos clusters con una sola celda, en caso de que no esté conectada fuertemente con ninguna otra, o clusters con únicamente dos celdas, en caso de que sólo se conecten con una ruta entre ambas. Por ser éstas menos interesantes en el análisis de la zona metropolitana en general, y por ser difícilmente visualizables sin añadir ruido a los grupos más importantes, se grafica un número fijo de clusters sobre el mapa, con colores distintos y ordenados por el número de elementos que los compongan.

Es posible especificar un parámetro al algoritmo que controla el paso de inflación del proceso. Esto permite controlar la granularidad con la que se obtienen los clusters, pudiendo dejar grupos grandes como uno solo o subdividiéndolos en clusters más

pequeños. En la figura 4-3 se muestra un mapa de clusters con el valor de inflación predeterminado (2.0), mostrando – en rojo – un gran cluster que cubre una gran parte de la ciudad, sugiriendo una gran cantidad de movimiento entre estas zonas por distantes que se encuentren. Otros clusters sí se distinguen, identificando los pueblos de Ocoyoacan y Santiago Tianguistenco. Se distinguen también las zonas de Cuajimalpa, Huixquilucan, Atizapán e incluso dentro del cluster mayor aparece uno muy pequeño en extensión que corresponde geográficamente al Aeropuerto Internacional Benito Juárez.

Sin embargo, es interesante saber cómo se descompone el cluster más grande en grupos más pequeños, por lo que se puede ajustar el valor de inflación y obtener una mayor granularidad. En la figura 4-4 se muestran los mapas de clusters con el mismo conjunto de datos pero con valores de inflación 3.5 y 4.0, respectivamente.

Conforme el valor de inflación crece, puede verse en el mapa como el tamaño de los clusters disminuye, algunos nuevos aparecen, algunos cambian en orden de importancia y otros desaparecen al disminuir su número de elementos y dejar de aparecer en la lista de los diez mayores.

Con esta mayor granularidad, se distinguen también zonas como Xochimilco, Tlalpan, Gustavo A. Madero y Tlalnepantla, así como un cluster que incluye gran parte del oriente de la ciudad.

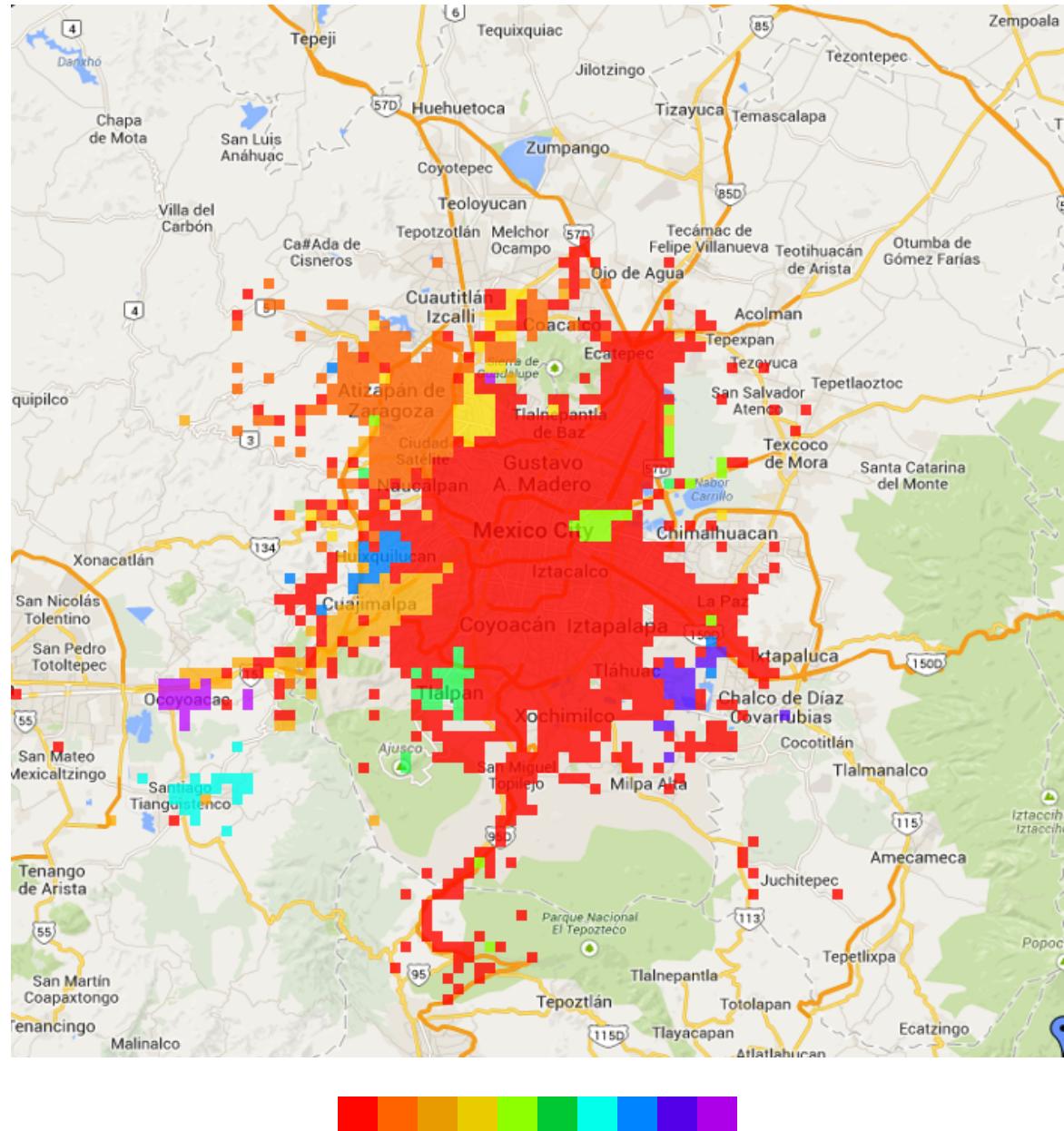
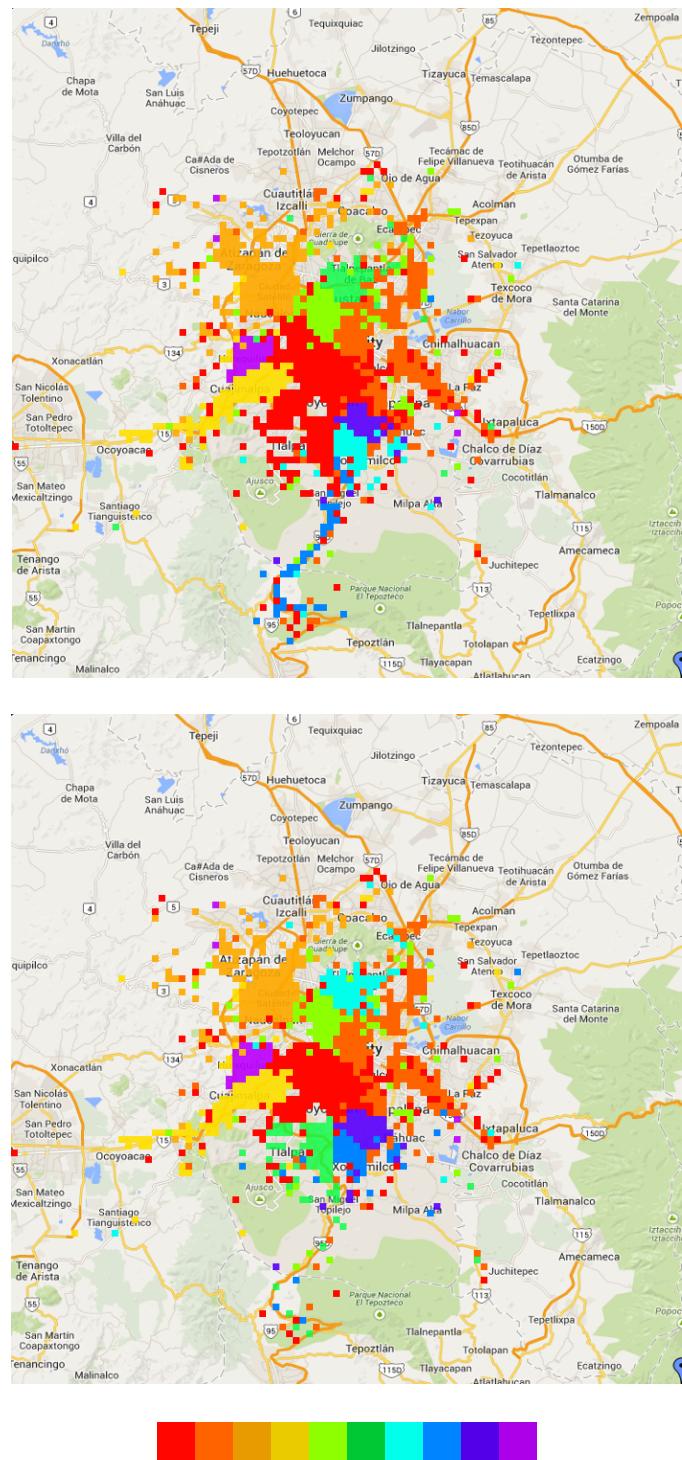


Figura 4-3 – Visualización de los diez clusters más importantes con valor de inflación

2.0 comprendiendo el periodo de 18 de junio a 4 de diciembre de 2013.



**Figura 4-4 – Mapas de los diez clusters más importantes, con valor de inflación 3.5
(arriba) y 4.0 (abajo) en el periodo de 18 de junio a 4 de diciembre de 2013**

4.5 Video

Por estar definido un grafo para cualquier periodo de tiempo, es posible generar una visualización para cada uno en una secuencia. De esta manera, la visualización incorpora la dimensión del tiempo y es posible observar como cambia la estructura de los datos con la hora o el día.

Para generar una secuencia de grafos eficientemente, y evitar hacer una consulta a la base de datos por cada periodo, se obtienen todas las tweets para los usuarios deseados y se ordenan únicamente por orden cronológico. Como en el algoritmo 4-1, se requiere para cada *tweet* saber cuándo fue observado por última vez su autor y en dónde. Entonces, se ordenaron los datos por usuario y se utilizaban variables auxiliares para guardar estos datos, pero por la necesidad de dividir el conjunto de datos en cuadros de video, las *tweets* se obtienen en orden cronológico y se requiere otra manera de registrar esta información para todos los usuarios. Se utiliza un diccionario que asocia el identificador de un usuario a un lugar y un tiempo. El procedimiento para generar una secuencia de grafos con una resolución de tiempo dada se describe en el algoritmo 4-2.

En la figura 4-5 se muestran algunos cuadros de una visualización de mapas de calor, donde cada cuadro incluye *tweets* de un periodo de 4 horas. Puede verse claramente como baja la actividad durante la noche y cómo se mantiene durante el día. Se crearon y se encuentran disponibles en la *web* tres videos, uno por cada método de visualización: mapas de calor en <https://vimeo.com/97008438>, mapas de rutas en <https://vimeo.com/97008774>, y mapas de clusters en <https://vimeo.com/97008998>.

```

lista_usuarios := los usuarios que se deseen considerar

fecha_inicio, fecha_final := el periodo de tiempo que se deseé considerar

resultado_query := seleccionar latitud, longitud, autor y fecha_de_creación de la
                     tabla de tweets donde autor esté en lista_usuarios y
                     fecha_de_creación esté entre fecha_inicio y fecha_final
                     ordenadas por fecha_de_creación

duracion_cuadro := longitud de tiempo real que se considera en cada cuadro del video

info_usuario := diccionario vacío

cuadro := fecha_inicio

para cada tweet en resultado_query:

    celda := determinar_celda(tweet.longitud, tweet.latitud)
    si tweet.fecha pertenece a cuadro:
        ultima_celda, ultima_fecha := info_usuario[tweet.autor]
        si celda != ultima_celda y
            5 minutos < tweet.fecha - ultima_fecha < 12 horas:
                agregar “ultima_celda → celda” a vértices
        si tweet.fecha no pertenece a cuadro:
            serializar y guardar vértices
            info_usuario := diccionario vacío
            cuadro := cuadro + duracion_cuadro
            vértices := conjunto vacío
            info_usuario[tweet.autor] := (celda, tweet.fecha)

```

Algoritmo 4-2 – Construcción de una secuencia de grafos dividiendo el periodo de tiempo en cuadros de longitud fija.

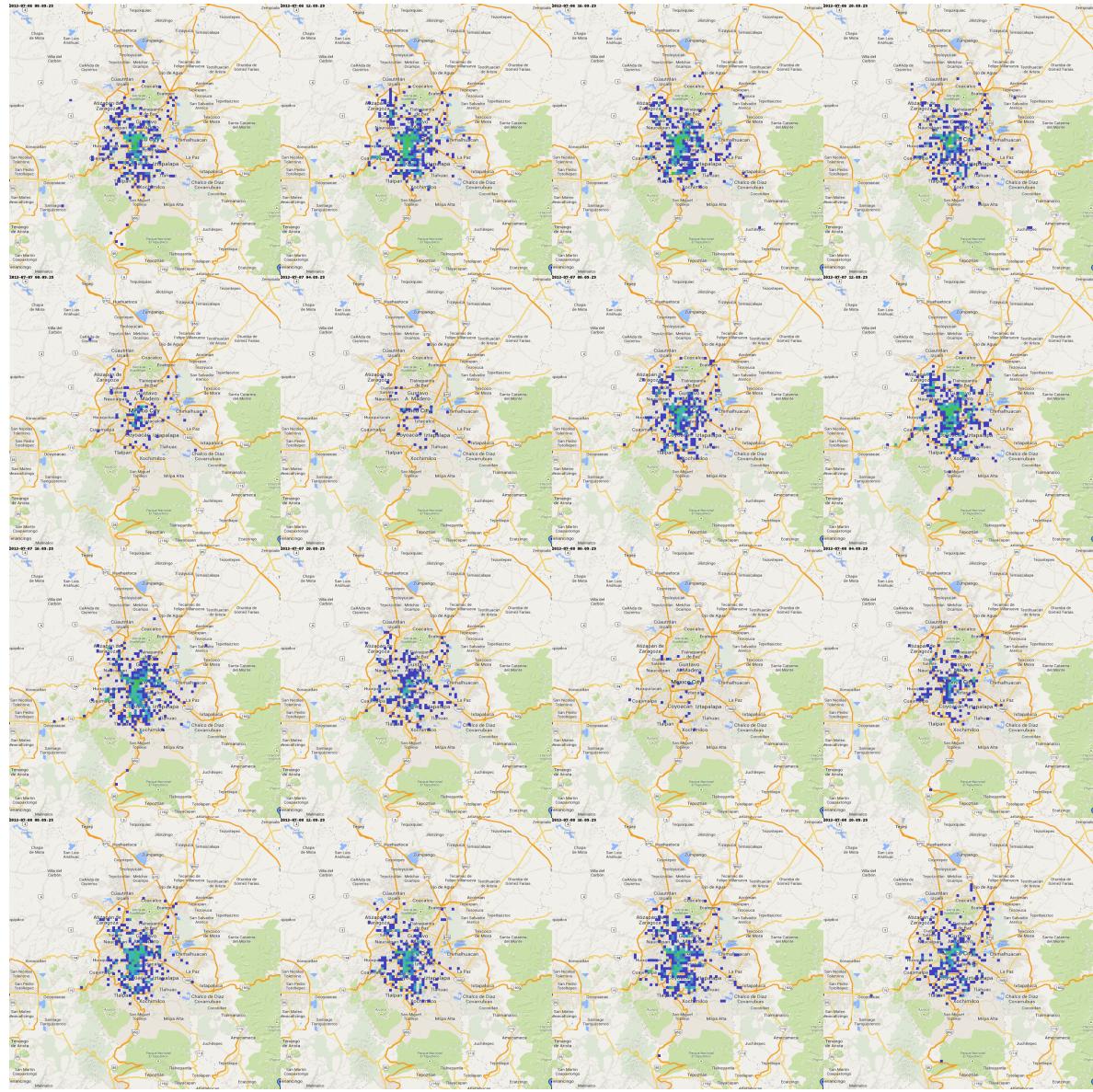


Figura 4-5 – Secuencia de 16 cuadros consecutivos de mapas de calor, cada uno incluyendo *tweets* de un periodo de 4 horas.

5. CONCLUSIONES

Es claro que es posible y útil interpretar las publicaciones de Twitter como mediciones indirectas de la ubicación de las personas y obtener información acerca de su movimiento.

Se demuestra que pueden detectarse y diferenciarse claramente zonas del área metropolitana como carreteras, distintas poblaciones y el aeropuerto. Por medio del clustering se pueden distinguir las regiones de la ciudad que son más autocontenido y autosuficientes que otras. Todo esto se logra utilizando datos públicos y de acceso gratuito, eliminando el costo de realizar mediciones directas o de despliegue de sensores. Además, dado que se están utilizando únicamente datos que fueron publicados activa y voluntariamente, no hay una transgresión a la privacidad de los usuarios.

El trabajo presenta una aplicación adicional a las descritas como antecedentes, demostrando la gran diversidad de usos que se puede dar a los datos de Twitter y su potencial como un poderoso sensor de distintos aspectos de la vida de sus usuarios. El extraer y presentar la información de movimiento de una manera fácilmente legible para las personas pude resultar tremadamente útil para la toma de decisiones, no solamente desde el nivel de planeación urbana, sino como guía para los mismos habitantes de un área que pueda influir en su comportamiento diario.

Se propone una representación de este movimiento que puede generarse automáticamente, incluso en tiempo real conforme se publican *tweets* nuevas, y que puede ser examinada posteriormente con distintos enfoques o métodos. Se describen tres distintas maneras de visualizar estas representaciones, cada una de ellas aportando un punto de vista distinto de

la estructura de los datos, y que pueden ser de gran ayuda para interpretarlos e identificar problemas a resolver.

Este tipo de métodos tendrán un rango de aplicaciones más amplio, mayor precisión, robustez y confiabilidad conforme avanza la adopción de tecnologías celulares – en particular el uso de *smartphones* – y conforme cambia la cultura de los usuarios y se vuelvan más dispuestos a compartir información en sitios de redes sociales de manera pública.

BIBLIOGRAFÍA

- Bollen, Johan, Huina Maoa y Xiaojun Zengb (2011). "Twitter mood predicts the stock market". *Journal of Computational Science*, V. 2, N. 1, pp. 1-8
- Chen, Ray y Marius Lazer (2011). "Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement", <<http://cs229.stanford.edu/proj2011/ChenLazer-SentimentAnalysisOfTwitterFeedsForThePredictionOfStockMarketMovement.pdf>> [Consulta: diciembre 2013]
- Cheng, Zhiyuan, James Caverlee y Kyumin Lee (2010). "You are where you tweet: a content-based approach to geo-locating twitter users". *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759-768
- Christakis, Nicholas A. Y James H. Fowler (2010). "Social Network Sensors for Early Detection of Contagious Outbreaks", <<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0012948>> [Consulta: diciembre 2013]
- Davis, Clodoveu A. Jr., Gisele L. Pappa, Diogo Rennó Rocha de Oliveira, Filipe de L. Arcanjo (2011). "Inferring the Location of Twitter Messages Based on User Relationships". *Transactions in GIS*, V. 15, N. 6, pp. 735-751
- Dermibas, Murat, Murat Ali Bayir, Cuneyt Gurcan Akcora, Yavuz Selim Yilmaz y Hakan Ferhatosmanoglu (2010). "Crowd-sourced sensing and collaboration using twitter." *World of Wireless Mobile and Multimedia Networks (WoWMoM)*, pp. 1-9
- Fan, Rui, Jichang Zhao, Yan Chen y Ke Xu (2013). "Anger is More Influential Than Joy: Sentiment Correlation in Weibo", <<http://arxiv.org/abs/1309.2402>> [Consulta: diciembre 2013]
- Naveed, Nasir, Thomas Gottron, Jérôme Kunegis y Arifah Che Alhadi (2011). "Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter", <<http://tw.rpi.edu/media/latest/WebSciPAper50.pdf>> [Consulta: diciembre 2013]
- Sakaki, Takeshi, Makoto Okazaki y Yutaka Matsuo (2010). "Earthquake shakes Twitter users: real-time event detection by social sensors". *Proceedings of the 19th international conference on World wide web*, pp. 851-860
- Takahashi, Tetsuro, Shuya Abe y Nobuyuki Igata (2011). "Can Twitter Be an Alternative of Real-World Sensors?" *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments Lecture Notes in Computer Science*, V. 6763, pp 240-249
- Tavares, Gabriela y Aldo Faisal (2013). "Scaling-Laws of Human Broadcast Communication Enable Distinction between Human, Corporate and Robot Twitter Users", <<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0065774>> [Consulta: diciembre 2013]
- van Dongen, Stijn (2000). "Graph Clustering by Flow Simulation." *PhD thesis, University of Utrecht*