

USO PRODUCTIVO DE BIG DATA Y REDES SOCIALES EN EL SECTOR
TURISMO

Octubre, 2014

Pedro Manuel Lichtle Fragoso
Secretaría de Turismo*

Juan Carlos Sánchez Salinas
Secretaría de Turismo**

Elaborado con la colaboración del Grupo de Trabajo sobre BIG DATA del Instituto Nacional de Estadística y Geografía (INEGI), integrado por Gerardo Leyva, Alfredo Bustos, Juan Muñoz, Silvia Fraustro y Abel Coronado.

Resumen

En este documento se describen algunos instrumentos que se han utilizado recientemente sobre el uso productivo de la información de fuentes no convencionales y también de las redes sociales en el sector turismo. Este tipo de procesos agilizan la construcción de modelos empíricos para explicar el comportamiento individual de los turistas.

El objetivo es conocer los movimientos de un perfil de usuario de Twitter, representativo de un turista promedio, con una prueba piloto en las entidades de Puebla y Guanajuato. La información de Twitter se integró con la colaboración del INEGI. El ejercicio consiste en determinar el lugar de origen de los turistas durante los días 1,2 y 3 de febrero de 2014; y contrastar este lugar de origen, que surge de Twitter, con estadísticas oficiales de Puebla y Guanajuato. El estudio representa una alternativa a los altos costos de las encuestas a nivel estatal para darle seguimiento al flujo doméstico de turistas.

La serie de Documentos de Investigación Estadística y Económica presenta resultados preliminares de investigación realizados en la Secretaría de Turismo con el propósito de generar intercambio y debate de ideas para el desarrollo del sector turismo. El contenido de los Documentos de Investigación Estadística y Económica, así como los argumentos vertidos, son responsabilidad exclusiva de los autores y no reflejan necesariamente los de la Secretaría de Turismo.

* Director General de Integración de Información Sectorial. Email: plichtlef@sectur.gob.mx

** Director de Análisis Regional. Email: jsanchezs@sectur.gob.mx

1. Introducción.

En este trabajo se describen algunos de los instrumentos que se han utilizado recientemente sobre el uso productivo Big Data y también de las redes sociales en el sector turismo. Si bien las actividades económicas y sociales requieren del análisis de datos, en la actualidad es indispensable considerar simultáneamente el volumen, la velocidad y la variedad de los datos sobre el análisis del comportamiento social con diversos objetivos y alcances. Por ejemplo, existen instrumentos que realizan extracción de información en línea automáticamente y la transforman en una base de datos con estructura definida. Este tipo de procesos agilizan la construcción de modelos empíricos para explicar el comportamiento individual.

En el contexto del desarrollo del “e-government”, “open-government” y de Big Data, la presente revisión de literatura busca evidencia empírica sobre cómo los gobiernos pueden beneficiarse de este tipo de plataformas y también de las redes sociales más populares a nivel mundial. La extracción de datos no se detiene con el Internet: mensajes de texto, teléfonos celulares y geo-ubicaciones, registros electrónicos, todos son parte de la nueva generación de información. En este sentido, el objetivo de este trabajo es obtener evidencia del uso productivo de la información y contribuir al desarrollo y uso de este tipo de instrumentos en el sector turismo de México.

Hal Varian (2009) estudiaba las atribuciones predictivas de “Google Trends”¹ y utilizaba modelos ajustados sobre las tendencias de variables como las ventas al menudeo, venta de automóviles, venta de casas e incluso movimiento de viajeros. De acuerdo con el autor, como los viajes en la actualidad se planean comúnmente por internet, los datos que provienen de la búsqueda sobre destinos turísticos pueden ser útiles en la predicción de las visitas a ese destino turístico particular.

En este orden de ideas, Tilly (2013) considera a la confianza como la variable de mayor envergadura para el uso de un servicio electrónico sobre turismo. El contenido generado por los usuarios genera mayor confianza que la información de agentes de viajes, sitios web oficiales de turismo y medios de comunicación. Los viajeros perciben el contenido generado por los usuarios de las redes sociales como una aproximación a la experiencia en el destino turístico, sin embargo, la cuestión que sigue en discusión es la posibilidad de aproximar el comportamiento de los turistas con la información que proviene tanto de Big Data como de redes sociales.

Anteriormente, los turistas solamente buscaban y utilizaban información de manera presencial en agencias de viajes, libros y folletos de viaje hechos por promotores de los destinos turísticos. La demanda de viajes se focalizaba en los paquetes turísticos clásicos programados por operadores locales, con itinerarios fijos y comunicación con un grado bajo de retroalimentación. La característica en el uso productivo de la información fue un acceso restringido a los destinos turísticos, flujo de información estático y un turista pasivo en la localización, recomendación y valoración de los servicios turísticos. Además, la decisión de

¹ Proporciona un índice del volumen de consultas de Google por ubicación geográfica y por categoría.

a dónde viajar surgía a partir de recomendaciones y comentarios de personas cercanas, sin considerar información generada de forma masiva.

Actualmente, los turistas buscan y utilizan información de manera remota en servicios digitales turísticos como sitios web, aplicaciones de todo tipo y comunidades en redes sociales. La demanda de viajes se encuentra diversificada en cuanto a alojamiento, actividades recreativas e incluso opciones gastronómicas, con itinerarios flexibles y personalizables, y un grado alto de retroalimentación. La información concerniente al turismo tiene un mayor volumen, variedad y velocidad. La característica que se supone esencial en el uso productivo de la información es un acceso diversificado a los destinos turísticos, flujo de información dinámico y un turista activo en la comunicación emergente sobre los servicios turísticos. Por ende, la decisión de a dónde viajar surge a partir de recomendaciones y comentarios que se pueden ver reflejados en la información contenida en Big Data y las redes sociales de los turistas.

Considerando este cambio de modelo de búsqueda y uso de información, las tendencias en los turistas, tanto nacionales como internacionales, se enfocan en el uso de fotos, mapas, búsquedas interactivas de restaurantes, actividades y atracciones. En este sentido, el llamado “turismo experiencial” tiene base en las emociones y adaptación a las necesidades del cliente, así como en la superación de expectativas del viaje, con el complemento de comunicar los mensajes de una forma fácil y con un lenguaje sencillo porque las redes sociales así lo demandan.

En suma, la confianza que los turistas tienen en la información contenida en Big Data y las redes sociales sobre los destinos y los servicios turísticos, y el cambio de modelo de búsqueda y uso de información son la plataforma para generar la posibilidad de aproximar el comportamiento de los turistas con la información de datos no estructurados y redes sociales.

El presente documento contiene las siguientes secciones: 2 Datos no estructurados y redes sociales; 3 Uso productivo de la información; 4 Big Data como fuente de información; 5 Twitter como fuente de información; 6 Conclusiones.

2. Datos no estructurados y redes sociales.

En primer lugar, los datos no estructurados se definen como datos abiertos y de gran volumen, accesibles para los usuarios en línea, regularmente son gratuitos y de formato abierto o indefinido. En el sector turismo se pueden encontrar datos no estructurados sobre la geo-localización de los usuarios, fotos, registros en todo tipo de hoteles y búsquedas de actividades recreativas y atracciones turísticas.

En segundo lugar, las redes sociales son un servicio en línea entre personas que, por ejemplo, comparten intereses y actividades. Un servicio de red social consiste en una representación de cada usuario (perfil), sus vínculos sociales, y una variedad de servicios adicionales.² En el sector turismo las redes sociales se utilizan para transmitir experiencias y sensaciones que vive el turista, y reconocimiento de los destinos sobre las actividades y servicios ofrecidos.

En este caso, se deben considerar las siguientes limitaciones sobre el uso productivo de la información tanto de datos no estructurados como de las redes sociales: los datos distan mucho de ser claros y contundentes y no hay un registro sistemático y estable temporalmente. Sin embargo, los problemas referidos ordinariamente son parte de las batallas que enfrentan los investigadores del comportamiento social.

Además, el mayor reto no está en recoger los datos de Internet, sino en la siguiente etapa: el procesamiento. Este proceso se desenvuelve en tres etapas, la extracción de la información deseada, depuración de la misma y conversión en información útil para la elaboración de estadísticas.³ Por ejemplo, un hotel a menudo publicará la información sobre sus servicios en más de un sitio web, también se publica en varios sitios de agregación. Esto significa que después de la recolección de los datos, un proceso en contra de la duplicación de información tiene que ser implementado.

3. Uso productivo de la información.

Recientemente, la investigación en ciencias sociales se concentra en el perfeccionamiento de modelos matemáticos para predecir las tendencias sociales y económicas. Sin embargo, la disponibilidad de los datos digitales abre nuevas opciones, tales como la tecnología de búsqueda en Internet, teléfonos móviles, correo electrónico y redes sociales. Los nuevos instrumentos ofrecen registros detallados de los comportamientos y contrarrestan los problemas clásicos de generación de estadísticas como el alto costo de los censos y encuestas a nivel nacional, así como la falta de registros administrativos precisos. Cabe resaltar que las encuestas contienen sesgos en los resultados por la posibilidad de encontrar subjetividad en los entrevistados, además de presentar tendencias de opinión y no el comportamiento real de los entrevistados.

² La mayoría de los servicios de redes sociales (como Facebook, Twitter, Google+) se basan en el uso de la red en tiempo real y proporcionan las interfaces para que los usuarios interactúen a través de medios como el correo electrónico y la mensajería instantánea.

³ Heerschap (2014).

Si bien existen recomendaciones para la medición del turismo en la escala nacional por parte de la Organización Mundial del Turismo, aceptadas internacionalmente, hoy en día el debate se centra en la generación de estadísticas regionales y sub-nacionales. Por ejemplo, se reconoce la falta de registros administrativos en algunas zonas geográficas para medir los flujos domésticos en los países, o en áreas de libre flujo de personas, así como lo costoso que resulta levantar encuestas para obtener información; lo anterior dificulta la medición del turismo.

Derivado de ello, la investigación sobre el uso productivo de la información de datos no estructurados y redes sociales, así como de nuevos métodos de estimación, adquieren gran relevancia para la generación de estadísticas en el sector turismo. Siguiendo a Heerschap (2014), los datos de Internet se pueden recoger en tiempo real y son más completos que los datos recogidos por las encuestas tradicionales. Sin embargo, para evaluar la calidad de los datos en internet, siempre es prudente realizar una comparación con los datos de otras fuentes disponibles como las estadísticas oficiales.

Un caso de investigación en turismo es la realizada por Wu (2013), donde se utiliza la tecnología de consultas en buscadores de Internet. Los autores consideran que este tipo de tecnología ha hecho posible la obtención de información a un costo mínimo y de manera prácticamente instantánea. Además, señalan que la información recabada es valiosa porque revela las características individuales del consumidor y la posibilidad de realizar una transacción económica, por lo tanto, el conocimiento de este tipo de información se puede utilizar para predecir la demanda y la oferta futura.

La información utilizada en el trabajo de Wu (2013) proviene del volumen de consultas de búsqueda en Internet relacionadas a bienes raíces de “Google Trends” con informes semanales y mensuales sobre estadísticas de consulta para diversas industrias. En este caso, la consulta permite a los usuarios encontrar información correspondiente a una frase específica, por ejemplo: “precio de la vivienda”.

La experiencia del trabajo anterior puede aplicarse para búsquedas específicas dentro del sector turismo. Por ejemplo, consultar información de usuarios con frases que contengan las palabras con mayor uso entre turistas: “hotel”, “motel”, “restaurante”, “atracción”, “diversión”, etc. Y por ende, el uso productivo de la información podría surgir del conocimiento del perfil del visitante y generar políticas públicas de atracción para este tipo de segmento en un destino turístico.

Los teléfonos móviles y las redes sociales se encuentran estrechamente relacionados porque una terminal inteligente “smartphone” contiene la capacidad de computación más avanzada y conectividad de base. Por ejemplo, existe el estudio de *Big Data y Turismo*⁴ para conocer mejor el sector turístico español. El propósito del estudio fue explorar las posibilidades que la información de Big Data puede ofrecer al sector y especialmente a la industria hotelera, incorporar macrodatos sobre los estudios de mercado de la actividad electrónica de turistas extranjeros anónimos.

⁴ Oliver (2014).

Este estudio pudo reunir y cruzar datos de dos empresas distintas: Telefónica Móviles España y BBVA. La primera es un operador integrado de telecomunicaciones líder en España y la segunda es una institución financiera privada con presencia internacional. A partir de los datos obtenidos se presenta información sobre: i) origen principal de los visitantes; ii) duración de la estancia por países; iii) desplazamiento entre dos destinos turísticos; iv) días y zonas donde prefieren alojarse los visitantes extranjeros; y v) gasto medio diario y gasto acumulado a lo largo de toda la estancia.

Las recomendaciones vertidas por el estudio se enfocan a aumentar la captación de clientes y determinar en qué países es recomendable focalizar la acción comercial, determinar las áreas de las ciudades donde se realizan las transacciones comerciales y garantizar un producto atractivo y adaptado a las verdaderas necesidades de los clientes. Como resultado, el uso productivo de la información es la generación de modelos de negocio que aumenten la atracción del destino turístico.

Adicionalmente, Spencer (2013) estudia la relación entre la ubicación de las fotografías de la página “Flickr” y las tasas de visita en distintos lugares del mundo. Este es el primer estudio sobre el terreno de la utilización de datos procedentes de las redes sociales para predecir tasas de visita en 836 ciudades de 31 países. Los autores realizan una comparación entre la densidad de las fotografías con geo-localización (coordenadas de latitud y longitud) y la tasa de ocupación media estimada con datos locales de atracción turística.

De acuerdo a los autores, la densidad de las fotografías genera un valor aproximado de la tasa de ocupación media en una función potencia,⁵ y señalan que la hipótesis central del trabajo es la siguiente: las imágenes podrían indicar la atracción de visitantes y, además, podrían grabarse para el futuro y conocer cuáles son las opciones de las personas como visitantes y proporcionar información útil del lugar.

Los autores concluyen que la información de las redes sociales (en su caso Flickr), no solamente puede romper con el problema del costo de los datos empíricos, sino también pueden generar predicciones eficaces. Pero también las redes sociales y sus complementos, pueden proporcionar información reveladora para responder preguntas acerca del porqué y en dónde las personas buscan recreación. Aquí, el uso productivo de la información llega hasta la aproximación de tasas de ocupación en varias ciudades alrededor del mundo.

Como se observa, no existe un consenso sobre el uso productivo de la información de datos no estructurados y redes sociales, incluso buscar un consenso sobre su uso en el sector turismo aún es una tarea de mayor dificultad. Con respecto al objetivo del presente documento, a continuación se realiza una breve revisión de los instrumentos utilizados en Big Data y Twitter.

⁵ $Y = Y_0 X^\beta$; donde $\beta = 0.698$, $Y_0 = 13.7$

En México, existe una investigación pionera sobre el uso productivo de la información de redes sociales, llevada a cabo por la Estrategia Digital Nacional.⁶ El objetivo fue responder las siguientes preguntas de investigación: i) ¿Existe alguna relación entre la dinámica social durante un desastre natural y los registros de llamadas telefónicas?; ii) ¿La información de las antenas telefónicas podría ser utilizada, en el futuro, para mejorar la respuesta ante un desastre natural como las inundaciones de Tabasco en el año 2009?; iii) ¿Los hallazgos obtenidos por los registros de las antenas contrastan con la información que proveen dependencias oficiales?; y iv) ¿Cómo afectan las inundaciones a los patrones usuales de comunicación en la región?

La información obtenida en este tipo de estudios podría servir como base para determinar zonas que necesiten apoyos inmediatos o el envío de recursos adicionales para centros de refugio cercanos a la zona de abastecimiento y también para generar localizaciones eficientes de los centros de distribución. Además, analizar los lugares de desplazamiento poblacional y aquellos donde se localizan los albergues temporales ayuda a evaluar si fueron localizados en un lugar correcto, así como a calcular mejor los insumos y recursos que serían necesarios hasta que las personas regresen a sus casas.

4. Big Data como fuente de información.

De acuerdo con Ronald Jansen,⁷ el uso de la nueva era de información en bases de datos de gran escala “Big Data” es la reunión de tres grandes tendencias a nivel internacional: i) transacciones en línea (compra, venta, transporte); ii) interacciones en línea (grupos sociales, promociones vía remota); y iii) procesamiento de información (técnicas de procesamiento, almacenamiento, visualización). En consecuencia, para analizar la gran variedad y la velocidad cada vez mayor de las nuevas bases de datos se debe considerar en primer término las características de la información, después los beneficios tangibles para el analista, y por último, las fuentes de información.

Cuadro 1. Consideraciones para el análisis de Big Data.

Características de la información	Beneficios para el analista	Fuentes de información
Altamente distribuida	Datos generados automáticamente	Dispositivos móviles
Poco estructurada	Datos oportunos	Transacciones digitales
Gran volumen	Potencialmente relevante	Redes sociales

Fuente: Ronald Jansen, *Big Data and Official Statistics*, presentado el 16 de junio de 2014.

⁶ *Desplazamiento poblacional y desastres naturales: análisis de datos, antenas celulares y Big Data durante las inundaciones de Tabasco en el año 2009*, presentado el 16 de junio de 2014.

⁷ Chief, International Trade Statistics, United Nations Statistics Division.

El especialista considera que este tipo de información tiene utilidad en varios sectores económicos, incluyendo al ramo turístico. Sin embargo, también considera que existen grandes retos para continuar con el análisis de Big Data. Los retos pueden agruparse en el desarrollo de metodologías de cálculo, problemas de privacidad y legalidad, desarrollo de capital humano y desarrollo de asociaciones para el análisis. Tomando como referencia este último, Ronald Jansen considera que el mayor auge se ha presentado en los proveedores y las fuentes de datos, porque son los primeros agentes que intervienen en el proceso de la cadena de generación de información.

Es importante destacar que los grandes proveedores como buscadores en la red (Google, Bing, Yahoo) y los intermediarios de los usuarios como las redes sociales (Facebook, Twitter, Flickr) pueden ser el punto de partida para generar agrupaciones con beneficios para el análisis de Big Data. A partir de la generación de información se pueden generar conexiones con tres agentes que también interactúan dentro y hacia afuera de la cadena de asociaciones. A continuación se presenta un esquema sobre la interacción de las asociaciones.

Diagrama 1. Desarrollo de asociaciones para el análisis de Big Data.



Fuente: Ronald Jansen, *Big Data and Official Statistics*, presentado el 16 de junio de 2014.

- Oficinas nacionales de estadística: tienen el conocimiento sobre los estándares en validez de datos, y las metodologías de inferencia para el uso de los datos con significancia estadística.
- Institutos de investigación y proveedores de tecnología: conocen el procesamiento para la transmisión y divulgación de la información, así como el uso de infraestructura de punta en las tecnologías de la información.
- Universidades y comunidades científicas: oferentes del capital humano necesario para la generación de modelos y fuente de innovación.

Otras fuentes de información. En Spencer (2013) se citan otras experiencias en diferentes países:

- Israel. Registro de patrones de movimiento entre un grupo de turistas que utilizaron voluntariamente un sistema de localización GPS.
- Suecia. Uso de un sistema que complementa la aplicación de encuestas con sistemas de localización GPS e imágenes aéreas para estimar el número de población que acude a un evento específico.

Lovelace (2014) señala que la información geográfica de los sitios de redes sociales tiene un gran potencial para mejorar los modelos geográficos de comportamiento, sobre todo si el volumen de los mensajes de las redes sociales geo-etiquetados sigue aumentando. Sin embargo, advierte que la información debe usarse sólo como un complemento a las fuentes de datos más consistentes o cuando los conjuntos de datos oficiales no están disponibles.

Se podría argumentar que la información en línea no es estadísticamente representativa, que constituye una pequeña minoría de la población con un accesorio inusual o dispositivo digital. Los tweets geo-referenciados utilizados por Lovelace (2014), por ejemplo, constituyen alrededor del 2% de la producción de información en Twitter y no se puede suponer que sea representativa de todos los tweets.

Por lo tanto, los académicos han sido cautelosos en este tipo de cuestiones: se utilizan en modelos con información geográfica voluntaria y especialmente cuando los datos se utilizan para explicar patrones de viaje irregulares y complejos.

Este autor utiliza un modelo de interacción espacial de los visitantes a 15 museos de 179 zonas administrativas en Reino Unido, para explicar el comportamiento de los flujos hacia los museos. Los modelos de predicción de comportamiento, de acuerdo con Lovelace (2014) son:

- i) viaje multipropósito: encadenar los viajes, por ejemplo, la cadena de viaje de casa-escuela-trabajo-tienda-trabajo-gimnasio-bar-casa, también puede ser representado como un simple viaje de casa al trabajo;
- ii) viaje estacional, la variación semanal y diurna en el momento y la frecuencia de los viajes; e
- iii) información actualizada por el usuario, adopción de los “smartphones”, por ejemplo, información de tráfico en tiempo real de Google o directamente a los sistemas de navegación en el coche por satélite.

5. Twitter como fuente de información.

Una de las redes sociales con mayor penetración en el mundo es Twitter, la cual permite enviar mensajes de texto de corta longitud (con un máximo de 140 caracteres), que se muestran en la página principal del usuario. También los usuarios pueden agrupar mensajes sobre un mismo tema mediante el uso de palabras o frases iniciadas mediante el uso de una “#”(almohadilla) conocidas como “hashtag”. Adicionalmente, dentro de Twitter existen diversas aplicaciones que brindan entretenimiento e información.

El uso de ésta herramienta, posibilita varios beneficios para el analista: costos bajos para un volumen mayor de información; acceso rápido y fácil de un número considerable de usuarios; y con información contextual, es decir, se considera la región geográfica de la extracción de los datos.

5.1 Actividad del usuario en Twitter.

Si se considera a las redes sociales como un sistema donde los usuarios comparten opiniones y experiencias, el efecto red se ocupa de manejar y transportar los mensajes hacia una gran masa de información y sobre todo se produce una posible réplica de comportamientos. En este caso, los usuarios participan en los espacios sociales en línea para obtener tres beneficios principalmente: i) adquisición de información; ii) identificación o pertenencia a una comunidad; y iii) diversión y entretenimiento.

Debido a lo anterior, la interacción de los turistas en sus redes sociales puede ser un complemento de análisis para conocer un comportamiento “promedio” en algún espacio físico como un destino turístico donde se ofrecen distintos tipos de diversión y entretenimiento. Por lo tanto, el objetivo de un análisis de Twitter podría enfocarse a conocer los movimientos de un “perfil de usuario” de redes sociales representativo de un “turista promedio” con una cuenta activa y en uso de twitter con acceso vía móvil.

5.2 Actividad del usuario en Twitter en México.

En México, el 95% de los usuarios de Twitter tienen entre 18 y 49 años.⁸ Este porcentaje se asemeja al perfil de turistas en destinos de ciudad y playa, en donde el 87% de los turistas tiene entre 18 y 55 años de edad.⁹ Por consiguiente, el mayor porcentaje de turistas son potencialmente usuarios de esta red social. Al tener un acercamiento del perfil entre los usuarios de redes sociales con respecto al perfil del turista en México, se puede generar un vínculo o encadenamiento de ambos y por lo tanto, generar información más fiable a la hora de comparar sus comportamientos.

Adicionalmente, el estrato socioeconómico de los usuarios de Twitter es mayoritariamente medio y alto. En correspondencia, los deciles de hogares con mayores ingresos tienden a dedicar una mayor proporción de sus ingresos en esparcimiento. En este sentido, la realización del viaje para vacacionar y realizar compras absorbe los mayores gastos realizados por el turista, seguidos de los gastos personales.¹⁰

Finalmente, debe destacarse que en México el uso de internet y redes sociales va en aumento. El número de usuarios de internet para acceder a redes sociales se ha incrementado hasta llegar a 15 millones en 2013. La penetración de internet es de aproximadamente 46 millones de usuarios a nivel nacional. En consecuencia, el crecimiento del uso de estas herramientas en México es extraordinario, así como su crecimiento potencial en los próximos años.

⁸ Consulta Mitofsky (2011), p.6.

⁹ SECTUR (2010), p.7.

¹⁰ SECTUR, Encuesta del Gasto en Turismo en Hogares, México, 2011, p.30.

5.3 Uso productivo de la información de Twitter en el sector turismo.

Hoy en día el uso de las Tecnologías de la Información y la Comunicación (TIC's), así como de nuevos métodos de estimación, adquieren gran relevancia para la generación de estadísticas en el sector turístico. En este sentido, la propagación del uso de las redes sociales en el sector turismo abre una ventana de oportunidades para subsanar la falta de registros administrativos y brindar información para la toma de decisiones en el sector.

5.4 Caso de estudio: Twitter y flujos domésticos en Puebla y Guanajuato.

En respuesta a los beneficios y ventajas descritas sobre la posibilidad de utilizar productivamente la información de Twitter, la Secretaría de Turismo en colaboración con el INEGI¹¹ realizó el siguiente ejercicio para aproximar el origen de los turistas domésticos a nivel estatal en Puebla y Guanajuato.

De una colección de 60 millones de tuits registrados desde la última semana de enero de 2014 hasta la primera semana de julio de 2014,¹² se seleccionaron todos los tuits georreferenciados en el territorio nacional. Posteriormente, se realizó una selección geográfica en los polígonos que corresponden a las entidades de Puebla y Guanajuato. Además, se realizó una selección temporal para los días 1, 2 y 3 de febrero de 2014, considerado como un fin de semana largo o coloquialmente como un puente en México.¹³

Consecutivamente, se ordenaron los tuits por tuitero, eliminando duplicados para obtener la relación de tuiteros con actividad en Puebla y Guanajuato durante ese fin de semana largo. En suma, se obtuvo un total de 7,955 tuiteros que tuitearon en las entidades de Guanajuato (48%) y Puebla (52%) durante un fin de semana largo.¹⁴ Posteriormente, de la base de datos original de 60 millones de tuits, se identificaron todos los tuits de la relación de los 7955 tuiteros con actividad en Puebla y Guanajuato durante ese fin de semana largo, para poder identificar desde qué entidad federativa emitían tuits durante todo el semestre, dando como resultado 827,424 tuits.

Estos 827,424 tuits se ordenaron por tuitero y cronológicamente para identificar cada cambio de entidad federativa por tuitero, y así determinar las secuencias de cambio con destino Puebla o Guanajuato en donde además la estancia en estos destinos no fuera mayor a 15 días. De este resultado se contabilizaron las visitas por cada estado de procedencia. Después, los resultados se plasmaron en mapas (Mapa 1 y Mapa 2) y se

¹¹ Resultado de un análisis preliminar de una muestra de tuits dentro del contexto de una prueba de concepto de Big Data.

¹² Es una muestra de tuits generados con georreferenciación dentro de la República Mexicana, recolectados de Twitter entre la última semana de enero y la primera de julio de 2014.

¹³ De acuerdo con el Decreto del 23 de agosto de 2013 por el que se establece el Calendario Oficial en México: el primer lunes de febrero es día no laboral porque se conmemora el 5 de febrero, como aniversario de la Batalla de Puebla. Diario Oficial de la Federación.

¹⁴ En un segundo ejercicio se podría considerar un periodo mayor, como la temporada de verano.

realizó una comparación con las estadísticas generadas por los observatorios turísticos correspondientes de Puebla y Guanajuato.

En los mapas, la coloración tiende al color rojo cuando existe un número menor de turistas provenientes de esa entidad y tiende al color verde cuando existe un número mayor de turistas provenientes de esa entidad.¹⁵ En este caso, también se ordenó jerárquicamente a las entidades de origen por el número de visitas de tuiteros (de estancia no mayor a 15 días).

Al realizar la comparación del orden jerárquico de las primeras 10 entidades de donde provienen los turistas que visitan Puebla y Guanajuato (considerando el ejercicio de tuits) y las primeras 10 entidades de donde provienen los turistas (considerando las estadísticas de los observatorios turísticos); se observa lo siguiente:

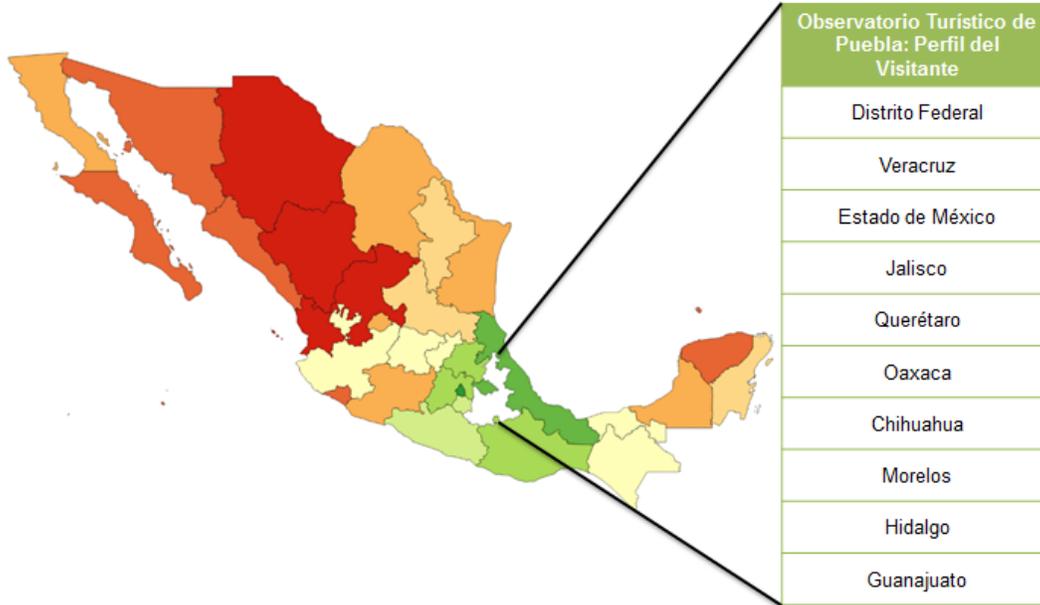
- En Puebla, las 10 principales entidades de donde provienen turistas domésticos acumulan el 95% de los tuits analizados durante el 1, 2 y 3 de febrero de 2014; y corresponden a la tendencia mostrada por el observatorio turístico de Puebla en 9 de 10 entidades.¹⁶
- En Guanajuato, las 10 principales entidades de donde provienen turistas domésticos acumulan el 96% de los tuits analizados durante el 1, 2 y 3 de febrero de 2014; y corresponden a la tendencia mostrada por el observatorio turístico de Puebla en 9 de 10 entidades.¹⁷

¹⁵ La segmentación utilizada para la distribución de los tuits fue por percentiles y las observaciones se estratificaron en 9 estratos.

¹⁶ El coeficiente de correlación en Puebla, considerando la composición para el porcentaje de cada entidad de origen del ejercicio de tuits y el porcentaje de cada entidad del observatorio turístico es 0.71.

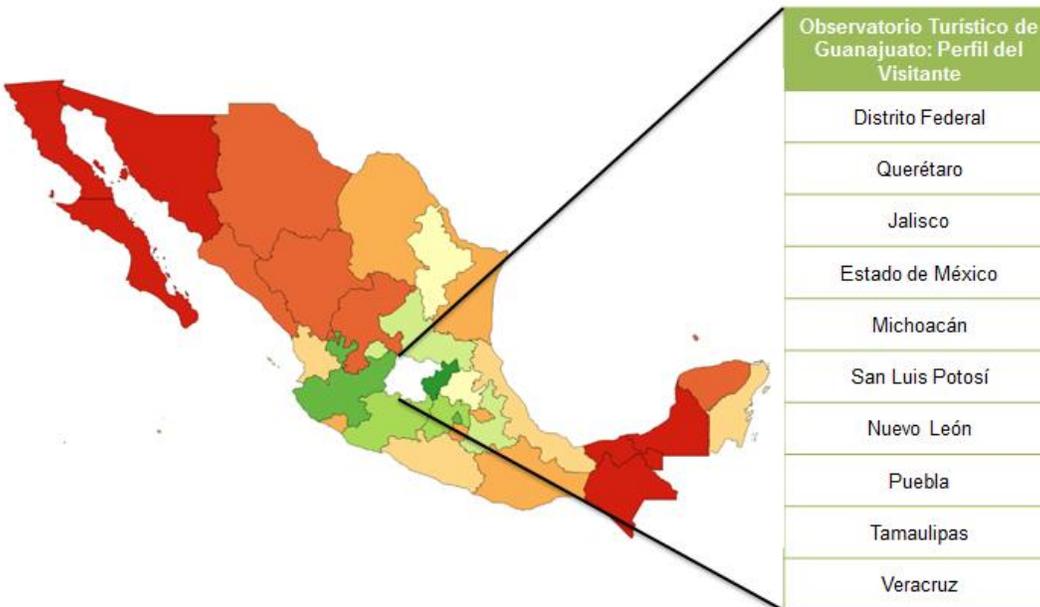
¹⁷ El coeficiente de correlación en Guanajuato, considerando la composición para el porcentaje de cada entidad de origen del ejercicio de tuits y el porcentaje de cada entidad del observatorio turístico es 0.82.

Mapa 1. Twitter y flujos domésticos en Puebla.



Fuente: Elaboración de SECTUR e INEGI con base en información de Twitter y del Observatorio Turístico de Puebla.

Mapa 2. Twitter y flujos domésticos en Guanajuato.



Fuente: Elaboración de SECTUR e INEGI con base en información de Twitter y del Observatorio Turístico de Guanajuato.

Por ejemplo, en el Mapa 1 las entidades de color verde intenso (Distrito Federal, Veracruz y Tlaxcala) corresponden a las entidades con mayor número de turistas que visitaron Puebla en el periodo de observación; mientras las entidades de color rojo intenso (Coahuila, Durango y Zacatecas) corresponden a las entidades con menor número de turistas que visitaron Puebla.

Lo mismo ocurre en el Mapa 2: Querétaro, Jalisco y Distrito Federal (color verde intenso) corresponden a las entidades con mayor número de visitantes en Guanajuato; y Baja California, Baja California Sur y Sonora (color rojo intenso) son las entidades con menor número de visitantes en Guanajuato durante el periodo de observación.

Los resultados pueden ser matizados identificando las rutas de acceso tanto para Puebla como para Guanajuato. Pueden existir barreras naturales (orografía) que evitan que los turistas de una entidad se trasladen a otra y viceversa. Además, puede existir diferente desarrollo de infraestructura para una mejor movilidad en algunas entidades con respecto a otras, por ejemplo, conectividad aérea.

En primer lugar, considerando la orografía del país, tanto Guanajuato como Puebla se encuentran en el eje neovolcánico lo que puede representar una barrera natural para los turistas que viajan en automóvil, debido al costo de acceso y posiblemente mayor tiempo de recorrido para llegar. En consecuencia, las entidades con un mayor número de turistas en ambos casos corresponden a las entidades cercanas.

En segundo lugar, considerando la conectividad aérea del país, se observa que las grandes ciudades tienen una centralidad en las rutas por lo que puede representar una diferencia de infraestructura para acceder a Puebla o Guanajuato de entidades lejanas del norte o sureste del territorio. Sin embargo, Tijuana tiene un aeropuerto importante a nivel de conectividad nacional y se observa que Baja California no es una entidad con el menor número de visitantes para Puebla. Por consiguiente, los resultados no se pueden generalizar y se requiere de una revisión con mayor precisión en cada caso.

Para efectos prácticos, se observa que el área de influencia turística en Puebla y Guanajuato se circunscribe en las entidades aledañas, de color verde en cada uno de los mapas. Lo anterior puede explicarse porque se observó un periodo de tres días, con una restricción para realizar viajes con una distancia mayor, existe la posibilidad de incrementar la selección temporal en periodos vacacionales con una mayor duración y posiblemente esto disminuirá el error de predicción, y por consiguiente, incrementará el coeficiente de correlación del orden jerárquico entre el ejercicio de tuits y el observatorio turístico de cada entidad.

Adicionalmente, se considera importante realizar un análisis de similitud entre los flujos domésticos para entidades ubicadas en la misma región. Por un lado, si tenemos flujos similares, podrían realizarse proyectos conjuntos en búsqueda de corredores turísticos entre las entidades. Por otro lado, si tienen flujos contrapuestos, existe la oportunidad de promover líneas de producto afines a cada entidad en cuestión y mejorar su afluencia de turistas diversificando su oferta hacia otras entidades donde no se encuentre presencia de origen de turistas.

En este sentido, el ejercicio podría replicarse en cada una de las entidades federativas y observar con mayor detalle las rutas de acceso para todos y cada uno de las entidades; para así mejorar las hipótesis sobre las barreras naturales y de infraestructura a las que se

enfrenta el turista doméstico. Incluso puede servir para identificar mejores rutas de acceso y mejorar las condiciones de infraestructura en algunos casos.

El ejercicio anterior puede ayudar a mejorar el entendimiento y seguimiento del flujo doméstico de turistas y también sirve como una alternativa a los altos costos de las encuestas a nivel nacional. No hay que perder de vista que la información que surge de Twitter no tiene un costo por sí misma, el costo proviene de la extracción, recolección y resguardo de la misma. En conclusión, el uso productivo de Twitter en el turismo puede proveer de los siguientes beneficios a los analistas: i) costos bajos para un gran volumen de información; ii) acceso rápido y en tiempo real de un número considerable de usuarios; e iii) información contextual, considerando la región geográfica de la extracción de los datos.

6. Conclusiones.

En un contexto de escasez de recursos financieros y en la búsqueda de generar información para una mejor toma de decisiones para los sectores público, privado y de la sociedad civil, se hace necesario acudir a fuentes de información susceptibles de aportar elementos de juicio para el diseño de políticas públicas orientadas al turismo.

Lo anterior ha de tomar en cuenta el uso de metodologías y algoritmos adecuados para el manejo de volúmenes importantes de información, así como los nuevos métodos gráficos y estadísticos para dar a conocer los hallazgos y la validación de los resultados obtenidos. En conclusión, la revisión del uso productivo de la información tanto de Big Data como de redes sociales constituye un proceso que exige contar con instituciones eficaces en las que exista una estrecha cooperación entre los gobiernos y las organizaciones privadas.

En el caso de estudio sobre Twitter, existen ventanas de oportunidad para alcanzar una mayor correlación entre las 10 principales entidades de donde provienen turistas domésticos, tanto en Puebla como en Guanajuato, con la tendencia mostrada por el observatorio turístico de las entidades. Por ejemplo, un mayor número de días de seguimiento a los tuiteros y también contemplar información de fechas más recientes en los observatorios.

En general, este documento es un primer paso para el análisis de Big Data y redes sociales en el turismo de México. La metodología puede mejorarse en un futuro con la integración de nuevas herramientas de acceso a la información y con la posible aportación de las observaciones, opiniones y sugerencias de actores que deseen incursionar en este tipo de análisis. En conclusión, el sector turismo puede ser un campo fértil para aprovechar el desarrollo del análisis de Big Data.

Bibliografía.

- [1] Castro, Carlos (2014), "*Datos para el Desarrollo. Desastres Naturales Tabasco 2009*". Presentación en el Seminario Internacional: Big-Data para la Información Oficial y la Toma de Decisiones. INEGI. CIDE. INFOTEC, 16-17 junio 2014.
- [2] Consulta Mitofsky (2011), "*Perfil de usuarios de redes sociales en Internet (Facebook y Twitter)*", Diciembre 2011.
- [3] Elsevier (2014), "*Open Government, Open Data and Digital Government*", Editorial, Government Information Quarterly, No. 31, 4-5, 2014.
- [4] Hal Varian & Hyunyoung (2009), "*Predicting the Present with Google Trends*", Google Inc.
- [5] Heerschap, Nico; Ortega, Shirley; Priem, Alex and Orrermans (2014), "*Innovation of tourism statistics through the use of new big data resources*", The Hague, The Netherlands, 27 March 2014.
- [6] Liran Einav & Jonathan D. Levin (2013), "*The Data Revolution and Economic Analysis*", NBER Working Paper No. 19035.
- [7] Lovelace, Robin; Nick Malleon, Kirk Harland and Mark Birkin (2014), "*Geotagged tweets to inform a spatial interaction model: a case study of museums*", School of Geography, University of Leeds, Leeds, UK.
- [8] Oliver, Víctor; García Enrique y otros (2014), "*Big Data y Turismo: Nuevos Indicadores para la Gestión Turística*", Barcelona, Telefónica y RocaSalvatalla.
- [9] Organisation for Economic Co-operation and Development (2012), "*OECD Workshop on e-government indicators*", Public Governance and Territorial Development Directorate, Paris, December 2012.
- [10] Pablo Barberá (2013), "*NYU Politics Data Lab Workshop: Scraping Twitter and Web Data Using R*", Department of Politics, New York University.
- [11] Ronald Jansen (2014) "*Big Data and Official Statistics*". Presentación en el Seminario Internacional: Big-Data para la Información Oficial y la Toma de Decisiones. INEGI. CIDE. INFOTEC. 16-17 junio 2014.
- [12] Secretaría de Turismo (2010), "*Perfil y Grado de Satisfacción del Turista*", CESTUR.
- [13] Secretaría de Turismo (2011), "*Encuesta del Gasto en Turismo en Hogares*".
- [14] Secretaría de Turismo del Gobierno del Estado de Guanajuato (2013). Perfil del visitante al Estado de Guanajuato 2013. Disponible en: <http://www.observatorioturistico.org/>
- [15] Secretaría de Turismo del Gobierno del Estado de Puebla (2013). Perfil del visitante, Diciembre 2013. Disponible en: <http://www.observatoriopuebla.org>.
- [16] Sencer A. Wood, Anne D. Guerry, Jessica M. Silver & Martin Lacayo (2013), "*Using social media to quantify nature-based tourism and recreation*". Scientific Reports. Stanford, CA, USA.
- [17] Tilly, Roman; Fischbach, Kai; and Schoder, Detlef (2013), "*Assessing The Potential Of Social Media To Reflect Global Tourism*" ECIS, Research in Progress. Paper 17.
- [18] Ubaldi, Barbara (2013), "*Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives*", OECD Working Papers on Public Governance, No. 22, OECD, Publishing.
- [19] Wu, Lynn & Erik Brynjolfsson (2013), "*The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales*". The National Bureau of Economic Research. May 2014.

La serie de Documentos de Investigación Estadística y Económica presenta resultados preliminares de investigación realizados en la Secretaría de Turismo con el propósito de generar intercambio y debate de ideas para el desarrollo del sector turismo. El contenido de los Documentos de Investigación Estadística y Económica, así como los argumentos vertidos, son responsabilidad exclusiva de los autores y no reflejan necesariamente los de la Secretaría de Turismo.
