Universidad Nacional Autónoma de México

Doctorado en Ciencias (Física)

Estudio de patrones de comportamiento humano en ciudades a partir de la actividad en Puntos de Interés. Un enfoque de redes y sistemas complejos.

T E S I S

que para optar el grado de:

Doctor en Ciencias (Física)

P R E S E N T A :

Carlos Francisco Betancourt Moreno

tutor

Dr. José Luis Mateos Trigos
Instituto de Física

Ciudad Universitaria, CDMX, septiembre 2023

A familiares, compañeros y amigos.

A Héctor Luis, María Cristina y Héctor.

A Karen. Sobre todo a Karen.

A toritas y toritos, mis camaradas.

Agradecimientos

Quisiera expresar mi sincero agradecimiento al Doctor José Luis Mateos, mi asesor, cuya orientación, apoyo y dedicación han sido fundamentales para la realización de este trabajo. Su paciencia y acompañamiento fueron invaluables a lo largo de este arduo proceso. Aprecio la confianza que depositó en mí y la libertad que me brindó para explorar y desarrollar mis ideas. La firmeza y sinceridad con la que expresó sus observaciones fueron la clave para no perder el rumbo de esta investigación. Igualmente, su sentido del humor fue, en muchas ocasiones, una bocanada de aire fresco en medio del cansancio y las dificultades que llegaron a presentarse.

Al Doctor Alejandro Pérez Riascos debo agradecerle su acompañamiento permanente pero, sobre todo, durante los momentos más difíciles de la investigación. Cuando las cosas parecían no avanzar, siempre estuvo ahí para ofrecer su apoyo y respaldo. Gracias por los sabios consejos y la infinita paciencia.

Agradezco al Doctor Luis Olivares Quiroz por las múltiples sesiones de discusión. Muchas veces, su tutoría fue un impulso importante para mi trabajo.

Al Doctor Denis Boyer por formar parte de mi comité tutor durante todos estos años. Gracias por sus sugerencias y críticas constructivas.

A todo el jurado que revisó este trabajo: Dr. Carlos Gershenson, Dr. Hernán Larralde, Dr. Francisco Sevilla y Dr. Lev Guzmán. Muchas gracias por el tiempo y cuidado dedicados a la revisión de mi escrito y por sus valiosas observaciones.

A la Universidad Nacional Autónoma de México y, especialmente, al Instituto de Física en el que conté con un espacio de trabajo. Todos los días que pasé ahí marcaron esta importante etapa de mi vida.

Al Posgrado en Ciencias Físicas de la UNAM. Fue un honor y un privilegio haber pertenecido a este programa, tanto en la maestría como en el doctorado. Gracias a todo el personal administrativo y a la coordinación que, con gran gentileza, me apoyaron en todo momento.

Al Consejo Nacional de Humanidades, Ciencias y Tecnologías por el apoyo económico sin el cual no hubiera sido posible este trabajo (beca 587060).

A la Dirección General de Asuntos del Personal Académico que, mediante el proyecto PAPIIT IN116220, permitieron que esta investigación se realizara.

Índice general

Índice de figuras

Índice de tablas

Parte del trabajo reportado en esta tesis doctoral se publicó el 25 de marzo de 2023 [1] en la revista Scientific Reports, que publica SPRINGER NATURE GROUP. Este artículo se reproduce íntegramente en el Apéndice A de esta tesis. La referencia completa es:

Francisco Betancourt, Alejandro P. Riascos & José L. Mateos, Temporal visitation patterns of points of interest in cities on a planetary scale: a network science and machine learning approach. *Scientific Reports* **13**, 4890 (2023).

`https://doi.org/10.1038/s41598-023-32074-w`

`https://www.nature.com/articles/s41598-023-32074-w.epdf`

Introducción

Las ciudades son un tema apasionante. Cada una tiene dinámicas y vida propias. Desde las ciudades de los palacios hasta las que nunca duermen. Ciudades gemelas y ciudades del borde. Las de la eterna primavera y las del eterno verano. Ciudades monstruo, ciudades rojas, ciudades blancas o ciudades de la moda. Entre las ciudades sultanas, airosas, prohibidas, heroicas y eternas, la lista es interminable. Cada una tiene una historia, una personalidad y un temperamento. A pesar de sus defectos y virtudes, cada una contribuye a contar la historia humana. Este trabajo es un esfuerzo más por aproximarnos al fascinante mundo de las ciudades esperando, como gente de ciencia, encontrar algo coherente.

En esta tesis se estudia la estructura de un sistema de cientos de ciudades alrededor del mundo. Este sistema está definido a partir de la dimensión temporal de la actividad humana. A qué hora comemos, a qué hora trabajamos, cuándo salimos y volvemos a nuestra casa, en qué momento nos reunimos con nuestros pares; todo esto depende en buena medida de en dónde nos tocó vivir. Muchas de nuestras rutinas y decisiones individuales parecen estar fortaleciendo una dinámica multitudinaria que se va cristalizando en un patrón urbano, al que aportamos pero que a la vez nos moldea como parte de un colectivo. El principal resultado de este estudio es que, al sumar las dinámicas individuales de miles de habitantes, emerge una huella que distingue a cada ciudad como sistema único. Sin embargo, en esta huella encontramos efectos de dos factores que influyen en nuestra identidad. Por un lado, al formar parte de un mismo fenómeno humano universal, todas tienen una marca común. Por otro lado, al ser fruto de historias de intercambios e interacciones, muchas tienen similitudes que esperan por ser explicadas. A partir de la similitud entre patrones temporales se creó una red que conecta urbes de 87 países en todos los continentes, lo que recoge una gran diversidad humana. Analizando esta red se encontraron tanto los aspectos comunes como los aspectos distintivos.

Esta investigación resulta interesante porque integra de manera novedosa fuentes y métodos que previamente estaban separados. Esto ha resultado en un estudio comparativo amplio que combina el análisis de la actividad temporal con la identificación de similitudes entre ciudades y regiones. Esta combinación ha permitido abordar un número significativamente mayor de ciudades en comparación con investigaciones anteriores. Además, ha superado desafíos previos al emplear áreas urbanas funcionales para filtrar datos de zonas urbanas sin recurrir a divisiones político-administrativas o umbrales arbitrarios. Es relevante porque a partir de información pública se desa-

rrollaron métodos y programas que pueden ser utilizados en conjuntos de datos más grandes y proyectos más ambiciosos. Este enfoque destaca por su utilidad práctica y su potencial para aplicaciones en diversos contextos.

Este trabajo se sitúa en la intersección entre la nueva ciencia de las ciudades, la física de redes y la física de sistemas complejos, con el propósito de contribuir al estudio de los sistemas urbanos a nivel global. En primera instancia, se desarrollaron programas para procesar una extensa base de datos de check-ins de la red social Foursquare, extrayendo los registros correspondientes a cada ciudad. Luego, se organizó la información con el fin de obtener, en cada caso, la distribución temporal de los check-ins por hora dentro de una ventana semanal. Mediante técnicas de teoría de la información y aprendizaje automático, se compararon estas distribuciones entre todas las ciudades, dando lugar a la construcción de una red de similitud. Se exploraron las propiedades de esta red y se identificaron comunidades utilizando el método de maximización de la modularidad. Finalmente, se analizaron las propiedades temporales de cada comunidad, con base en la distribución geográfica de las ciudades que las conforman. Este enfoque demuestra cómo las herramientas y enfoques provenientes de estas disciplinas pueden facilitar la comprensión de la complejidad y dinámica de las ciudades, contribuyendo al análisis de un fenómeno sumamente relevante para la sociedad.

En el primer capítulo se examinan las bases de la ciencia de las ciudades. Abordamos las teorías que tratan de la forma en que las ciudades se organizan, así como las interacciones entre los diversos componentes que conforman los sistemas urbanos. Se hace un repaso de los principales representantes y los resultados más relevantes en este novedoso enfoque. Ese capítulo se nutre principalmente de textos recientes [2–4] que, desde la disciplina de la física estadística, han abordado la dinámica, evolución, surgimiento y estructura de las ciudades como fenómeno humano universal. El objetivo del capítulo es presentar, junto con su marco conceptual, la Base de Datos Global de Núcleos Urbanos [5] que se utilizó en este capítulo para definir nuestro objeto de estudio: un sistema de áreas urbanas funcionales [6].

En el segundo capítulo se profundiza en la teoría de grafos y la ciencia de redes complejas [7, 8], examinando cómo estas herramientas matemáticas y computacionales han sido utilizadas para entender mejor la estructura y el funcionamiento de las ciudades. La exposición que se hace del tema busca ser alternativa a la que se encuentra en los libros más populares sobre la materia. Además, se complementan los conceptos con el estudio de sistemas reales y se proporcionan los códigos computacionales para la reproducción de los resultados. Esta exposición es fruto del curso *Temas Selectos de Termodinámica y Física Estadística - Ciencia de Redes* impartido desde 2019 en la Facultad de Ciencias de la UNAM, a cuyas notas y simulaciones se hace referencia durante todo el capítulo.

En el tercer capítulo predomina un enfoque de ciencia de datos. Se presenta una base de datos de la red social basada en ubicaciones Foursquare, a partir de la cual se define el sistema de interés para este estudio: los puntos de interés ubicados en 632 ciudades de todo el mundo y los registros, llamados check-ins, que se realizaron en

cada uno. Se explora esta base de datos, la forma en que fue obtenida y sus principales características. Además se presentan algunos resultados de su análisis que permiten entender las dinámicas de las ciudades y extraer conocimientos fundamentales para tomar decisiones informadas sobre políticas urbanas.

Finalmente, en el cuarto y último capítulo, se presentan los resultados del análisis de un fenómeno urbano particular: regularidades en los patrones de comportamiento temporal de los habitantes de muchas ciudades del mundo. En este capítulo se define una red de similaridad de patrones temporales con la ayuda de la teoría de la información, específicamente de la divergencia de Kullback-Leibler. Se analizan las propiedades de esta red de similaridad, se desarrolla un método para elegir un umbral de similitud crítico que recoja las características esenciales del sistema y se utiliza la teoría de redes complejas para identificar grupos de ciudades a partir de sus patrones temporales de actividad. Este método se complementa con algoritmos de aprendizaje automático no supervisado. Se concluye que la dimensión temporal de la interacción entre personas y puntos de interés en las ciudades depende, por un lado, de factores humanos comunes a todas las urbes y, por otro, de factores específicos, culturales e históricos, correspondientes a regiones geográficas concretas.

Al final se presentan las conclusiones, los apéndices y, anexo, el artículo en el que se publican los principales resultados de esta investigación.

1 La ciencia de las ciudades

1.1. Las ciudades como objeto de estudio

Las ciudades han sido un fenómeno recurrente en la historia universal [9]. Han surgido en prácticamente todos los lugares habitados por seres humanos [10]. Actualmente, la mayoría de las personas viven en entornos urbanos y éstos se han convertido en el escenario principal de muchas actividades fundamentales para la vida de las naciones: economía, política, comercio, investigación, etc. El crecimiento urbano está experimentando un aumento sin precedentes. Según Naciones Unidas [11], el 55 % de la población mundial en 2018 vivía en ciudades y se espera que llegue a 68 % en 2050.

Las ciudades son un fenómeno relativamente reciente desde una perspectiva evolutiva [10]. Su registro arqueológico coincide con el surgimiento del sedentarismo, la agricultura, la burocracia y las sociedades políticamente asimétricas [9]. La importancia social, política y económica de las ciudades como núcleos de interacción social humana radica en que el intercambio de conocimiento genera ideas que resultan en innovación, crecimiento económico y rendimientos crecientes [10, 12, 13]. Además, las ciudades se han convertido en el principal escenario para la generación de innovaciones sociales, institucionales y tecnológicas [12, 13]. Por lo anterior, las ciudades pueden verse como reactores sociales que fomentan la creatividad y la imaginación humana, lo que las vincula con aspectos positivos [3].

El desarrollo urbano también conlleva múltiples desafíos. Problemas como el congestionamiento vehicular, la contaminación, las afectaciones a la salud de sus habitantes, la delincuencia, la desigualdad social y económica, entre otros, son comunes en la mayoría de las ciudades, independientemente de su tamaño [14]. Es difícil encontrar alguna ciudad, ya sea grande o pequeña, que no enfrente uno o varios de estos problemas. Hoy en día, alrededor de mil millones de personas en todo el mundo viven en barrios informales y sin acceso adecuado a servicios básicos [11]. Aproximadamente el 20 % de la población de cualquier ciudad no puede costear los precios de vivienda del mercado, lo que da lugar a situaciones de falta de hogar o campamentos sin servicios mínimos. Esta problemática afecta tanto a países ricos como a los menos desarrollados [11].

Se proyecta que en las próximas décadas se alcanzará el máximo histórico en la población mundial, lo que conducirá a la creación de ciudades de proporciones nunca antes vistas [3]. Hacia finales de este siglo, se espera que cada una de las 10 ciudades

más grandes del mundo supere los 50 millones de habitantes, e incluso podrían llegar a alcanzar los 70 u 80 millones. Este pronóstico indica que dichas ciudades superarán el tamaño actual de Tokio, que hoy es la zona metropolitana más poblada con 40 millones de habitantes [3]. Si este crecimiento se mantiene sin atender las problemáticas asociadas con la vida en las ciudades, las consecuencias humanas pueden ser incalculables.

Aprovechar los aspectos positivos de las ciudades es clave para el futuro. Por eso, una comprensión científica del fenómeno urbano es tan importante. Frente a esto, en los últimos años ha surgido grupos que intentan enfocar a las ciudades desde un punto de vista científico y con la contribución de múltiples disciplinas [2–4, 10, 15–18]. Estos grupos, además, tienen fuertes influencias de las ciencias de la complejidad [19]. En pocas palabras, está emergiendo un nuevo paradigma. Este nuevo paradigma, a veces llamado ciencia urbana o ciencia de las ciudades, pretende entender los procesos fundamentales que conducen, dan forma y sostienen a las ciudades y la urbanización.

Este nuevo enfoque se ha dado a la tarea de recoger resultados que se han obtenido en estudios previos sobre las ciudades. De forma interdisciplinaria, se trata de incorporar conclusiones obtenidas en campos como la arquitectura [20], la geografía [21–23], el urbanismo [9, 24], la ingeniería [25], la economía [26–31], sociología y antropología [32–34].

El objetivo de los siguientes apartados de este capítulo es desarrollar las bases y los principales resultados de la ciencia de las ciudades [2, 3, 10, 35, 36].

1.2. Ciudades como sistemas complejos

Desde los años 60 del siglo XX se ha aplicado un enfoque sistémico al estudio de las ciudades [37]. Este enfoque estudia a los fenómenos como sistemas de elementos y considera las interacciones y relaciones entre ellos como parte fundamental para su comprensión [38]. Se basa en la premisa de que los elementos de un sistema están interconectados y que el sistema como totalidad tiene propiedades y comportamientos que no se pueden entender analizando sus partes individuales de forma aislada. Las mismas propiedades de los elementos están determinadas por sus interacciones. Todo esto suele resumirse de forma coloquial en la frase "el todo es más que la suma de las partes" [39]. Como las interacciones entre elementos puede producir comportamientos complejos en formas difíciles de predecir, a estos comportamientos del sistema como un todo se les suele llamar emergentes [19]. En este sentido la emergencia, el surgimiento de comportamientos emergentes, es una relación no trivial entre las propiedades a escala *microscópica* (al nivel de los elementos) y a escala *macroscópica* (al nivel del sistema) [40]. Además, desde el enfoque sistémico se busca comprender cómo los elementos y la totalidad se influyen mutuamente generando bucles de causalidad entre el todo y las partes [38].

Aplicar un enfoque sistémico a las ciudades implica considerar que las conexiones y la interdependencia juegan el papel más importante. Este enfoque parte de que las pro-

piedades de las ciudades son sistémicas [24, 32]. Dicho de otro modo, en las ciudades aparecen comportamientos colectivos a partir de las interacciones de distintos tipos de elementos. Las ciudades tienen propiedades que van más allá de las propiedades de los espacios arquitectónicos y de los actos individuales de sus habitantes, lo que las hace entidades con un comportamiento y una vida propia [41]. En las ciudades surgen propiedades emergentes. Por ejemplo, en las ciudades surgen procesos cognitivos y comportamientos colectivos distintos a los de los individuos fuera de ellas. Igualmente, en las ciudades han surgido innovaciones en los social, político, económico, y demás aspectos de la vida. También pueden verse como fenómenos emergentes el crimen, la segregación, los asentamientos irregulares y el incremento en los costos. Estos, aunque no son fenómenos exclusivos de la vida en las ciudades, en ellas adquieren una relevancia especial. Igualmente, en las ciudades se generan sorpresas y las urbes están expuestas a las catástrofes; en ellas pueden surgir movimientos sociales contestatarios cuyas consecuencias van mucho más allá de la fronteras de una sola ciudad.

Se puede emplear una perspectiva sistémica a las ciudades en muchas formas. Una es viendo a los individuos como elementos que interactúan en los espacios urbanos. A ese nivel, la interrelación e interdependencia de los elementos se comprueba en el hecho de que el comportamiento de las personas está fuertemente influido por aquellos con quienes interactúan [33, 42]. La relación entre la totalidad y los elementos, en este caso, se manifiesta en que, por un lado, las ciudades son el espacio en el que suceden y se moldean los comportamientos y las vidas de millones de personas; y por otro, las contribuciones individuales son las que dan sentido, vigencia y vida a la ciudad. Los habitantes no son sin su ciudad, pero la ciudad tampoco es sin sus habitantes [41]. También puede verse a las ciudades como relaciones entre espacios o regiones con distintas características. Por ejemplo, pensar en las ciudades como compuestas de lugares de trabajo y lugares de residencia. En este caso, las relaciones entre elementos podrían ser los flujos de personas, bienes e información de un lugar a otro [2]. Otra forma de comprender la realidad urbana como un sistema es mediante la vinculación de los distintos aspectos de la realidad social como elementos [10]. En este caso, por poner un ejemplo, cambios en la economía, la política o en la tecnología provocan cambios en la dinámica y forma de las ciudades. Así ocurrió cuando la disminución súbita de los costos de transporte que se experimentó en el siglo XX provocó un crecimiento de las áreas urbanas sin precedentes, la aparición de los suburbios y el surgimiento de ciudades lejos de los recursos naturales, situación que antes era imposible [43].

Cuando las propiedades del sistema se derivan de propiedades e interacciones entre los elementos, se dice que surgen estructuras "de abajo hacia arriba". Muchos de los fenómenos en las ciudades se pueden analizar como emergencia de patrones en escalas superiores a lo local, en donde operan los individuos. Es el caso, por ejemplo, de los patrones de segregación en ciudades predichos a partir de comportamientos individuales mediante modelos de agentes [44, 45]. Un razonamiento similar lleva a la idea de que una ciudad es, en sí misma, un fenómeno emergente resultado de interacción a un nivel inferior. Muchas de las propiedades de las ciudades pueden verse como propiedades emergentes de la aglomeración humana: los sistemas culturales y tecnológicos que la humanidad ha desarrollado no provienen únicamente de la capacidad

cognitiva individual, ni de la suma de capacidades individuales, sino del conocimiento distribuido y mantenido en redes sociales, en los que grupos más grandes mantienen mayor acumulación de tales conocimientos [10, 34, 46]. Esto es consistente con que muchas características socioeconómicas de los asentamientos humanos dependen del tamaño de la población [47–49].

Las ciudades mismas pueden considerarse elementos de un sistema superior. Este planteamiento es muy utilizado en estudios urbanos y plantea que entre ciudades se dan interacciones físicas, sociales, económicas y de proximidad geográfica [41]. Algunos ejemplos de este procedimiento se observan en [50–52]. Este enfoque es relevante pues en este trabajo se sigue este camino, definiendo un sistema de ciudades cuyas relaciones se miden a partir de la similitud de sus patrones temporales como se verá en el capítulo 4.

Una pionera de los estudios de las ciudades como sistemas fue Jane Jacobs [24]. Para Jacobs, las ciudades son un claro ejemplo de sistemas que expresan una complejidad organizada [53, 54], como lo plasma en el capítulo "Qué tipo de problema es una ciudad" [24]. Las ideas sobre la complejidad organizada introducidas por Weaver [53] que Jacobs aplicó a las ciudades, se han convertido en el estudio de lo que hoy llamamos sistemas complejos. Un sistema complejo se puede definir [19,40] como aquel en el que grandes redes de componentes sin un control central y con reglas de operación simples dan lugar a comportamientos colectivos complejos, procesamiento de información y adaptación mediante evolución o aprendizaje. Jacobs sentó las bases de la perspectiva sistémica y de complejidad para abordar a las ciudades [3]. Las ciencias de la complejidad, ocupadas de los sistemas complejos, han complementado a la teoría de sistemas y han significado una nueva etapa del estudio de los sistemas urbanos [55].

Las ciencias de la complejidad proporcionan un método para identificar los mecanismos detrás de algunas propiedades empíricas de las ciudades [2], como son su evolución, su dinámica temporal, los flujos en las ciudades, y cómo las ciudades y los procesos que conllevan cambian en la medida que las ciudades y sus poblaciones crecen. Desde esta corriente se describe a las ciudades como sistemas adaptativos [56] y esta idea es dominante hoy ya que inspira a gran parte de la arquitectura y planificación contemporáneas hacia ciudades sostenibles con planificación y diseño orientados a las personas [57].

Las posibilidades combinatorias de las acciones de todos los agentes que constituyen una ciudad (personas, familias, vecindarios, empresas, organizaciones, regiones, etc.) son incontables. Como muchos sistemas complejos, superan nuestra capacidad de describir cualquier estado particular del sistema por lo que los escenarios futuros son demasiado numerosos para generar y evaluar en detalle. En este sentido, es imposible planificar a las ciudades de forma exhaustiva [3]. Las ciudades no se crean a conciencia ni se estructuran según fines voluntarios. Su organización depende de comportamientos y decisiones que se toman en todas las escalas, desde lo local y más rutinario, hasta lo global y las decisiones de políticas públicas. Pero son tomadas por individuos y grupos humanos, por lo que se dice que la organización de una ciudad es altamente descentralizada [3].

La física estadística trata sobre el vínculo entre las propiedades e interacciones microscópicas y el comportamiento macroscópico emergente. Se han desarrollado muchas técnicas y conceptos para comprender esta relación entre los sistemas y sus elementos, sobre todo cuando éstos son muchos. El vínculo entre las interacciones microscópicas y el comportamiento colectivo emergente es una razón importante que impulsó a los físicos a pensar que estas herramientas y conceptos podrían ser "exportados" a sistemas socioeconómicos complejos. Las herramientas de la física estadística pueden aportar a la construcción de un conocimiento científico que apunte a las regularidades de los sistemas urbanos [4]. Es en el marco de estas reflexiones que surge la intención de desarrollar una ciencia de las ciudades. Ésta debe ser capaz de identificar y explicar las propiedades emergentes a través de las distintas escalas, incorporando métodos y conceptos que faciliten la comunicación entre disciplinas e integrando conocimientos de otras disciplinas como la economía, la sociología, la geografía, la ecología, antropología, arqueología e historia [10]. Una de las herramientas más importantes para el estudio de los sistemas complejos es la teoría de redes o grafos que se introduce en el siguiente apartado.

1.3. Ciudades como redes

Como se dijo en el apartado anterior, la nueva ciencia de las ciudades está fuertemente influida por grupos de física estadística y física de sistemas complejos. Estos grupos han propuesto cambiar el enfoque de estudio de las ciudades, pasando de un análisis centrado en las ubicaciones a otro centrado en las interacciones. Para ellos, es conveniente pensar a las ciudades como sistemas de comunicación, interacción, comercio e intercambio; en resumen, varios grupos e investigadores coinciden en que hay que pensar en las ciudades como redes [2–4, 10, 36, 58–60].

Los sistemas se definen generalmente como entidades organizadas compuestas de elementos y sus interacciones. La idea de una red está muy relacionada con esta definición. Según Barabási [7] detrás de cada sistema complejo hay una intrincada red que codifica las interacciones entre las componentes del sistema y no se puede comprender al sistema si no se logra una comprensión profunda de la red que está detrás. No es casualidad que la mayoría de los exponentes de esta ciencia de ls ciudades dediquen un capítulo de sus libros a la ciencia de las redes [2–4]. Este trabajo no es la excepción, ya que el capítulo 2 está dedicado a ese tema.

Batty [2], uno de los principales precursores de la ciencia de las ciudades, defiende que con las herramientas de la teoría de redes se pueden deducir los resultados en el estudio de las ciudades. Desde esta perspectiva, las ubicaciones son el resultado de las relaciones humanas y no al revés [2]. Los espacios urbanos son el resultado de las relaciones que se representan través de redes (sociales, comerciales, de flujos de información, energía, etc.) que, a su vez, se interrelacionan entre sí. Estas redes pueden ser de flujos de personas, bienes, energía, materia e información [10]; o interacciones entre diversos tipos de elementos como personas, empresas o regiones. A estas redes se puede aplicar modelos físicos [8] para dar cuenta del funcionamiento y estructura

de las ciudades las leyes que parecen gobernar a las ciudades.

Aunque habrá un capítulo dedicado al tema, vale la pena mencionar que, en este caso, las redes pueden aportar a la comprensión de cada nivel de análisis de las ciudades como sistemas. Redes sociales entre individuos; redes entre empresas para representar flujos de bienes; redes de infraestructura sobre las que se dan los flujos de materia, personas, energía o recursos entre regiones; redes de sectores económicos, y redes de ciudades en las que cada ciudad interactúa con otras a nivel nacional o mundial [1, 51, 52, 61]. De este modo, las ciudades pueden analizarse como redes, pero se necesitan distintos tipos de redes, como son las espaciales [62, 63] o las sociales [7, 8, 64], para dar cuenta de sus procesos y evolución. Para comprender a las ciudades se les debe ver no sólo como lugares en el espacio sino como interacciones y flujos (de gente, de bienes, de información) y para entender interacciones y flujos se deben entender las redes [10, 65–67]. Este es el enfoque que se adopta en este trabajo

1.4. Ciudades y datos

El desarrollo de la tecnología en la última década ha producido gran cantidad de datos relacionada con la vida en las ciudades. Por ejemplo, el uso de tarjetas de crédito [68], dispositivos digitales para servicios de transporte como taxis [69], autobuses y metro [70], bicicletas [71] y redes de telecomunicación como celulares [72–78] o dispositivos GPS [79–82]. Esta disponibilidad de datos es el factor clave que ha revolucionado la investigación sobre las ciudades [4, 35].

El surgimiento de la ciencia de las ciudades ha sido posible gracias a la disponibilidad de nuevos datos sobre los fenómenos urbanos. En las últimas dos décadas se ha visto una explosión en las fuentes de información que dan luz sobre características y aspectos de la vida urbana a todos los niveles: comportamiento individual, barrios, áreas urbanas y sistemas de ciudades. Este surgimiento de grandes cantidades de datos, el Big Data urbano [10], ha implicado especialmente tres cambios: (i) nuevos métodos para analizar los datos existentes; (ii) nuevos métodos de generación de datos urbanos, y (iii) recopilación y difusión de datos en manos de múltiples actores, no solo gobiernos y empresas.

Un caso especial se ha dado en estudios de movilidad. Hoy en día, lo datos permiten monitorear la posición de las personas en todo momento y esto ha desencadenado una gran cantidad de estudios cuantitativos sobre la movilidad humana. Gracias a los datos disponibles, se ha podido demostrar que los desplazamientos de las personas Δr describen una distribución de probabilidad $P(\Delta r)$ tipo *cola larga* [83]. La importancia de $P(\Delta r)$ radica en su papel central en la modelación del comportamiento de viajes humanos, y muchos autores han estudiado esta cantidad. Brockmann y colaboradores [84] publicaron un artículo pionero utilizando datos de billetes en Estados Unidos [85], encontrando rastros de movimientos con distribución de probabilidad de cola larga en la longitud de los pasos. Otros estudios continuaron en la misma dirección encontrando comportamientos de cola larga en la movilidad humana a nivel nacional e interurbano [72, 78, 86–91], caracterizados por un decaimiento de ley de

potencia para desplazamientos largos.

Los datos disponibles para el estudio cuantitativo de ciudades abarcan muchas escalas temporales diferentes. Los dispositivos electrónicos personales proporcionan ubicación de personas en cada momento, en escalas de minutos u horas. Los sistemas de transporte proporcionan datos de movilidad a lo largo del día [70, 92]; muchos sensores y dispositivos miden variables como contaminación del aire a distintas horas del día. En el otro extremo, las encuestas y censos proporcionan información en escalas de años o décadas sobre uso de suelo o infraestructura, mientras que la digitalización de mapas y documentos históricos permiten estudiar las ciudades a escalas de siglos del crecimiento de las ciudades y su evolución. Incluso nuevos métodos en la arqueología están generando datos cuantitativos de ciudades de un pasado muy remoto, por ejemplo el uso de LiDAR ha generado datos para el estudio de los asentamientos humanos de la antigüedad [93].

Todo esto ha permitido extender los estudios comparativos de ciudades y asentamientos humanos a distintas épocas. También los asentamientos antiguos se investigan como redes de interacciones sociales en la llamada Teoría de Escalamiento de Asentamientos (Settlement Scaling Theory) [94] con la que se ha podido identificar cuáles de las propiedades que tienen hoy en día todas las ciudades, surgieron desde los primeros asentamientos y cómo ha sido su evolución. Algunas de estas propiedades se resumen en el siguiente apartado.

1.5. Regularidades de las ciudades

Las ciudades de todo el mundo y a lo largo de la historia varían mucho en tamaño, forma y organización. Sin embargo, hay regularidades empíricas que, como tendencias generales y no como leyes inviolables, están bien documentadas. En ocasiones se les llama universales urbanos, otras veces se agrupan como leyes de la geografía o firmas de la complejidad urbana. Estos "universales" urbanos incluyen organización en vecindarios [31, 95–97]; autosimilitud, invariancia espacial a través de diferentes escalas y fractalidad [98]; desigualdad social [99–101]; relaciones alométricas o de escalamiento entre distintas variables y la aglomeración humana [12, 13, 94, 102]; los flujos entre ciudades descritos por modelos gravitacionales de Tobler [103, 104]; las leyes de crecimiento proporcional de Gibrat y de distribución de los tamaños relativos de Zipf [4, 105–107], entre otros. La importancia de estas regularidades radica en que dan cuenta aspectos comunes a todas las ciudades cuya comprensión invita a un estudio científico [2–4,14]. De éstas, desarrollaremos un poco las leyes de escalamiento y la ley de Zipf, porque las utilizaremos al final de este capítulo para probar la importancia de la base de datos de áreas urbanas que se usó en este trabajo para definir nuestro sistema de interés.

1.5.1. Escalamiento

Algunas características de las ciudades cambian con su escala. La escala es generalmente medida por el tamaño de la población. Por ejemplo, las ciudades más grandes dentro de la misma nación suelen ser más densas y hacen un uso diferente de su infraestructura. Las ciudades más grandes también suelen ser más productivas económicamente, pero también más costosas. [13]. La aplicación de este marco en el análisis de ciudades, llamado *escalamiento urbano*, cuantifica muchas de sus características como su capacidad para crear "rendimientos crecientes a escala en actividades socioeconómicas" [13, 108, 109]. Abordar los problemas de las ciudades es generalmente un problema dependiente de la escala [3]. Según algunos autores [10, 13], este es un aspecto común de las ciudades: la capacidad de que los efectos se incrementen más que proporcionalmente con el cambio en las causas.

El escalamiento de una variable Y respecto a un parámetro de tamaño N se reduce a la siguiente forma

$$Y \sim Y_0 N^{\beta} \tag{1.1}$$

en donde Y_0 es una constante y β es el exponente de escala. Los tres regímenes de escalamiento posibles que se siguen de la evidencia empírica podrían resumirse como sigue:

- Algunas variables relacionadas con la actividad socioeconómica (por ejemplo el producto interno bruto o los casos de SIDA en la población) escalan de forma super-lineal con el tamaño de la población ($\beta > 1$). Esto significa que la cantidad per cápita de estas variables tiende a incrementarse con el tamaño de la ciudad en un efecto llamado rendimientos crecientes a escala. Típicamente, una ciudad grande genera más riqueza que dos ciudades de la mitad de la población juntas.

- Otras variables asociadas con los servicios individuales básicos (número de casas, consumo de agua, por ejemplo) escalan de forma lineal con el tamaño de la población ($\beta = 1$).

- Y las variables relacionadas con la infraestructura como los cables eléctricos o el número de estaciones de gas, por ejemplo, escalan de forma sub-lineal ($\beta < 1$). Esto implica que las grandes ciudades requieren menos infraestructura per cápita.

El escalamiento es un ejemplo de regularidad empírica de las ciudades lo suficientemente general como para justificar, según Bettencourt, el desarrollo de una teoría científica urbana. Para este autor, la teoría del escalamiento urbano proporciona un escalón para desarrollos teóricos más elaborados que involucren estadísticas, diversidad socioeconómica y crecimiento. Para Batty, el escalamiento y las leyes de potencia que las describen son la firma de los sistemas complejos y una muestra convincente de que sólo mediante las ciencias de la complejidad se puede abordar el problema urbano. Ribeiro y Rybski [12] revisan los trabajos que explican el escalamiento no lineal con algún modelo que va más allá de la caracterización empírica.

1.5.2. Ley de Zipf

Al ordenar las ciudades de mayor a menor según el tamaño de su población, la ciudad de mayor tamaño tendrá aproximadamente el doble de población que la segunda, el triple que la tercera y así sucesivamente. Este enunciado es comúnmente conocido como la ley de Zipf para ciudades y captura el hecho de que hay muchas más ciudades pequeñas que grandes a lo largo de varios órdenes de magnitud [110]. Este es un resultado clásico y da cuenta de una propiedad de sistemas de ciudades a nivel país o en el mundo entero. Si se define el rango de los elementos de un conjunto, como en el caso de las ciudades, como la posición que cada una ocupa al ser ordenadas según una variable específica, como la población (de mayor a menor), se observa un comportamiento universal de la forma:

$$P_r \sim r^{-\nu} \tag{1.2}$$

donde r se refiere al rango y P_r la población de la ciudad correspondiente a tal rango. El exponente ν toma un valor cercano a 1 [2–4],

El origen de la ley de Zipf observada para las poblaciones de las ciudades podría residir en los intercambios de población entre ciudades [4]. Bettencourt incluye la ley de Zipf en ciudades, junto con la ley de la gravedad de los flujos entre ciudades, la ley de crecimiento proporcional de las ciudades de Gibrat y una serie de otras observaciones y regularidades estadísticas relacionadas que caracterizan la influencia espacial, la migración o las características espaciales y funcionales de las ciudades de diferentes tamaños, como las leyes de la geografía.

1.6. Delimitación de las áreas urbanas

La definición de ciudad y sus fronteras es un tema complejo y ha sido tratado muchas veces antes [111]. En este trabajo, para los análisis que se llevarán a cabo en los capítulos siguientes, optamos por la definición de Área Urbana Funcional [6] que integra factores como infraestructura, población y economía. Particularmente se usó la Base de Datos de Centros Urbanos - Capa Global de Asentamientos Humanos [5] (GHS-UCDB su nombre en inglés Global Human Settlement Urban Centre Database) creada por el Centro Común de Investigación (JRC) de la Comisión Europea para estudiar las áreas urbanas de todos los países de forma consistente. A continuación se desarrolla la definición de Área Urbana Funcional y se describe la GHS-UCDB, además de que se reproducen con ella algunos resultados de interés para la ciencia de las ciudades como el escalamiento y las leyes de Zipf para la población y tamaño de las ciudades.

1.6.1. Área Urbana Funcional

Antes de describir la Base de datos de centros urbanos, en este apartado explicaremos la definición de Área Urbana Funcional desarrollada por la Unión Europea y la OCDE con la finalidad de hacer estudios comparativos de ciudades.

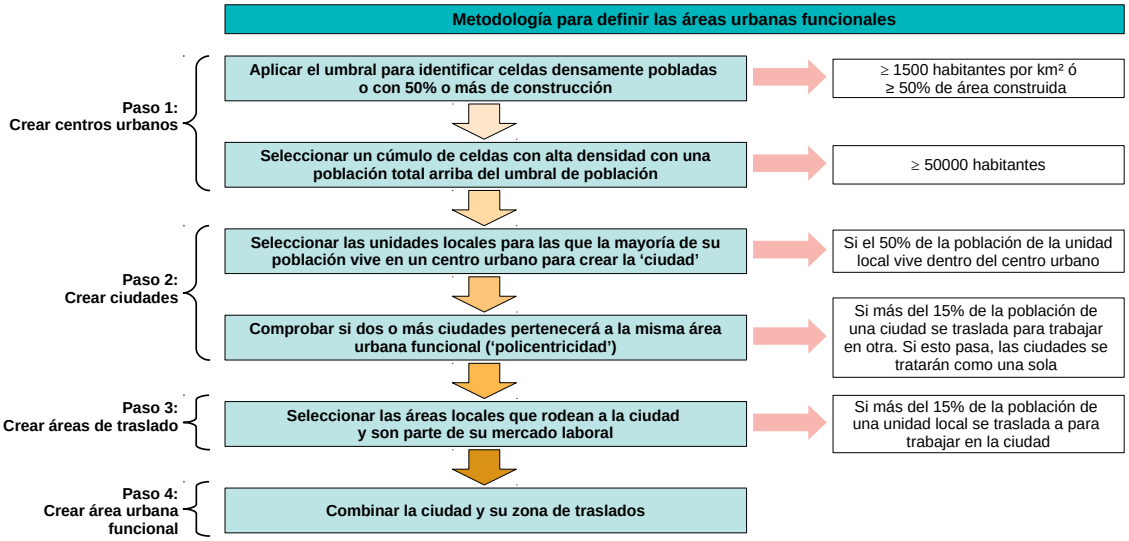


Figura 1.1: **Algoritmo para la definición de un área urbana funcional**. Elaboración propia a partir de la figura 5 de [6]

Un área urbana funcional se define en cuatro pasos: identificar un centro urbano, identificar la ciudad asociada al centro urbano, identificar la zona de desplazamiento asociada a la ciudad y, finalmente, combinar todas las anteriores. Ahora se explica con detalle cada paso.

Centro urbano. Para definir los centros urbanos se parte de dos fuentes, imágenes satelitales (Landsat y Sentinel) y datos censales o estadísticos, y se aplica el siguiente procedimiento:

- Se genera una capa con información del área de construcción mediante algoritmos de clasificación automática supervisada de imágenes y se obtiene el porcentaje de área construida por cada celda en una cuadrícula de 1km $\times$ 1km de resolución.
- Superponiendo a la cuadrícula de área construida la información censal y estadística, se estima la población de cada celda de 1km^2.
- Se identifican todas las celdas que satisfagan una de dos condiciones: una población mayor a 1,500 residentes o un porcentaje de construcción de al menos 50 %.
- Se generan cúmulos a partir de las celdas contiguas que cumplen con el punto anterior y se seleccionan aquellos que suman una población total mayor a 50,000 habitantes.
- Los bordes se suavizan mediante un filtro iterativo en el que a cada celda se le asocia el valor mayoritario de su vecindad, y los huecos (celdas que no forman parte del cúmulo pero están rodeadas de celdas del cúmulo) menores a 15 km^2 se rellenan.

Los cúmulos resultantes conforman los centros urbanos.

Ciudad. Una ciudad está compuesta por una o más unidades locales con al menos el 50 % de su población en un centro urbano. Una unidad local puede ser tanto administrativa como estadística. Ejemplos de unidades administrativas son municipios, distritos, vecindarios, áreas metropolitanas, distritos electorales o áreas de gobierno local, dependiendo del país. Las unidades estadísticas pueden ser áreas de enumeración, bloques de censo, distritos censales, entre otros. La mejor unidad local para esta definición es la unidad más pequeña para la cual se disponga de datos de desplazamiento de personas. Si el 15 % de las personas empleadas que viven en una ciudad trabajan en otra ciudad, estas ciudades se tratan como un mismo destino policéntrico en el siguiente paso.

Zona de desplazamiento. Una vez que todas las ciudades han sido definidas, las zonas de desplazamiento pueden identificarse siguiendo los siguientes pasos:

1. Se identifican como parte de la zona de desplazamiento de una ciudad todas las unidades locales en las que al menos el 15 % de sus residentes empleados trabajan en esa ciudad o destino policéntrico.

2. Se incluyen los enclaves, es decir, las unidades locales completamente rodeadas por otras unidades locales que pertenecen a una zona de desplazamiento o a una ciudad, y se excluyen las unidades locales no contiguas.

Área Urbana Funcional. Finalmente, el Área Urbana Funcional (AUF) consiste en la ciudad y su respectiva zona de desplazamiento. Puede suceder que, debido a una baja intensidad de flujos de desplazamiento diario, no haya una zona de desplazamiento diario. En este caso, hay una correspondencia entre la AUF y la ciudad.

Esta definición es compatible con el enfoque de este trabajo por dos razones principales. La primera es que nos interesa hacer estudios comparativos de muchos casos alrededor del mundo ya que partimos de la premisa de que hay aspectos comunes a todas las ciudades y es tarea de los estudios de complejidad urbana dar cuenta de ellos [14]. En segundo lugar, esta definición no pone en el centro los espacios físicos para definir a las ciudades para luego dar cuenta de los procesos e interacciones que ocurren en ella; al contrario, parte de los procesos e interacciones humanas para, en función de ellos, establecer los límites relevantes del sistema de interés que, la mayoría de las veces, trasciende las fronteras políticas y administrativas con las que se suele encuadrar el análisis de las ciudades. Este es, a nuestro juicio, un cambio clave en la dirección propuesta por Batty [2] de poner en el centro las interacciones.

Las unidades espaciales que resultan de este método no son perfectas [94] ya que su obtención recae en la identificación de áreas urbanizadas mediante imagen satelital a una resolución de 1-km, y la intersección de estas áreas con las unidades locales de conteo de población con mayor o menor resolución espacial en cada nación. Sin embargo, este método define y mide áreas urbanas funcionales de un modo consistente y tiene la ventaja de haber sido aplicado alrededor de todo el mundo, haciendo factible comparar relaciones entre población y área a una escala global. Como tal, represen-

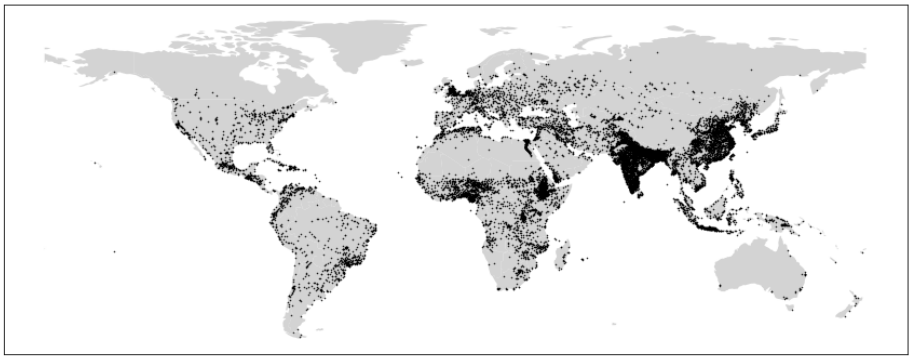


Figura 1.2: Ubicación de las 13,135 áreas urbanas funcionales contenidas en la Base de datos de Centros Urbanos del proyecto Global Human Settlement Layer. Los continentes, sin divisiones políticas, se muestran en color gris mientras que cada área urbana funcional se dibuja como un pixel negro.

ta un recurso importante para la investigación amplia de patrones y variaciones en sistemas urbanos.

1.6.2. Base de datos de centros urbanos

La Comisión Europea desarrolla el proyecto Global Human Settlement Layer para producir y analizar herramientas científicas y mapas globales de asentamientos humanos y densidad de población con el objetivo de "comprender la presencia humana en el Planeta Tierra"[1]. Una de estas herramientas es la Base de datos de Centros Urbanos (GHS-UCDB por sus siglas en inglés) que, a partir de la definición de AUF expuesta en el apartado anterior, contiene las fronteras y 160 campos de información para 13,135 áreas urbanas en las que viven 3,535,326,299 personas en todo el mundo. En la figura 1.2 se muestran todas las AUF sobre un mapa, cada una como un pixel negro, para ver su distribución espacial en todo el mundo.

La información de la GHS-UCDB está contenida en un archivo tipo shapefile (.shp) con todos los polígonos en el espacio de coordenadas longitud-latitud que abrimos y analizamos con la ayuda de la paquetería GeoPandas [112], de Python. Además, para cada polígono se incluye el nombre de la ciudad, su población, las coordenadas de su centroide, su área, el país y región en la que se encuentra y si se extiende más allá de las fronteras de una sola región dentro del mismo país (por ejemplo Nueva York, cuya área urbana funcional se distribuye en condados pertenecientes a los estados de New York y New Jersey, en Estados Unidos) o, incluso, en más de un país (por ejemplo Detroit, en Estados Unidos, cuya área urbana funcional se extiende hasta Ontario, Canadá, incluyendo a la ciudad de Windsor). Para ilustrar el hecho de que las fronteras funcionales de las ciudades, en la mayoría de los casos, van más allá de las fronteras administrativas, en la figura 1.3 se muestran tres AUF (línea

[1] `https://joint-research-centre.ec.europa.eu/scientific-tools-and-databases/global-human-settlement-layer-scientific-tool_en`

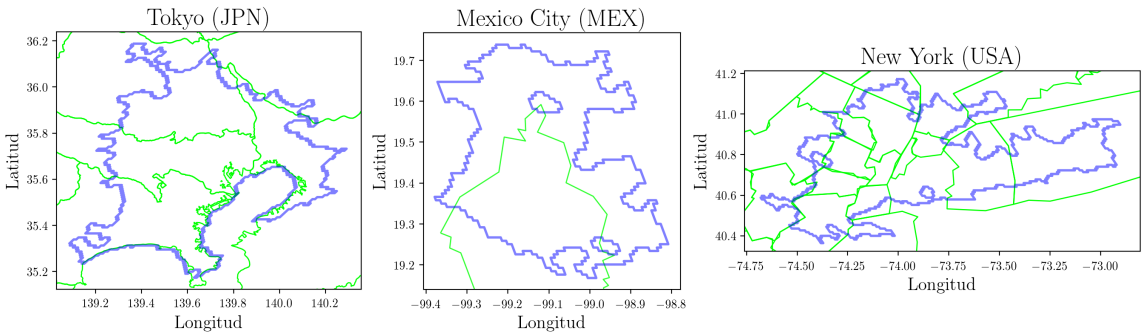


Figura 1.3: **Superposición del polígono de tres centros urbanos y las fronteras administrativas con las que coinciden**. En los tres casos, las fronteras administrativas se muestran en verde y el polígono del AUF se muestra en azul.

azul) superpuestas a las fronteras subnacionales correspondientes (líneas verdes). Se eligieron las ciudades de Tokio, Ciudad de México y Nueva York debido a que las tres abarcan más de un estado y múltiples entidades subestatales (condados en el caso de Estados Unidos, municipios y alcaldías en el caso de México). En la figura se hace evidente que, si se acotara cualquier análisis a las fronteras político administrativas, se perderían regiones considerables de las AUF.

Esta base de datos ha sido utilizada anteriormente en múltiples trabajos de ciencia de las ciudades [94]. Además de la información antes mencionada, incluye muchas variables ordenadas en las siguientes categorías:

- Geográficas: bioma, temperatura, tipo de suelo, precipitación, elevación, cuencas hidrográficas y clima.
- Socioeconómicas: población residente (años 1975, 1990, 2000 y 2015), superficie construida (años 1975, 1990, 2000 y 2015), clase de ingreso y grupo de desarrollo según el World Urbanization Prospect 2018 de las Naciones Unidas, emisiones nocturnas de luz (año 2015), producto interno bruto y el tiempo de traslado a la capital del país.
- Medioambientales: áreas verdes, concentración y emisión de partículas suspendidas (PM 2.5) y emisión de CO2.
- Sobre riesgo de desastres: exposición a inundaciones, estimación del riesgo de terremotos, exposición a marejadas ciclónicas e índice de olas de calor.
- Relacionadas con los objetivos del desarrollo sustentable: eficiencia de uso de suelo, porcentaje de espacios abiertos y porcentaje de la población que vive en zonas con alta presencia de áreas verdes.

Como muestra se seleccionaron algunas variables para las ciudades de Tokio, Ciudad de México y Nueva York que se muestran en la tabla 1.1. La primera elegida, UC_MN_LST, es muy importante en el contexto de nuestra reflexión. Se trata de la lista de unidades administrativas que se intersectan con el área urbana funcional;

Clave variable	Tokio	Ciudad de México	Nueva York
UC_NM_LST	Tokyo; Yokohama; Kawasaki; Saitama; Chiba; Setagaya; Nerima; Ota; Sagamihara; Edogawa; Adachi; Funabashi; Itabashi; Kawaguchi; Hachiōji; Suginami; Koto; Ichikawa; Katsushika; Machida; Fujisawa; Kashiwa; Shinagawa; Kita; Koshigaya; Tokorozawa; Kawagoe; Nak	Mexico City; Ecatepec; Nezahualcóyotl; Naucalpan; Tlalnepantla; Chimalhuacan; López Mateos; Cuautitlan Izcalli; Xico; Ixtapaluca; La Cumbre; El Hielo; San José Tejamanil	New York; Islip; Newark; Jersey City; Yonkers; Huntington; Paterson; Stamford; Elizabeth; New Brunswick
EL_AV_ALS	38.4455	2316.37	37.7714
P15	3.30287e+07	1.95596e+07	1.59507e+07
AREA	5318	2114	5384
B15	3664.91	1298.2	3678.08
INCM_CMI	HIC	UMIC	HIC
GDP15_SM	1.00776e+12	3.61621e+11	6.75538e+11
E_WR_T_14	15.5522	14.2108	13.031
E_BM_NM_LST	Temperate Broadleaf and Mixed Forests	Deserts and Xeric Shrublands	Temperate Broadleaf and Mixed Forests
E_KG_NM_LST	Mild temperate, fully humid, and Hot summer	Mild temperate with dry winter, and Warm summer	Mild temperate, fully humid, and Hot summer

Tabla 1.1: **Selección de variables contenidas en la GHS-UCDB para Tokio, Ciudad de México y Nueva York**. Se muestran la lista de nombres de las entidades administrativas UC_MN_LST, la elevación promedio en metros sobre el nivel del mar EL_AV_ALS, la población (2015) P15, la clase de ingresos según el World Urbanization Prospect 2018 ICM_CMI, el producto interno bruto (2015) GDP15_SM, la temperatura promedio anual (°C) E_WR_T_14, las clases de biomas E_BM_NM_LST y el clima (Köppen-Geiger) E_KG_NM_LST.

en el caso de la Ciudad de México, además de la entidad federativa homónima, se agregan 12 municipios pertenecientes a la entidad vecina, Estado de México. Estos municipios forman parte de lo que en otros contextos se conoce como Zona Metropolitana del Valle de México [113] que incluye 59 municipios de Estado de México y uno de Hidalgo, pero la definición de área urbana funcional es más restrictiva. También podemos ver el caso de Nueva York, que además de los cinco distritos que la componen (Bronx, Brooklyn, Manhattan, Queens y Staten Island), el área urbana funcional incorpora otras nueve unidades entre las que se encuentran la ciudad de Newark o la ciudad de Jersey, ubicada en el estado de Nueva Jersey. Finalmente, como otro ejemplo ilustrativo aunque nos sea menos familiar, se muestra el caso de Tokio al que, además de la unidad que le da nombre, se le suman otras 27 unidades.

Otras variables elegidas son la elevación promedio sobre el nivel del mar medida en metros (EL_AV_ALS), la población residente en el área urbana funcional en 2015 (P15) y la clase de ingresos del país al que pertenecen (ICM_CMI) según el World Urbanization Prospect 2018 de las Naciones Unidas que le da a México el nivel Up-

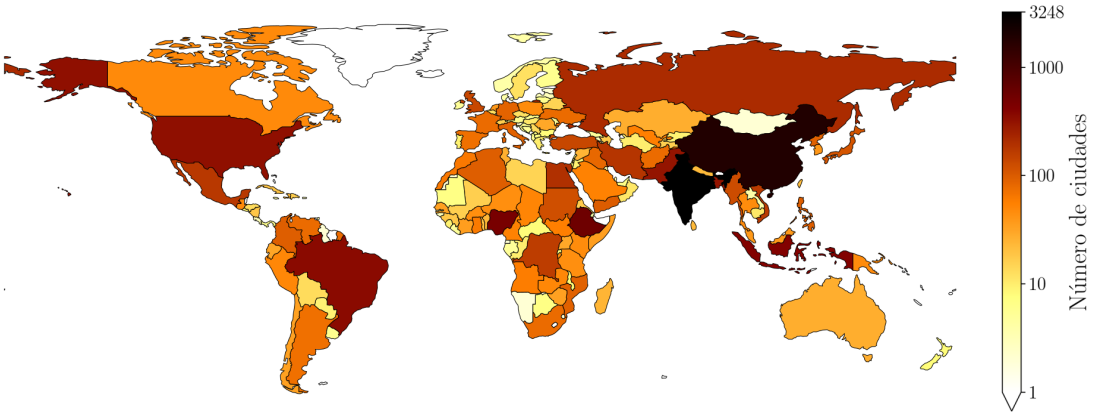


Figura 1.4: **Número de AUF por país.** En el mapa se muestra cada país y se indica el número de AUF que contiene con el color correspondiente al mapa de color que se muestra en la barra de la derecha. Se utilizó una escala logarítmica porque los valores cubren 3 órdenes de magnitud, desde 1 hasta 3248 para el caso de la India.

per Middle Income Country (UMIC) y a Japón y Estados Unidos el High Income Country (HIC). Se muestra también el producto interno bruto en 2015 (GDP15_SM), la temperatura promedio anual en grados centígrados (E_WR_T_14) y, finalmente, para dar cuenta del nivel de detalle y la profundidad de la información disponible, se muestran las clases de biomas que cruzan con el AUF (E_BM_NM_LST) y los climas de las regiones que cruzan con ella según la clasificación de Köppen-Geiger (E_KG_NM_LST). Se trata, según nosotros, de una fuente de información muy robusta para todo tipo de análisis.

1.6.3. Descripción general de la base de datos

En este apartado presentamos algunas propiedades generales de la base GHS-UCDB.

Las 13,135 AUF contenidas en la base de datos se ubican en 183 países. El país con más AUF es India con 3248, seguido por China con 1850. El tercer lugar, Etiopía, se aleja bastante de los primeros con apenas 557. Un análisis de la base de datos arroja que el número de AUF por continente es 7737 en Asia, 2805 en África, 1448 en América, 1059 en Europa y 86 en Oceanía. Esto implica que 58.9 % de las ciudades se encuentran en Asia y 21.4 % en África. Esto es consistente con lo mencionado en apartados anteriores respecto a que Asia y África habían experimentado el mayor crecimiento de ciudades en los últimos años. En la figura 1.4 se dibujaron todos los países del mundo y se señala con un gradiente de color el número de AUF que tiene cada uno. Dada la gran variedad de valores, se utilizó una escala logarítmica en un mapeo de colores que va del blanco para números pequeños al negro para números grandes, pasando por tonos amarillos, naranjas y cafés. Como complemento a esta figura, en la tabla 1.2 se muestran los primeros 15 países según el número de AUF que tiene cada uno. Entre ellos sólo hay 3 países americanos y uno europeo, reforzando la afirmación de que la población urbana se está concentrando cada vez más en África y Asia.

	País	Número de AUF	Continente
1	India	3248	Asia
2	China	1850	Asia
3	Ethiopia	557	Africa
4	Nigeria	483	Africa
5	Indonesia	393	Asia
6	Brazil	349	South America
7	United States	324	North America
8	Pakistan	301	Asia
9	Bangladesh	301	Asia
10	Russia	209	Europe
11	Egypt	190	Africa
12	Iran	182	Asia
13	Mexico	168	North America
14	Vietnam	163	Asia
15	Democratic Republic of the Congo	160	Africa

Tabla 1.2: **Número de AUF por país**. Los quince países con más AUF en su territorio. De ésos, 7 se encuentran en Asia, 4 en África, 3 en América y 1 en Europa.

Además, se hizo un análisis de las ciudades según su población en 2015. Las 20 AUF más pobladas se muestran en la tabla 1.3 mientras que en la figura 1.5 se vuelven a dibujar todas las áreas urbanas funcionales pero ahora distinguiendo cada una con un círculo cuya área es proporcional a la población de la ciudad y su color también indica la población según un mapeo de color arbitrario. Puede notarse que las ciudades más pobladas se encuentran en Asia, en donde destacan varios círculos de colores verdes y amarillos. Las poblaciones van de 50,000 (límite inferior establecido por la definición de AUF) hasta arriba de 40 millones para la ciudad china de Guangzhou (Cantón). Cairo, como la representante de las ciudades africanas, ocupa el puesto 11 con 19.7 millones de habitantes. Ciudad de México ocupa el primer lugar de las ciudades de América con una población de alrededor de 19.6 millones de habitantes en 2015. Le siguen la brasileña São Paulo con 19.1 millones, Nueva York con 15.9 millones y Los Ángeles con 14.3 millones. La única ciudad europea que figura en este ranking, ubicada en el lugar 20, es Moscú con una población de casi 14.1 millones en 2015. El dominio asiático es notable, no solo constituyen la mayoría de las áreas urbanas funcionales sino que allí se ubican las más grandes. Por otro lado, aunque en África también hay muchas AUF, no olvidar que ese continente ocupa el segundo lugar en ese rubro, éstas no son de las más grandes.

1.6.4. La base de datos como herramienta para la ciencia de las ciudades

Para mostrar la utilidad de la base de datos de centros urbanos de la capa global de asentamientos humanos (GSH-UCDB) en el estudio de las ciudades desde el punto de vista de los sistemas complejos, a continuación se presentan algunos resultados parciales sobre escalamiento, leyes de Zipf y comportamientos tipo ley de potencia que se pueden observar en el sistema global de megaciudades.

	Nombre	País	Región	Población (2015)
1	Guangzhou	China	Eastern Asia	4.058988e+07
2	Jakarta	Indonesia	South-Eastern Asia	3.631254e+07
3	Tokyo	Japan	Eastern Asia	3.302873e+07
4	Delhi [New Delhi]	India	South-Central Asia	2.665871e+07
5	Shanghai	China	Eastern Asia	2.447212e+07
6	Dhaka	Bangladesh	South-Central Asia	2.394235e+07
7	Mumbai	India	South-Central Asia	2.175588e+07
8	Quezon City [Manila]	Philippines	South-Eastern Asia	2.169114e+07
9	Kolkata	India	South-Central Asia	2.162029e+07
10	Seoul	South Korea	Eastern Asia	2.160084e+07
11	Cairo	Egypt	Northern Africa	1.973409e+07
12	Mexico City	Mexico	Central America	1.955956e+07
13	São Paulo	Brazil	South America	1.911434e+07
14	Beijing	China	Eastern Asia	1.798267e+07
15	New York	United States	Northern America	1.595067e+07
16	Osaka [Kyoto]	Japan	Eastern Asia	1.569280e+07
17	Bangkok	Thailand	South-Eastern Asia	1.473084e+07
18	Los Angeles	United States	Northern America	1.428172e+07
19	Istanbul	Turkey	Western Asia	1.411124e+07
20	Moscow	Russia	Eastern Europe	1.407736e+07

Tabla 1.3: **Áreas Urbanas Funcionales más pobladas**. Se muestran las 20 áreas urbanas funcionales más pobladas con el nombre del país en el que se ubican, la región y su población en 2015.

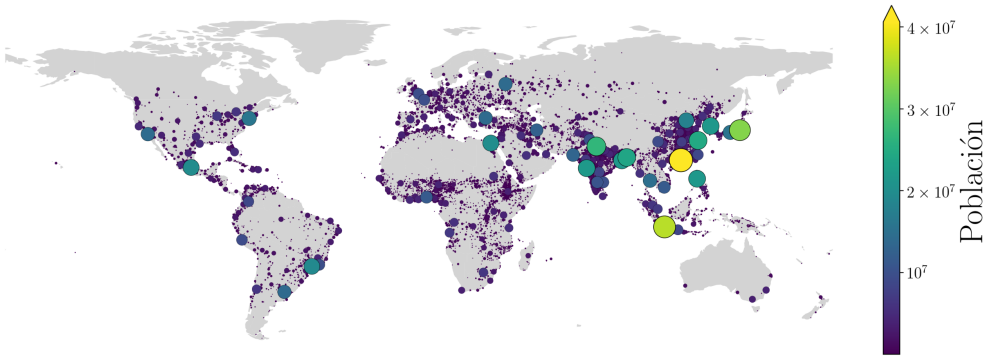


Figura 1.5: **Población de las AUF**. Cada ciudad está representada por un círculo cuyo centro se ubica en el centroide del área urbana funcional. El área del círculo es proporcional a la población y, para ayudar a distinguir, se coloreó cada círculo con un mapa de color que va del morado para valores bajos hasta el amarillo para valores altos. El AUF más poblada, fácilmente identificable por su color amarillo, es la correspondiente a la ciudad China de Guangzhou (Cantón) que tiene más de 40 millones de habitantes.

En primer lugar, se generaron gráficos de escalamiento para las ciudades de Estados Unidos y México. Se reproducen las gráficas presentadas por Ribeiro y Rybski [12] pero se sustituyó a Brasil con México. Se eligió analizar el área, el área construida y el Producto Interno Bruto (GDP por sus siglas en inglés), y los resultados se muestran en los páneles de la figura 1.6. Además, en cada caso se ajustó una ley de potencia con la finalidad de obtener el exponente y verificar que se encontrara en el régimen predicho. Efectivamente, el área total y el área de construcción obedecen una economía de escala, es decir, crecen más lentamente que una función lineal de la población.

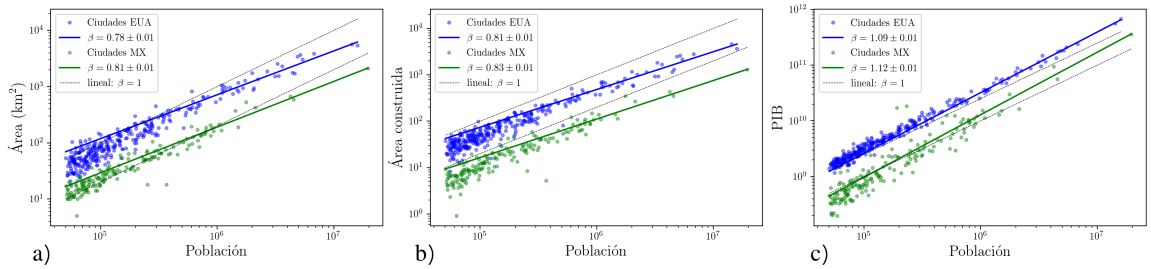


Figura 1.6: **Escalamiento urbano**. Se analiza, para las ciudades de Estados Unidos en azul y México en verde, (a) el área total del área urbana funcional como función de la población, (b) el área total de construcción como función de la población y (c) el producto interno bruto como función de la población. Además, en cada caso se muestra la tendencia lineal para poder comparar visualmente el comportamiento ajustado.

Por su parte, el Producto Interno Bruto exhibe un retorno creciente a escala o, de forma equivalente, crece de forma superlineal con el tamaño de la población. Esto puede interpretarse como que los individuos de la ciudad son más productivos mientras mayor sea la población de la ciudad. Los resultados son consistentes con lo que menciona la literatura expuesta anteriormente [2–4, 10, 12] y es solo una prueba del alcance y utilidad que tiene esta base de datos. Estos diagramas se realizaron con códigos propios con la ayuda de la paquetería Pandas para el análisis de datos con Python, y con ellos se pueden reproducir los resultados para las ciudades de cualquier país.

Siguiendo con la revisión de la base de datos a la luz de la ciencia de las ciudades, se generaron los diagramas de población contra el rango (diagramas de Zipf) de las ciudades de Brasil, Estados Unidos de América, India y China. Estos sistemas se caracterizan, según la bibliografía [2–4, 10] por exhibir un comportamiento tipo ley de Zipf, es decir, la variable decae como una potencia del rango. El resultado se muestra en la figura 1.7. Las poblaciones se normalizaron en cada caso para poder comparar los distintos sistemas, no hay que olvidar que los prefactores en escala logarítmica no hacen más que "subir" o "bajar" las gráficas sin variar su pendiente. Además, como referencia y sin un ajuste exhaustivo en cada caso, se agrega la ley de potencia con exponente igual a -1. Se puede verificar una vez más, salvo fluctuaciones, el comportamiento predicho por la teoría. En el caso de India y China, países con la mayor cantidad de centros urbanos, el comportamiento se extiende por más de dos órdenes de magnitud. Como las anteriores, esta gráfica se generó con códigos propios y se pueden agregar los sistemas de ciudades de cualquier país, siempre que tenga suficientes ciudades para verificar el comportamiento en más de un orden de magnitud y medio, esto es, que el país tenga un número en el orden de cientos de ciudades.

Finalmente, se analizó la distribución de probabilidad del área de las AUF, a nivel global y agrupándolas a nivel país. En la figura 1.8 se presentan los resultados. Se generaron las distribuciones de probabilidad de área para las AUF contenidas en los 40 países con más datos, mostradas en color, así como la distribución de probabilidad de área de todas las AUF de la base de datos para la que se usan puntos negros.

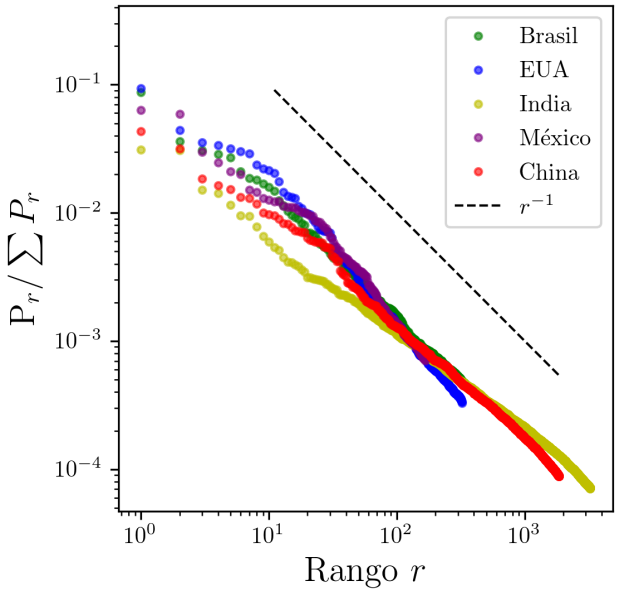


Figura 1.7: **Gráficos de Zipf para las poblaciones de las ciudades por país.** Se muestra el diagrama de población contra rango para los sistemas de ciudades de Brasil (verde), Estados Unidos de América (azul), china (rojo), India (amarillo) y México (púrpura). Además, la línea punteada equivale a la función r^{-1} para cada caso. Las poblaciones se normalizaron para poder comparar sistemas de distintos tamaños. El caso de México es consistente con el reportado por [114].

Para generar la distribución en cada caso se utilizó un bin logarítmico para agrupar los valores por área, procedimiento descrito por Clauset, Shalizi y Newman [115] para un análisis adecuado de las leyes de potencia. Por esta razón los puntos se ven equidistantes en la escala logarítmica. Como referencia se incluye la distribución de probabilidad de una ley de potencia con exponente -2 con la línea punteada. Sin hacer el cálculo exhaustivo, por no ser el objetivo de este trabajo, puede notarse que los valores para cada país fluctúa en torno al valor -2, mientras que la distribución considerando todos los datos se comporta de forma muy similar a la ley de potencia.

En conclusión, la información contenida en la base de datos GHS-UCDB da cuenta de muchas de las regularidades empíricas descritas en los primeros apartados de este capítulo, además tiene la ventaja de que los sistemas que define son consistentes con las propiedades que desde la ciencia de las ciudades se espera que tengan los sistemas; esto es, que los espacios urbanos sean la síntesis y la consecuencia de los flujos y las interacciones socioeconómicas de las personas. Por todo esto, esta base de datos se eligió para la delimitación de los sistemas de interés en los siguientes capítulos de este trabajo.

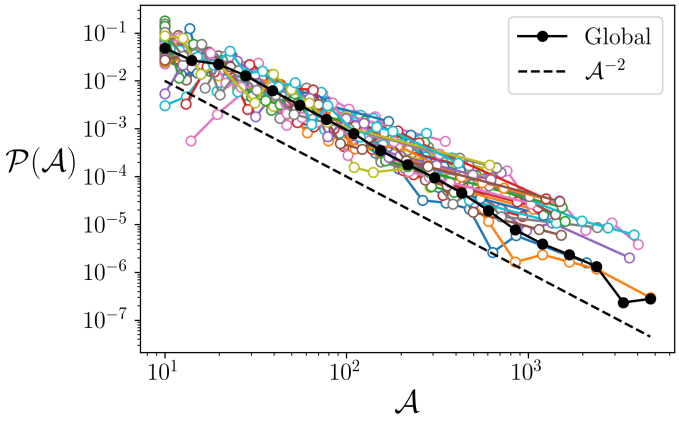


Figura 1.8: **Distribución de probabilidad de área.** Con distintos colores se muestran las distribuciones de probabilidad de área de las ciudades de 40 países. Cada gráfica representa la probabilidad $\mathcal{P}(\mathcal{A})$ de que, al elegir una ciudad aleatoria de cada sistema, ésta tenga un área $\mathcal{A}$. Los 40 países son aquellos que tienen más cantidad de ciudades. Con puntos negros se muestra la misma distribución de probabilidad, pero ahora tomando el sistema de todas las 13,135 áreas urbanas funcionales de la base de datos. Como referencia se grafica también, con línea negra discontinua, la ley de potencia con exponente -2.

1.7. Conclusiones

El estudio científico de las ciudades como sistemas complejos agrupa muchos y muy diversos trabajos y aportes. Una clasificación propuesta por Marta González y Diego Rybski [14] consiste en trabajos de movilidad urbana, escalamiento urbano, flujos en ciudades, análisis espaciales, ciudades y tecnologías de la información y evolución de las ciudades. Por su parte, Elsa Arcaute y José Ramasco [36] organizan los avances más recientes de la disciplina en los siguientes tópicos: leyes de escalamiento, morfología de las ciudades, organización jerárquica y movilidad urbana. Marc Barthelemy [4] afirma que estamos experimentando un momento crucial durante el cual se pueden someter a validación empírica los desarrollos puramente teóricos realizados en las últimas décadas. Según él, el objetivo de una ciencia de las ciudades se alcanzará cuando, considerando un caso específico, podamos decir qué sucederá y qué ingredientes es necesario introducir en un modelo para obtener información y predicciones más detalladas.

En esta introducción a la ciencia de las ciudades dejamos fuera muchos aspectos como la evolución de las ciudades o la modelación como un paso fundamental en la comprensión de las ciudades. Nos concentramos principalmente en aquellos aspectos que son relevantes para los resultados que se expondrán en el capítulo 4. Sin embargo, es importante concluir este capítulo insistiendo en que este trabajo pretende ubicarse como parte de esta corriente y aportar, aunque sea un poco, en uno de sus tópicos de interés: el estudio de los sistemas de ciudades y sus interacciones.

2 Ciencia de redes

En el capítulo anterior se mencionó que las ciudades pueden ser entendidas como redes de interacciones, de muchos tipos y entre varias clases de elementos. El estudio de las redes se hace mediante la teoría de grafos, un formalismo matemático muy adecuado para el estudio de sistemas, en especial de los sistemas complejos. Aprovechando la gran cantidad de datos disponibles en las últimas décadas para el estudio de cualquier fenómeno, se ha desarrollado un campo de estudio llamado Ciencia o Teoría de Redes [7, 8], que se enfoca en la aplicación de la teoría de grafos a múltiples sistemas y en particular a las redes complejas.

En este capítulo, se presentan las bases y definiciones de la teoría de grafos, así como los principales modelos de redes complejas utilizados en el estudio de sistemas urbanos. Además, se introduce la modularidad como medida de la estructura de comunidades en las redes y se presenta el algoritmo de detección de comunidades que se utilizará en los siguientes capítulos.

2.1. Introducción

En tiempos recientes, las redes han adquirido una gran relevancia en el análisis de sistemas complejos, puesto que mediante la representación de los elementos del sistema a través de nodos y sus interacciones por medio de enlaces, las propiedades matemáticas de las redes han permitido una mejor comprensión de estos sistemas. La teoría de grafos es el formalismo matemático que permite un análisis cuantitativo de la estructura de las redes. Esta teoría se enfoca en problemas como la conectividad, las características de cada nodo y la forma en que los nodos se agrupan en comunidades, entre otros. La teoría de grafos ha sido utilizada en campos como la física, la química y la biología, así como en problemas relevantes para las ciencias sociales. En las últimas décadas, se ha extendido su uso al análisis de sistemas complejos, tales como la estructura de la célula, el internet, las redes sociales, las redes de comunicaciones y las redes de colaboraciones científicas, entre otros. De esta manera, se puede hablar del estudio de redes complejas.

A continuación se desarrollarán las nociones básicas de la teoría de grafos o de redes. En este capítulo se le llamará de forma indistinta.

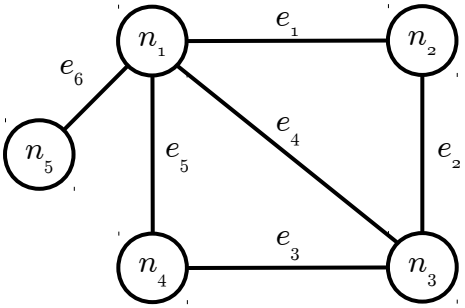


Figura 2.1: Red simple con $N = 5$ nodos y $L = 6$ enlaces. $\mathcal{V} = \{n_1, n_2, n_3, n_4, n_5\}$, $\mathcal{E} = \{e_1, e_2, e_3, e_4, e_5, e_6\}$

2.2. Conceptos básicos

2.2.1. Definición de una red

Una red, en su forma más simple, es una colección de puntos unidos por líneas. En la terminología de este campo, los puntos son llamados vértices o nodos mientras que las líneas son llamadas enlaces. Muchos objetos de interés en la física, la biología, ciencias sociales, etc., pueden ser pensados como redes y este enfoque se ha vuelto muy popular en los últimos años porque permite revelar propiedades de los sistemas que son difíciles de comprender si estudiamos a los elementos o las relaciones particulares por separado.

En su forma más general, una red $\mathcal{G}$ consiste de un conjunto $\mathcal{V}$ de N nodos y un conjunto $\mathcal{E}$ formado por L enlaces. En ocasiones, además de los conjuntos $\mathcal{V}$ y $\mathcal{E}$, la red puede contener información adicional asociada a los nodos y enlaces. Con esta definición, el concepto de red es muy amplio y permite describir muchos tipos de sistemas y situaciones variadas.

Con frecuencia, un primer paso para analizar la estructura de una red es hacer una imagen de ella. Visualizar la red puede ser muy útil en el análisis de los datos, permitiendo ver al instante características estructurales importantes que sería difícil extraer directamente de los datos. Es lo que haremos a continuación para seguir exponiendo las bases matemáticas de la teoría de grafos. Sin embargo, es importante aclarar que la visualización directa de las redes es útil únicamente cuando se habla de unos cuantos cientos de nodos y cuando la proporción de enlaces por nodo es relativamente baja; en otro caso es difícil lograr una visualización sistemática de las propiedades de las redes. Muchos de los sistemas de interés hoy en día alcanzan cientos o hasta miles de millones de elementos, lo que significa que su visualización directa no ayudaría mucho y se necesitan otras técnicas para el estudio sistemático de su estructura, como se verá más adelante.

Hay muchas formas de representar matemáticamente a una red. De la propia definición se derivaría una en la que, suponiendo que hay N nodos y se etiquetan con los

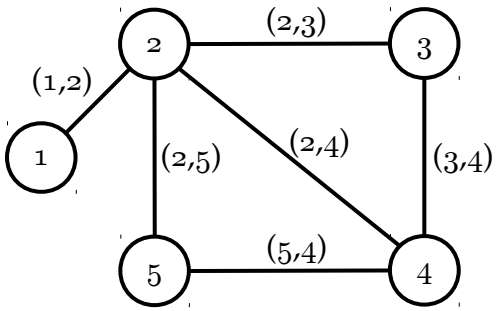


Figura 2.2: Red simple con $N = 5$ nodos y enlaces $\mathcal{E} = \{(1,2),(2,3),(2,4),(2,5),(3,4),(5,4)\}$.

números enteros 1, 2, ..., N. No importa cuál nodo tiene qué etiqueta, sólo que todas las etiquetas sean distintas para referirse a cada nodo sin ambigüedades. Con estas etiquetas, denotando al enlace entre los nodos i, j como (i, j), toda la estructura de la red podría reconstruirse a partir del número de nodos y la lista de los enlaces. Por ejemplo, el sistema de la figura 2.1 podría describirse ahora como:

$$N = 5 \; ; \; \mathcal{E} = \{(1,2),(2,3),(2,4),(2,5),(3,4),(5,4)\} \tag{2.1}$$

y el resultado de este etiquetado se muestra en la figura 2.2 Hay que notar que en esta definición no se menciona nada sobre si cada una de las parejas contenidas en $\mathcal{E}$ deba estar formada por nodos distintos, o si una pareja puede aparecer más de una vez en $\mathcal{E}$, si estas parejas se consideran ordenadas o su orden no es relevante, etc. Incluir o no cada una de estas condiciones dará lugar a redes con distintas propiedades, como veremos a más adelante La representación dada en 2.1 es conocida como lista de enlaces o lista de adyacencia. Es la más adecuada para guardar la información de las redes en computadora, pero no es la más conveniente para el desarrollo de muchas de las propiedades matemáticas de las redes que nos interesa exponer en este capítulo.

Otra forma de representar a una red es aprovechando el hecho de que existe una correspondencia biunívoca entre las redes de N nodos y las matrices cuadradas de dimensión $N \times N$ no negativas. Muchas de las propiedades de interés en la red están contenidas en las propiedades algebraicas del espacio $M_{N\times N}$. Partamos de una matriz 4×4 particular para exponer la regla de correspondencia con el espacio de las redes. Iniciemos con una matriz:

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0.2 & 0 \\ 1.5 & 0 & 2 & 0 \\ 1 & 1.8 & 0 & 0.5 \\ 0 & 0 & 0 & 2 \end{bmatrix} \tag{2.2}$$

y asociemos una red a partir de la siguiente regla:

- La entrada $[\mathbf{A}]_{ij} = a_{ij}$ proporciona el valor asociado al enlace que sale de i y llega a j.

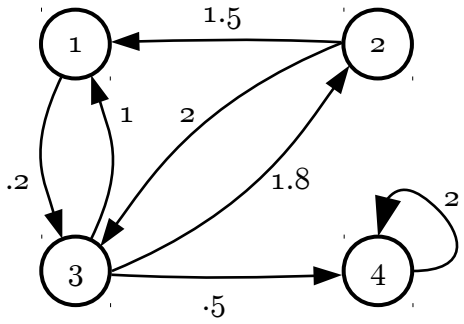


Figura 2.3: Red correspondiente a la matriz de la ecuación 2.2 según la regla de que la entrada $[\mathbf{A}]_{ij}$ corresponde al valor del enlace que sale de i y llega a j.

Con esta regla, a la matriz 2.2 le correspondería la red de la figura 2.3. Para verificarlo bastaría decir que, de la regla anterior, se sigue que la j-ésima entrada del i-ésimo vector fila de la matriz proporciona el valor del enlace que va de i a j. De que la primera fila de la matriz es $(0, 0, .2, 0)$ se sigue que del nodo 1 no salen enlaces mas que uno dirigido al nodo 3 con valor .2. De forma simétrica, la entrada i-ésima del j-ésimo vector columna proporciona el valor del enlace que llega a i desde j. En el ejemplo, observando que el vector correspondiente a la primera columna es $(0, 1.5, 1, 0)$, se verifica que al nodo 1 llegan enlaces provenientes del nodo 2 y 3 con valores 1.5 y 1 respectivamente.

Mediante esta regla se puede asociar una red a cualquier matriz cuadrada no negativa, y en sentido inverso, toda red, entendida como un conjunto de nodos y relaciones ponderadas entre todas las parejas nodos, tiene una matriz cuadrada no negativa asociada; a esta matriz asociada se le llama *matriz de adyacencia* de la red. Si entre dos nodos i, j hay un enlace e con valor asociado distinto de cero se dice que los nodos son adyacentes y, de la misma forma, se dice que el enlace es adyacente a los nodos.

Se define un camino en la red como una secuencia ordenada de nodos y enlaces con valor asociado distinto de cero, $\mathcal{P} = \{n_1, e_1, n_2, e_2, \ldots, n_k, e_k, n_{k+1}, ...\}$ que satisface que cada enlace es adyacente a los nodos anterior y siguiente en la secuencia. Particularmente, se define un camino entre los nodos n_i y n_j como un camino que inicia en n_i y termina en n_j. La condición de que cada enlace sea adyacente a los nodos anterior y siguiente, en la secuencia, permite sustituir las etiquetas e_k por las entradas de la matriz de adyacencia correspondientes, y reescribir el camino como:

$$\mathcal{P} = \left\{n_1, a_{n_1,n_2}, n_2, a_{n_2,n_3}, n_3, \ldots, n_{k-1}, a_{n_{k-1},n_k}, n_k, \ldots\right\} \tag{2.3}$$

La existencia de caminos entre dos nodos de una red tiene una consecuencia importante en las propiedades de la matriz de adyacencia: si existe un camino entre los nodos i y j podemos afirmar que hay una secuencia de enlaces con valores distintos de cero que permiten llegar de i a j, y en términos de la matriz de adyacencia, que

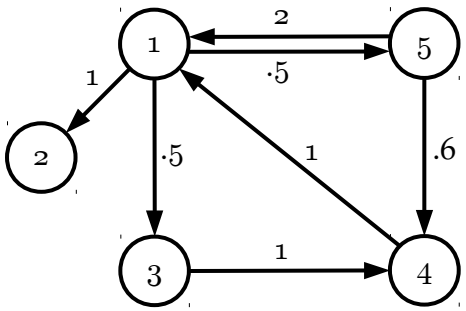


Figura 2.4: Red correspondiente a la matriz 2.6

hay un conjunto de entradas de ésta que satisfacen:

$$a_{i,n_2}a_{n_2,n_3}...a_{n_{l-1}n_l}a_{n_l,j} > 0 \tag{2.4}$$

para algunos $n_2,\ n_3,\ \ldots,\ n_l$, y si se quieren considerar todos los caminos en la red, habría que hacer la misma operación con todas las combinaciones posibles para los $l-1$ nodos intermedios (etiquetados para este fin, sin pérdida de generalidad, como $n_2, n_3, \ldots, n_l$) y verificar si el producto 2.4 es distinto de cero. Dado que si el camino existe, el producto es distinto de cero, y si no existe, el producto es igual a cero, una forma de considerar todas los caminos entre los nodos i y j es sumando sobre todos los nodos de la red para cada índice $n_2, n_3, \ldots, n_l$. Esto, además, sería una forma de sumar las contribuciones de todos los caminos de longitud l que conectan a los nodos i, j. Esta suma adquiere la conocida forma:

$$\sum_{n_2=1}^{N}\sum_{n_3=1}^{N}...\sum_{n_l=1}^{N} a_{i,n_2}a_{n_1,n_2}...a_{n_{l-1}n_l}a_{n_l,j} = \sum_{n_2,n_3,...,n_l} a_{i,n_2}a_{n_1,n_2}...a_{n_{l-1}n_l}a_{n_l,j} \tag{2.5}$$

que no es otra que la expresión de la entrada i, j de la potencia l de la matriz de adyacencia

$$[\mathbf{A}^l]_{i,j}$$

Para ejemplificar, tomemos el caso más sencillo $l = 2$ para los nodos 1 y 4 de la red mostrada en la figura 2.4 que corresponde a la siguiente matriz de adyacencia:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 2 & 0 \\ 2 & 0 & 0 & 0 & .6 \end{bmatrix} \tag{2.6}$$

Se trataría de buscar todos los caminos de longitud 2 que van del nodo 1 al 4. Un reacomodo de los nodos que no modifica la estructura de la red (fig.2.5) permite ver que la suma sobre todos los nodos proporciona los caminos entre 1 y 4, considerando los valores asociados a los enlaces, es decir:

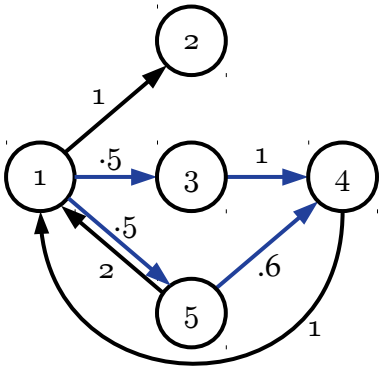


Figura 2.5: Reacomodo de los nodos de la red de 2.4 en la que se pueden ver los caminos que van del nodo 1 al nodo 4, a decir, $\{(1,5),(5,4)\}$ y $\{(1,3),(3,4)\}$

$$\begin{aligned}\sum_{n=1}^{5} a_{1,n}a_{n,4} &= a_{1,1}a_{1,4} + a_{1,2}a_{2,4} + a_{1,3}a_{3,4} + a_{1,4}a_{4,4} + a_{1,5}a_{5,4} \\ &= (0)(0) + (.5)(0) + (.5)(1) + (0)(0) + (.5)(.6) \\ &= .5 + .3 \\ &= .8\end{aligned}$$

Los dos caminos que conectan a 1 y 4, ponderando cada uno con los valores de los enlaces que contienen, da un total de .8. En el caso particular de que los pesos de los enlaces funcionen como probabilidades de salida, es decir, que los pesos de los enlaces que salen de cada nodo sumen 1 (lo cual significaría que la red describe una cadena de Markov), esta cantidad es muy importante pues proporciona la probabilidad de llegar de uno nodo a otro, con un número determinado de pasos.

Otro caso de interés es cuando todos los enlaces tienen un valor booleano, esto es, 0 o 1 (cero si no hay enlace y 1 si sí lo hay); en este caso, la cantidad $[\mathbf{A}^l]_{ij}$ es igual al número de caminos de longitud l que conectan a i con j.

$$[\mathbf{A}^l]_{ij} = \#\text{ de caminos de longitud } l \text{ entre } i \text{ y } j \tag{2.7}$$

En la misma línea de los dos últimos párrafos, continuemos estableciendo condiciones sobre las redes y observando qué consecuencias tienen tales condiciones en la matriz de adyacencia.

Hay muchos sistemas en los que la relación entre dos elementos no tiene una dirección característica, o lo que es lo mismo, que un enlace (i,j) entre los nodos i,j, implica la relación de i con j y también de j con i. Este tipo de enlaces describen relaciones simétricas, por ejemplo coincidencia de personas en un lugar (si la persona i coincidió con j en algún lugar, j coincidió con i en el mismo lugar) o amistades en Facebook (si la cuenta i tiene una amistad con j, dado que la amistad se establece cuando ambas cuentas están de acuerdo, entonces j tiene una amistad con i). Esto tiene como consecuencia que la matriz de adyacencia sea simétrica, dado que

$$[\mathbf{A}]_{ij} = a_{ij} = a_{ji} = [\mathbf{A}]_{ji} \tag{2.8}$$

Aunque está implícito en lo que se ha dicho, vale la pena mencionar que si un enlace entre i y j no necesariamente implica un enlace entre j e i, además de que la matriz de adyacencia no es simétrica, se tienen enlaces con una dirección dada. Ejemplos de sistemas descritos por este tipo de enlaces podrían ser las transferencias de dinero entre dos cuentas bancarias, las importaciones o exportaciones entre países o la relación de "seguir" entre cuentas de Twitter; en los primeros dos casos, los montos de dinero y las exportaciones/importación son en general diferentes, en el último caso, que una cuenta i siga a otra j no implica que j también siga a i.

Hay sistemas en los que la relación de un elemento consigo mismo no es concebible, o no está bien definida. Esto implica que los elementos en la diagonal de la matriz de adyacencia sean todos cero,

$$[\mathbf{A}]_{ii} = a_{ii} = 0 \tag{2.9}$$

Como se mencionó más arriba, hay sistemas para los que las relaciones entre elementos no admiten ser descritos por números reales sino valores binarios, presencia o ausencia, 1 ó 0. Es el caso de conocimiento entre personas (i puede conocer, o no, a j) o si la pregunta es si entre dos teléfonos celulares ha habido comunicación (sí la ha habido, o no, no hay más posibilidades). En estas situaciones, la matriz de adyacencia está formada únicamente por ceros y unos:

$$[\mathbf{A}]_{ij} = \begin{cases} 1 & \text{si hay relación entre } i \text{ y } j \\ 0 & \text{si no} \end{cases} \tag{2.10}$$

En oposición a este tipo de enlaces, están aquellos que describen relaciones que admiten magnitudes; es el caso general del que partimos y se pueden describir los mismos fenómenos pero si se incorpora una magnitud a la relación: se puede conocer mucho o poco a una persona, entre dos teléfonos celulares pudo haber muchos o pocos minutos de llamadas, por ejemplo. En tal caso se dice que las redes tienen pesos o son redes ponderadas. Ya se dijo que los pesos pueden ser probabilidades en el caso en que la red describa un proceso estocástico, pero otra posibilidad es que los valores que puede tomar la relación entre dos nodos sea un valor discreto. El número de caminos entre dos poblados, el número de mensajes de Whatsapp entre dos números telefónicos, el número de encuentros que han tenido dos personas en una semana. En tal caso se habla de redes con multienlaces y la matriz de adyacencia debe estar formada por puros números enteros.

Con esto habríamos agotado todos los tipos de redes de interés para describir una gran variedad de sistemas, con sus consecuencias sobre la matriz de adyacencia. Muchos fenómenos de interés pueden ser descritos por redes llamadas simples, que son aquellas en los que los enlaces no tienen dirección (redes no dirigidas), entre dos nodos hay uno o ningún enlace (redes sin pesos) y no se permiten auto-enlaces; en tal caso, las matrices de adyacencia son simétricas, formada por ceros y unos, y con puros ceros en la diagonal. El caso de mayor interés para este estudio, que se presenta en el capítulo 4, es el de una red con pesos que mediante un proceso de umbralización (Batty) se convierte en una red simple. Por esto, vale la pena establecer los conceptos correspondientes a una red general (dirigida, con pesos, con multienlaces, auto-enlaces,

etc.) y, cuando sea necesario, aclarar que algunas propiedades están definidas sólo para redes simples. Otras posibilidades no exploradas en esta investigación incluyen relaciones definidas por un signo (por ejemplo, relaciones mutualistas o parasitarias entre especies de un ecosistema) o, recientemente, se están incorporando al análisis mediante redes, sistemas para los que se pueda establecer una relación entre más de dos nodos; a veces se denomina como hiperenlace a la descripción de estas relaciones, e hipergrafos a las redes formadas por hiperenlaces.

2.2.2. Grado, distancias y centralidades

Continuaremos exponiendo más propiedades de las redes que dan información sobre los sistemas de estudio y sus elementos. Se define el *grado* de un nodo como el número de enlaces que tiene. Si la red es simple es un valor único y si es una red dirigida se debe distinguir entre el grado de entrada y de salida, como el número de enlaces que llegan o salen del nodo, respectivamente. Esta propiedad también se puede obtener a partir de la matriz de adyacencia, recordando que las filas proporcionan los enlaces que salen de los nodos y las columnas los que llegan. Así el grado de salida del nodo i, denotado por k_i^{out}, se obtiene mediante:

$$k_i^{out} = \sum_j \mathbf{A}_{ij}$$

y el de entrada,

$$k_j^{in} = \sum_i \mathbf{A}_{ij}$$

Cuando la red es simple, ambos valores son iguales y el grado del nodo i se denota simplemente k_i

$$k_i = \sum_i A_{ij} = \sum_j A_{ij} \tag{2.11}$$

El grado de un nodo es una primera medida sobre el papel que tiene un nodo en el sistema. Si se considera importantes o influyentes a los nodos que concentran muchos enlaces, el grado de un nodo es una medida de su importancia o su centralidad. Del grado de los nodos se puede obtener mucha información de la red. Por ejemplo, si se toma el promedio del grado sobre todos los nodos

$$\langle k \rangle = \frac{1}{N} \sum_i k_i$$

se obtiene una descripción de qué tan densa es la red, en el sentido de cuántos enlaces por nodo hay. Esta medida es más clara si en ella se sustituye k_i usando la ecuación 2.11, para redes simples,

$$\langle k \rangle = \frac{1}{N} \sum_i k_i = \frac{1}{N} \sum_i \sum_j \mathbf{A}_{ij} = \frac{2L}{N} \tag{2.12}$$

La última igualdad se obtiene considerando que sumar todos los elementos de la matriz de adyacencia de una red simple significa contar todos los enlaces de la red

dos veces; una vez como enlace llegada y otra como enlace de salida en cada nodo. En el caso de redes dirigidas desaparece el número 2. La ecuación 2.12 se comprende más claramente si se toman sus valores extremos. Una red sin enlaces tiene grado promedio nulo, mientras que una red con todos los nodos conectados entre sí tiene $L_{max} = N(N-1)/2$ que, sustituyendo en 2.12 da como resultado $\langle k \rangle = N - 1$. En tal caso, el número promedio de conexiones de cada nodo es $N - 1$, es decir, cada nodo está conectado con los $N - 1$ nodos restantes. De modo que el grado promedio es cercano a cero si el número de enlaces en bajo y cercano a $N-1$ cuando el número de enlaces de la red es alto. Como observa Barabasi [7], la mayoría de las redes que describen fenómenos reales son "sparse" (traducido como escasas o ralas, en español), que quiere decir que el número de enlaces de la red es mucho menor al número de parejas de nodos, es decir, que $L \ll L_{max}$ y por lo tanto

$$\langle k \rangle \ll N - 1 \tag{2.13}$$

La otra propiedad sistémica que se puede obtener de los grados de los nodos es la distribución de probabilidad de grado, o *distribución de grado* p_k, que se define como la probabilidad de que, al elegir un nodo de la red de forma aleatoria, éste tenga grado k. Esencialmente es un histograma de frecuencias de los valores del k en la red, normalizado dividendo cada valor entre el número total de nodos. Como se verá, la forma de la distribución de grado puede decir mucho sobre las propiedades del sistema.

Más arriba se definió la longitud de un camino como el número de enlaces que contiene. Un camino geodésico entre dos nodos es el camino más corto entre ellos, es decir, un camino tal que no existe otro más corto entre los dos nodos [8]. La distancia entre dos nodos i, j se define como la longitud de un camino geodésico entre ellos. En términos de la ecuación 2.7 sería el mínimo valor de l que hace a la entrada i, j de la potencia l de la matriz de adyacencia distinta de cero:

$$d(i,j) = \text{mín}\left\{ l : \left[\mathbf{A}^l\right]_{ij} > 0 \right\} \tag{2.14}$$

Si no hay un camino que una a los dos nodos, este mínimo valor no existe y por lo tanto la distancia no está definida.

El diámetro de la red es la máxima distancia entre cualesquiera dos nodos de la red, o lo que es lo mismo, la longitud del más grande camino más corto entre los nodos de la red

$$\text{diam}(\mathcal{G}) = \text{máx}\left\{ d(i,j) | \, i,j \in \mathcal{V} \right\} \tag{2.15}$$

En redes simples, el diámetro sólo está definido si existe un camino entre cualquier pareja de nodos de la red, lo que además define a una red conexa.

El diámetro es una medida muy útil de la estructura de la red. Existe entre las redes correspondientes a sistemas sociales y naturales la tendencia de que su diámetro es relativamente pequeño, comparado con su tamaño; esta tendencia se conoce como propiedad de mundo pequeño y quedará más clara cuando se describa el modelo de red aleatoria. Otra medida útil relacionada con la distancia entre los nodos de una red

es la distancia promedio entre los nodos de la red, esto es, el promedio de la longitud de las geodésicas entre todas las parejas de nodos de la red.

$$\langle d \rangle = \frac{1}{N(N-1)} \sum_{i,j} d(i,j) \tag{2.16}$$

Existen otras medidas de los nodos que, de forma similar al grado, pueden dar información sobre el papel que juegan éstos en la red. Se decía que el grado en sí mismo es una forma de jerarquizar a los nodos de una red según el número de enlaces que concentran; nodos con muchos enlaces serían importantes mientras que nodos con pocos enlaces son menos importantes. Esto lleva a definir la centralidad de grado de los nodos que, básicamente, es el grado de los nodos pero, para que tenga un valor independiente del tamaño de la red, se normaliza dividiéndolo entre el valor máximo posible del grado de un nodo [64]. Además de la de grado, existen muchas otras medidas de centralidad que permiten jerarquizar a los nodos dependiendo de su importancia en la red como función de alguna propiedad útil. Por ejemplo, se puede medir la importancia de un nodo a partir de su cercanía al resto de los nodos, medida como el inverso de la suma de las distancias entre dicho nodo y todos los demás. Esto define la *centralidad de cercanía* del nodo i:

$$C_c(i) = \frac{1}{\sum_{j=i}^{N} d(i,j)}$$

que, como en el caso de la de grado, también se puede dividir entre el máximo valor posible que puede tomar que es 1/(N-1) [64] lo que significaría que

$$C_c^*(i) = \frac{N-1}{\sum_{j=i}^{N} d(i,j)} \tag{2.17}$$

Otra medida de centralidad es la *centralidad de intermediación* (betweenness centrality) que consiste en contar el número de geodésicas de la red que pasan por determinado nodo. En ese sentido, le asocia importancia a un nodo en función de qué tanto se ubica en el camino más corto entre otros nodos, volviéndolo relevante en términos de mediación en los flujos de información o materia sobre la red. Si se define $g_i(u,v)$ como el número de geodésicas entre los nodos u y v que pasan por i, y $g(u,v)$ el número total de geodésicas entre los nodos u y v, la centralidad de intermediación se calcula mediante:

$$C_b(i) = \sum_{u,v} \frac{g_i(u,v)}{g(u,v)}$$

que se puede normalizar dividiendo entre el máximo valor que puede tomar para una red con N nodos [8]

$$C_b^*(i) = \frac{1}{N^2 - N + 1} \sum_{u,v} \frac{g_i(u,v)}{g(u,v)} \tag{2.18}$$

Hasta aquí se han mencionado tres medidas de centralidad que miden la relevancia o importancia de un nodo en la red a partir de la función estructural que desempeña:

si concentra enlaces, si es más cercano en promedio al resto o si es intermediador entre las parejas de nodos de la red. Sin embargo, este es un pequeño conjunto de las posibles medidas de centralidad que se pueden definir. Para poner esto en contexto, en la documentanción de NetworkX [1], paquetería de python para el análisis de redes, se da cuenta de 42 medidas de centralidad disponibles como métodos de la librería; sea cual sea la función estructural que se quiera medir en una red, dados los objetivos o la naturaleza específica de un estudio, se puede definir una medida de centralidad adecuada.

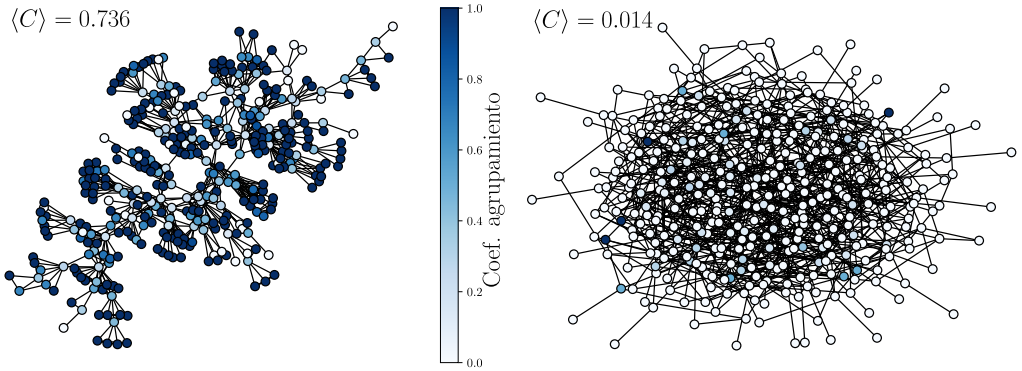
2.2.3. Coeficiente de agrupamiento

La vecindad de un nodo es el conjunto de nodos con los que está enlazado, también llamados sus nodos adyacentes o vecinos. A partir de ella se puede definir el *coeficiente de agrupamiento* de un nodo, a veces llamado coeficiente de acumulación o en inglés clustering coefficient, como la fracción que representan los enlaces entre los nodos de su vecindad respecto al número máximo de enlaces que podría haber entre ellos; dicho de otro modo, se define el coeficiente de agrupamiento de i como la división del número de enlaces entre nodos vecinos de i y las parejas entre esos mismos nodos vecinos. Si el grado del nodo i es k_i, el número de parejas que se pueden formar es $k_i(k_i - 1)/2$ y el coeficiente de agrupamiento es [7]:

$$C_i = \frac{2L_i}{k_i(k_i - 1)} \tag{2.19}$$

con L_i el número de enlaces entre los nodos de la vecindad de i. Como su nombre lo indica, el coeficiente de agrupamiento o de acumulación es una medida de cuánto se acumulan los enlaces en un subconjunto de la red dado por la vecindad de un nodo. Si el coeficiente es alto para un nodo, la mayoría de sus vecinos son, a su vez, vecinos entre sí; por el contrario, si el coeficiente es bajo para un nodo, sus vecinos casi no son vecinos entre sí, generando estructuras tipo "estrella" como caso extremo, en la que no hay enlaces entre los nodos de la vecindad. Si un nodo tiene un solo vecino, es decir grado $k = 1$, el coeficiente se define como cero. El coeficiente de agrupamiento sirve para identificar subconjuntos de nodos de la red entre los que hay una mayor densidad de enlaces, formando estructuras conocidas como cúmulos. En la figura se muestra una red en la que el coeficiente de agrupamiento se representa con el color de los nodos: puede observarse que hay grupos de nodos claros, que dan cuenta de la presencia de cúmulos, mientras que hay partes de la red con baja presencia de cúmulos. El coeficiente de agrupamiento promedio de una red mide la tendencia de los nodos de la red a formar cúmulos. Nuevamente se encuentra una regularidad entre las redes sociales y biológicas que es su tendencia a formar cúmulos, lo cual produce coeficientes de agrupamiento relativamente grandes [7]. De nuevo, el significado de "un valor alto" quedará más claro cuando se compare con el coeficiente de agrupamiento predicho por el modelo de red aleatoria en la siguiente sección.

[1] https://networkx.org/documentation/stable/reference/algorithms/centrality.html



a) Red real b) Red aleatoria

Figura 2.6: Coeficiente de agrupamiento de las redes. a) Se visualiza el coeficiente de agrupamiento de una red real de colaboración en ciencia de redes [116] disponible en `networkrepository.com`. El coeficiente de agrupamiento de cada nodo se representa mediante un color de una paleta arbitraria, y el coeficiente de agrupamiento promedio de esta red es 0.736. b) Se generó una red aleatoria, como se describe en el siguiente apartado, con las mismas características que la red real, es decir, mismo número de nodos y mismo número de enlaces, pero el mecanismo con el que se asocian los enlaces es aleatorio. El coeficiente de agrupamiento de esta red es mucho menor, de 0.014.

Hasta aquí se han desarrollado y mencionado las bases matemáticas para estudiar y caracterizar sistemas a partir de su descripción como redes. En la búsqueda de modelos que recojan las propiedades de las redes reales, la ciencia de redes, o ciencia de redes complejas [7, 8], parte de tres modelos principales. En el siguiente apartado se describen estos tres modelos importantes por su relevancia histórica y porque de cada uno se obtienen propiedades presentes en las redes reales.

2.3. Modelos de redes complejas

En esta sección se desarrollan tres de los modelos más relevantes en el estudio de las redes complejas que se utilizan para reproducir las características de fenómenos sociales, naturales, digitales, etc.

2.3.1. Modelo de redes aleatorias de Erdös y Rényi

El modelo de Erdös-Rényi de redes aleatorias ayuda a conocer las propiedades de redes que se construyen mediante algún mecanismo aleatorio simple. En su versión estándar, el modelo se basa en dos parámetros fundamentales: el número de nodos en la red, N, y la probabilidad de conexión entre dos nodos cualesquiera, p. En este modelo, se construye una red aleatoria tomando los N nodos y conectando cada pareja con una probabilidad p. Cada conexión se establece de forma independiente y con la misma probabilidad, lo que significa que es posible que un nodo no tenga ninguna conexión en absoluto. La teoría de Erdös-Rényi es muy útil porque nos permite estudiar las propiedades estadísticas de las redes aleatorias que se explicarán

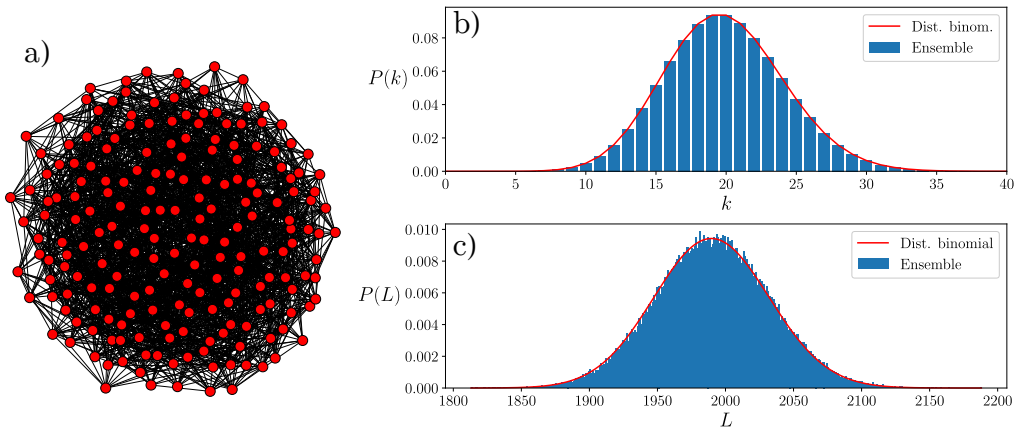


Figura 2.7: Propiedades de la red aleatoria. a) realización de red aleatoria con parámetros $N = 200$, $p = .1$. b) Distribución P_L obtenida mediante un ensamble de 100,000 realizaciones y con 2.20. c) Distribución P_k obtenida mediante un ensemble de 100,000 realizaciones y con 2.22. Los códigos para obtener esta imagen están disponibles en `https://curso-redes-f-ciencias-unam.github.io/ciencia-de-redes/index.html`

brevemente a continuación.

Dada una red de Erdös-Rényi con N nodos y probabilidad p, $\mathcal{G}_{ER}(N,p)$, la primera pregunta pertinente es cuántos enlaces tendrá. Para responder hay que considerar que, por tratarse de un modelo aleatorio, cada red generada mediante este mecanismo es una realización; las propiedades del modelo son en realidad las propiedades en un *ensemble* de realizaciones con los mismos parámetros N y p, de modo que lo procedente es obtener una distribución de probabilidad de L, el número de enlaces de la red. Esta distribución es la de una variable aleatoria dada por la probabilidad de ocurrencia de L enlaces con probabilidad p, p^L, junto con la probabilidad de no ocurrencia de $N(N-1)/2 - L$ enlaces con probabilidad $1-p$, $(1-p)^{N(N-1)/2-L}$, resultado que puede ocurrir tantas veces como el número de combinaciones de $N(N-1)/2$ en L; estos tres factores dan lugar a una distribución binomial de la forma

$$P_L = \left(\begin{array}{c} \frac{N(N-1)}{2} \\ L \end{array} \right) p^L (1-p)^{\frac{N(N-1)}{2}-L} \tag{2.20}$$

cuyo valor esperado es bien conocido

$$\mathrm{E}\left[L\right] = \sum_L L P_L = p\frac{N(N-1)}{2} = pL_{\text{máx}} \tag{2.21}$$

Otra información relevante del modelo es la distribución de grado esperada[2], y el grado promedio esperado, para el modelo $\mathcal{G}_{ER}(N,p)$ de Erdös-Rényi. Para obtenerlos debemos preguntarlos la probabilidad de que un nodo de una realización tenga grado

[2]se habla de "distribución de grado esperada" debido a que, estrictamente, cada realización del modelo tiene una distribución de grado, y para obtener una distribución de grado característica del modelo habría que promediar la distribución de grado de un ensamble.

k, P_k. Con un razonamiento completamente análogo al utilizado para p_L, se obtiene que esta probabilidad es el resultado de multiplicar tres factores: la probabilidad de que el nodo tenga k enlaces con probabilidad p, la probabilidad de que tenga $N-1-k$ no-enlaces con probabilidad $1-p$, y el número de combinaciones en las que esto puede ocurrir, a decir, las combinaciones de $N-1$ en k. Vuelve a ser una distribución binomial de la forma:

$$P_k = \left(\begin{array}{c} N-1 \\ k \end{array} \right) p^k (1-p)^{N-1-k} \tag{2.22}$$

Con esta distribución de grado se puede obtener el valor esperado del grado promedio, nuevamente, utilizando las propiedades de la distribución binomial

$$\mathrm{E}\left[\langle k \rangle\right] = \sum_k k P_k = p(N-1) \tag{2.23}$$

Finalmente, se puede obtener el coeficiente de agrupamiento esperado de un nodo de una red aleatoria de Erdös-Rényi. Una forma de interpretar las ecuaciones 2.21 y 2.23 es que, si la probabilidad de que un enlace entre dos nodos ocurra es p, el número esperado de enlaces en cualquier contexto es una fracción p del máximo posible; si el máximo número de enlaces en la red es $N(N-1)/2$, el valor esperado es p veces ese valor; si el máximo número de enlaces que puede tener un nodo particular es $N-1$, el valor esperado del grado de un nodo es p ese valor. Para calcular el coeficiente de agrupamiento partimos de que un nodo i tiene grado k_i y debemos obtener L_i. Al tratarse de una realización, únicamente podemos obtener su valor esperado $\mathrm{E}\left[L_i\right]$ que, siguiendo el razonamiento anterior, será p veces el valor máximo que pueda tomar. L_i es el número de enlaces entre los k_i vecinos de i así que su valor máximo son las $k_i(k_i-1)/2$ parejas de nodos que forman dichos vecinos. Con esto podemos afirmar que $\mathrm{E}[L_i] = pk_i(k_i-1)/2$ y, sustituyendo en 2.19, se obtiene

$$\begin{aligned} \mathrm{E}\left[C_i\right] &= \frac{2\mathrm{E}\left[L_i\right]}{k_i(k_i-1)} \\ &= p\frac{k_i(k_i-1)}{2}\frac{2}{k_i(k_i-1)} \\ &= p \end{aligned} \tag{2.24}$$

que es independiente del nodo i. Esto lleva a que el agrupamiento promedio esperado de la red sea

$$\mathrm{E}\left[\langle C \rangle\right] = p \tag{2.25}$$

Se había dicho antes que existe una tendencia entre las redes reales de tener un alto coeficiente de agrupamiento y un grado bajo (ecuación 2.13). Combinando las ecuaciones 2.25 y 2.23 se tiene que, para la red Erdös-Rényi,

$$\mathrm{E}\left[\langle C \rangle\right] = \frac{\mathrm{E}\left[\langle k \rangle\right]}{N-1} \tag{2.26}$$

de donde se obtiene que, si se quiere una red aleatoria "rala", esta tendrá un coeficiente de agrupamiento bajo, y viceversa, si se quiere un coeficiente de agrupamiento alto

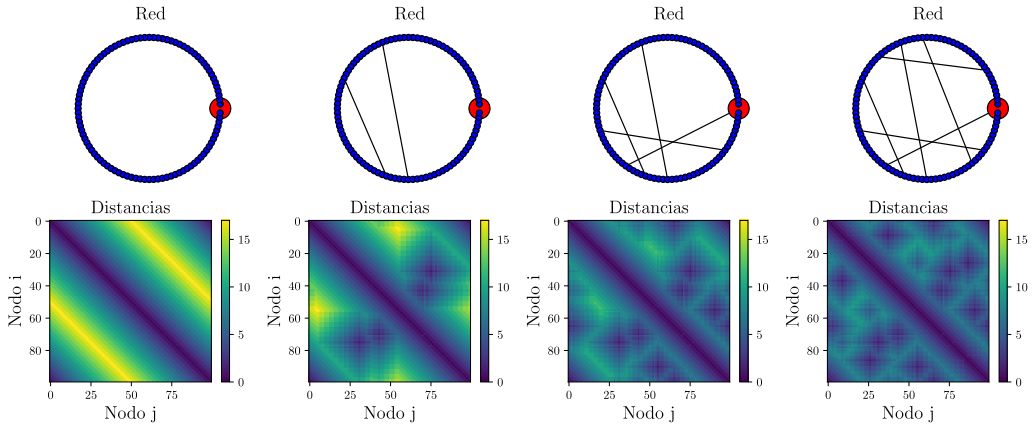


Figura 2.8: Primera aproximación al modelo de Watts-Strogatz. Se inicia con un anillo de N = 100 nodos y se agregan líneas para que cada nodo tenga 6 vecinos cercanos. Se modifican 2, 4 y 6 enlaces del anillo original. En esta implementación se elige de forma aleatoria un enlace y se sustituye por otro asignado, también de forma aleatoria. En cada caso se muestra la matriz de distancias entre los nodos; la estructura inicial de la matriz de adyacencia, cuando la red es un anillo, se debe a que cada nodo es cercano a su vecindad y lejano a los nodos del extremo opuesto el anillo, el nodo 0 se distingue en rojo para apoyar a la interpretación. Las distancias entre nodos van de 0 a 16.67, la sexta parte de 100. Conforme se van cambiando enlaces, en este caso se cambiaron de dos en dos, la estructura de la matriz de distancias se modifica por completo, de modo que con solo 6 enlaces modificados predominan las distancias menores a 10. Los códigos para obtener esta imagen están disponibles en `https://curso-redes-f-ciencias-unam.github.io/ciencia-de-redes/index.html`

el grado promedio debe ser alto. Esto hace al modelo de Erdös-Rényi un mal modelo para las redes reales. Esto sugiere que los mecanismos subyacentes al establecimiento de los enlaces en las redes reales distan mucho de ser aleatorios. Esto es lo que lleva a los científicos de las redes a buscar otros modelos.

El modelo de Erdös-Rényi sirve para cuantificar la llamada "propiedad de mundo pequeño" que establece que las distancias características de una red, el diámetro y la distancia promedio, son mucho menores que el tamaño de la red. Para el caso de la red aleatoria, las distancias características cumplen que [8]

$$d \sim \frac{\log(N)}{\log(\langle k \rangle)} \tag{2.27}$$

que significa que las redes aleatorias sí recogen la propiedad de mundo pequeño tan común en los sistemas reales.

2.3.2. Modelo de redes de mundo pequeño de Watts y Strogatz

El modelo de redes de mundo pequeño de Watts y Strogatz [117] genera redes que satisfacen la propiedad de mundo pequeño, presente en muchos fenómenos reales. Sin embargo, falla al describir la distribución de grado y el coeficiente de clustering promedio de muchas redes del mundo real. Este modelo propone un mecanismo para

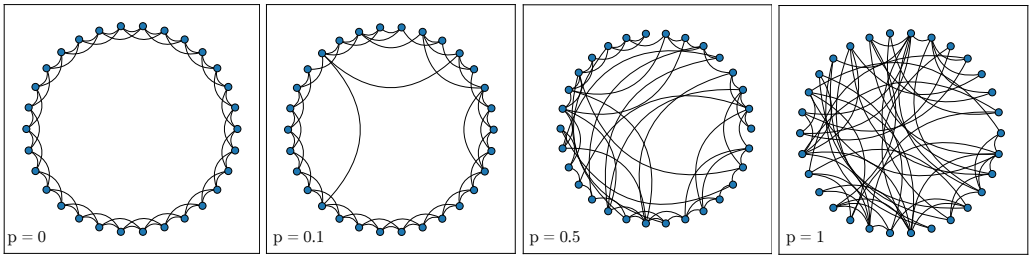


Figura 2.9: Modelo de Watts-Strogatz. Se inicia con un anillo de N = 30 nodos y se agregan enlaces para que cada nodo tenga 4 vecinos cercanos. Para $0p1$ se reubican una proporción p del total de enlaces, en este caso 60, de tal manera que se pasa de una red muy estructurada para $p = 0$ a una de aleatoria cuando $p = 1$. En el camino se obtiene una red con características del extremo estructurado y el extremo aleatorio. Los códigos para obtener esta imagen están disponibles en `https://curso-redes-f-ciencias-unam.github.io/ciencia-de-redes/index.html`

formar redes con dos de las propiedades que tienen muchas redes reales; este modelo surge de la observación de que muchas redes reales comparten un alto coeficiente de agrupamiento y la presencia de caminos cortos entre nodos (es decir, la capacidad de llegar rápidamente desde un nodo a cualquier otro nodo en la red a través de un pequeño número de pasos).

Para lograr esto, Watts y Strogatz propusieron un mecanismo simple. En éste, una red se construye inicialmente con una estructura regular, tipo anillo, en la que cada nodo está conectado directamente a sus vecinos más cercanos. Luego, se modifican algunas conexiones incorporando enlaces aleatorios a la red que acortan la distancia entre nodos opuestos en el anillo. Estas conexiones aleatorias se dan con una probabilidad determinada por un parámetro p. Si p es muy pequeño, entonces la red seguirá siendo muy regular y no tendrá muchos caminos cortos. Si p es muy grande, la mayoría de los enlaces serán aleatorios y la red adquirirá las propiedades de una red aleatoria que se vieron en el apartado anterior, distancias cortas pero agrupamiento bajo. El trabajo de Watts y Strogatz demuestra que hay un régimen intermedio entre las redes regulares, con un alto coeficiente de agrupamiento pero distancias grandes, y las redes aleatorias con distancias cortas pero bajo coeficiente de agrupamiento; en este régimen coinciden las características deseadas de cada extremo: alto coeficiente de agrupamiento y distancias pequeñas en promedio.

Este modelo ha sido muy influyente en el estudio de las redes sociales, la biología y otras áreas de la ciencia, ya que ayuda a explicar cómo las redes reales pueden ser tan efectivas para la comunicación y la coordinación. Este modelo recoge dos de las propiedades presentes en muchas redes reales, pero no atiende la presencia de hubs en la red, es decir, un pequeño grupo de nodos con un grado muy alto en comparación con el resto tales que tienen una fuerte influencia en la estructura de la red.

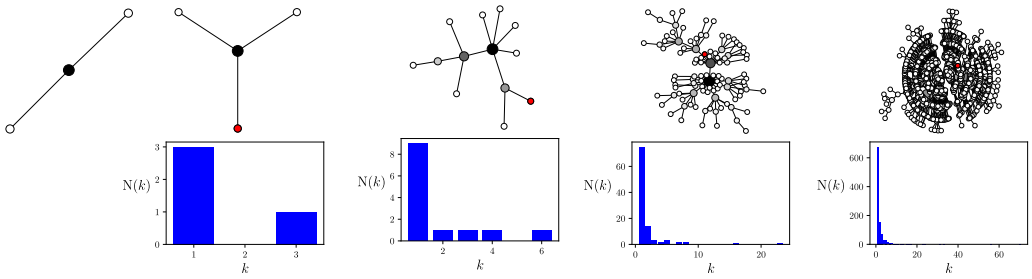


Figura 2.10: Implementación del modelo de Barabási-Albert. Se inicia con una semilla de 3 nodos. En cada paso del proceso se agrega un nuevo nodo mediante el mecanismo de enlace preferente con una probabilidad de enlace proporcional al grado de los nodos presentes en la red. En este caso, el nuevo nodo se distingue con el color rojo y el grado de los nodos previos de distingue con un color blaco para grado 1 y negro para el grado máximo de la red. Se muestra el resultado de agregar 1, 10, 100 y 1000 nuevos a la misma semilla. En cada caso también se incluye la distribución de grado, en la que se puede ver cómo van predominando los nodos con grado bajo, aunque la propia escala del eje horizontal nos permite saber que hay algunos nodos con grados muy grandes, aunque sean una minoría. Los códigos para obtener esta imagen están disponibles en `https://curso-redes-f-ciencias-unam.github.io/ciencia-de-redes/index.html`

2.3.3. Redes con independencia de escala y modelo de Barabási-Albert

Las redes con independencia de escala (también conocidas como "scale-free networks" en inglés), a veces llamadas redes libres de escala, son un conjunto de redes complejas, descubierto por Barabasi, Albert y Jeong [118] al revelar por primera vez la estructura de la world wide web, que se caracteriza por tener una distribución de grado que sigue una ley de potencia y una longitud de camino promedio proporcional a N.

En una red con independencia de escala, algunos nodos tienen un grado muy alto (es decir, tienen muchas conexiones con otros nodos), mientras que la mayoría de los nodos tienen un grado bajo. Esta distribución de grado sigue una ley de potencias, lo que significa que la probabilidad de encontrar un nodo con un grado k sigue una función de la forma

$$P_k \sim k^{-\lambda} \tag{2.28}$$

Caracterizar a las redes con independencia de escala fue muy importante porque reflejan la estructura de muchas redes del mundo real, como las redes sociales, las redes de colaboración científica, las redes de transporte y las redes neuronales. En estas redes, algunos nodos tienen una influencia desproporcionada en el comportamiento global de la red, lo que significa que es importante comprender cómo interactúan estos nodos "centrales" con el resto de la red.

El modelo de Barabási-Albert [119] es uno de los modelos más conocidos de redes con independencia de escala y tuvo la virtud de ser el primero en generar una red con tal propiedad. En este modelo, se construye una red a partir de un pequeño número de nodos y se van agregando nuevos nodos de manera secuencial. Cada nuevo

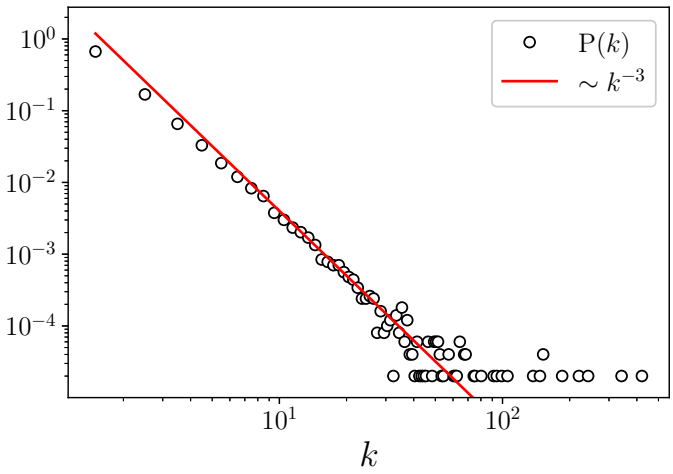


Figura 2.11: Distribución de grado del modelo de Barabási-Albert después de 50,000 iteraciones. A la misma semilla mostrada en la figura 2.4, se le agregaron nodos hasta tener 50,000 y se obtuvo su distribución de probabilidad de grado. Se utilizan escalas logarítmicas tanto en el grado (eje horizontal) como en la probabilidad (eje vertical) para notar que se aproxima a una ley de potencia con exponente $\gamma = -3$. Los códigos para obtener esta imagen están disponibles en `https://curso-redes-f-ciencias-unam.github.io/ciencia-de-redes/index.html`

nodo se conecta a un número determinado de nodos existentes, con una probabilidad proporcional a su grado en ese momento. Este proceso de crecimiento y conexión preferencial da lugar a una distribución de grado que sigue una ley de potencias.

Este modelo es importante en el contexto de teorías como el efecto Mateo y otros mecanismos de crecimiento desigual, como "los ricos se hacen más ricos", porque ilustra cómo los procesos de conexión preferencial pueden dar lugar a la formación de redes complejas que exhiben características de independencia de escala, como la distribución de grado de los nodos. En términos simples, la conexión preferencial se refiere a la idea de que los nodos más conectados en una red tienen más probabilidades de atraer nuevas conexiones que los nodos menos conectados. Esto además plantea un paralelismo con sistemas que exhiben distribuciones tipo ley de potencia como la distribución de Pareto que se encuentra en muchos sistemas complejos, incluyendo la distribución de riqueza y la distribución de tamaño de las ciudades.

2.4. Modularidad y detección de comunidades

Muchas veces los elementos de un sistema, que son representados por los nodos de una red, tienen características que no están determinadas por la estructura de la red. Es decir, a diferencia de las medidas de centralidad de los nodos, que se obtienen a partir del lugar o el papel que juega un nodo en la estructura de enlaces de la red, estamos hablando de propiedades de los elementos del sistema mismo como pueden ser, en el caso de redes sociales, grupos de edad, etnicidad, profesiones, religión, etc. Muchas

veces hay una coincidencia entre estas comunidades, definidas a partir de propiedades "externas" a la red, y módulos de la red, definidos a partir de la presencia significativa de enlaces al interior de los cúmulos en comparación con los enlaces entre nodos de diferentes comunidades. Para medir esta relación entre comunidades y módulos se ha definido [8] la modularidad de la red como:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{2.29}$$

Donde:

- Q es la modularidad de la red.
- m es el número total de enlaces en la red.
- A_{ij} es el elemento ij de la matriz de adyacencia de la red.
- k_i y k_j son los grados de los nodos i y j, respectivamente.
- c_i y c_j son las comunidades a las que pertenecen los nodos i y j, respectivamente.
- $\delta(c_i, c_j)$ es la delta de Kronecker, que es 1 si $c_i = c_j$ y 0 en caso contrario.

Esta fórmula mide la diferencia entre la fracción de enlaces que se encuentran dentro de las comunidades y la fracción esperada si los enlaces se distribuyeran aleatoriamente. Por lo tanto, cuanto mayor sea la modularidad, mejor coincidirán las comunidades con los módulos de la red.

Para seguir el camino inverso, es decir, a partir de propiedades topológicas de la red obtener un conjunto de comunidades que coincidan con los módulos de la red y que, presumiblemente, correspondan a grupos de elementos del sistema con características distintivas, se puede utilizar el método de Louvain [120]. Este método consiste en un algoritmo que busca una partición de los nodos de la red que, al tomarlos como comunidades en la fórmula de la modularidad, ésta alcance un valor máximo dentro del conjunto de todas las particiones posibles de los nodos de la red.

El algoritmo se compone de dos fases iterativas:

1. Fase de agregación: En esta fase, el algoritmo busca comunidades en la red a nivel local, agregando nodos en comunidades cada vez más grandes, lo que maximiza la modularidad de la red. Para hacer esto, el algoritmo comienza asignando cada nodo a su propia comunidad, y luego, en cada iteración, examina cada nodo y evalúa si moverlo a otra comunidad aumentaría la modularidad de la red. Si es así, el nodo se mueve a la comunidad que maximiza la modularidad.

2. Fase de refinamiento: En esta fase, el algoritmo examina cada comunidad encontrada en la fase anterior y la trata como un nodo en una nueva red. Luego, repite la fase de agregación en esta nueva red, buscando comunidades más finas en cada nivel y agregando nodos en comunidades cada vez más pequeñas. Este proceso se repite hasta que no se puede mejorar la modularidad de la red.

El resultado final del método de Louvain es una jerarquía de comunidades en la red, que puede ser representada en forma de un dendrograma. Cada nodo en el dendrograma representa una comunidad, y las ramas del dendrograma indican cómo se agrupan las comunidades a diferentes niveles de granularidad.

El método de Louvain es muy eficiente y escalable para redes de gran tamaño, lo que lo hace popular en muchas aplicaciones prácticas. Además, ha demostrado tener un buen rendimiento en la detección de comunidades en comparación con otros métodos [121]. Este tema será importante en el estudio de la red urbana desarrollado en el capítulo 4.

2.5. Redes y flujos en el marco de la ciencia de las ciudades

En el capítulo anterior se describieron los rasgos generales de la ciencia de las ciudades. En este apartado se describen algunos métodos para analizar las ciudades como sistemas a partir de redes.

2.5.1. Umbralización de los flujos en ciudades

Batty [2] propone un método para estudiar los flujos en ciudades, por ejemplo de movilidad de personas o de tráfico, mediante redes. Este método consiste en convertir una matriz de origen-destino para las regiones de una ciudad en una red simple o dirigida, mediante el uso de un umbral. El método consiste en los siguientes pasos:

1. Definir una matriz de flujos de origen-destino: esta matriz describe la cantidad de personas, bienes, información, etc., que se mueven desde una ubicación de origen a una ubicación de destino en la ciudad. Esta matriz puede ser construida a partir de datos de transporte, telefonía celular, redes sociales, etc.

2. Aplicar una función de umbralización: se aplica una función de umbralización a la matriz de flujos de origen-destino. Esta función transforma la matriz de flujos en una matriz de conexiones binarias, en la que una conexión entre dos ubicaciones existe si el valor del flujo entre ellas supera un cierto umbral. El valor del umbral se determina empíricamente y puede variar según el contexto y los objetivos del análisis.

3. Construir la red: se construye la red a partir de la matriz de conexiones binarias obtenida en el paso anterior. Cada ubicación se representa como un nodo en la red, y las conexiones entre las ubicaciones se representan como enlaces entre los nodos.

Este método permite construir una red a partir de los flujos en una ciudad, lo que puede ser útil para analizar la estructura y la dinámica de la ciudad desde una perspectiva de redes complejas. Las redes resultantes comparten las propiedades mencionadas aquí para la mayoría de los sistemas complejos: propiedad de mundo pequeño, gran

coeficiente de agrupamiento, existencia de hubs, etc. Además, estas redes presentan una transición, tanto para el tamaño de la máxima componente conectada como para el número de enlaces de la red, similar en varios sentidos a una transición de fase en términos de la física estadística. El estudio de estas transiciones será relevante en el capítulo 4.

2.5.2. Redes de calles

Otro enfoque de redes que puede ser utilizado para estudios de ciudades es el de las redes de calles, propuesto por Batty [2] y desarrollado también por Barthelemy [4] y otros autores [122]. Este enfoque se basa en la idea de que las calles de una ciudad pueden ser consideradas como una red, en la que las intersecciones son los nodos y las calles son los enlaces. La topología de esta red puede ser analizada utilizando herramientas de la teoría de grafos y las redes complejas, lo que permite comprender mejor la estructura y la dinámica de las ciudades.

En este enfoque se han propuesto varios indicadores para medir la estructura de las redes de calles, como la longitud media del camino más corto entre dos nodos, la centralidad de intermediación de los nodos, la densidad de la red, etc. Estos indicadores pueden ser utilizados para caracterizar la eficiencia, la accesibilidad y la resiliencia de la red de calles de una ciudad. Además, este enfoque también se ha utilizado para estudiar la evolución de las redes de calles a lo largo del tiempo, y cómo esto afecta a la forma y la función de las ciudades. Por ejemplo, se ha observado que las ciudades más antiguas suelen tener redes de calles más irregulares y menos eficientes, mientras que las ciudades más nuevas suelen tener redes de calles más regulares y eficientes.

Todo lo anterior hace que el enfoque de las redes de calles sea una herramienta útil para entender la estructura y la dinámica de las ciudades, y para informar la planificación urbana y el diseño de políticas públicas.

2.5.3. Conexión preferencial como mecanismo activo en ciudades

Partiendo de que las ciudades son redes de interacciones, desde la ciencia de las ciudades se plantea que el tamaño de las ciudades y el escalamiento pueden deberse a mecanismos de conexión preferencial. Por un lado, la teoría de la conexión preferencial de Barabási-Albert postula que las redes complejas, incluyendo las redes urbanas, se organizan a través de un mecanismo de crecimiento por conexión preferencial, en el que los nodos más conectados atraen más enlaces. En el caso de las ciudades, esto significa que las ciudades más grandes y más conectadas tienen más probabilidades de atraer nuevas inversiones y residentes, lo que a su vez refuerza su tamaño y conectividad. Según esta teoría, el tamaño de una ciudad y su escalamiento son consecuencias directas de su capacidad para atraer y retener recursos.

Por otro lado, las ciudades se benefician de economías de escala, en las que el aumento de la población y la actividad económica generan mayores oportunidades para la

innovación y el crecimiento. Sin embargo, este crecimiento no es homogéneo en todas las partes de la ciudad, y algunas zonas pueden ser más propensas al crecimiento y al desarrollo que otras. Esto significa que, en lugar de depender exclusivamente de factores externos como la conexión preferencial, el tamaño de una ciudad y su escalamiento también están influenciados por factores internos que afectan la distribución espacial de la actividad económica y la innovación.

Combinando ambos mecanismos, las ciudades son sistemas complejos que se benefician tanto de su capacidad para atraer recursos externos como de su propia capacidad interna para generar innovación y crecimiento. De esta manera, el tamaño y el escalamiento de una ciudad son el resultado de una interacción dinámica entre factores internos y externos, que pueden actuar como catalizadores o frenos del crecimiento urbano.

3 Datos de la actividad humana en ciudades

3.1. Introducción

Como se mencionó en el capítulo 1, este trabajo parte de una propuesta metodológica en la que las ciudades son entendidas más como interacciones, relaciones y vínculos entre distintos elementos, que como las zonas o regiones delimitadas por fronteras administrativas. En este capítulo se arrancará el estudio de las ciudades a partir de una interacción muy concreta y de la que tenemos mucha información: la interacción entre personas y puntos de interés tal y como se manifiestan en los check-ins de la red social basada en ubicaciones Foursquare. Decimos entonces que los check-ins son una manifestación particular de las interacciones que se dan entre personas y que conforman las ciudades; como síntesis de esas interacciones sociales, económicas, etc., en los términos de Batty [2], tenemos los puntos de interés, cuya dinámica es un reflejo de los procesos que ocurren en las ciudades.

En el capítulo se describen las distintas fuentes de información que existen para el estudio de las ciudades y se profundiza en una de estas fuentes: las redes sociales basadas en ubicaciones. Posteriormente se exploran conjuntos de datos de cuatro redes sociales basadas en ubicaciones: Foursquare, Brightkite, Gowalla y Weeplaces; y se desarrollan con detalle las propiedades de la base de datos de Foursquare que se utilizará en adelante para nuestros análisis. Finalmente, se definen los sistemas de interés para los resultados posteriores: los puntos de interés de Foursquare ubicados en las áreas urbanas funcionales (definidas en el capítulo 1) dentro de las que se realizaron más de 10,000 check-ins. Esto da como resultado un sistema de 632 áreas urbanas funcionales sobre las que más adelante se hacen análisis espaciales y temporales.

3.2. Datos para el estudio de las ciudades

La disponibilidad de datos sobre ciudades está aumentando constantemente [2, 4]. Diversas agencias y empresas, tanto estatales como privadas, publican vastas cantidades de datos a diferentes escalas espaciales y temporales. Las fuentes de estos datos suelen ser nuevas tecnologías como los teléfonos inteligentes y otros dispositivos personales. Con los recientes avances en tecnologías de información y comunicación, el

enfoque de investigación ha complementado a las fuentes tradicionales, principalmente encuestas y fuentes de datos gubernamentales [81, 123–125], con nuevas y variadas fuentes de datos. Ahora es posible seguir las trayectorias y patrones de actividad individuales a partir de llamadas telefónicas [72, 126–131], servicios de intercambio de ubicaciones [87, 132–135]y microblogging [68], datos extraídos de sistemas integrados de transporte público, registros GPS de taxis [69], coches privados, bicicletas [71]. Para la mayoría de las fuentes de datos, las posiciones espaciales son bastante confiables y esta información se puede utilizar para estudiar qué hacen las personas, cómo, cuándo y hasta por qué lo hacen.

Las diversas fuentes de información permiten estudiar la actividad humana en las ciudades desde la escala de horas (detección de horas pico [136], cálculo de horas de sueño [92], etc.), hasta las escalas de tiempo más larga mediante la digitalización, por ejemplo, de documentos históricos que permiten monitorear la evolución a largo plazo de la infraestructura en las ciudades [4].

La disponibilidad de datos que permitan contrastar y validar teorías y con los que se pueda evaluar el desempeño de un modelo, es un factor muy importante para el desarrollo del estudio científico de las ciudades. Siempre que no haya problemas de privacidad, sería deseable que cualquiera pudiera descargar estos conjuntos de datos. Esto garantiza la reproducibilidad de los resultados y acelera el proceso de investigación. Afortunadamente, este punto de vista ahora es compartido por una proporción cada vez mayor de grupos y publicaciones [4].

En la última década se ha popularizado lo que algunos autores denominan información geográfica voluntaria [137], que es información que las personas proporcionan voluntariamente y que está disponible en internet para que cualquiera la conozca, modifique o utilice para distintos fines, desde individuales (búsqueda de sugerencias, por ejemplo) hasta colectivos y públicos. Esta información se incorpora cada vez más al estudio de aspectos urbanos como como movilidad y actividad humana. Un tipo de información geográfica voluntaria son los Puntos de Interés que, como su nombre lo indica, son sitios que resultan atractivos por alguna razón y, en consecuencia, son muy frecuentados, y esto puede constatarse mediante el uso de trazas de GPS [137, 138], pero también aparecen en portales y páginas públicas como google sites [81] o redes sociales como Facebook o Twitter, porque las personas los mencionan, los comparten, los recomiendan, etc.

3.3. Redes sociales basadas en ubicaciones

Dentro de las aplicaciones y sitios web en los que las personas comparten y consultan puntos de interés, existen un conjunto llamado redes sociales basadas en ubicaciones (LBSN por sus siglas en inglés) que están íntimamente relacionadas con la recolección, clasificación y divulgación de los puntos de interés. Los usuarios de estas redes sociales pueden registrar un sitio si lo consideran un punto de interés (POI por ser las siglas de Point of Interest) independientemente de si es un restaurante, un bar, un parque, un sitio de trabajo o una casa particular, por ejemplo. Una vez que alguien

registra el sitio en la red social dando información sobre la categoría, la dirección, los horarios, y demás datos de referencia, cualquier persona puede encontrar ese POI en la aplicación. Estos lugares son identificables dentro de la plataforma a partir de sus coordenadas y una ventaja es que el usuario puede encontrarlas a partir de las categorías disponibles (restaurantes, entretenimiento, salud, etc.) o a partir de la proximidad física, utilizando la geolocalización de los dispositivos. Como cualquier red social, si ésta logra tener muchos usuarios, será deseable para los establecimientos comerciales aparecer en ella y los propios dueños comenzarán a registrase; por otro lado, si hay muchos lugares registrados, la LBSN será más interesante para la gente y eso atraerá más usuarios. Comienza así un ciclo de retroalimentación positiva en el que más usuarios provoca más sitios registrados, y más sitios atraen más usuarios, retroalimentación descrita por Srnicek [139] bajo el nombre de "efecto de red" de las plataformas digitales.

La importancia de las redes sociales basadas en ubicaciones para el estudio del comportamiento humano radica en que, entre los lugares visitados por un usuario de una LBSN, hay algunos en los que decide compartir con otros su experiencia. Esto puede ser por razones personales de cualquier índole pero resaltan una dimensión importante dentro de las determinaciones por las que las personas actúan: el gusto por ciertos lugares, el deseo y la motivación de compartir con el resto su experiencia. En concreto, las redes sociales basadas en ubicaciones proporcionan información valiosa sobre las interacciones de dos clases de entes: las personas y los puntos de interés de las ciudades, por lo que han sido utilizadas para estudios de movilidad urbana y comportamiento humano que pueden resultar de mucho interés para quienes hacen políticas públicas o para las empresas que buscan ofrecer un servicio a la población local.

La actividad en las LBSN genera información espacial y temporal de muchas personas y lugares. Particularmente, cuando una persona hace un registro comparte, entre otras cosas, su posición en un momento específico. Por esta razón las LBSN son una herramienta muy útil y han sido ampliamente utilizadas como fuente primaria de datos para estudio de comportamiento humano [140–142], interacciones sociales [143] y redes de encuentro [69, 82, 88, 144–147]. Esta cantidad masiva de datos contiene el comportamiento de muchos usuarios y proporciona una oportunidad de explorar comportamientos colectivos a gran escala. Por ejemplo, el análisis de los metadatos de los registros de diferentes poblaciones (por ejemplo, personas en diferentes países) permiten estudiar diferencias en el comportamiento entre ellas [148].

El número de usuarios que utilizan LBSN es considerable, atrayendo a plataformas que originalmente no ofrecían servicios basados en ubicaciones a hacerlo; Facebook y Twitter[1], por ejemplo, han incorporado funcionalidades de una red social basada en ubicaciones, Foursquare, para que sus usuarios puedan compartir y consultar reseñas de puntos de interés. Los datos de las redes sociales basadas en ubicaciones no son, en general, públicos. Sin embargo, diversos grupos han obtenido y publicado datos de Gowalla, Brightkite, Weeplaces y Foursquare. Estos cuatro conjuntos de datos tienen

[1] `https://help.twitter.com/es/using-twitter/tweet-location`

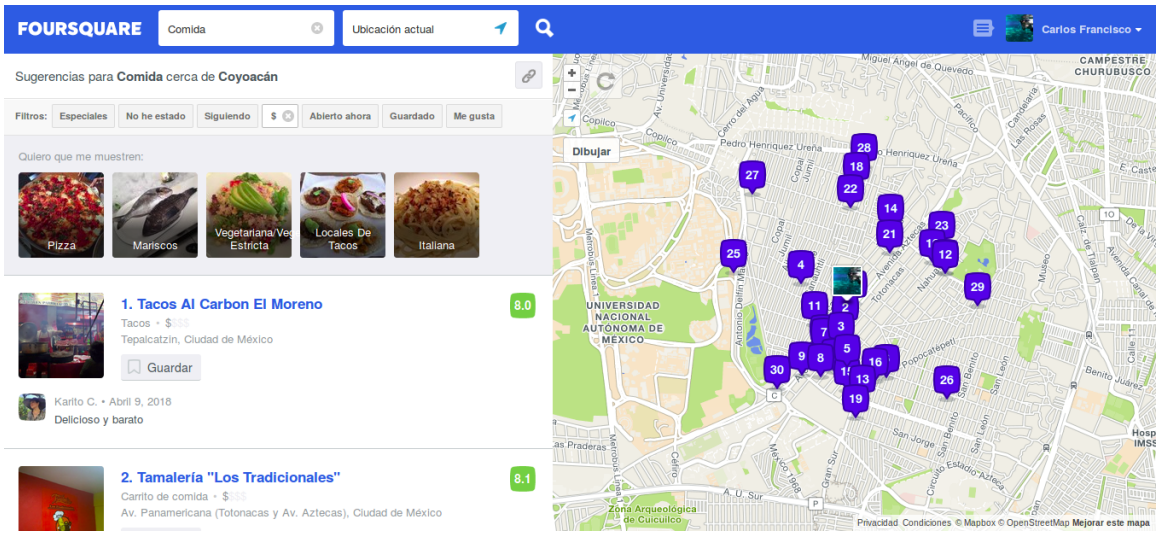


Figura 3.1: Plataforma de Foursquare City Guide. La interfaz de Foursquare incluye un buscador que permite filtrar los puntos de interés según su categoría y por proximidad a partir de la ubicación GPS del dispositivo (Ubicación actual) o de alguna ubicación den el mapa de la derecha. Además se incluyen muchos criterios de búsqueda relacionado con los horarios, la experiencia previa del usuario, entre otros. A partir de esos criterios se muestran los puntos de interés en la parte izquierda, junto con la calificación promedio, la dirección, un estimado del precio y algunos de los comentarios destacados. Cada check-in es un comentario como el que se ve en la imagen.

características distintas respecto al número de datos, el periodo de tiempo que cubren y su cobertura geográfica. A continuación se detalla cada una y se explica la razón por la cual se decidió utilizar el conjunto de datos de Foursquare.

3.3.1. Foursquare

Foursquare City Guide [149] es una red social basada en ubicaciones. Fue creada en 2009 por Dennis Crowley y Naveen Selvadurai en Nueva York, EUA. Desde sus orígenes, se trata de una aplicación que permite a sus usuarios encontrar sitios de interés y, mediante gamificación, estimularlos para visitar lugares, hacer reseñas, calificar, comentar, así como interactuar entre sí como en cualquier red social: formar parte de un círculo social, enviar mensajes, reaccionar a publicaciones, etc. Además, si se usa desde un dispositivo móvil se incorporan funciones de geolocalización mediante GPS y otras; la aplicación cuenta con un mapa que muestra lugares cercanos a la ubicación del dispositivo y los clasifica en categorías para facilitar su búsqueda por parte el usuario (ver Fig. 3.1). Los conceptos clave de este trabajo, punto de interés y check-in, adquieren en Foursquare un sentido muy preciso Un punto de interés es un lugar que la comunidad de la red social registra y, tras ser verificada por el equipo de Foursquare, se incorpora a la red social y su mapa para ser consultada según la categoría de la que se trate. Cualquier usuario puede registrar un lugar si lo con-

sidera un punto de interés[2] y quienes lo verifican son los superusuarios que, según la plataforma de desarrolladores de Foursquare, se trata de "miembros dedicados y apasionados de nuestra comunidad que ayudan a mantener organizados los lugares de Foursquare detrás de escena [...] Los superusuarios crean y editan lugares, organizan y mantienen nuestra base de datos en general, reportan errores y proporcionan comentarios de calidad"[3]. Un punto de interés en Foursquare es cualquier tipo de sitio, independientemente de si se trata de un restaurante, un bar, un parque, un lugar de trabajo o una casa, por ejemplo, siempre y cuando alguien proporcione información sobre la categoría, la dirección, el horario de atención, entre otras. Los lugares públicos son visibles para cualquier usuario mientras que sitios con ciertas categorías, como Residencial o Casa, no son visibles en la red social y solo se puede acceder a ellos con el identificador único que la plataforma proporciona. Absolutamente todos los sitios verificados cuentan con un identificador.

Por su parte, un Check-in es la visita confirmada de un usuario a un punto de interés. Foursquare define los check-ins activos y los pasivos; los primeros se logran mediante la interacción que el usuario puede hacer con ellos, como comentarios, fotografías, recomendaciones, y que quedan registrados en la red social para su consulta; las plataformas generan distintos mecanismos para garantizar que cada check-in represente una visita real[4], sin embargo esto no siempre es posible [150]. Un check-in pasivo ocurre cuando se detecta, vía la geolocalización, que un dispositivo registrado en la plataforma visita un punto de interés, incluso si el usuario no abre la aplicación para hacer el check-in.

Los check-ins de Foursquare son interacciones espacio-temporales entre usuarios y puntos de interés. Proporcionan mucha información sobre los intereses de las personas, las características de los sitios, los patrones de comportamiento en las ciudades, etc., razón por la cual, los metadatos de Foursquare valen tanto. De hecho, en la actualidad Foursquare City Guide es uno de varios productos de la empresa Foursquare Labs, Inc. [151], que provee muchas fuentes de información para analistas, desarrolladores y vendedores. Algunos ejemplos son *Foursquare Places POI Data*[5], un dataset con más de 100 millones de puntos de interés en más de 190 países y 50 territorios, clasificados en más de 900 categorías, y que reporta 2.4 millones de actualizaciones a POIs cada mes; también el conjunto *Foursquare visits*[6], con más de 14 mil millones de check-ins generados en sus aplicaciones desde 2009. Para desarrolladores y vendedores existen herramientas como APIs, marketing basado en ubicaciones (Location-Based Marketing), Proximity Targeting, métricas sobre visitas e interacción de usuarios con los sitios, entre muchas otras. Asimismo, los datos de check-ins y POIs de Foursquare han dado lugar a múltiples investigaciones científicas y acadé-

[2] https://foursquare.atlassian.net/servicedesk/customer/portal/20/topic/f1dc53ca-614a-42dc-8f10-9ce221ef4a12/article/932872676

[3] https://developer.foursquare.com/docs/places-data-schema

[4] https://foursquare.atlassian.net/servicedesk/customer/portal/20/topic/9c4995d8-939f-434c-9ea1-6b89b05f3114/article/1929675008

[5] https://location.foursquare.com/products/places/

[6] https://location.foursquare.com/products/visits/

micas [86–88, 142, 143, 150, 152–156].

Actualmente, según la página de Foursquare City Guide, hay más de 50 millones de personas utilizando esa aplicación [7]. La demografía de los usuarios es la siguiente [157]:

- Comunidades urbanas: 64 %, pueblos y comunidades rurales: 36 %.
- Pueblos medianos: 28 %, ciudades grandes: 26 %.

Esto significa que hay usuarios de Foursquare tanto de comunidades urbanas como de pueblos y comunidades rurales. Aunque el porcentaje de usuarios urbanos es mayor, como era de esperar, la diferencia no es tanta; la proporción de usuarios rurales es aproximadamente 1 de cada 3. Esto significa que los datos de Foursquare reflejan no solo una composición urbana, sino que también refleja características de pueblos pequeños o rurales. Por otro lado, la distribución de género y edad es la siguiente [158]:

- Hombres: 52 %, mujeres: 48 %
- Edades: de 18 a 24 años: 19 %, de 25 a 34 años: 32 %, de 35 a 44 años: 20 %, de 45 a 54 años: 14 %, de 55 a 64 años: 10 %, de 65 años o más: 6 %.

Debe resaltarse que el porcentaje de hombres y mujeres es prácticamente el mismo. Por lo tanto, los datos de Foursquare pueden aplicarse tanto a mujeres como a hombres. En cuanto a la distribución por edad, notamos que alrededor del 50 % de los usuarios tienen entre 25 y 44 años. Esto es lo esperado ya que las personas a esa edad, los adultos jóvenes, son los que muestran más movilidad y actividad. Es evidente que la movilidad, actividad e interés en informar a través de check-ins en la red social tienden a disminuir con la edad. Sin embargo, la distribución es amplia.

En cuanto al tipo de lugares visitados y marcados como puntos de interés, vale la pena mencionar que las categorías son muy variadas. En Foursquare hay más de 1100 categorías de sitios, distribuidas en diez grupos principales de categorización de POIs de Foursquare [159]: arte y entretenimiento, negocios y servicios profesionales, comunidad y gobierno, comida y bebida, eventos, salud y medicina, lugares famosos y al aire libre, tiendas minoristas, deportes y recreación, viajes y transporte. Esto apunta a una gran diversidad de intereses que se capturan en el conjunto de datos.

Otros datos relevantes sobre Foursquare son los siguientes [160, 161]:

- Foursquare Places tiene más de 100 millones de POIs en 247 países y territorios, a partir del 4 de enero de 2023.
- 100M+ POIs globales
- 14 mil millones de check-ins verificados por usuarios
- 1100+ categorías de lugares
- 1 mil millones + fotos, consejos, opiniones.

[7] `https://es.foursquare.com/cities/mexico-city`

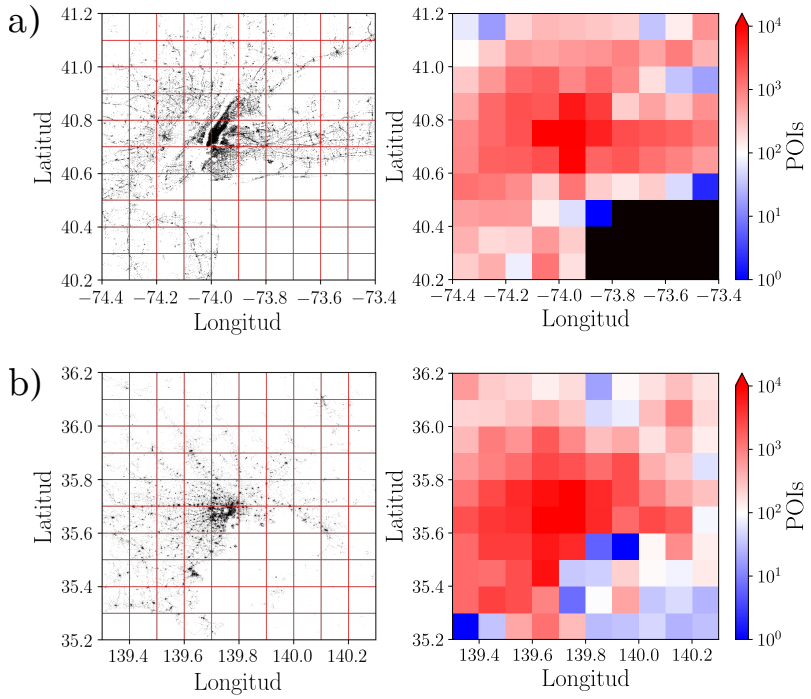


Figura 3.2: Conteo de puntos de interés de Foursquare. Todos los puntos de interés tienen unas coordenadas de la forma (lon, lat), que van de $-180°$ a $180°$, y de $-90°$ a $90°$. Esto permite agrupar los puntos de interés en "cuadrados" de décimas de grado. Enla figura se muestra el resultado de partir el territorio de esta forma para las regiones de (a) Nueva York y (b) Tokio. Posteriormente, se cuenta el número de puntos de interés que hay en cada elemento de la cuadrícula, y a ese número se le asigna un color a partir del mapeo de color que aparece en los páneles de la derecha. Es importante utilizar una escala logarítmica debido a la gran variedad de valores que se pueden encontrar, desde ninguno (color negro) o 1 (color azul) hasta más de 10,000 (color rojo). Esto se hizo para todos los puntos de interés del mundo.

Si bien se necesita contratar los servicios de Foursquare para acceder a los datos de POIs y check-ins, existen algunos conjuntos de datos públicos disponibles. A continuación describimos de forma general la base de datos publicada por Yang y colaboradores [143, 150, 162], aunque en el siguiente apartado se explicará con detalle el método para su obtención y sus características.

Esta base de datos contiene dos conjuntos: uno de casi 3.4 millones de POIs y otro de un poco más de 90 millones de check-ins realizados por más de 11 millones de usuarios en esos POIs. Para visualizar la distribución espacial del los POIs, los agrupamos según sus coordenadas. Se hizo un histograma dividiendo los intervalos de la longitud y la latitud, $[-180°, 180°]$ y $[-90°, 90°]$ respectivamente, en subintervalos de $.1°$ y contando el número de POIs en cada "cuadrado" en el espacio de coordenadas. En la figura 3.2 se presenta el procedimiento para obtener la densidad de los puntos de interés a nivel planetario. En los páneles superior e inferior de la izquierda se muestran los POIs de la región donde se encuentran las ciudades de Nueva York y

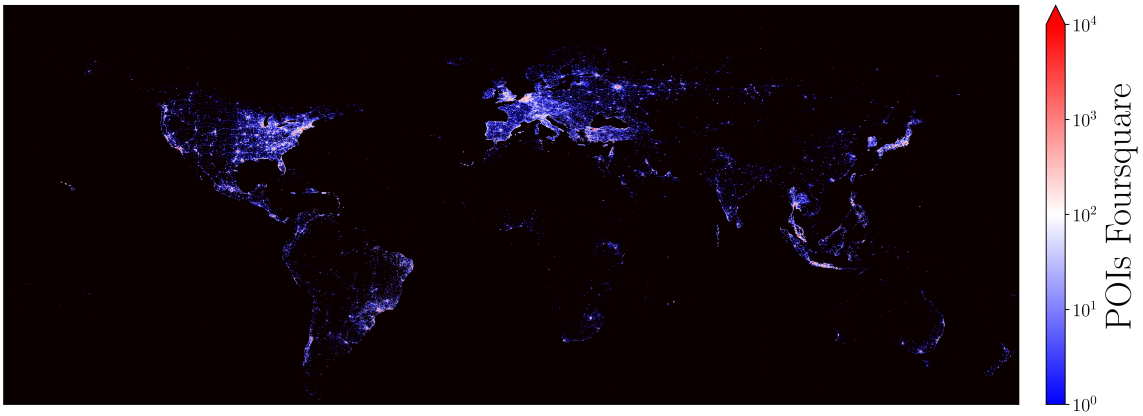


Figura 3.3: Distribución de puntos de inerés de Foursquare a nivel global. En la figura se muestra el resultado de hacer lo descrito en la figura 3.2, con la misma resolución y escala, pero en todo el mundo. Se identifica mucha actividad en las regiones normales para este tipo de datos (Europa, Estados Unidos, Japón), pero lo que destaca es la presencia de bastante actividad en países de los que no se suele encontrar muchos datos: Turquía, Rusia, Sudamérica, Asia continental, incluso Oceanía.

Tokio, respectivamente. Se grafican en el eje x las longitudes y en el y las latitudes de los puntos sin ninguna proyección geográfica. También se muestran las líneas correspondientes a las décimas de grado, tanto en la longitud como en la latitud. A la derecha se muestran, mediante una escala de colores arbitraria, el número de POIs que hay dentro de cada cuadrado, y en color negro se muestran los cuadrados en los que no hay ningún POI. Esto mismo se hizo para todas las regiones entre -180 y 180 grados de longitud y $-90°$ a $90°$, dando como resultado la figura 3.3.

Esto nos permitió identificar regiones con datos y aquéllas con la mayor concentración de POIs. En la figura 3.3 se muestra la distribución espacial obtenida de esta forma; consiste en 3600×1410 cuadrados cubriendo las latitudes entre $-56.6°$ y $84.6°$ y las longitudes de $-180°$ a $180°$; las latitudes excluídas se explican por la falta de POIs en esas regiones. Se usó el mismo mapa de color que en 3.2 para el número de sitios y se eligió una escala logarítmica debido a la amplia variación de los valores. La cobertura global de este conjunto de datos es claro así como la existencia de regiones con gran cantidad de sitios alcanzando 70,128 POIs en una unidad de área definida antes. La gráfica es el resultado de proyectar las coordenadas del globo terráqueo en un rectángulo, lo cual implica distorsiones espaciales. Aún así, los límites de los continentes son claramente identificables a pesar de que no se dibujaron fronteras de ningún tipo. Muchos POIs se aprecian en Norte América y Europa, pero regiones con mucha actividad también se observan en América del Sur, Medio Oriente y Asia Oriental. Aunque con mucho menos intensidad, hay regiones activas en África y Oceanía.

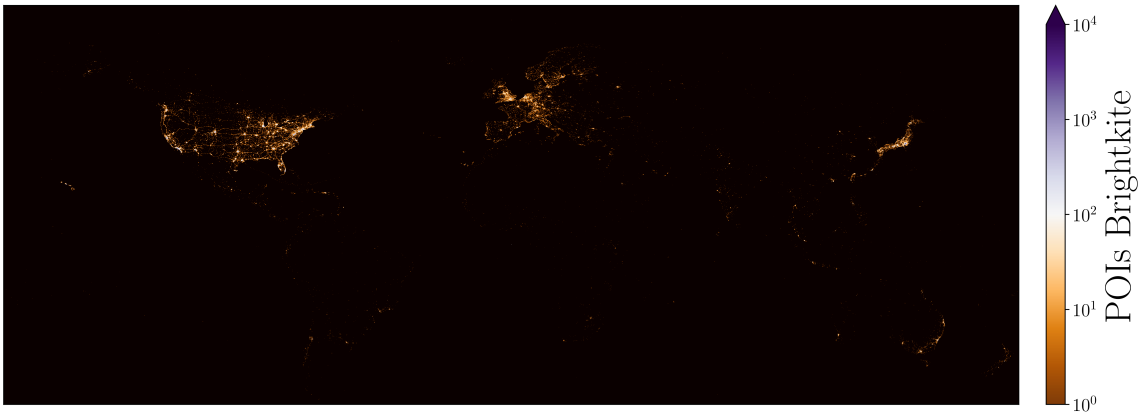


Figura 3.4: Distribución de puntos de interés de Brightkite a nivel global. La figura se generó mediante el mismo procedimiento descrito en 3.2 aunque la paleta de color en este caso va del color cobre cuando hay un único POI en el área en cuestión, a morado cuando hay 10 mil POIs o más, utilizando una escala logarítmica debido a la amplia variedad de valores. Se pueden identificar tres principales regiones de actividad intensa; Estados Unidos de América, Europa y Japón. Se perciben regiones más pequeñas de actividad en Australia, Asia y Sudamérica.

3.3.2. Brightkite

Brightkite fue una red social basada en ubicaciones que estuvo disponible hasta 2011[8]. Fue creada en 2007 por Brady Becker, Martin May y Alan Seideman y en 2010 contaba con 2 millones de usuarios[9]. En ella, como en cualquier LBSN, los usuarios hacían check-ins en sitios reales. Los usuarios registrados podían conectar con sus amigos y también conocer a nuevas personas con base en los lugares que habían visitado. Una vez que alguien registraba un check-in, que podían ser notas y fotos, otros podían comentar esa publicación.

Cho y colaboradores [163, 164] recolectaron información de 4,747,280 check-ins realizados en 772,966 puntos de interés por 51,406 usuarios entre el 21 de marzo de 2008 y el 18 de octubre del 2010. El 60 % de los check-ins fueron hechos en POIs ubicados en Estados Unidos de América. El resto se divide en más de 100 países encabezados por Japón con el 9 % de los check-ins, Reino Unido con 4 % y Australia con 2 %. En la figura 3.4 se muestra la distribución de POIs de esta red social generada con el mismo procedimiento descrito en 3.2 para Foursquare. En este caso la mayoría de los POIs se concentran en los Estados Unidos de América, Europa y Japón.

3.3.3. Gowalla

Gowalla es una LBSN lanzada en 2007 y que en 2010 contaba con 600 mil usuarios[10]. En 2011 fue adquirida por Facebook, dejó de estar disponible en 2012 pero fue re-

[8] https://en.wikipedia.org/wiki/Brightkite
[9] https://techcrunch.com/2010/02/26/brightkite-2-million-users/
[10] https://en.wikipedia.org/wiki/Gowalla

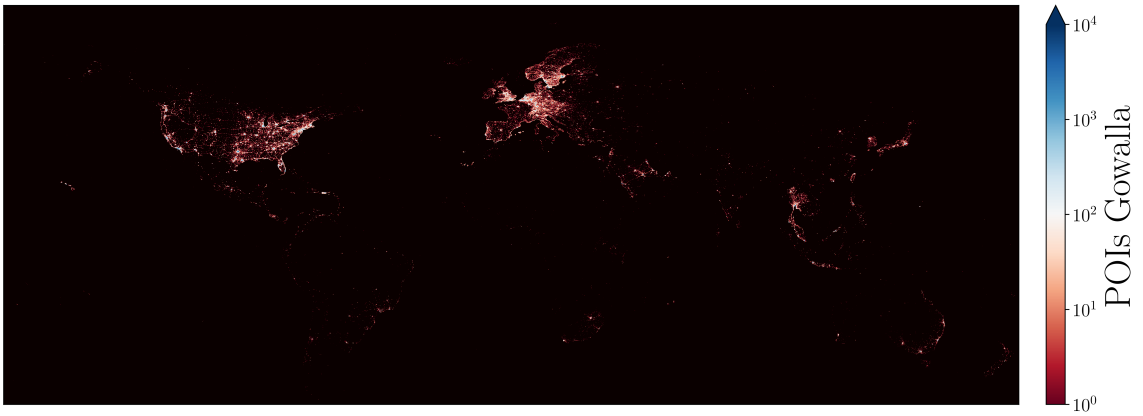


Figura 3.5: Distribución de puntos de interés de Gowalla a nivel global. La figura se generó mediante el mismo procedimiento descrito en 3.2 aunque la paleta de color en este caso va del color rojo cuando hay un único POI en el área en cuestión, a azul cuando hay 10 mil POIs o más, utilizando una escala logarítmica debido a la amplia variedad de valores. Además de Estados Unidos de América, Europa y Japón, se suman como regiones de gran actividad Tailandia e Indonesia. Se perciben regiones más pequeñas de actividad en Australia, Asia, Sudamérica, India y Turquía.

lanzada en marzo de 2023[11] como *Gowalla Exploration*[12]. De la misma forma que en Foursquare City Guide, en la aplicación de Gowalla se incorpora un mapa con los sitios registrados, pero en su nueva versión es posible ver a las amistades y su actividad en tiempo real.

Liu y colaboradores [133, 134, 165] obtuvieron mediante APIs perfiles de usuarios, amistades, información de puntos de interés y 36,001,959 de check-ins realizados en 2,844,076 POIs por 319,063 usuarios entre el 9 de octubre de 2009 y el 4 de julio de 2011. El 50 % de los check-ins fueron hechos en 1,149,891 POIs localizados en Estados Unidos. Los siguientes países con más check-ins son Suecia con el 11 %, Alemania con 5 %, Tailandia y Reino Unido con 4 % cada uno y Arabia Saudita con 3 %.

En la figura 3.5 se muestra la distribución de POIs mediante el mismo procedimiento descrito antes. A las típicas áreas de actividad intensa (Estados Unidos, Europa y Japón) se suman zonas de Asia como Tailandia e Indonesia, y pequeñas regiones en India, Turquía, Australia y Brasil.

3.3.4. Weeplaces

Weeplaces era una página web en la que cualquier usuario podía inscribirse mediante sus credenciales de alguna LBSN (Facebook Places, Foursquare, Gowalla, etc.) y, vinculándose con las APIs de las distintas plataformas, ofrecía servicios para organizar, guardar y compartir sus datos (check-ins), además de diferentes formas de visualizar la actividad del usuario y diversos juegos para estimular que los usuarios tuvieran más

[11] `https://wwwhatsnew.com/2023/03/10/gowalla-vuelve-...`
[12] `https://www.gowalla.com/`

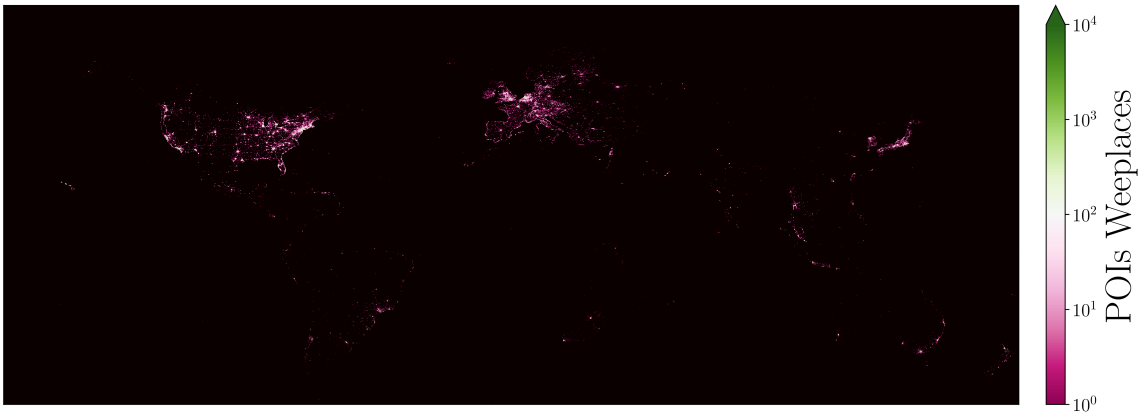


Figura 3.6: Distribución de puntos de interés de Weeplaces a nivel global. La figura se generó mediante el mismo procedimiento descrito en 3.2 aunque la paleta de color en este caso va del color morado cuando hay un único POI en el área en cuestión, a verde cuando hay 10 mil POIs o más, utilizando una escala logarítmica debido a la amplia variedad de valores. Además de Estados Unidos de América, Europa y Japón, se suma Indonesia como región de actividad considerable, además de actividad moderada en Brasil, Australia y el Caribe.

actividad[13]. Actualmente la página no sigue activa, pero en su perfil de Facebook se pueden encontrar algunos de los productos y análisis que ofrecían como son el análisis de actividad nocturna diferenciada por género, visualización en forma de video de las actividades recientes de un usuario o mapas personalizados para compartir en blogs y redes sociales[14].

Liu y colaboradores [133, 134, 165] recolectaron, mediante la página de Weeplaces, 5,303,599 check-ins realizados por 15,799 usuarios en 971,309 POIs, entre noviembre del 2003 y junio del 2011. Dado que los check-ins y POIs no están etiquetados con ninguna información salvo las coordenadas, nuestro trabajo consistió en agruparlos por país dando cuenta de que el 57 % de los check-ins fueron realizados en alguno de los 390,631 POIs ubicados en Estados Unidos, que representan en 50 % del total de los check-ins del conjunto. Los países que le siguen en cantidad de check-ins son Reino Unido con el 5 % de los datos y Canadá, Alemania, Países Bajos, Australia y Francia con el 3 % cada uno. En la figura 3.6 se muestra la distribución de POIs a nivel global de la que se concluye que la mayoría de la actividad registrada ocurrió esencialmente en las mismas regiones que Brightkite y Gowalla.

3.3.5. Resumen

La información de las cuatro bases de datos disponibles se resume en la tabla 3.1. En ella se muestran el total de check-ins, puntos de interés, número de usuarios activos, las fechas del primer y el último check-in registrado y el total de días transcurridos

[13]Un ejemplo en: `https://www.youtube.com/watch?v=OKRVKggRsDo&feature=youtu.be`

[14]`https://www.facebook.com/profile.php?id=100069316311838&locale=sq_AL&paipv=0&eav=Afa1Gc8szvfVN4euJnIf7ac9O8cQbVhEK1zaiG01u8C0bRWoo6hXTD5PLrj6HZ6Jp6I`

entre el primero y el último check-in para las redes Foursquare (4S), Brightkite (BK), Gowalla (GW) y Weeplaces (WP). La de Foursquare es la base con más datos, cuyos más de 90 millones de check-ins casi triplican a los de Gowalla, que es la que le sigue con poco más de 36 millones; ocurre lo mismo con los puntos de interés y los usuarios. Considerando el número de check-ins y el tiempo de cobertura de cada conjunto de datos, el de Foursquare alcanza un promedio de 135,410 check-ins por día, seguido de Gowalla con 56,884 check-ins por día, mientras que Brightkite y Weeplaces apenas alcanzan 5,050 y 1,897 check-ins por día.

Dataset	Check-ins	POIs	Usuarios	Primer check-in	Último check-in	Días
4S	90047754	11179790	2733324	2012-04-03 18:00:06	2014-01-29 16:44:25	665
BK	4747280	772966	51406	2008-03-21 20:36:21	2010-10-18 18:39:58	940
GW	36001678	2844076	319063	2009-10-09 18:37:55	2011-07-04 22:54:28	633
WP	5303599	971309	15799	2003-11-02 06:04:47	2011-06-29 13:38:36	2796

Tabla 3.1: **Resumen de conjuntos de datos de LBSN**. Se presentan las propiedades generales de las bases de datos de Foursquare (4S) proporcionada por Yang y su grupo [143, 150, 162], de Brightkite publicada por Cho y colaboradores [163, 164], y Gowalla y Weeplaces obtenida por Lui y su grupo [133, 134, 165].

En lo que respecta al tiempo de cada base de datos (figura 3.7), Weeplaces es la que cubre un periodo mayor (casi ocho años entre 2003 y 2011) mientras Foursquare es la que contiene datos más recientes (entre 2012 y 2014).

Respecto a la cobertura espacial de las cuatro bases de datos, se puede hacer un análisis cualitativo observando los "mapas de calor" de cada una (figuras 3.3, 3.4, 3.5 y 3.6) ya que se hicieron con los mismos parámetros para poder comparar, aunque los mapeos de color sean distintos para distinguirlas. En todas destacan Estados Unidos, Europa y Japón, sin embargo es muy claro que Foursquare proporciona información de muchas otras regiones de las que no suele haber gran cantidad de datos, particularmente América del Sur, Europa del Este y Asia, además de algunas regiones en África.

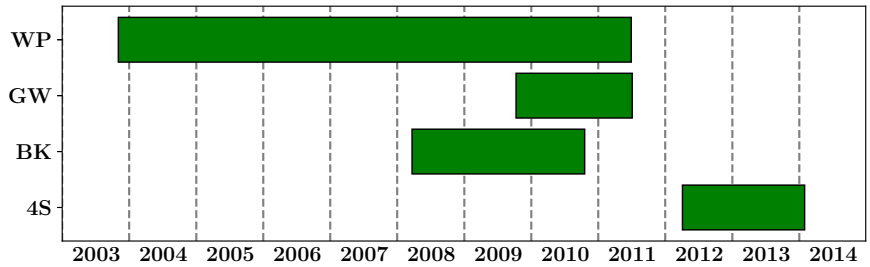


Figura 3.7: **Cobertura temporal de los cuatro conjuntos de datos**. Se representa gráficamente el periodo de tiempo que cubren los datos de cada una de las bases de datos consideradas en esta revisión, a decir, Weeplaces (WP), Gowalla (GW), Brightkite (BK) y Foursquare (4S).

	País	Foursquare	Gowalla	Brightkite	Weeplaces
1	United States of America	1990327	1149891	465806	390631
2	Indonesia	1198611	23707	823	6618
3	Brazil	1159258	14262	1094	20761
4	Turkey	1098373	1952	637	1210
5	Russia	546532	8857	1568	2942
6	Japan	519409	40257	113059	18807
7	Malaysia	493453	39594	2289	10887
8	Mexico	408434	15122	1430	4944
9	Thailand	353444	152583	622	9905
10	Philippines	219097	9751	548	2308
11	Spain	212161	33917	6282	9530
12	United Kingdom	210777	123868	33154	37535
13	Italy	197007	36544	6879	11335
14	Chile	195226	2469	1444	4557
15	Germany	142347	169617	16909	26473
16	Netherlands	126822	36636	11745	19603
17	South Korea	118365	16445	640	6338
18	Canada	118200	63706	11599	25876
19	France	108269	25291	8596	22478
20	Belgium	105894	43117	2259	5912

Tabla 3.2: **Número de POIs por país en cada LBSN**. Se presentan los 20 países con más puntos de interés en Foursquare, red social con más puntos de interés totales, y el número de puntos de interés de cada una de las 4 bases de datos consideradas. Como elemento auxiliar al valor numérico, se utiliza una paleta de colores que va del blanco a morado, pasando por distintos tonos de azul, asociando colores claros con los valores bajos y los colores fuertes con los valores altos.

Con la intención de hacer también un análisis cuantitativo de la extensión espacial de los datos de cada base de datos, se contaron los puntos de interés y los check-ins registrados en cada LBSN agrupados por país. En la tabla 3.2 se presenta el número de puntos de interés registrados por país en cada una de las bases de datos, para los 20 países con más POIs de Foursquare. Para poder hacer una comparación general, además del número de POIs, cada celda se rellena con una paleta de color que va del blanco para el valor mínimo al morado para el valor máximo (en este caso, los 1,990,327 POIs de Foursquare ubicados en Estados Unidos de América). Esto nos permite no solo confirmar que Estados Unidos tiene el máximo de puntos de interés en cada red social, sino que con excepción de Foursquare, este país concentra la mayoría de POIs en cada caso. Por su parte, la base de datos de Foursquare tiene información considerable de más países. En la selección de 20 países se puede ver que todos tienen más de 100 mil POIs en Foursquare, mientras que solo 4 cumplen esta condición en Gowalla, 2 en Brightkite y sólo Estados Unidos en Weeplaces. Comparando a Foursquare con el resto, ésta contiene 3 países con más POIs que Estados Unidos (el máximo) de Gowalla, 7 países con más POIs que Estados Unidos en Brightkite y 8 con más POIs que Estados Unidos en Weeplaces. En suma, la base de datos de Foursquare es la que da información de mayor cantidad y variedad de países en todo el mundo.

En la figura 3.3 se hace el mismo análisis pero con los check-ins, a decir, el conteo de check-ins por país para cada una de las bases de datos consideradas. Los países

	País	Foursquare	Gowalla	Brightkite	Weeplaces
1	Turkey	17500113	5889	1663	3216
2	United States of America	12778097	17871063	2857602	3045347
3	Brazil	9991354	68857	21781	132659
4	Indonesia	7765315	117918	15240	23517
5	Malaysia	4926145	368560	20681	44195
6	Japan	4784080	298940	417437	128148
7	Russia	4291601	58225	5821	23216
8	Mexico	3981409	65863	12854	17741
9	Thailand	2633608	1548247	4094	42043
10	Chile	2209981	10269	9075	25166
11	Philippines	1998063	55613	21923	9756
12	United Kingdom	1271622	1465370	203767	251242
13	Spain	1083153	184833	39402	33581
14	Singapore	871619	0	0	0
15	Italy	867931	206606	33924	43726
16	Netherlands	689177	377056	60312	147005
17	Germany	623759	1970651	80241	152252
18	Canada	599268	597384	84037	166285
19	Belgium	571648	418985	20564	29478
20	France	527439	171652	43332	135348

Tabla 3.3: **Número de check-ins por país en cada LBSN**. Se presentan los 20 países con más check-ins en Foursquare, red social con más registros totales, y el número de registros de cada una de las 4 bases de datos consideradas. Como elemento auxiliar al número, se utiliza una paleta de colores que va del blanco a morado, pasando por distintos tonos de azul, asociando colores claros a los valores bajos y colores fuertes a los valores altos.

se ordenan según el número de check-ins en Foursquare por ser la red de la que más datos hay. Para facilitar el análisis, como antes, se utilizan la misma paleta de colores que va de los colores claros (azules) para valores pequeños a los colores fuertes (morados) para valores altos. En este caso, el máximo conteo de check-ins se da en Estados Unidos para la plataforma Gowalla, con 17,871,063 registros. En las bases de Gowalla, Brightkite y Weeplaces, Estados Unidos concentra la mayor cantidad de check-ins, mientras que en Foursquare es Turquía el país con mayor volumen de datos. Visualmente, destaca el dominio que tienen los Estados Unidos en las LBSN que no son Foursquare, lo que se manifiesta como una fila coloreada en el segundo renglón, mientras que la gran cantidad de registros de Foursquare en muchos países se manifiesta como una columna coloreada que contrasta con los tonos claros del resto de bases de datos. Trece de los veinte países seleccionados tienen, en Foursquare, más de un millón de check-ins, mientras que solo 4, 2 y 1 países cumplen esta condición en Gowalla, Brightkite y Weeplaces, respectivamente. Frente a la cantidad de check-ins ocurridos dentro de los Estados Unidos en Brightkite y Weeplaces, hay ocho países contienen más check-ins en Foursquare.

Por todo lo anterior, para los análisis realizados en este trabajo se optó por utilizar únicamente la base de datos de Foursquare, pues tiene la ventaja de contener más datos, en más países, y para un periodo más reciente. Sin embargo, las otras pueden utilizarse para validar o contrastar resultados. Particularmente, si se requiriera un periodo de observación más largo, Weeplaces sería la más relevante, mientras que si se realizara un estudio concentrado en los Estados Unidos de América, Gowalla sería

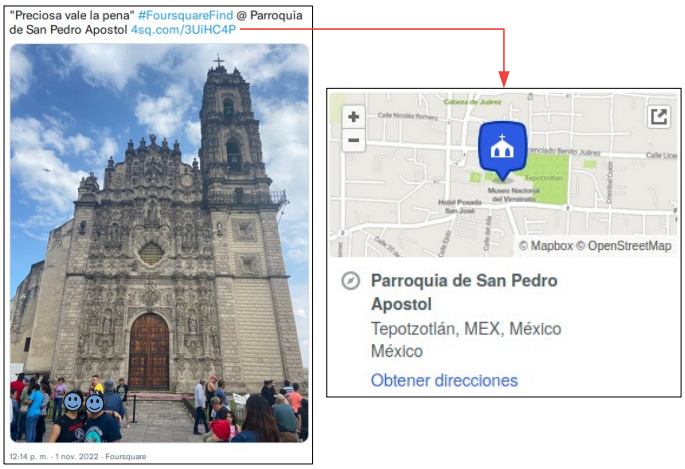


Figura 3.8: Sincronización de las cuentas de Foursquare y Twitter. Si se vinculan las cuentas de Foursquare y Twitter, cada vez que se hace un check-in (comentario, fotografía, etc.) en Foursquare, se genera un twit automatizado con una estructura fija: "comentario" #FoursquareFind @ Nombre_del_sitio 4sq.com/POI_ID. Además, el link lleva a la página de Foursquare desde donde se puede obtener información de la ubicación del lugar (dirección, ciudad, país) además de la categoría y las coordenadas geográficas a partir del mapa de OpenStreetMap.

la mejor opción. En el siguiente apartado se desarrolla con más detalle el procedimiento desarrollado por Yang [143, 150] para recolectar los datos de Foursquare, las características específicas de esta red social que la dotan de ventajas y desventajas para nuestro trabajo, así como algunos otros detalles de la información que proporciona. Posteriormente, se utiliza la base de datos de áreas urbanas funcionales para acotar nuestros análisis a los puntos de interés ubicados dentro de éstas y los check-ins realizados en ellos.

3.4. Descripción de la base de datos de Foursquare

Como se dijo en el apartado anterior, para poder hacer un estudio comparativo de actividad humana en ciudades, se decidió usar la base de datos de Foursquare publicada por Yang *et al.* [143, 150]. Recordemos que ésta contiene 90,047,754 check-ins hechos por 2,733,324 usuarios, entre abril de 2012 y enero de 2014, en 11,179,790 de puntos de interés. A continuación se describirá con detalle la metodología con la que fue obtenida y el procesamiento que se realizó en este trabajo.

La red social basada en ubicaciones de Foursquare, Foursquare City Guide, permite a sus usuarios vincular su perfil con sus cuentas de otras redes sociales, particularmente Twitter y Facebook[15]. Si fuera el caso, cada vez que el usuario hace un check-in en Foursquare, se genera en la otra plataforma una publicación automática con la información del registro. En el caso de Twitter, se publica un twit con un formato

[15] `https://foursquare.atlassian.net/servicedesk/customer/portal/20/topic/fa67aa59-dc9b-45d8-9dfe-201a8576da16/article/937132046`

estándar que contiene el comentario escrito entre comillas, el hashtag #Foursquare-Find, el nombre del sitio junto al signo @ (arroba) y el hiperenlace a la página del sitio en la plataforma de Foursquare. Además, si el usuario compartió una imagen, esta se incorpora al twit (ver figura 3.8 a). Adicionalmente, en la página de Foursquare del sitio se puede acceder, entre toda la información antes mencionada, a un mapa de OpenStreetMap con las coordenadas del lugar (figura 3.8 b). Foursquare, como prácticamente todas las páginas comerciales, impiden la extracción masiva de su información. Mediante estos registros automatizados en Twitter y su servicio de API, hasta hace poco tiempo, se podía obtener mucha información sin violar las políticas de ninguna plataforma. De este modo, los autores obtuvieron primero los 90,047,754 de check-ins con información del usuario, del sitio y la hora y fecha del registro.

Posteriormente, con el conjunto de lugares en los que los check-ins ocurrieron, Yang y su equipo colectaron la información de la página web de Foursquare. Esta información es pública, sólo hay que acceder al link proporcionado en el twit; bastaría un programa que acceda a la lista de hipervínculos y extraiga los datos del punto de interés. Se puede asociar el ID del POI con las coordenadas del sitio (mapa de OpenStreetMap), el código del país y la categoría del lugar según la clasificación de Foursquare. Esto explica cómo se construyó, por un lado el conjunto de datos de los check-ins a partir de Twitter, con información sobre el sitio, el usuario y la hora precisa a la que se hizo; y, por otro lado, el conjunto de datos de los POIs de Foursquare, a partir de su propio portal, con el ID público del sitio, las coordenadas, el país en el que se encuentra y la categoría. Con estos dos conjuntos, vinculados por los ID de los sitios, se puede recrear la actividad espacio-temporal de muchos usuarios junto con las propiedades del entorno en el que dicha actividad tuvo lugar. Además, Yang y su equipo incorporan una "red social" definida vía la relación de mutuo seguimiento entre las cuentas de Twitter. Esta red social no son las amistades de Foursquare sino que es una definición dada por este mismo grupo con la finalidad de predecir el surgimiento de enlaces; en este trabajo no exploramos tal red de amistades. A continuación se expone nuestro análisis detallado de los dos conjuntos, el de puntos de interés y el de check-ins.

3.4.1. Puntos de interés

El procesamiento inició con el análisis del conjunto de datos de los puntos de interés de Foursquare que llamaremos D_{POI}. El conjunto está contenido en un archivo de texto sin formato de 704.8 MB con 11,180,160 filas y 5 columnas. Se abrió con la ayuda de la paquetería Pandas [166], de Python [167], para explorar sus características generales. En la tabla 3.4 se muestran algunas filas aleatorias.

La primera columna de D_{POI}, POI ID, corresponde al identificador alfanumérico del punto de interés; esta columna contiene 11,180,160 valores distintos, cada uno correspondiente a un POI y es el que aparece en la url del sitio en Foursquare. Para acceder a la página de Foursquare del sitio, basta con agregar la clave POI ID al url https://es.foursquare.com/v/ ya que así se construyó la base de datos. En la figura 3.9 se muestra la información obtenida de la página de Foursquare de los sitios de la tabla 3.4. En el panel (a) se muestra el sitio público con ID

ID POI	Latitud	Longitud	Categoría	Código de país
3fd66200f964a52000e61ee3	40.729209	-73.998753	Post Office	US
4c4108b3a48d9c740a780c40	-6.908247	107.621338	Asian Restaurant	ID
4d939ce4a148d7ce8c851dcd	52.173667	4.525192	Home (private)	NL
4e9d0c6bc2ee2113cb4f00e6	10.270668	122.853241	Diner	PH
4f808720e4b009277f0bcbc1	47.228883	0.025210	Home (private)	FR
5036b8c6e4b0024b22fa3b24	25.623438	-100.299279	Home (private)	MX
⋮	⋮	⋮	⋮	⋮

Tabla 3.4: **Puntos de interés en la base de datos**. Algunas líneas del conjunto de datos de los Puntos de interés, D_{POI}, seleccionadas aleatoriamente. Incluye el identificador del sitio en Foursquare, las coordenadas, la categoría y el código del país. Esta información se obtiene de la página web de Foursquare, como se muestra en la figura 3.8 b

3fd66200f964a52000e61ee3 cuya categoría es "Post office" y que se encuentra en Estados Unidos, específicamente en la ciudad de Nueva York. En (b) está el punto de interés cuyo ID es 4c4108b3a48d9c740a780c40, ubicado en Bandung, Indonesia, y cuya categoría es "Asian Restaurant". El recuadro (c) contiene la información de un POI no público con ID 4d939ce4a148d7ce8c851dcd y categoría "Home (private)", ubicado en Países Bajos y al que el usuario que lo registró le llamó "Home Sweet Home". Debido a su categoría, la página no muestra la ubicación exacta de este sitio ya que Foursquare muestra únicamente las coordenadas de aquellos POIs públicos. De hecho, sólo los sitios públicos aparecen en el mapa, a los sitios de carácter privado sólo puede accederse mediante el ID. El POI mostrado en (d) con identificador 4e9d0c6bc2ee2113cb4f00e6 es público y se ubica en Filipinas. Tiene categoría "Diner". Finalmente, (e) y (f) son POIs no públicos con categoría "Home (private)", razón por la cual no tienen ubicación precisa aunque su nombre sí se muestre. Están en Francia y Monterrey-México.

Las segunda y tercera columnas de D_{POI} son la latitud y longitud, respectivamente, descritas por un número flotante con 6 cifras decimales que obedece el estándar World Geodetic System 1984 (WGS84), lo cual corresponde a una precisión de menos de un metro[16]. Con estos datos se puede realizar un análisis espacial muy preciso de los POIs y de la actividad de los usuarios. Esta información nos permitió, en primer lugar, encontrar la distribución espacial de de los POIs mostrada en la figura 3.3, separar los POIs por regiones y, como se verá más adelante, agrupar los POIs por áreas urbanas.

La cuarta columna contiene el nombre de la categoría del POI según la clasificación de Foursquare descrita en la sección 3.3.1; el dataset contiene 519 categorías distintas, de las cuales "Home (private)" es la más común con 1,310,012 sitios, seguida de "Residential Building (Apartment / Condo)" con 354,858; la tercera más popular es "Office" con 317,149; el cuarto lugar lo ocupa "Building" con 255,121 sitios; los siguientes son "Café", "Restaurant", "Bar" y "Hotel" con 188436, 153027, 145878 y 138476 sitios, respectivamente.

[16] `https://wiki.openstreetmap.org/wiki/Precision_of_coordinates`

Figura 3.9: **Puntos de interés en Foursquare**. Cada panel corresponde a un punto de interés mostrado en la tabla 3.4. **(a)** POI público con categoría "Post office" ubicado en Nueva York, Estados Unidos; **(b)** POI público en Bandung, Indonesia, con "Asian Restaurant"; **(c)** POI no público con categoría "Home (private)" en Países Bajos y al que el usuario nombró "Home Sweet Home"; **(d)** POI público en Filipinas, con categoría "Diner"; **(e,f)** POIs no públicos con categoría "Home (private)" ubicados en Francia y Monterrey-México.

Finalmente, en la quinta columna de D_{POI} se indica el código del país del POI según el estándar ISO 3166-1 de dos letras; el dataset contiene 253 códigos distintos, doce de los cuales no están incluidos en la lista de la Organización Internacional para la Estandarización (ISO, por sus siglas en inglés). No hay que perder de vista que, en su plataforma, Foursquare hace referencia a POIs en más de 190 países y 50 territorios. De lo anterior se concluye que el conjunto de datos contiene información de cada país del mundo. Una inspección a los datos arroja que de los 253 códigos de país, los 84 con 5000 POIs o más representan el 98.92 % de los datos. En la tabla 3.5 se muestran los 15 países con más puntos de interés. Es de destacarse que hay países de 4 continentes en el ranking, y se tienen países que representan una gran variedad en términos lingüísticos, culturales, geográficos, económicos, etc. En estos 15 países se encuentran el 80 % de los lugares de D_{POI}.

3.4.2. Check-ins

Un análisis análogo se hizo para el dataset de los check-ins que llamaremos D_{4S}. Este conjunto de datos está contenido en un archivo de texto sin formato de 4.7GB con 90,048,627 líneas, cada una de las cuales corresponde a un check-in, es decir, la interacción entre un usuario y un punto de interés en un momento específico. En la tabla 3.6 se muestran algunas líneas. La primera columna de la tabla es el número de

País		POIs		Check-ins		Usuarios
Código	Nombre	Número	%	Número	%	Número
US	United States	1990327	17.80	12778097	14.19	426341
ID	Indonesia	1198611	10.72	7765315	8.62	361193
BR	Brazil	1159258	10.37	9991354	11.10	261079
TR	Turkey	1098373	9.82	17500113	19.43	592630
RU	Russia	546532	4.89	4291601	4.77	122268
JP	Japan	519409	4.65	4784080	5.31	81293
MY	Malaysia	493453	4.41	4926145	5.47	127390
MX	Mexico	408434	3.65	3981409	4.42	147563
TH	Thailand	353444	3.16	2633608	2.92	82765
PH	Philippines	219097	1.96	1998063	2.22	60197
ES	Spain	212161	1.90	1083153	1.20	67638
GB	United Kingdom	210777	1.89	1271622	1.41	77949
IT	Italy	197007	1.76	867931	0.96	52394
CL	Chile	195226	1.75	2209981	2.45	53714
DE	Germany	142347	1.27	623759	0.69	45574

Tabla 3.5: Datos de Foursquare por país. Los datos de D_{POI} se agruparon a partir de la columna del código de país. Se muestran los primeros 15 países ordenados según el número de puntos de interés que contienen. Además se muestra la cantidad de usuarios activos y de check-ins realizados en ese país, así como el porcentaje que representan esos datos respecto al total de registros.

identificación del usuario de Foursquare/Twitter anonimizado por Yang y su grupo. Esta columna contiene 2,733,324 valores distintos, de donde se deduce el número de usuarios cuya actividad está recogida en esta base de datos. La segunda columna de D_{4S} corresponde código alfanumérico del POI, cuyo identificador es el mismo que en D_{POI}. Por este motivo, el conteo de valores distintos que contiene esta columna arroja el total de 11,180,160 puntos de interés que conforman D_{POI}. La tercera columna es el tiempo universal coordinado (UTC) en el que el check-in ocurrió y la cuarta columna es la corrección del UTC (*Timezone offset*) correspondiente a la zona horaria en la que se encuentra el POI. Para obtener la hora local en la que ocurrió el check-in, basta con sumar a la hora UTC el número de minutos indicados por esta corrección. Obteniendo la fecha y hora local de todos los check-ins del dataset encontramos que el primero fue registrado a las 18:00:06 del martes 3 de abril de 2012 y el último a las 16:44:25 del miércoles 29 de enero de 2014, tal y como se expuso en la tabla 3.1.

Como ya se ha dicho, ambos datasets, D_{POI} y D_{4S}, se vinculan mediante la columna POI ID. Uniendo las dos tablas con la ayuda de la paquetería Pandas, pudimos completar la información que la base de datos nos proporciona de cada país, esto es, el número de POIs, el número de check-ins y el número de usuarios que tuvieron actividad dentro de sus fronteras. A pesar de que el orden varía, los mismos países que concentran la mayoría de los POIs contienen el mayor número de check-ins (ver tabla 3.5). Si se suma el número de usuarios que tuvieron actividad en cada país se obtiene un total de 3,357,212 que contrasta con los 2,733,324 usuarios diferentes que aparecen en D_{4S}; esto da cuenta de que hay usuarios que hacen check-ins en distintos países y, por tanto, aparecen representados en más de un país. Del total de usuarios,

ID usuario	ID POI	UTC	Despl. zona horaria
50756	4f5e3a72e4b053fd6a4313f6	Tue Apr 03 18:00:06 +0000 2012	240
48358	4b5afc00f964a5206edd28e3	Thu Apr 19 05:16:35 +0000 2012	-180
227676	4c39a8bda52cb71314a55f26	Wed May 02 08:37:03 +0000 2012	180
139033	4bdcc9803904a59325c74f9e	Mon May 14 15:53:00 +0000 2012	-420
18356	4e28893fa809ec0663ef4660	Mon May 28 21:40:39 +0000 2012	-180
172930	4f52299ee4b03be3628cc0f2	Tue Jun 12 08:01:33 +0000 2012	420
⋮	⋮	⋮	⋮

Tabla 3.6: Algunas líneas del conjunto de datos de check-ins, D_{4S}, seleccionadas de forma aleatoria. El conjunto contiene un identificador anonimizado para el usuario de Foursquare/Twitter, el identificador alfanumérico del punto de interés, asignado por Foursquare, el tiempo coordenado universal (UTC) en el que se realizó el check-in y el desplazamiento de zona horaria, en minutos, correspondiente a la localización del punto de interés; esta última es el número de minutos que se tienen que agregar al tiempo UTC para obtener la hora local.

el 14.02 % hizo check-ins en más de un país.

En el mismo sentido, uniendo ambas tablas se puede asociar cada check-in con las coordenadas del punto de interés en el que se registró, completando la información espacio-temporal de los usuarios de Foursquare. Con esta información se hizo el histograma de dos dimensiones que se muestra en la figura 3.10. Esta representación de la distribución espacial de los check-ins se hizo tal y como se explicó en 3.2, pero ahora ponderando cada par de coordenadas con el número de check-ins que ocurrieron en el POI. Se perciben las mismas características que tiene la distribución de POIs en 3.3: muchas áreas con un único check-in por cuadrado de décima de grado, otras con cantidades en el orden de millones de check-ins. La región con la mayor concentración de check-ins tiene 1,664,655. Utilizando ambos datasets se genera la tabla 3.5.

3.4.3. Categorías

Habiendo vinculado el contenido de D_{POI} con el de D_{4S}, y agrupando los datos por país, se puede hacer una inspección más profunda de la información de las categorías que proporciona esta base de datos. Esta es una dimensión clave de las redes sociales basadas en ubicaciones, particularmente Foursquare, ya que nos permite caracterizar el comportamiento humano a la luz del tipo de actividades e intereses de las personas que habitan en cada país. En este pequeño apartado se presentan los principales resultados de este análisis.

En el apartado 3.4.1 se presentaron las categorías más comunes entre los puntos de interés de este conjunto de datos y en el apartado anterior obtuvimos las propiedades espaciales de los check-ins a partir de la información espacial de los POIs en los que fueron registrados. Haciendo lo propio con la información de las categorías, se obtuvieron las categorías más comunes pero ahora según el número de check-ins registrados en puntos de cada una. La categoría más común vuelve a ser “Home (private)” con 6,758,340 check-ins (7.5 % del total), pero ahora “Mall”, “Other Great Outdoors”, “Of-

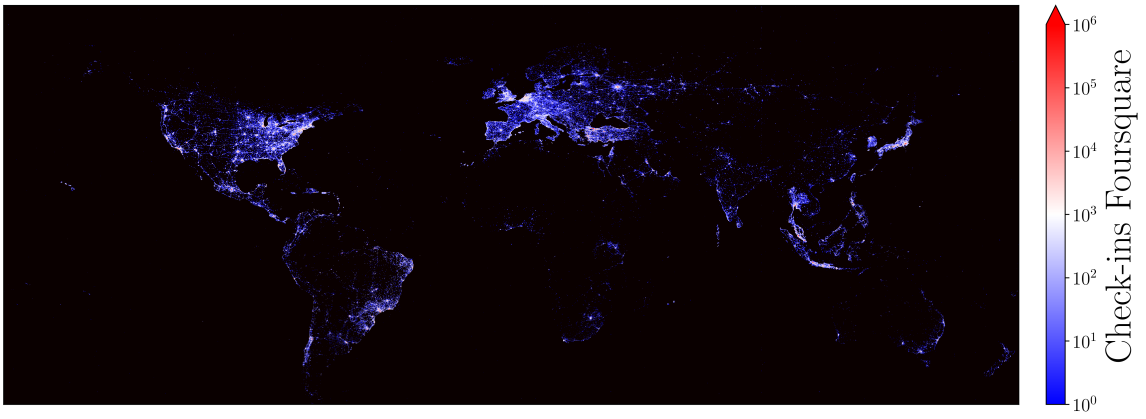


Figura 3.10: Distribución de check-ins de Foursquare a nivel global. En este caso se siguió exactamente el mismo procedimiento descrito en 3.2 pero ahora se contaron no los puntos de interés sino los check-ins realizados en cada región. Se trataría entonces del mismo histograma que 3.3 pero ahora cada punto está ponderado con el número de check-ins que tiene. Eso hace aumentar los valores dos órdenes de magnitud más.

fice" se colocan en el segundo, tercer y cuarto lugar con 4,304,865 (4.78 %), 2,550,782 (2.83 %) y 2,492,375 check-ins (2.77 %), respectivamente. El orden de las categorías a nivel global, en relación con POIs y check-ins, no es necesariamente el mismo en cada país presente en la base de datos. Saber qué categoría es la más común por país nos da una primera aproximación a las características culturales de sus habitantes pues, como se sugirió antes, los check-ins nos hablan no sólo de los lugares que visitan las personas, sino de aquellos lugares que motivan a la gente a compartir su experiencia. Con esto en mente, cerramos esta sección exponiendo las categorías más comunes en cada país.

En la tabla 3.8 se presentan las 10 categorías más comunes, según el número de puntos de interés, para los cinco países con más sitios en D_{POI}. En todos los casos, la mayoría de los puntos de interés pertenecen a la categoría "Home (private)". En Estados Unidos se encuentra que el segundo lugar lo ocupa "Office" mientras que en Indonesia, "Indonesian Restaurant". Brasil, Turquía y Rusia tienen en común que la categoría con más puntos de interés, después de los hogares/privados, es "Residential Building". En adelante, cada país presenta particularidades que hay que tomar en cuenta para describir a cada población.

Por su parte, las categorías más comunes según el número de check-ins, para los países con más check-ins de D_{4S}, se muestran en la tabla 3.8. Esta información es relevante para describir el comportamiento humano a nivel país pues nos dice el tipo de sitios en los que la población se ve motivada a compartir sus visitas con su círculo social. Como se verá más adelante, esto también puede hacerse a nivel área urbana para aproximarnos más al análisis de los sistemas urbanos que son objeto de estudio de este trabajo.

	United States		**Indonesia**		**Brazil**		**Turkey**		**Russia**	
	Categoría	POIs	Categoría	POIs	Categoría	POIs	Categoría	POIs	Categoría	POIs
1	Home (private)	141516	Home (private)	237112	Home (private)	202286	Home (private)	215904	Home (private)	68019
2	Office	53125	Indonesian Restaurant	50620	Residential Building (Apartment / Condo)	50410	Residential Building (Apartment / Condo)	74896	Residential Building (Apartment / Condo)	59530
3	Gas Station / Garage	51341	Office	36914	Office	36004	Café	28862	Office	16754
4	Building	46809	Building	31891	Building	26232	Office	27935	Building	13136
5	American Restaurant	44581	Asian Restaurant	30285	General Entertainment	24361	Building	25867	Housing Development	9805
6	Fast Food Restaurant	38258	College Classroom	24940	Brazilian Restaurant	21782	Turkish Restaurant	22815	Salon / Barbershop	8420
7	Automotive Shop	34783	Café	20720	Salon / Barbershop	18976	Housing Development	22421	Café	8373
8	Residential Building (Apartment / Condo)	33933	Residential Building (Apartment / Condo)	20247	Bar	17740	Factory	18365	Bank	8187
9	Pizza Place	32143	Mosque	18471	Bank	16573	Coworking Space	18319	Other Great Outdoors	7643
10	Church	31706	Food Truck	18281	Road	16352	Salon / Barbershop	14892	Bus Station	6471

Tabla 3.7: **Categorías por cantidad de POIs por país**. A partir de la clasificación de los POIs según el país en el que se ubican, se cuentan cuántas veces aparece cada categoría y se presentan las 10 más comunes en cada caso. Se decidió mostrar esta información para los cinco países con más datos en D_{POI}.

3.5. Puntos de interés y check-ins a nivel ciudad

Nuestro tema de interés es el comportamiento de personas en las ciudades por lo que clasificar los POIs y los check-ins por país no es suficiente. Para lograr el siguiente nivel de análisis se usó el Global Human Settlement Dataset, descrito en la sección 1.6.2, para agrupar los POIs por áreas urbanas funcionales. Los puntos de interés son fruto de las interacciones y los intereses de las personas, fenómenos ambos que no se acotan a divisiones políticas. Las divisiones políticas dividen nuestras unidades de análisis que son las ciudades. La mayoría de las veces las ciudades abarcan dos o más distritos, condados, estados, municipios, etc., dependiendo de la clasificación que se use en cada país. En la figura 3.11 se demuestra cómo, haber optado por una delimitación administrativa, hubiera dejado fuera puntos de interés que forman parte de la misma unidad.

En la figura 3.11 se presenta el tratamiento que se dio a los puntos de interés: mediante scripts de Geopandas, paquetería de Python para el análisis de datos con información geográfica, se seleccionaron aquellos POIs que quedan dentro de las áreas urbanas funcionales en todo el mundo. Geopandas permite seleccionar aquellos puntos (coordenadas) que quedan dentro de ciertos polígonos (fronteras), y definir nuevos conjuntos de datos acotados por las fronteras de cada centro urbano. La imagen muestra seis páneles correspondientes a las ciudades de París (Francia), Londres (Reino Unido), Tokio (Japón), Los Ángeles (Estados Unidos), Nueva York (Estados Unidos) y Ciu-

	Turkey		**United States**		**Brazil**		**Indonesia**		**Malaysia**	
	Categoría	Check-ins	Categoría	Check-ins	Categoría	Check-ins	Categoría	Check-ins	Categoría	Check-ins
1	Other Great Outdoors	1329405	Home (private)	850978	Home (private)	1149045	Home (private)	934692	Home (private)	434297
2	Café	998356	Coffee Shop	375292	Mall	424226	Mall	572166	Mall	406922
3	Home (private)	941626	Office	369102	Office	367951	Indonesian Restaurant	225606	Malaysian Restaurant	276007
4	Mall	936770	American Restaurant	368623	Residential Building (Apartment / Condo)	330432	Office	176331	Asian Restaurant	129660
5	Café	673797	Airport	368513	University	271582	Airport	157611	Residential Building (Apartment / Condo)	114200
6	Neighborhood	629515	Bar	329931	Other Great Outdoors	241156	High School	154135	Building	107499
7	University	594774	Grocery Store	287095	Neighborhood	238630	Asian Restaurant	149428	Café	91126
8	Residential Building (Apartment / Condo)	583372	Gym	275350	Building	208784	Building	148640	Office	89938
9	Restaurant	458165	Gas Station / Garage	208591	Road	190688	University	137687	Fast Food Restaurant	89641
10	Bar	369866	Hotel	208239	High School	182109	Coffee Shop	135512	Indian Restaurant	87033

Tabla 3.8: **Categorías por cantidad de check-ins por país**. Dado que cada check-in se realizó en un punto de interés, se puede asociar con cada uno una categoría. Se presentan las diez categorías más comúnes según el número de check-ins para los cinco países que concentran más registros en D_{4S}.

dad de México (México); en cada caso, los polígonos de las áreas urbanas funcionales, dibujados con azul, separan aquellos POIs que pertenecen a la ciudad (puntos negros) de aquellos que quedan excluidos del sistema (puntos rojos). Para resaltar el hecho de la insuficiencia de las fronteras administrativas y políticas, en cada caso se muestran en verde los polígonos de tales unidades, mostrando que utilizarlas hubiera significado dejar fuera parte importante de los puntos de interés de cada ciudad. Con este método podemos estar seguros de que estamos seleccionando POIs agrupados en áreas con interacciones sociales y económicas fuertes, lo que hace también que los check-ins ocurridos en ellos formen parte de ese mismo conjunto de interacciones relevantes.

De las 13,135 áreas urbanas obtenidas contenidas la base GHS, 6,463 tienen, al menos, un punto de interés de la base de datos. Los POIs que se ubican dentro de estas ciudades representan el 74 % de D_{POI} y los check-ins realizados en ellas en son el 82 % de D_{4S}.

Las treinta y un ciudades con más registros se muestran en la tabla 3.9. Nuevamente, encontramos una gran diversidad en términos culturales, sociales, geográficos y hasta religiosos, ya que se tienen ciudades de Turquía, Indonesia, Malasia, Japón, México, Tailandia, Rusia, Brasil, Filipinas, Estados Unidos, Chile, Singapur, Kuwait, Gran Bretaña, Perú y Corea; dando a este sistema gran relevancia para el estudio de dinámicas urbanas. En cada caso se tiene el número de puntos de interés, el total de check-ins realizados en ellos, y junto con la información de las áreas urbanas funcionales se podrían relacionar estas variables con el área, la población, entre otras. En

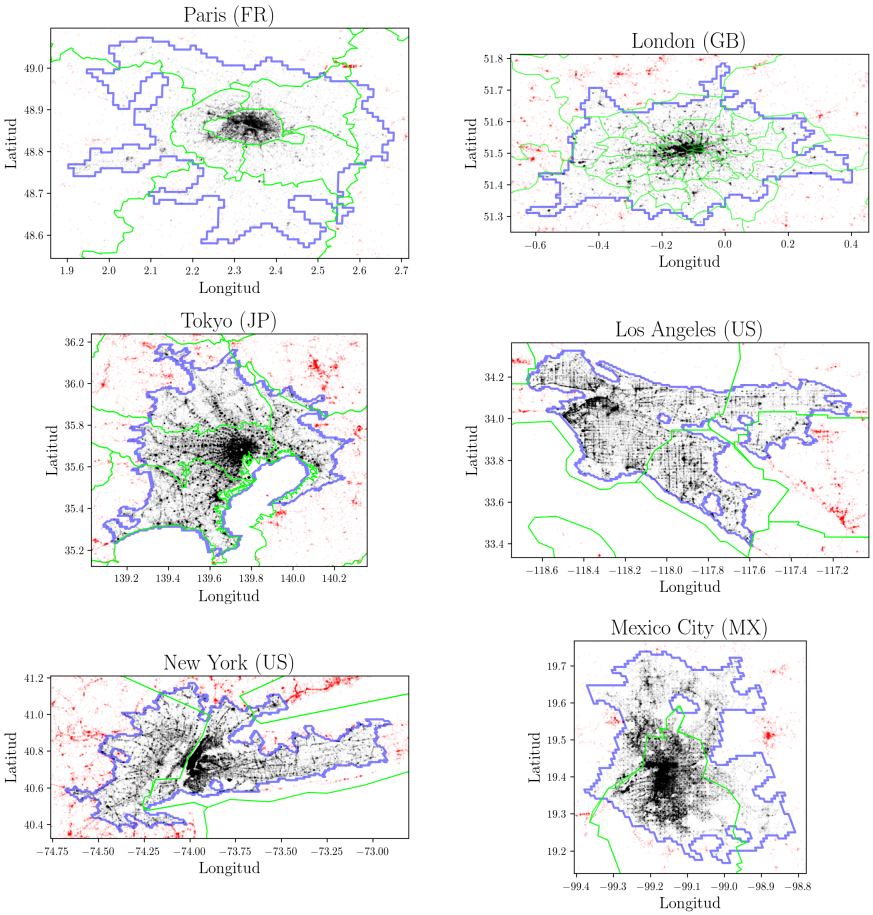


Figura 3.11: POIs en áreas urbanas funcionales de París, Londres, Tokio, Los Ángeles, Nueva York y Ciudad de México. En la figura se muestra el resultado de filtrar los puntos de interés ubicados dentro del área urbana funcional (puntos negros) de aquellos que están fuera (puntos rojos) y en azul se muestra el polígono obtenido del GHS Database [168]. En verde se muestran los polígonos de las fronteras administrativas subnacionales correspondientes a cada caso.

este caso, con la intención de revisar la variedad de nuestros datos, se calcularon la densidad de puntos de interés (puntos de interés promedio por kilómetro cuadrado) y los check-ins promedio por persona en cada ciudad. En lo que respecta al número de check-ins, el primer lugar lo ocupa la ciudad turca de Estambul con 7,343,552 que supera por mucho a la ciudad indonesia de Yakarta, que tiene 2,908,026. En cuanto al número de puntos de interés, se invierte el orden pues el primer lugar lo ocupa Yakarta con 398,154 y el segundo Estambul con 334,517. Llaman mucho la atención las ciudades turcas de Estambul y Esmirna (Izmir en inglés) pues en ellas hay más de 200 POIs/km^2 y se hicieron más de .5 check-ins por persona. De un análisis equivalente al hecho para países se concluye que el 46.62 % de los usuarios hizo check-ins en al menos dos ciudades distintas. En ese sentido, este dataset proporciona información no sólo de la dinámica urbana sino de la dinámica interurbana, no explorada en este trabajo.

Ciudad	País	POIs	Check-ins	POIs / km^2	Check-ins PC
Istanbul	TR	334517	7343552	249.640	0.520404
Jakarta	ID	398154	2908026	79.488	0.080083
Kuala Lumpur	MY	200199	2558716	150.752	0.403605
Tokyo	JP	191162	2405876	35.946	0.072842
Mexico City	MX	137655	1804660	65.116	0.092265
Bangkok	TH	169220	1741638	65.896	0.118231
Moscow	RU	160868	1643199	85.477	0.116726
São Paulo	BR	137494	1478616	68.576	0.077356
Izmir	TR	74237	1443711	210.303	0.516692
Quezon City [Manila]	PH	123941	1343955	61.055	0.061959
New York	US	137841	1311179	25.602	0.082202
Santiago	CL	89115	1276748	123.771	0.201507
Ankara	TR	59610	1021287	158.537	0.340152
Bandung	ID	119357	933686	117.709	0.114121
Singapore	SG	72981	827715	83.027	0.119596
Surabaya	ID	110852	756628	63.308	0.090586
Los Angeles	US	91011	698605	16.157	0.048916
Osaka [Kyoto]	JP	66574	667631	21.081	0.042544
Yogyakarta	ID	83635	609749	53.785	0.119592
Rio de Janeiro	BR	55765	570441	40.794	0.058220
Belém	BR	39125	490162	143.842	0.234762
Chicago	US	55319	485198	14.444	0.071565
Saint Petersburg	RU	56800	482715	107.780	0.112237
Kuwait City	KW	43547	473904	91.485	0.149590
London	GB	48778	430524	26.168	0.044801
Lima	PE	34989	404408	39.942	0.043646
Denpasar	ID	54399	393148	130.453	0.209318
Bursa	TR	25214	392454	120.067	0.233577
Manaus	BR	34633	390084	132.693	0.193399
Medan	ID	46528	387255	62.961	0.097886
Seoul	KR	79306	382646	32.383	0.017714

Tabla 3.9: **Datos de Foursquare por ciudad**. Se muestran el número de puntos de interés y de check-ins registrados dentro de los 31 centros urbanos con más datos. Adicionalmente, se utilizó la información de la base GHS para calcular el número promedio de POIs por kilómetro cuadrado y el número de check-ins per cápita (PC).

Viendo la cantidad ciudades con información disponible en esta base de datos, de la que la tabla 3.9 es sólo una muestra, se necesita un criterio para determinar qué ciudades serán consideradas en los análisis posteriores y que nos permita comparar dinámicas espaciales y temporales del comportamiento humano y del entorno urbano en las ciudades. Por eso, se cerrará este capítulo definiendo este criterio y los sistemas de interés que de él se derivan.

3.5.1. Definición de los sistemas de interés

Considerando que hay 6,463 centros urbanos con al menos un punto de interés de Foursquare en la base de datos, se requiere un criterio que determine las ciudades que serán tomadas en cuenta para posteriores análisis. Este criterio puede darse a partir de los puntos de interés o de los check-ins. Se decidió tomar en cuenta los check-ins porque son los que dan cuenta de la dinámica espacio temporal de la actividad de las personas, y se eligió una cantidad que permitiera, a nuestro juicio, hacer estadística sobre los

datos y que mantuviera la gran diversidad geográfica y cultural que caracteriza a este conjunto de datos.

Por lo anterior se definirá el sistema de interés para este estudio como *los puntos de interés en las áreas urbanas funcionales dentro de cuyas fronteras se hayan realizado 10,000 check-ins o más*. Con esta definición se podría reconstruir el procedimiento de preparación de los datos de la siguiente forma:

1. Se parte de los datasets D_{4S} y D_{POI}. El primero es un conjunto de check-ins obtenido por Yang y colaboradores [162] a partir de la búsqueda automatizada en Twitter; el segundo explorando las páginas de Foursquare a partir de los POIs en los que se hicieron los check-ins.

2. Se definen los datasets $D_{POI}(c)$ para cada ciudad c filtrando D_{POI} con la ayuda de las áreas urbanas funcionales contenidas en la base de datos global de asentamientos humanos de la capa de centros urbanos (GHS [168]) mediante programas propios generados con Geopandas-Python.

3. Se definen los datasets $D_{4S}(c) = D_{4S} \cap D_{POI}(c)$ filtrando el conjunto D_{4S} con la lista de POIs contenida en cada conjunto $D_{POI}(c)$. Esto se hace con la ayuda de programas propios generados con Pandas-Python.

4. Se selecciona el conjunto de ciudades

$$\mathcal{C} = \left\{ c \;\middle|\; N\left(D_{4S}(c)\right) \geq 10^4 \right\}$$

   donde, en este caso, $N\left(D_{4S}(c)\right)$ denota la cardinalidad o número de check-ins del conjunto $D_{4S}(c)$ o, dicho de otro modo, el número de check-ins registrados en puntos de interés ubicados dentro de las fronteras de c.

5. Se definen los conjuntos $\mathcal{D}_{4S} = \{D_{4S}(c) \;\forall\; c \in \mathcal{C}\}$ y $\mathcal{D}_{POI} = \{D_{POI}(c) \;\forall\; c \in \mathcal{C}\}$.

A continuación se analizan las propiedades generales de $\mathcal{C}$, $\mathcal{D}_{4S}$ y $\mathcal{D}_{POI}$.

Este procedimiento da lugar a un conjunto $\mathcal{C}$ formado por 632 ciudades ubicadas en 87 países. En la figura 3.12 se muestra la distribución de ciudades por país. Los países con más ciudades son: Estados Unidos con 94, Brasil con 85, Turquía con 63, Rusia con 42, Indonesia con 39, México con 37, Japón con 26, Malasia con 19, Tailandia con 14 y Gran Bretaña con 14. Nuevamente, la variedad cultural y geográfica del conjunto es notable. En estas ciudades se ubican el 63 % de los POI totales (7,026,688), y el 76 % del total de check-ins (68,356,896). En otras palabras, más de tres cuartas partes del total de check-ins de la base de datos se realizaron en áreas urbanas con más de 10,000 check-ins.

Otra representación de estas 632 ciudades se muestra en la figura 3.13; en ella, cada ciudad es un círculo cuyo color representa el número de check-ins que ocurrieron dentro de sus fronteras. Esta representación nos permite, una vez más, dimensionar la importancia de este conjunto en términos de cobertura y diversidad de ciudades que incluye.

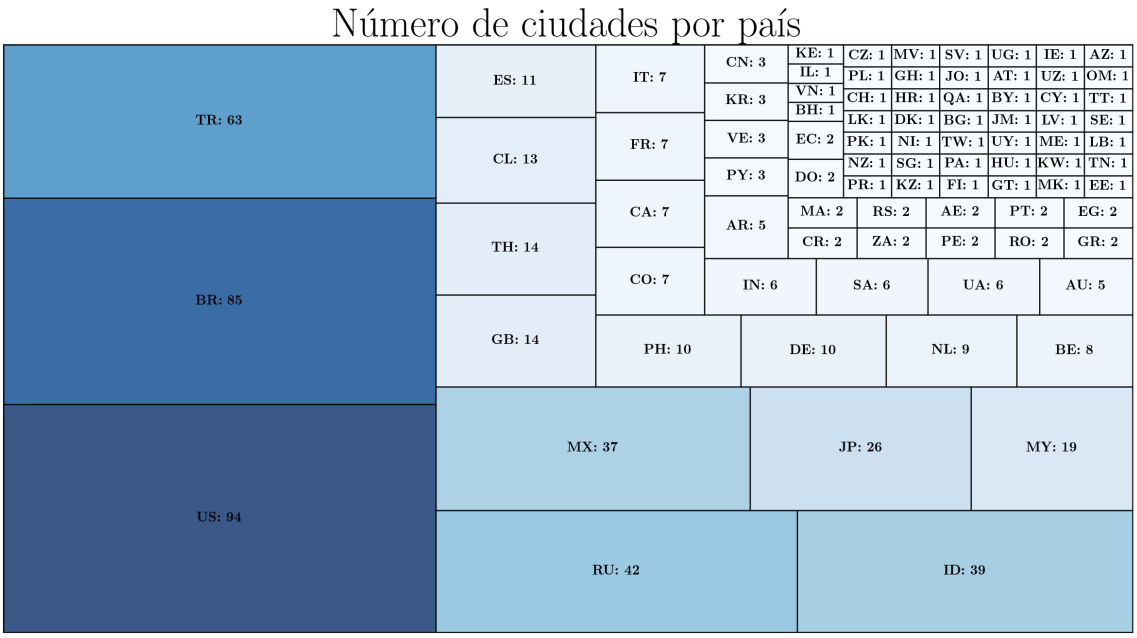


Figura 3.12: **Áreas Urbanas Funcionales con más de 10,000 check-ins de Foursquare por país**. Después de seleccionar las ciudades con más de 10,000 check-ins de la base de datos, nos quedamos con 632 áreas urbanas funcionales, distribuidas en 87 países. Destacan Estados Unidos de América, Brasil y Turquía, sin embargo alrededor de la mitad de los 87 países tienen un área urbana funcional que cumple con la condición considerada aquí. Esto hace que la muestra de ciudades sea muy rica e interesante para su análisis.

Con los conjuntos $\mathcal{D}_{4S}$ y $\mathcal{D}_{POI}$ se pueden estudiar, por fin, las propiedades espaciales del entorno urbano, la dinámica espacio-temporal de la actividad humana y características urbanas en términos de las categorías de los puntos de interés ubicados en las áreas urbanas funcionales. Cerraremos este capítulo aproximándonos a la heterogeneidad espacial de los puntos de interés y los check-ins de Foursquare y a las categorías de los POIs, dejando la dinámica temporal de la actividad humana para el siguiente capítulo.

En la figura 3.14 se resume la información espacial de POIs y check-ins en las 12 ciudades con más datos de $\mathcal{C}$. Se optó por una representación, alternativa a las anteriores, en la que a cada POI le corresponde un punto y el número de check-ins realizados en ese POI se indica con el color. Se eligió una paleta que cubre todo el espectro desde el azul para un solo check-in hasta rojo para aquellos POIs con $10^{4.5} \simeq 31623$ check-ins o más. De esta representación se pueden sacar dos conclusiones principales. La primera es la heterogeneidad que presentan las ciudades en términos de la actividad en Foursquare: no hay un patrón espacial claro, algunas ciudades tienen un centro de mucha actividad y periferias de menor actividad, pero otras (Estambul, Ciudad de México, Esmirna, Tokio) con múltiples puntos de actividad distribuidos en el espacio. La otra conclusión, más importante para este trabajo, es la emergencia de un entorno complejo a partir de los check-ins de Foursquare: se encuentra una gran mayoría de sitios con muy baja actividad (uno o unos cuantos check-ins), algunos

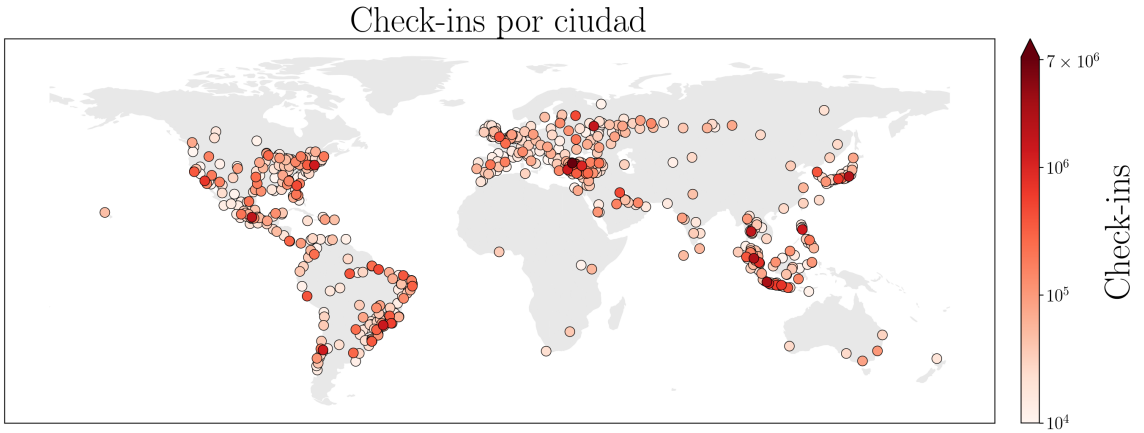


Figura 3.13: **Check-ins en áreas urbanas funcionales**. Se muestran las 632 ciudades que contienen más de 10,000 check-ins con el número de check-ins indicado en el color del círculo. Se aprecian tres grandes regiones: América, Europa y Asia con muchos círculos rojos. Pero se ven círculos blancos en varias regiones como África, India y Oceanía. Al haberse tomado como unidad las áreas urbanas funcionales se pueden estudiar muchas regiones que no suelen incluirse en este tipo de fuentes.

lugares con actividad mediana (decenas y centenas de check-ins), y unos pocos sitios con actividad extremadamente alta (miles y decenas de miles de check-ins); esta es una propiedad típica en sistemas complejos, como se mencionó en el capítulo 1, y cuyo análisis estadístico comprueba un comportamiento tipo ley de potencia.

Finalmente, en las tablas 3.10 y 3.11 se muestran las 10 categorías más comunes, para las cinco ciudades con más datos, en términos del número de POIs y del número de check-ins, respectivamente.

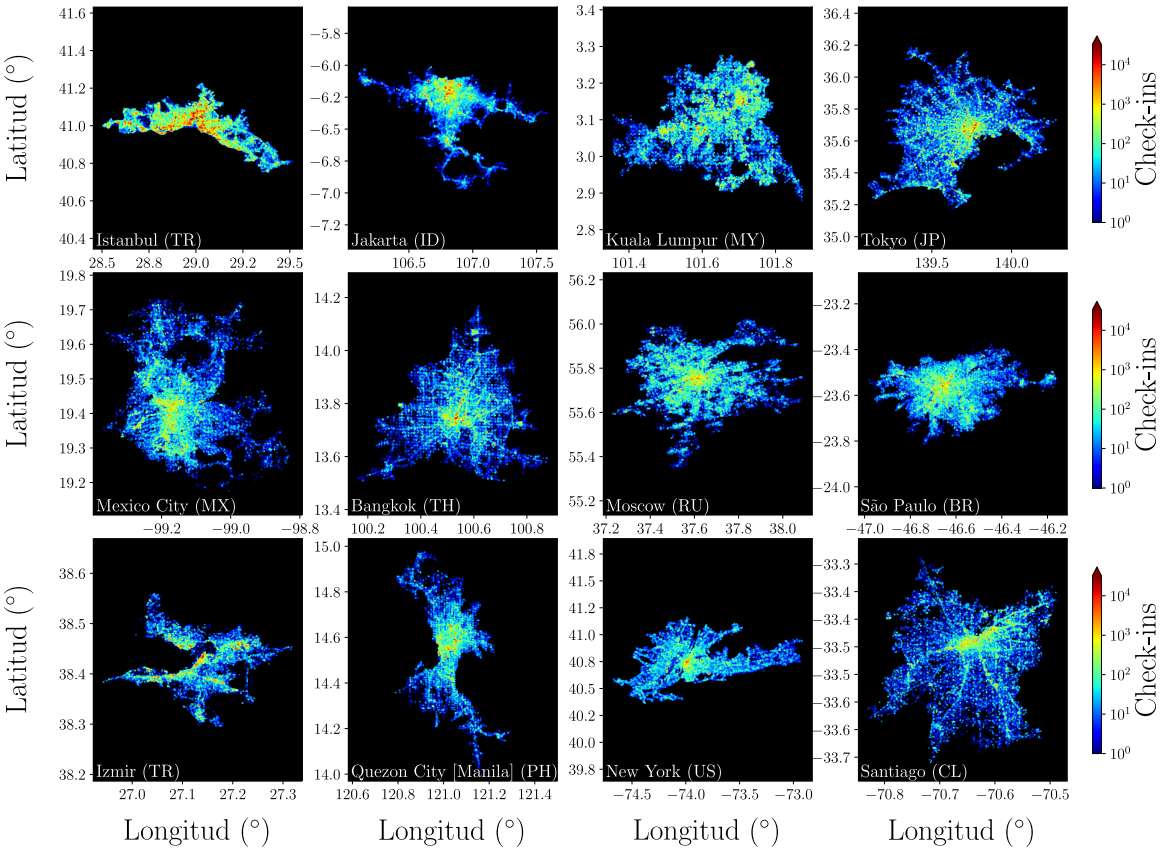


Figura 3.14: **Check-ins por punto de iterés en las 12 ciudades con más registros**. Se muestran los puntos de interés y el número de check-ins registrados en ellos, para las doce ciudades con más datos. Cada punto es un POI y se dibuja en sus coordenadas, el número de check-ins se representa en el color del punto a partir de una paleta que cubre todo el espectro de colores; azul para 1 check-in y rojo para $10^{4.5} \simeq 31623$ check-ins o más. Estos valores fueron elegidos convenientemente para una visualización clara. Se muestran las primeras doce ciudades de la tabla 3.9.

	Jakarta (ID)		**Istanbul (TR)**		**Kuala Lumpur (MY)**		**Tokyo (JP)**		**Bangkok (TH)**	
	Categoría	POIs	Categoría	POIs	Categoría	POIs	Categoría	POIs	Categoría	POIs
1	Office	15610	Residential Building (Apartment / Condo)	22820	Malaysian Restaurant	7931	Ramen / Noodle House	9854	Thai Restaurant	8422
2	Indonesian Restaurant	15376	Office	14092	Office	6929	Convenience Store	9655	Residential Building (Apartment / Condo)	7138
3	Building	12805	Building	9061	Residential Building (Apartment / Condo)	6791	Sake Bar	5515	Office	5120
4	Asian Restaurant	9995	Café	9009	Asian Restaurant	6376	Café	5130	Building	5101
5	College Classroom	7442	Coworking Space	7222	Building	6271	Chinese Restaurant	4919	Ramen / Noodle House	4141
6	Food Truck	6648	Turkish Restaurant	6681	Chinese Restaurant	5711	Italian Restaurant	4087	Coffee Shop	3682
7	Residential Building (Apartment / Condo)	6645	Salon / Barbershop	5303	College Classroom	4339	Grocery Store	3679	Asian Restaurant	3580
8	Road	6445	Factory	5231	Automotive Shop	3854	Bar	3658	College Classroom	3481
9	Mosque	5868	Housing Development	4589	Café	3072	Restaurant	3239	Road	2696
10	Café	5647	Restaurant	4449	Road	2287	Office	3019	Bank	2639

Tabla 3.10: **Categorías por cantidad de POIs por ciudad**. Diez categorías de Foursquare más comunes en las 5 ciudades con más puntos de interés. Las categorías se ordenan en función del número de POIs de cada una.

	Istanbul (TR)		**Jakarta (ID)**		**Kuala Lumpur (MY)**		**Tokyo (JP)**		**Mexico City (MX)**	
	Categoría	check-ins	Categoría	check-ins	Categoría	check-ins	Categoría	check-ins	Categoría	check-ins
1	Other Great Outdoors	672613	Home (private)	341758	Mall	263850	Train Station	754617	Office	104028
2	Mall	468978	Mall	306516	Home (private)	211033	Subway	96285	Mall	90545
3	Café	464858	Office	89092	Malaysian Restaurant	150655	Ramen / Noodle House	93465	Home (private)	81073
4	Home (private)	366181	Building	74538	Residential Building (Apartment / Condo)	77222	Japanese Restaurant	71831	Mexican Restaurant	59768
5	Neighborhood	345430	Indonesian Restaurant	64437	Office	63808	Convenience Store	59435	University	57638
6	University	334126	University	58435	Café	59253	Mall	57144	Coffee Shop	53746
7	Restaurant	243223	Church	58076	Asian Restaurant	58640	Grocery Store	43960	Gym	48628
8	Residential Building (Apartment / Condo)	243101	Multiplex	57103	Building	58419	Coffee Shop	38566	Building	37761
9	Café	194575	High School	48801	Indian Restaurant	57192	Electronics Store	38482	Bar	36002
10	Office	193160	Airport	48687	Chinese Restaurant	46551	Platform	31549	Residential Building (Apartment / Condo)	35108

Tabla 3.11: **Categorías por cantidad de check-ins por ciudad**. Diez categorías de Foursquare más comunes en las 5 ciudades con más puntos de interés. Las categorías se ordenan en función del número de check-ins realizados en POIs de cada una.

4 Patrones temporales de comportamiento humano en ciudades

4.1. Introducción

En este capítulo se reproducen y amplían los resultados publicados en el artículo "Temporal visitation patterns of points of interest in cities on a planetary scale: A network science and machine learning approach" [1]. Primero se presentan las regularidades temporales de la actividad humana en las ciudades. Para ello, se filtraron los check-ins realizados en los POIs dentro de cada área urbana funcional, seleccionando las 632 que contienen más de 10000 check-ins. Se trabajó con los check-ins de estas 632 ciudades, agrupándolos por hora del día, día de la semana y agregando todas las semanas de los 22 meses de observación. De esta manera, se generó una huella temporal para cada ciudad, lo que permitió obtener un patrón temporal específico. Posteriormente, se analizan las diferencias en los patrones temporales de las 632 regiones. A partir de estas diferencias y una medida de la similitud entre cualesquiera dos ciudades, se agruparon las ciudades en macroregiones mediante el uso de métodos de la teoría de redes, la teoría de la información, la detección de comunidades y agrupamiento aglomerativo. Finalmente, se interpretan los resultados a la luz de las características socio-culturales de las regiones encontradas. En resumen, este capítulo presenta un análisis detallado de la actividad humana en las ciudades, enfocándose en la regularidad temporal de la misma y en la agrupación de las ciudades en macroregiones, lo que permite obtener una visión general de la distribución geográfica y características de la actividad humana en las distintas áreas urbanas funcionales

4.2. Dimensión temporal de la interacción personas-lugares

Las ciudades pueden entenderse como interacciones entre personas y sitios. Como se dijo en el capítulo 1 dedicado a la ciencia de las ciudades, es más conveniente entender a éstas como interacciones entre distintos elementos, en lugar de solo regiones geográficas, y los sitios y las personas son elementos más destacados. En nuestro caso de interés, estas interacciones entre personas y sitios se traducen en los check-ins de Foursquare. Una característica importante de los datos obtenidos de estos check-ins

en redes sociales basadas en ubicaciones es que nos permiten estudiar la dimensión temporal de la actividad de grandes cantidades de personas en las ciudades. Este tema es relevante tanto para la comprensión teórica de las ciudades como sistemas complejos, como para las aplicaciones concretas en planeación urbana o infraestructura, ya que es necesario conocer los horarios de actividad de las personas en las ciudades, las diferencias en horarios por regiones o sectores de la población, el tipo de actividades que se realizan en qué horarios, entre otras variables.

Se necesitan métodos para estudiar a las ciudades como un todo [169]. Encontrar los patrones emergentes de la actividad humana es clave para atender los grandes problemas que implica la vida en las ciudades, tales como los tiempos de traslado, el tráfico, las multitudes y sus consecuencias tanto positivas como negativas, el suministro de energía y la calidad del aire, entre otros. En particular, las características temporales de la actividad humana son importantes porque de ellas derivan otros fenómenos de gran relevancia como las redes de contacto entre personas [88] o los efectos de ráfaga o *burstiness* [170], cruciales para comprender fenómenos de propagación de información, enfermedades, etc.

La dimensión temporal de la actividad humana en ciudades ha sido investigada desde varios enfoques y a distintas escalas [171]. Por ejemplo, investigaciones enmarcadas en la biología han investigado patrones temporales individuales a partir de ritmos circadianos y ciclos de sueño; otras, desde la geografía, se han preocupado por patrones temporales colectivos de la actividades sociales en regiones geográficas [171–173]. Gracias a esto, se sabe que el conjunto de factores que influyen en los patrones temporales de las personas es grande, y entre ellos están los factores ambientales como el clima, la intensidad y la cantidad de radiación solar [92, 174], así como culturales como costumbres o hábitos colectivos (horarios de comida, de descanso, de convivencia, etc.). Estos factores que influyen en los patrones temporales individuales también importan para los patrones temporales de ciudades y regiones [171].

Los patrones temporales colectivos se han analizado a partir de muchas muy variadas fuentes de información. Entre ellas se puede encontrar el estudio de correos electrónicos [170, 175], registros de llamadas telefónicas [92, 176], datos de fuentes oficiales sobre crímenes o accidentes [123] o, como es nuestro caso, análisis de actividad en Puntos de Interés y Redes Sociales Basadas en Ubicaciones [171, 172]. Si bien hay muchos trabajos que abordan el tema de la emergencia de patrones temporales en ciudades, nuestra contribución es presentar el estudio de las huellas temporales de la actividad humana en 632 ciudades de 87 países para un análisis comparativo.

4.3. Las huellas temporales

Como se ha reportado en muchos sistemas complejos, los elementos de las ciudades se sincronizan generando patrones que algunos grupos han llamado los latidos de las ciudades [123]. Estos latidos están moldeados por al menos dos fuerzas: los ciclos biológicos circadianos de las personas, que impacta fuertemente en los ciclos de sueño-vigilia [92], y factores sociales y económicos que, también se ha dicho antes,

tienen una importancia capital en el surgimiento y evolución de las ciudades. Otros factores ambientales geográficos como las horas de luz en las distintas épocas del año, la temperatura y otros aspectos climáticos también pueden ser relevantes.

Un primer paso en el estudio de ciudades es identificar las similitudes de los sistemas de interés. Luego se puede centrar la atención en las particularidades de cada sistema individual. La existencia de mecanismos comunes a todas las ciudades es una premisa importante en el estudio de la complejidad urbana [14] y en los enfoques de la ciencia de las ciudades explicados en el capítulo 1. En ese sentido, la actividad humana en general está organizada alrededor de ciclos diarios que gobiernan todas las interacciones sociales y económicas [123]. Al mismo tiempo, otros ciclos semanales operan haciendo contrastar la actividad humana entre distintos días, por ejemplo lunes y viernes, a lo que ocurre entre semana y los fines de semana [86, 123]. Esto, más que a determinaciones biológicas, se debe a costumbres y tradiciones sociales e históricas. Éstas no se encuentran presentes en la misma medida en todos los países del mundo, y es sólo un ejemplo de lo que puede hacer distinta a una ciudad de otra. Siguiendo estas pautas, se utilizaron los check-ins de la base de datos para analizar las similitudes y diferencias en nuestro sistema de 632 ciudades.

Como se explicó en el apartado 3.5, los check-ins se agruparon a partir de las áreas urbanas funcionales, de las cuales 632 tiene más de 10,000 check-ins. En cada caso, se tiene información de gran cantidad de interacciones que vinculan a un usuario con un punto de interés y la hora exacta en la que ocurrieron. El comportamiento temporal se puede analizar a distintas escalas: por horas del día, por horas de la semana, por días de la semana, mes, etc. A continuación se presentan los resultados al agrupar los check-ins por hora de la semana. Esto significa que se agruparon todos los check-ins que fueron realizados cierto día de la semana y a cierta hora, con independencia del mes y año, y se generó un histograma de frecuencias para 168 horas (las horas de una semana completa). Consideramos estos histogramas como una especie de huella temporal de la actividad humana en la región de estudio. Haciendo esto, claros patrones circadianos emergen en todas las regiones. En la figura 4.1 se muestran los patrones de comportamiento de las 16 regiones con más check-ins de la base de datos. Algunas regularidades saltan a la vista: hay un tráfico de check-ins bajo durante las noches y alto durante el día. Esto coincide con el comportamiento reportado por Yang y su equipo [153] para un subconjunto de estos mismos datos a escala global. A este nivel de análisis (área urbana funcional) el patrón persiste y se puede comparar con lo encontrado en otros fenómenos como el crimen y los accidentes [123] y las horas de descanso [92]. Sin embargo, algunas diferencias en los patrones resultan notables: la parte del día que concentra la mayor actividad, el número de máximos locales de actividad por día, el cambio en los patrones de actividad entre semana y en fin de semana; en breve, la forma de la huella temporal varía de ciudad a ciudad.

En resumen, los check-ins manifiestan una propiedad que emerge de millones de personas que sincronizan sus rutinas diarias y semanales. Esta huella temporal, o latido, se ha encontrado en muchos otros fenómenos [78, 86, 123, 142, 143, 148, 153, 177–180].

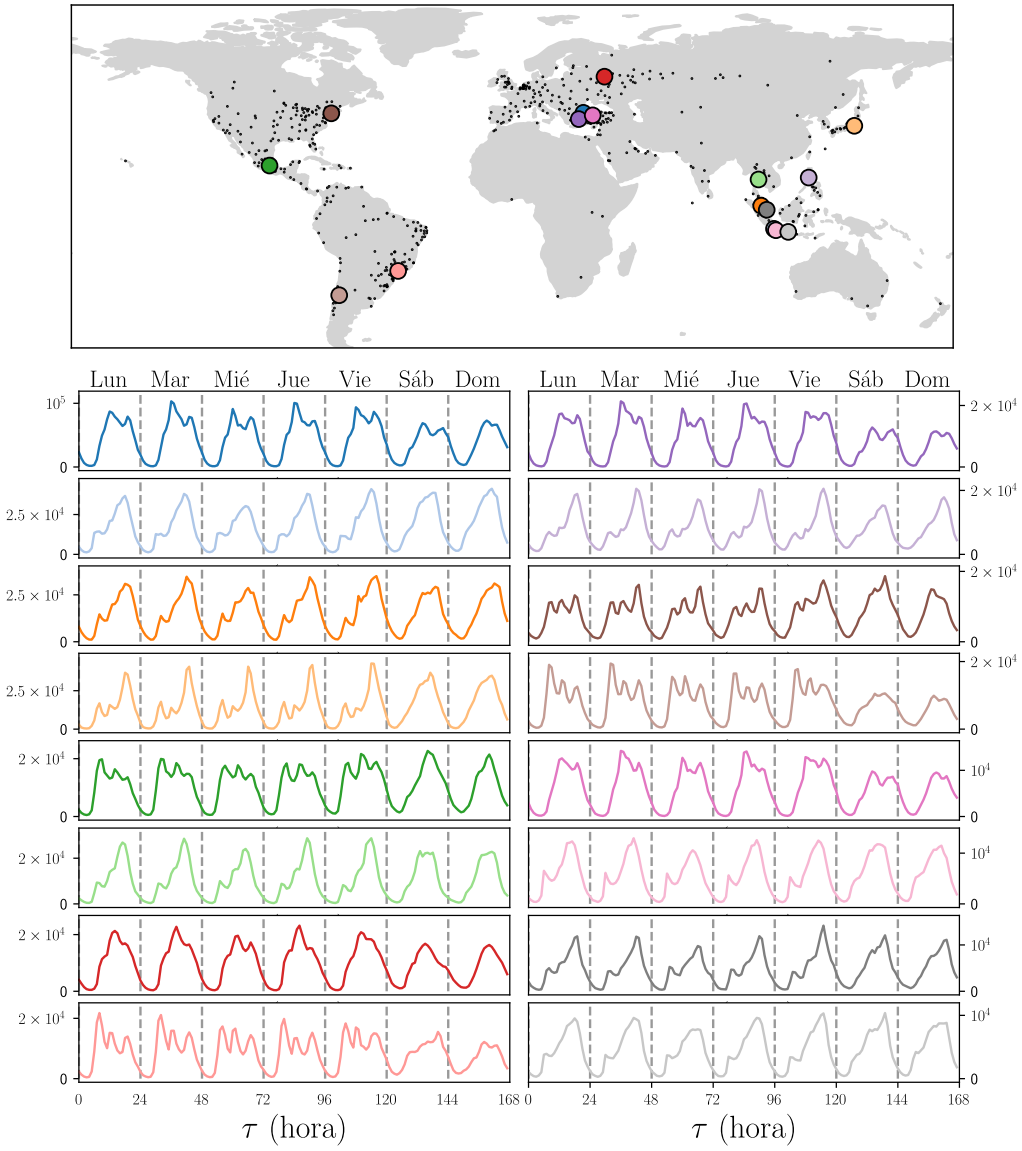


Figura 4.1: Histogramas de frecuencias de check-ins por hora en las 16 ciudades con más check-ins. Los check-ins realizados en los puntos de interés de las 16 ciudades con más datos, según la tabla 3.9, se agruparon por el día y la hora local. Las cuentas corresponden al número de check-ins por cada hora de la semana, desde la primera hora del lunes hasta la última del domingo. Como se puede observar, las frecuencias tienen valores distintos debido a que en cada región se hizo una cantidad de check-ins distinta. Sin embargo, algunas regularidades saltan a la vista: pocos check-ins durante la noche y muchos durante el día, formas similares de lunes a viernes, etc. Las diferencias serán el objeto de los siguientes apartados.

4.3.1. Similitudes y diferencias de la huellas

Se procedió a comparar todas las ciudades seleccionadas. Para poder hacerlo se necesitaba una huella independiente de las dimensiones de cada sistema, en este caso reflejado en el número de check-ins. Esto se hizo dividiendo cada histograma de frecuencia de check-ins por hora entre el número total de registros realizados en esa región. Se obtuvo así una distribución de probabilidad: la probabilidad de que un check-in tomado al azar haya sido realizado a una hora τ, con τ de 1 a 168. Esta distribución de probabilidad es, por construcción, independiente del número de check-ins, usuarios y POIs en cada ciudad, permitiendo un análisis comparativo entre ciudades.

En la figura 4.2a) se grafican las distribuciones de las 632 ciudades. Esto permite verificar, en todas, el contraste entre la baja actividad nocturna y la alta actividad diurna que se conoce como ciclo sueño-vigía [92], una de las componentes de los ritmos circadianos. En todas emergen patrones circadianos al agrupar los check-ins por su hora local de ocurrencia. Esta tendencia común se hace evidente mediante el promedio de actividad sobre la muestra de las 632 ciudades mostrada con una línea negra en la figura 4.2b), lo que sería el comportamiento promedio en torno al que cada ciudad fluctúa. Es claro que estas fluctuaciones son pequeñas durante la noche, pero durante el día es evidente que la media no es una buena aproximación para casi ninguna ciudad. No solo eso, sino que se alcanzan a ver desviaciones de todo tipo: algunas se alejan considerablemente en las primeras horas de la mañana, otras por la noche y otras al medio día. Para aclarar esto, en la figura 4.2b) se muestra también la desviación estándar junto con la media. Así como se hizo para la media, por cada hora de la semana τ se toman como muestra los valores de la probabilidad $P_i(\tau)$ para todas las ciudades, y de esa muestra se calcula la desviación estándar. Se tiene entonces una desviación estándar para cada τ, que en la gráfica aparece como un área gris alrededor de la distribución promedio. No sólo se confirma que la dispersión de los valores es bajo en las noches y grande en el día, sino que se identifican ligeras variaciones entre los días y, además, dos dinámicas completamente distintas para días de la semana y fines de semana [86, 123].

4.4. Comparación entre ciudades

Este trabajo combina el análisis temporal de la actividad humana y la similitud entre ciudades, fenómeno que se ha abordado previamente tomando otras dimensiones de la vida en las ciudades [140, 181–186].

4.4.1. Medida de la similitud de los patrones temporales

Se necesita algún criterio para dar cuenta no sólo de los grandes contrastes que hay entre ciudades sino también de las diferencias más sutiles entre ellas. La forma en que en cada ciudad se modifica el comportamiento durante los fines de semana y los días de la semana y, en este último caso, diferencias entre lo que ocurre los lunes

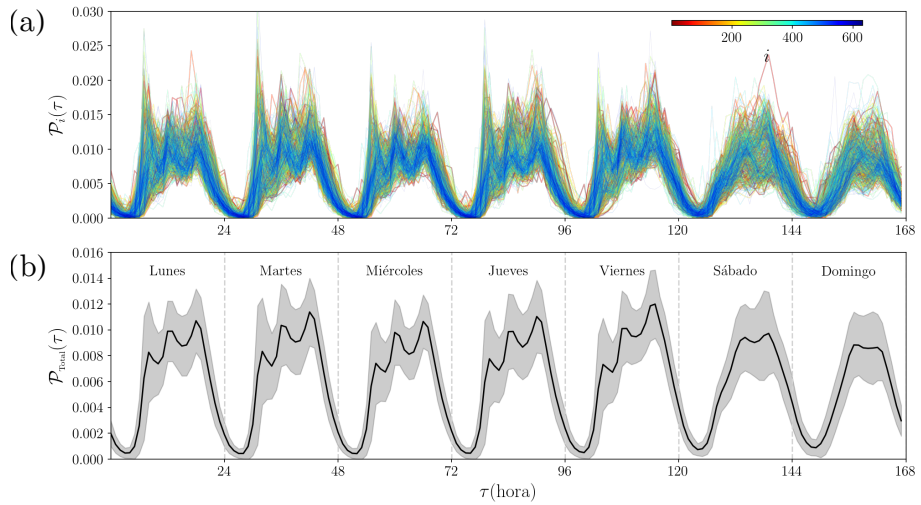


Figura 4.2: Análisis temporal de los check-ins en 632 ciudades. (a) Probabilidad $\mathcal{P}_i(\tau)$ de check-ins al tiempo τ en $N = 632$ ciudades denotadas por $i = 1, 2, \ldots, N$. Todos los check-ins en cada ciudad se agruparon por la hora local a la que fueron hechos. El conteo de las frecuencias se normalizó sobre la semana para obtener cada $\mathcal{P}_i(\tau)$. El tiempo τ puede tomar valores de 1 (la primera hora del lunes) a 168 (la última hora del domingo), y los resultados para cada ciudad i (codificada en la barra de color) se presenta con líneas continuas. (b) La probabilidad $\mathcal{P}_{\text{Total}}(\tau)$ obtenida para todos los check-ins en las 632 ciudades se grafica con una línea continua. En este caso, la desviación estándar $\sigma(\tau)$ para todos los check-ins en todas las ciudades en (a) se evaluó al tiempo τ y define las variaciones de estas probabilidades. Los resultados se muestran como una región definida para $\mathcal{P}_{\text{Total}}(\tau) \pm \sigma(\tau)$ presentada como una región gris.

y los viernes, por poner algunos ejemplos ejemplo. Además, quisiéramos distinguir aquellas ciudades para las que la actividad se concentra en la mañana de otras en la se concentra en la tarde. Para lograr esto se optó por un criterio que permite comparar distribuciones de probabilidad por pares: una versión simétrica de la divergencia de Kulback-Leibler [187].

La variedad de resultados observados para las probabilidades $\mathcal{P}i(\tau)$ en la Figura 4.2a) motiva la exploración de un criterio para comparar la actividad temporal entre dos ciudades particulares i, j. Para ello, utilizamos la divergencia de Kullback-Leibler, también conocida como entropía relativa, definida por [187, 188]

$$\mathcal{D}_{KL}(i, j) = \sum_{\tau} \mathcal{P}_i(\tau) \log \left[\frac{\mathcal{P}_i(\tau)}{\mathcal{P}_j(\tau)} \right], \tag{4.1}$$

donde la suma en el tiempo τ va desde 1 hasta 168 (todas las horas en una semana) y $i, j = 1, 2, \ldots, N$. La divergencia de Kullback-Leibler satisface nuestro interés en comparar las distribuciones temporales de los check-ins, pero su definición no genera una medida simétrica ya que $\mathcal{D}_{KL}(i, j)$ es diferente de $\mathcal{D}_{KL}(j, i)$. Esto provocaría que incluso si una ciudad i se determina como similar a otra ciudad j, j no necesariamente sería similar a i. Sin embargo, el promedio de la divergencia de Kullback-Leibler entre

los pares (i, j) y (j, i)

$$\mathcal{D}_{KLS}(i,j) \equiv \frac{\mathcal{D}_{KL}(i,j) + \mathcal{D}_{KL}(j,i)}{2} \tag{4.2}$$

es simétrico[1]. Ésta última expresión toma un valor cercano a cero si las distribuciones son similares, cero si son idénticas, y, para el caso general, no tiene una cota superior. Entonces, esta cantidad simétrica es adecuada para describir la similitud entre las distribuciones temporales de las ciudades.

Se obtuvo D_{KLS} para todas las parejas de ciudades en nuestra muestra. En la figura 4.3a) se presenta la densidad de probabilidad de los valores D_{KLS} entre ciudades mientras que en el panel interior se muestra la matriz $D_{KLS}(i, j)$ para todas las parejas de ciudades. El intervalo entre 0 y 0.6078, el máximo valor entre todas las parejas, se mapeó con una paleta de colores divergente; los colores rojos representan un valor de D_{KLS} pequeño, lo que significa una mayor similitud entre ciudades, y los colores blancos y azules representan menor similitud. La misma paleta de colores se usó para el área debajo de la densidad de probabilidad para completar una imagen de cómo se comporta nuestro sistema en términos de esta métrica de similitud. La mayoría de las parejas de ciudades tienen un valor de D_{KLS} alrededor de 0.1.

De un análisis específico de las parejas se obtiene que el valor máximo para dos ciudades distintas, 0.6078, se da entre la ciudad brasileña de Sapucaia do Sul y Meru, en Malaysia (fig. 4.3b). En el extremo opuesto, el mínimo valor alcanzado es 0.0023 entre Ankara e Izmir, ambas en Turquía. En la figura 4.3c) se comparan ambas distribuciones; el parecido entre estas dos ciudades es notable considerando que están a más de 520 km de distancia y unas 7 horas de viaje en auto. Adicionalmente, una revisión de los datos arroja que de los 82,285 usuarios con actividad en Ankara y los 90,923 usuarios con actividad en Izmir, sólo 11,141 tuvieron actividad en ambas ciudades; esto representa sólo el 6.87 % del total de usuarios con actividad en alguna de las dos ciudades, por lo que la similitud en sus patrones no se explica por usuarios en común sino por comportamientos urbanos comunes.

4.4.2. Red de similitud de patrones temporales

Partiendo de que dos ciudades cuyo valor D_{KLS} es pequeño tienen patrones temporales similares, se utilizó la teoría de redes para el análisis de las similitud entre ellas. Se definió una red en la que los nodos representan a ciudades y los enlaces la relación de similitud entre sus patrones temporales. Los nodos de dos ciudades lo suficientemente parecidas se enlazan, mientras que si las ciudades no son tan parecidas en términos de su huella temporal, no hay enlace entre ellas; sin embargo es necesario definir lo que

[1]Otras versiones simétricas de la divergencia de Kullback-Leibler se reportan en [189], incluida la divergencia de Jensen-Shannon

$$\mathcal{D}_{JS}(i,j) = \frac{1}{2}\mathcal{D}_{KL}(i,m) + \frac{1}{2}\mathcal{D}_{KL}(j,m)$$

con $m = \frac{1}{2}(i + j)$ el promedio de las dos distribuciones.

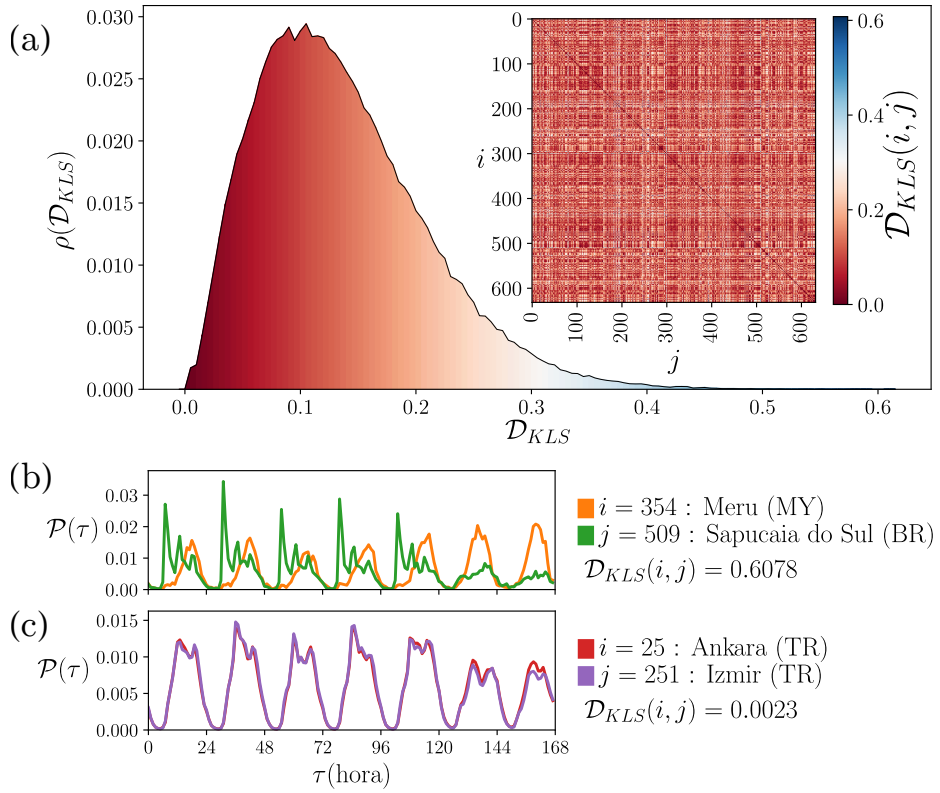


Figura 4.3: Comparación de actividad temporal entre ciudades. (a) Análisis estadístico de las distancias simétricas de Kullback-Leibler $\mathcal{D}_{KLS}$ de la comparación de las distribuciones temporales de check-ins en parejas de ciudades. Los valores se obtienen de la ecuación (4.2) para todas las parejas de ciudades $i,\ j\ =\ 1,\ldots,\ 632$. La densidad de probabilidad $\rho(\mathcal{D}_{KLS})$ se obtiene utilizando conteos de intervalos con $\Delta\mathcal{D}_{KLS}\ =\ 0.005$. La matriz con todos los elementos $\mathcal{D}_{KLS}(i,j)$ se presenta como un panel interior. (b) Las dos densidades de probabilidad más diferentes del conjunto son de las ciudades Sapucaia do Sul en Brasil y Meru en Malasia, con $\mathcal{D}_{KLS} = 0.6078$; el valor máximo encontrado. (c) Las dos ciudades más similares de nuestra muestra son Ankara e Izmir, ambas en Turquía, con $\mathcal{D}_{KLS} = 0.0023$, el valor mínimo no nulo.

se considera suficientemente parecidas. Para ello es necesario un umbral por debajo del cual los valores de D_{KLS} son tan pequeños como para considerar a las ciudades similares, y por encima no. La matriz de adyacencia de la red se define a partir de la siguiente condición:

$$A_{ij}(H) = \begin{cases} 1 & \text{si } \mathcal{D}_{KLS}(i,j) < H \\ 0 & \text{si no} \end{cases} \tag{4.3}$$

que depende de H, el valor del umbral. Considerar a esta una matriz de adyacencia significa que las ciudades i, j se enlazan en la red si $D_{KLS}(i,j) < H$ y en caso contrario, no hay enlace entre ellas. De la simetría de D_{KLS} se sigue la simetría de A_{ij}, así que se trata de una red simple.

Con esto en mente, se hizo el análisis de las propiedades de la red en función del valor de este umbral y los resultados se muestran en la figura 4.4. La representación

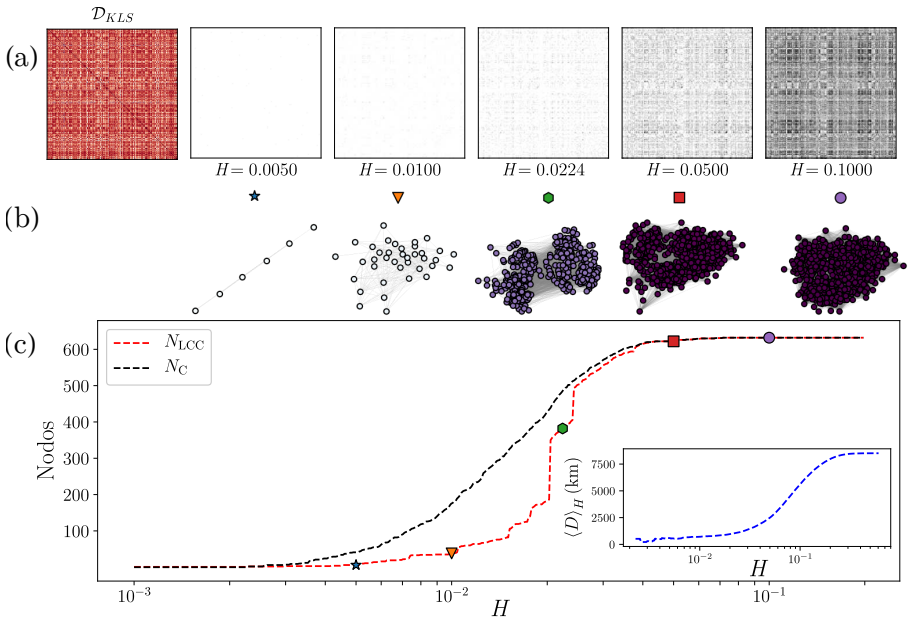


Figura 4.4: **Redes de similitud en función del umbral H.** (a) Matrices de adyacencia $\mathbf{A}(H)$ con elementos dados por la ecuación (4.3) utilizando los valores $H \in \{0.005, 0.01, 0.02, 0.04, 0.0723\}$. Las entradas binarias $A_{ij}(H)$ se representan en blanco para 0 y en negro para 1. (b) Representación gráfica de la más grande componente conectada de la red correspondiente a las matrices de adyacencia de (a). (c) Número de nodos en la más grande componente conectada N_{LCC} y número de nodos enlazados con otro N_C, para cada red como función de H en el intervalo $0.001 \leq H \leq 0.32$. El panel interior muestra $\langle D \rangle_H$, el promedio de distancia geográfica entre nodos adyacentes calculado utilizando la distancia geográfica entre las ciudades como función de H.

matricial de $D_{KLS}(i, j)$ de la figura 4.3 se convierte en una matriz de unos y ceros con la condición 4.3 de modo que los colores se reducen a blanco para valores encima del umbral y negros para valores por debajo del umbral. Cada matriz de unos y ceros (Fig. 4.4a) se convierte en una red cuya más grande componente conectada se muestra en la figura 4.4b. Resulta obvio que para un valor de H lo suficientemente cercano a 0, en este caso $H < 0.0023$, la red está formada únicamente por nodos aislados, es decir, no habría enlaces. Es igual de obvio que para valores de H por encima de 0.6078 habría un enlace entre todas las parejas de nodos, se trata de una red completamente conectada. La transición entre una situación y otra es la que dará lugar a una elección conveniente de H que genere una red que recoja la naturaleza de la similitud entre las ciudades que estamos analizando. En la figura 4.4c) se muestra la evolución en ese intervalo de H, desde los nodos aislados hasta una red compuesta por únicamente una componente. En el eje horizontal se representa el valor de H en una escala logarítmica y en el eje vertical el número de nodos de la más grande componente conectada (N_{LCC}) y el número de nodos enlazados con al menos otro nodo (N_C). Esta representación permite ver que el número de nodos con conexiones (N_C) crece suavemente, resultado que es coherente con los presentados por Batty en redes de flujos en ciudades [2]. El tamaño de la más grande componente

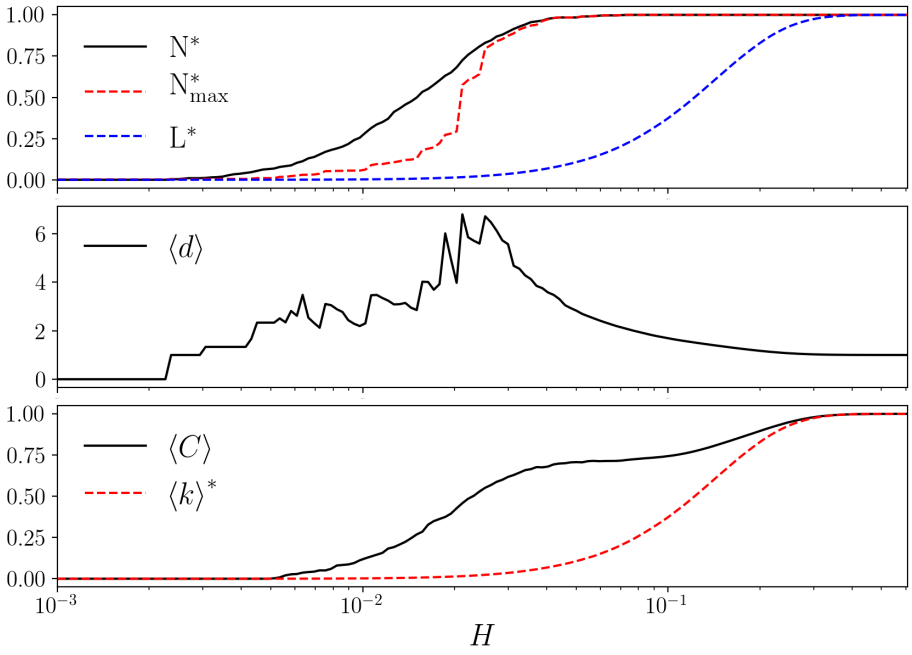


Figura 4.5: **Evolución de las propiedades de la red de similitud entre ciudades**. Se muestra la evolución de varias propiedades de la red como función de H. En el panel superior se muestra la fracción de nodos conectados N^* en negro, la fracción de nodos pertenecientes a la más grande componente conectada $N^*_{\max}$ en rojo y la fracción de enlaces de la red L^* en azul. En el panel central se muestra la evolución de la distancia promedio entre nodos, $\langle d \rangle$, definida en la ecuación 2.14. En el panel inferior se muestran el coeficiente de agrupamiento promedio, definido en la ecuación 2.19, y el grado promedio normalizado $\langle k \rangle^*$.

conectada no evoluciona suavemente debido a que su crecimiento es fruto de la unión de otras componentes de distintos tamaños. También se observa que hay un intervalo alrededor de 10^{-2} en el que conviven varias componentes de distintos tamaños que, mientras crece H, son absorbidas por la componente gigante. Finalmente, nos interesa hacer notar que la componente gigante, cuyo tamaño se representa por una línea roja punteada, contiene más del 90 % de los nodos en $H = 0.031$, el 99 % en $H = 0.052$, y se vuelve única en $H = 0.072$. Las propiedades de la máxima componente sobre este inervalo ($H \in (.031, 0.072)$) se concentrarán los siguientes análisis.

En la figura 4.5 se reporta la evolución de las propiedades de la red de similitud entre ciudades. Se calcularon las fracciones de nodos conectados, N^*, y pertenecientes a la más grande componente conectada, $N^*_{\max}$, que se obtuvieron dividiendo en cada caso el número de nodos entre el tamaño del sistema, es decir, 632. Esto da valores entre cero y uno que se muestran en el panel superior (líneas negra y azul). Una vez más se calculó la fracción de enlaces L^*, dividiendo el número de enlaces que tiene la red para cada valor de H entre las parejas de nodos que conforman el sistema, es decir, $N(N-1)/2$. Esto también da un valor entre 0 y 1 que se muestra en el mismo panel (línea roja). Estas tres propiedades varían entre 0 y 1 conforme H aumenta, pero L^* llega al 1 para valores mayores, entre 0.3 y 0.4. Esto se explica porque, si bien

todos los nodos pueden estar conectados y la red formada por una única componente, tal situación puede lograrse aún con fracciones relativamente pequeñas de enlaces. De hecho, esta es la propiedad de "escases" de muchos sistemas reales definida en la ecuación 2.13.

Se calculó la distancia promedio de la red (ec. 2.16) como función de H. El resultado se muestra en el panel central de la figura 4.5. En ella puede observarse que antes de todos los nodos se conecten en una única componente, la distancia promedio alcanza su máximo valor de 7. Cuando la red se conecta por completo, alrededor de $H = 0.7$, disminuye a un valor cercano a 3. $\langle d \rangle$ continúa disminuyendo hasta alcanzar el valor de 1 cuando todos los nodos están conectados entre sí. Podemos concluir que, en lo que respecta a la distancia entre nodos, la red cumple la propiedad de mundo pequeño (ec. 2.27) en el intervalo $H \in (.031, 0.072)$. Para completar este análisis, en el panel inferior de la figura se muestran tanto el coeficiente de agrupamiento promedio (ec. 2.19) como el grado promedio normalizado. Puede observarse que el coeficiente $\langle C \rangle$ se mantiene superior a 0.5, que es un valor grande comparado con el predicho por el modelo de Erdös-Renyi para un sistema de este tamaño. Con esto se completa el análisis de las propiedades de la red introducidas en el capítulo 2. A partir de la evolución de estas propiedades se eligió el valor $H = 0.0723$ en el que no hay cambios abruptos en las métricas del sistema y es el valor en torno al cual todos los elementos del sistema se incorporan a una única componente.

4.4.3. Detección de comunidades

Dada la red de similitud temporal, usamos el algoritmo Louvain [120] descrito en el apartado 2.4 para la detección de comunidades. Se encontraron cinco comunidades. El resultado se muestra en la figura 4.6a). El número de nodos en cada comunidad $\mathcal{C}_s$ (con $s = 1, 2 \ldots, 5$) se denota por $\mathcal{N}_s$ y, a partir de este análisis, obtenemos grupos de ciudades con $\mathcal{N}_1 = 148$, $\mathcal{N}_2 = 136$, $\mathcal{N}_3 = 133$, $\mathcal{N}_4 = 113$ y $\mathcal{N}_5 = 102$. En la figura 4.6b) se representan con líneas finas y grises las probabilidades $\mathcal{P}_i(\tau)$ mostradas en la figura 4.2a) para los grupos de ciudades definidos por cada comunidad $\mathcal{C}_s$. Cada panel contiene $\mathcal{N}_s$ curvas. Además, incluimos el análisis estadístico considerando todos los check-ins en cada grupo; los resultados se muestran con líneas gruesas y de color. Cuando se agrupan de esta manera, las curvas observadas dentro de cada comunidad $\mathcal{C}_s$ son similares, lo que evidencia que tienen, esencialmente, el mismo comportamiento.

Las curvas promedio en la figura 4.6b) muestran que la comunidad $\mathcal{C}_1$, cuya media se muestra en azul, tiene un comportamiento de lunes a viernes caracterizado por tres máximos durante todo el día (a las 8, 12 y 18 horas). Durante los fines de semana, este patrón se rompe y se observa un único máximo alrededor de las 20 horas del sábado y al mediodía del domingo. La comunidad $\mathcal{C}_2$, con la línea anaranjada, se caracteriza por un máximo pronunciado a las 13 horas y un segundo máximo relativo a las 19 horas, de lunes a viernes. Este comportamiento se mantiene los fines de semana, pero con menos check-ins. La comunidad $\mathcal{C}_3$, con una curva verde, es la que tiene menos contraste entre el comportamiento de lunes a viernes, y los fines de semana.

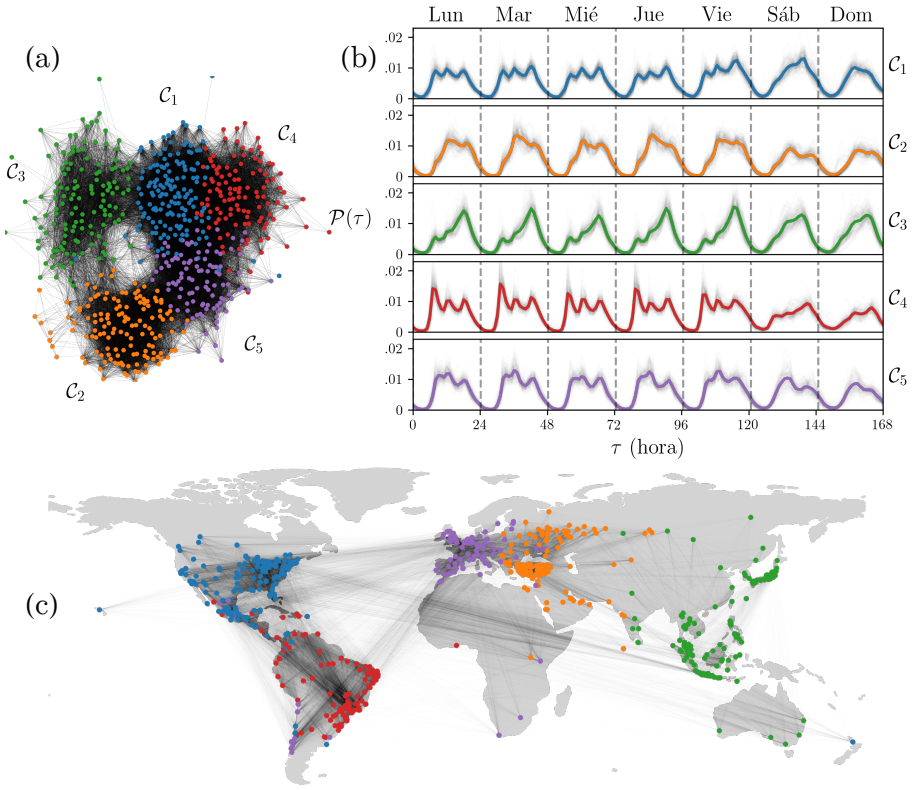


Figura 4.6: Patrones de actividad humana identificados mediante detección de comunidades en redes de similitud. (a) Estructura de comunidad de la red de similitud generada con $H = 0.0723$, se detectaron cinco comunidades $\mathcal{C}_1$, $\mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4$, $\mathcal{C}_5$ utilizando el algoritmo de Louvain y se representan con diferentes colores. (b) Probabilidad $\mathcal{P}(\tau)$ para la actividad temporal de las ciudades en cada comunidad. Las líneas grises delgadas presentan las curvas $\mathcal{P}_i(\tau)$ representadas en la Fig. 4.2(a), mientras que la línea gruesa coloreada representa el análisis estadístico de todos los check-ins en las ciudades de la comunidad. (c) Representación geográfica de la red en (a). En este caso, cada nodo es una ciudad representada en un mapa mundial.

Cada día, el máximo se encuentra a las 18 horas. Aun así, de lunes a viernes hay un pequeño máximo local a las 8 horas que desaparece en el fin de semana. $\mathcal{C}_4$, con la media presentada en rojo, tiene las mismas características que $\mathcal{C}_1$ pero con tamaños relativos diferentes; en este caso, el primer máximo diario domina sobre el resto. La comunidad $\mathcal{C}_5$ es la más pequeña y su comportamiento promedio se representa en púrpura. En este caso, la curva sufre cambios a lo largo de la semana. Mientras que hay un patrón con máximos a las 9, 12 y 19 horas, y un valle a las 15 horas, de lunes a viernes, los tamaños relativos no siempre son iguales; el lunes dominan el primero y segundo máximo, el martes el segundo máximo es el más grande, el miércoles los tres son prácticamente del mismo tamaño, el jueves los dos primeros máximos casi se fusionan y dominan sobre el tercero, y el viernes el primero casi desaparece mientras que el segundo domina. Finalmente, el sábado hay un cambio que da lugar a dos máximos a las 13 y 20 horas, también presentes el domingo aunque más pequeños.

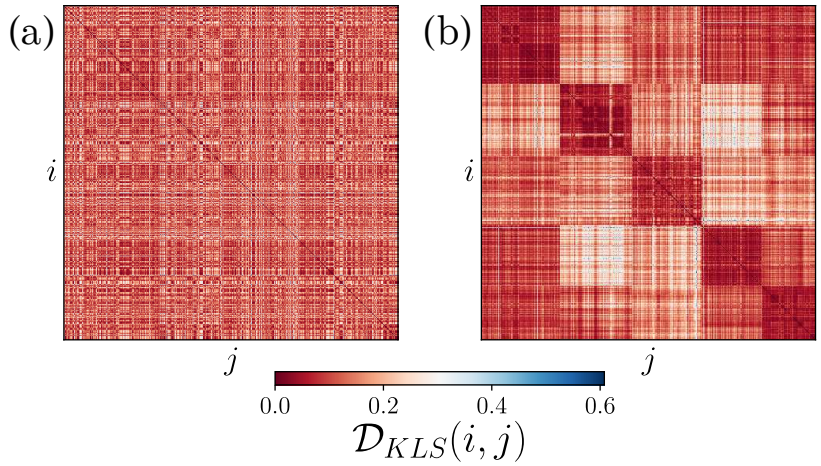


Figura 4.7: **Matriz de similitud entre ciudades ordenadas por comunidad**. Después de identificar las comunidades en la red mediante el algoritmo de Lovain para la maximización de la modularidad, se pueden reordenar los índices de las ciudades para que la matriz $\mathcal{D}_{KLS}(i,j)$, presentada en la figura 4.3a), revele los módulos encontrados. (a) Matriz $\mathcal{D}_{KLS}(i,j)$ en la que los índices se asignan a partir del orden alfabético de los nombres de las 632 ciudades. (b) Matriz $\mathcal{D}_{KLS}(i,j)$ en la que los índices se asignan, en primer lugar, a partir de la comunidad a la que pertenecen y, en segundo lugar, del orden alfabético de sus nombres.

Con las coordenadas de las ciudades correspondientes a cada nodo, se pudo generar un diseño geográfico e identificar macroregiones que agrupan a cada conjunto. No hay que olvidar que las comunidades que se visualizan en la figura 4.6a), en un espacio abstracto, son grupos de ciudades ubicadas en el espacio. Ya que cada nodo está asociado con sus coordenadas (el centroide del polígono correspondiente), también podemos representar las comunidades por colores en un mapa (ver figura 4.6c), en el que regiones geográficas claras exhiben similitud temporal para la actividad humana. De este modo, los patrones temporales descritos antes se corresponden con una escala intermedia entre las ciudades individuales y el sistema global de ciudades.

Las comunidades encontradas en el sistema se pueden usar para dar un nuevo orden a los elementos. Durante todo el análisis realizado a las distribuciones de check-ins, y la comparación por pares de éstas, se utilizó un orden alfabético para asignar los índices. Tanto en las curvas de la figura 4.2a) como en la representación gráfica de la matriz $\mathcal{D}_{KLS}(i,j)$ mostrada en la figura 4.3a), el índice 1 asignó a la ciudad Abu Dhabi (Emiratos Árabes), el 2 a Acapulco (México), el 3 a Accra (Ghana), y así hasta el 632 para la ciudad turca de Sanliurfa. Esta última se escribe oficialmente Şanlıurfa y Pandas le asocia el último lugar ya que la letra turca Ş se coloca después de todas las letras latinas. Ya que los nombres y el comportamiento temporal de las ciudades no tienen relación, ninguna de las dos figuras muestra una estructura de comunidades. Sería prácticamente lo mismo asignarles un orden aleatorio. Por el contrario, si los índices se asignan respetando, en primer lugar, el número de comunidad a la que cada ciudad pertenece, una estructura de grupos emerge. En la figura 4.7 se ilustran ambos métodos de asignación de los índices; en (a) el orden alfabético y en (b) el orden por

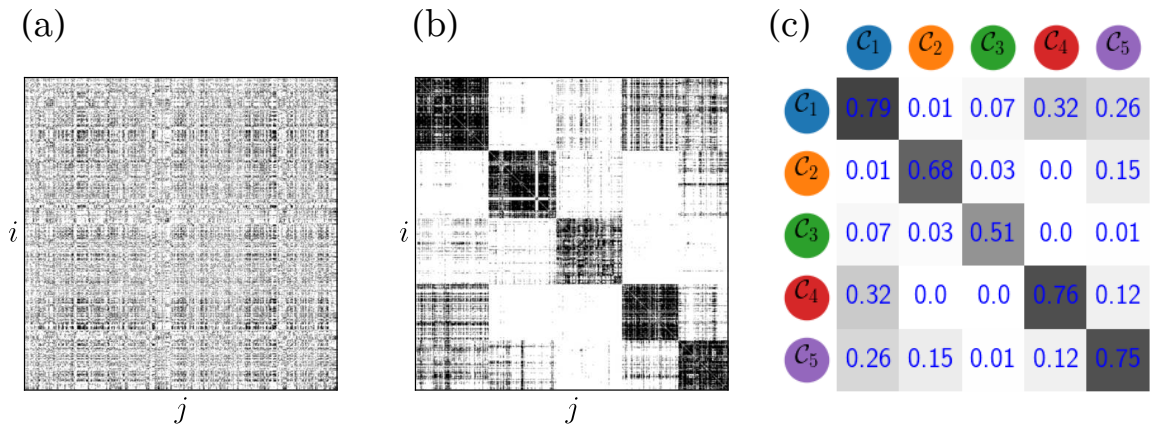


Figura 4.8: **Matriz de adyacencia y fracción de enlaces inter e intra comunidades**. Se aplica la definición 4.3 cuando los índices de las ciudades se asignan (a) en orden alfabético y (b) agrupando por comunidad. (c) Se compara el número de enlaces intra e intercomunidades con el máximo valor posible que pueden alcanzar. Esto da una idea del rendimiento de nuestra partición o la modularidad de la red tomando estas comunidades.

comunidades. La aparición, en este último caso, de módulos cuadrados en la diagonal de la matriz son el efecto de este nuevo ordenamiento. Lo mismo puede hacerse para la matriz de adyacencia definida en 4.3: mientras el orden alfabético no muestra ninguna estructura de cúmulos y se asemeja a un patrón aleatorio (figuras 4.4b) y 4.8a), después de ordenar las ciudades a partir de sus comunidades se identifica un patrón modular en la matriz (figura 4.8b).

Con las comunidades de la red también podemos comparar la fracción de enlaces intra-comunidad e inter-comunidad. Denotamos L_s como el número de enlaces entre nodos en la comunidad $\mathcal{C}_s$, y L_{st} como el número de enlaces entre un nodo en $\mathcal{C}_s$ y un nodo en $\mathcal{C}_t$. La fracción de enlaces inter-comunidad se calcula con $L_{st}/(\mathcal{N}_s\mathcal{N}_t)$ y la de enlaces intra-comunidad con $2L_s/(\mathcal{N}_s(\mathcal{N}_s-1))$. Los valores obtenidos nos permiten comparar el número de enlaces con el número total de enlaces posibles (figura 4.8c). Los valores para la fracción de enlaces intra-comunidad son 0.79, 0.68, 0.51, 0.76 y 0.75 para $\mathcal{C}_1,\ldots,\mathcal{C}_5$, respectivamente. En cuanto a la fracción de enlaces inter-comunidad, los valores están por debajo del 0.1 excepto entre las comunidades $\mathcal{C}_1$ y $\mathcal{C}_4$ (Norteamérica y Sudamérica) con 0.32, y 0.26 entre $\mathcal{C}_1$ y $\mathcal{C}_5$ (Norteamérica y Europa). Hay una fracción de 0.15 de nodos inter-comunidad entre $\mathcal{C}_2$ (Medio Oriente) y $\mathcal{C}_5$ (Europa), y 0.12 entre $\mathcal{C}_5$ y $\mathcal{C}_4$ (Sudamérica). Los resultados también muestran que $\mathcal{C}_3$ tiene poca similitud con las otras comunidades. Europa es una comunidad con muchos elementos en común con América y el Cercano Oriente. Asia Oriental, por otro lado, tiene pocos elementos en común con el resto del mundo en términos de patrones temporales de comportamiento humano. Estas características se observan en el mapa de la Fig. 4.6c), evidenciando aspectos culturales e históricos de cada región. Esta información se muestra en la matriz de la figura 4.8c). En ella, cada fila y cada columna representan a una comunidad y en los cruces la fracción de enlaces intra e inter-comunidad, según sea el caso. Es notable cómo los valores de la diagonal,

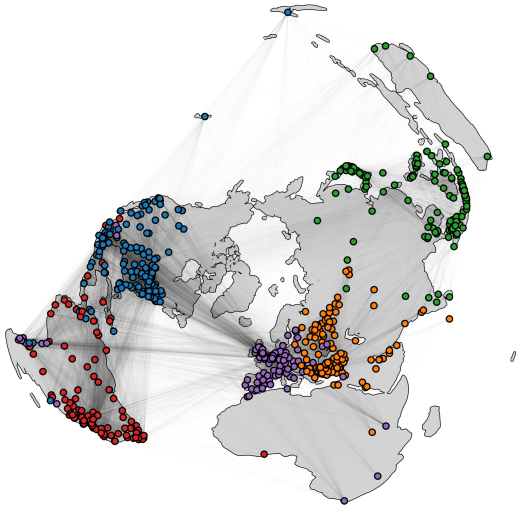


Figura 4.9: Otra proyección generada con GeoPandas (Python) para el análisis de los vínculos inter-comunidades.

correspondientes a la fracción de enlaces intra-comunidad, son altos mientras que los valores fuera de la diagonal, fracción de enlaces inter-comunidades, son relativamente bajos; estos valores se representan con una escala de grises.

Finalmente, vale la pena mencionar que, a diferencia de otros métodos de identificación de cúmulos, como la agrupación aglomerativa (discutida en la siguiente sección), la información sobre la similitud entre regiones específicas se conserva en los vínculos por pares. En la figura 4.6c), se observa una alta densidad de vínculos entre Norteamérica y el Reino Unido, así como entre Brasil y Portugal, regiones con relaciones culturales y lingüísticas. Una proyección geográfica alternativa en la figura 4.9 contribuye a visualizar las relaciones entre regiones. Los resultados también muestran que no todas las ciudades pertenecen a la misma comunidad que se esperaría según su región geográfica. Específicamente, hay 8 ciudades en Chile, 3 en México y 1 en Uruguay que son más similares a ciudades europeas que a las americanas.

4.4.4. Agrupamiento jerárquico aglomerativo

Como método alternativo se agruparon los datos mediante agrupamiento jerárquico por aglomeración (Agglomerative Hierarchical Clustering). Éste consiste en agrupar conjuntos de datos en cúmulos de distintas jerarquías que van desde uno único, de jerarquía superior y que agrupa a todos los datos, y los cúmulos de jerarquía más baja que corresponden a los propios datos individuales. Este método requiere una métrica para medir la distancia entre los conjuntos de datos y un criterio de enlace entre los grupos [190].

El procedimiento consiste en definir primero tantos cúmulos como datasets hay en la muestra. En nuestro caso, tendríamos 632 cúmulos para 632 ciudades individuales.

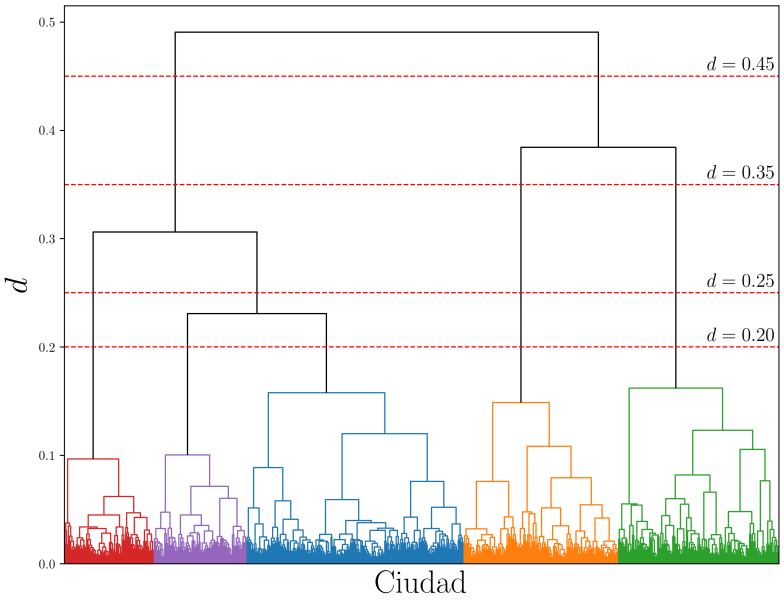


Figura 4.10: Dendrograma generado por el método de agrupamiento jerárquico aglomerativo

Posteriormente se unen los cúmulos para formar otros de jerarquía superior, consecuentemente más grandes. Estos grupos de jerarquía superior se generan a partir del criterio de enlace elegido y el procedimiento consiste en variar el valor de un parámetro d: si la distancia entre dos grupos i y j es tal que $d_{ij} < d$, esos dos se unen (se aglomeran). Este proceso se sigue hasta que todos los conjuntos de datos pertenecen a un único cúmulo. Así se obtiene una función del parámetro d que a cada valor le asocia grupos de datasets.

Esto puede graficarse en un plano cartesiano de la siguiente forma. En uno de los ejes se grafica el parámetro d. Cuando a un cierto valor d^* le corresponde la unión de dos cúmulos, esta unión se representa mediante un segmento de recta perpendicular al eje d que une los dos clusters, representados por rectas paralelas al mismo eje. El nuevo cluster formado se grafica como una nueva recta paralela a d.

El resultado es un diagrama con forma de árbol, conocido como dendrograma, en el que las hojas son todos los datos originales (cuando d=0) y la "raíz" es la unión de los cúmulos que singifica unir todos los datos en un solo conjunto. De esta forma, las rectas d = cte cruzan con el dendrograma y el número de cruces da el número de clusters correspondientes a ese valor. Se trata de un método jerárquico porque el resultado es un conjunto de cúmulos de distinta jerarquía, cada uno de los cuales contiene otros hasta llegar a las "hojas".

En la figura 4.10 se muestra el dendrograma generado con nuestros conjuntos de datos, las 632 distribuciones temporales de check-ins, mediante una métrica euclidiana

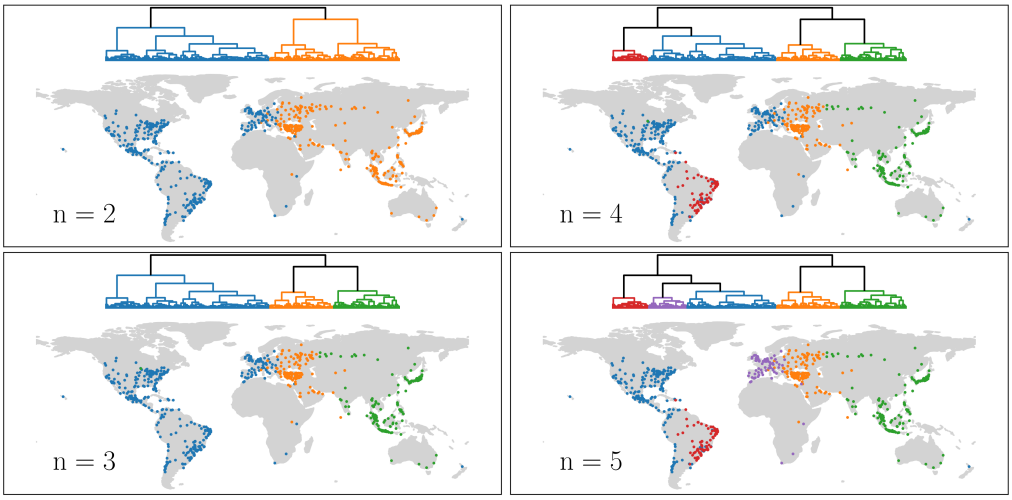


Figura 4.11: Dendrograma generado por el método de agrupamiento jerárquico aglomerativo para distintos valores del número de comunidades a encontrar. En cada caso se hacen coincidir los colores con las ubicaciones geográficas.

para la distancia entre conjuntos. Sobre el dendrograma se muestran distintos valores de d, a decir 0.45, 0.35, 0.25, 0.20 y 0.16, que dan lugar a 2, 3, 4, 5 y 6 clusters, respectivamente.

Como las hojas representan ciudades, el resultado es un conjunto de grupos de ciudades. En la figura 4.11 se distinguen con colores los grupos resultantes para cada d, tanto en las ramas del dendrograma como en los puntos del mapa. Cuando $d = .45$, $n = 2$ y los dos cúmulos se distinguen con azul y amarillo. La representación geográfica muestra una división de las ciudades en dos grandes regiones: en la azul se ubican predominantemente América y Europa mientras que en la amarilla Asia Central, Asia Oriental y Oceanía. Cuando $d = .35$ la rama amarilla se divide en dos: la más grande hereda el color amarillo y a la otra se le asignó un color verde. En la representación geográfica, el cluster amarillo contiene a las ciudades de Asia Central y el Este de Rusia; en el verde están las ciudades de la India y las ubicadas más al Este. Con $d = .25$ el cúmulo azul se divide en dos. El más grande hereda el azul y al pequeño le dimos un color rojo. Éste último contiene predominantemente a las ciudades de Brasil y algunas otras cercanas en Uruguay, Paraguay y Venezuela. Finalmente, el grupo azul se parte en dos cuando $d = .20$, separando ahora a las ciudades que se ubican al este del Océano Atlántico y que dibujamos en morado. Se trata de las ciudades de Europa. Los valores menores de d empeoran el desempeño del agrupamiento, por lo que no se incorporan en esta gráfica.

Finalmente, dado que cada conjunto de datos representa la actividad de los usuarios de Foursquare en una cuidad, los clusters generados mediante el agrupamiento jerárquico por aglomeración deben representar aspectos comunes en la actividad de los usuarios. En la figura 4.12 se muestra la representación geográfica de los 5 clusters

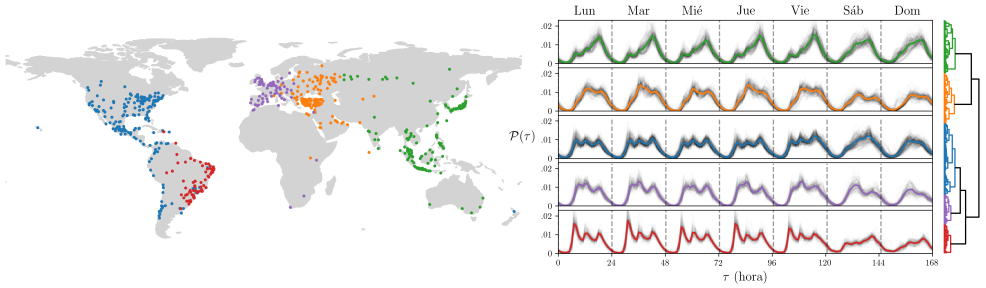


Figura 4.12: Dendrograma generado por el método de agrupamiento jerárquico aglomerativo, las comunidades encontradas mediante este método y el análisis estadístico de las huellas temporales correspondientes a cada comunidad.

para $d = 0.20$, valor que arroja la misma cantidad de grupos que el análisis de redes, y todos los histogramas de los patrones temporales reunidos en esos mismos cúmulos. En el mapa de la izquierda se muestran con más detalle las ciudades distinguidas con los mismos colores que antes y en el panel de la derecha los cinco grupos de histogramas correspondientes.

En estos grupos de histogramas se muestran las curvas de todas las ciudades pertenecientes a cada cúmulo en gris tenue, junto con la tendencia promedio con una línea gruesa de color. El color de la tendencia promedio es el mismo que el de las ciudades del cúmulo en la representación geográfica del mapa. El primer grupo, con línea verde, corresponde a las ciudades de India, Rusia oriental y toda Asia oriental. Tiene 140 ciudades. El segundo grupo, cuya línea promedio se muestra en amarillo y que corresponde a las ciudades de Asia central, tiene 135 ciudades. En el tercer grupo, con línea azul, se muestran las ciudades de América del Norte y la parte occidental de Sudamérica. Son 189 ciudades. El cuarto grupo, con color morado, corresponde principalmente a ciudades de Europa y contiene 81. Finalmente, el quinto grupo, con color distintivo rojo, corresponde a las ciudades de Brasil y cercanas. En total son 78 ciudades. En el extremo derecho de la figura se muestra una vez más el dendrograma pero ahora en una configuración horizontal para que el orden de los grupos en los histogramas coincidan con el orden de las ramas de cada cúmulo.

Los resultados obtenidos mediante este método son consistentes con los que se presentaron en el apartado anterior, dando cuenta de procesos de generación de estructuras superiores a los elementos individuales, en este caso ciudades, que por un lado distinguen a las ciudades a nivel global, pero por otro las agrupan en conjuntos que parecen compartir rasgos comunes en términos de historia, cultura y proximidad espacial.

Conclusiones

En este trabajo utilizamos los check-ins de Foursquare como indicadores de la actividad humana en las ciudades. Estos datos nos brindaron una gran cantidad de información sobre los intereses de las personas, las características de los lugares que visitan y los patrones de comportamiento en las ciudades, entre otros aspectos relevantes. En total, se analizaron 90,048,627 check-ins realizados por 2,733,324 usuarios de Foursquare, abarcando el periodo de abril de 2012 a enero de 2014. Esto permitió el estudio de 632 ciudades pertenecientes a 87 países, cada una con más de 10,000 check-ins registrados.

Durante el análisis, se examinaron tanto las propiedades espaciales de los puntos de interés y los check-ins, como las características temporales de los check-ins. Estas últimas, al ser interacciones espacio-temporales auto-reportadas entre personas y lugares, nos brindaron información valiosa sobre las actividades ocurridas en las ciudades. En particular, hicimos un análisis estadístico del número de registros por hora, lo que nos permitió obtener una visión más detallada de los patrones de actividad a lo largo del tiempo a partir de la distribución de probabilidad de check-ins por hora de la semana. Tomamos estas probabilidades como una huella de la actividad temporal de cada ciudad y aplicamos una versión simétrica de la divergencia de Kullback-Leibler para comparar todas las huellas. Esta medida nos permitió cuantificar la similitud o diferencia en los patrones de actividad temporal entre las ciudades.

Con base en esta información, definimos una red de similitud utilizando un valor umbral denotado como H. Este umbral nos permitió determinar si dos ciudades tenían una actividad similar o no. Posteriormente, exploramos el tamaño de la componente conectada más grande, así como varias métricas y propiedades de la red, como función de H. De manera interesante, encontramos que alrededor de $H = 0.02$, se produjo un umbral de percolación en el cual el tamaño de la máxima componente conectada experimentó un cambio abrupto. Esto significa que hubo un cambio significativo en la estructura de la red en términos de la conectividad de las ciudades. Además, observamos que para $H = 0.0723$, esta componente conectada incluyó a todos los nodos de la red, es decir, todas las ciudades estaban interconectadas directa o indirectamente. Entre estos dos valores de H, notamos cambios cualitativos en las propiedades de la red como el grado promedio, el coeficiente de agrupamiento, la distancia promedio entre nodos y el número de enlaces. Estos cambios dan cuenta de una transición o cambio significativo en la estructura de la red de similitud de ciudades.

Este método revela comportamientos colectivos emergentes entre ciudades que van

más allá de la proximidad física y la simple movilidad entre áreas urbanas. Permite identificar estructuras de mayor orden basándose en la similitud entre ciudades que están demasiado distantes como para que ésta se deba únicamente al flujo de personas entre ellas. Con el fin de capturar estas estructuras, utilizamos el algoritmo de Louvain para detectar comunidades en la red de similitud generada con un valor H de 0.0723. Como resultado de este análisis, se identificaron 5 grupos de ciudades que exhibían dinámicas temporales similares. Al localizar estas ciudades en un mapa, se observó un patrón notable: las comunidades identificadas correspondían a regiones fácilmente distinguibles en términos culturales, históricos, lingüísticos, entre otros aspectos relevantes. Estos resultados respaldan la noción de que los comportamientos y las actividades humanas están influenciados por factores más allá de la proximidad física y la movilidad geográfica, y que existen patrones culturales y regionales que influyen en las interacciones entre ciudades.

Adicionalmente, se usó un algoritmo de Machine Learning de agrupamiento aglomerativo no supervisado para clasificar las huellas de actividad de las ciudades. Las comunidades obtenidas por este método coinciden con las obtenidas mediante la detección de cúmulos en la red de similitud. Estas cinco comunidades corresponden con las regiones de Norteamérica, Sudamérica, Europa Occidental, Europa del Este junto a Oriente Medio y, finalmente, Asia Oriental junto al Sudeste Asiático. Estas *macro regiones* dan cuenta de un nivel de organización intermedio entre los sistemas urbanos individuales (las áreas urbanas funcionales) y el sistema de ciudades global.

Es interesante destacar que, más allá de las diferencias en la actividad humana entre áreas urbanas individuales y macroregiones, se observan patrones comunes en todas las ciudades analizadas. Estos patrones se reflejan en los máximos y mínimos que distinguen los horarios de actividad durante el día y la ausencia de actividad por la noche. Consideramos que estos patrones universales están relacionados con aspectos intrínsecos al comportamiento humano, especialmente el hecho de que somos una especie diurna. Que todas las ciudades muestren un mínimo de actividad en la noche, alrededor de las 2 y 3 de la mañana, es reflejo de estos patrones universales que, a su vez, están influenciados por los ritmos circadianos. A través de la aplicación de la ciencia de redes y algoritmos de Machine Learning, pudimos revelar comunidades con patrones de comportamiento distintos. Esto sugiere que estos patrones pueden ser universales con ciertas variaciones culturales, geográficas e históricas a nivel global. Estos hallazgos respaldan la idea de que existen aspectos fundamentales del comportamiento humano que influyen en la actividad diaria de las ciudades, independientemente de su ubicación geográfica o contexto cultural. Al comprender y reconocer estos patrones universales, podemos obtener una comprensión más profunda de la interacción entre los seres humanos y sus entornos urbanos, así como de las variaciones y particularidades que pueden surgir en diferentes contextos socioculturales.

En resumen, este estudio utilizó una gran cantidad de datos de redes sociales basadas en ubicaciones para analizar múltiples ciudades en todo el mundo y obtener los patrones temporales de la actividad de las personas en los puntos de interés. A través del enfoque de la ciencia de las ciudades, la teoría de redes complejas y la ciencia de

datos, se abordó el fenómeno del comportamiento humano en las ciudades. Una de las contribuciones de este trabajo es la propuesta de una metodología novedosa para el estudio de la sincronización de la actividad humana a nivel de ciudades y a nivel regional. Entre los principales aportes y perspectivas de este estudio se encuentran:

1. La comprensión integral de las bases de datos utilizadas en los estudios urbanos, centrándose en las redes sociales basadas en ubicaciones como una fuente rica de información sobre el comportamiento humano en las ciudades.

2. El reconocimiento de los puntos de interés como síntesis de interacciones humanas, que a su vez generan más interacciones. Esto resalta la importancia de considerar el espacio físico y las interacciones materiales como escenario y determinación de las relaciones sociales, elementos fundamentales en el estudio de los fenómenos urbanos.

3. La identificación de diferencias en el comportamiento de las personas en las ciudades, lo cual sugiere la influencia de factores como la historia, la cultura y las tradiciones. Estos hallazgos tienen implicaciones importantes para el diseño de políticas urbanas y la mejora de la calidad de vida en las ciudades.

4. El enfoque en las áreas urbanas funcionales como una perspectiva innovadora para el análisis de conjuntos de datos urbanos. Esta perspectiva permite obtener resultados novedosos y ofrece nuevas formas de entender la complejidad de los espacios urbanos.

5. El uso de fuentes de información abiertas y globales.

En conjunto, este estudio contribuye a ampliar nuestro conocimiento sobre el comportamiento humano en las ciudades y ofrece perspectivas interesantes para futuras investigaciones en el campo de la ciencia de las ciudades. Por ejemplo, dado que los puntos de interés proporcionan información muy valiosa acerca del tipo de actividades que realizan las personas, se pueden hacer análisis de distribución espacial y temporal de la actividad humana como función de las categorías de los sitios visitados. Esto podría complementar los estudios de flujos de personas en las ciudades con información acerca del tipo de actividad. ¿A dónde y qué hora la gente se traslada para trabajar, hacer ejercicio, comer o recrearse? ¿Existe alguna regularidad espacial o temporal en las ciudades del mundo relacionada con el tipo de actividades que se realizan? ¿Qué tipo de lugares son los más concurridos, por hora y día de la semana, en cada ciudad del mundo? ¿Existe alguna correlación entre estos aspectos y otras variables de las áreas urbanas funcionales tales como la densidad de población, el producto interno bruto, el nivel de desarrollo, la incidencia delictiva o la abundancia de áreas verdes? Todas estas preguntas, de hecho, forman parte de investigaciones en curso que, esperamos, pronto rindan frutos.

Finalmente, vale la pena mencionar que existe la perspectiva de aplicar estas herramientas al estudio de otras bases de datos más grandes y con más información, como parte de proyectos de investigación mucho más ambiciosos en los que se eche

mano de convenios con gobiernos y empresas para atender problemáticas específicas en ciudades.

Bibliografía

[1] Francisco Betancourt, Alejandro P. Riascos, and José L. Mateos, "Temporal visitation patterns of points of interest in cities on a planetary scale: a network science and machine learning approach," *Scientific Reports*, vol. 13, no. 3, p. 4890, 2023. https://doi.org/10.1038/s41598-023-32074-w.

[2] M. Batty, *The New Science of Cities*. Cambridge: MIT Press, 2013.

[3] Bettencourt, Luís M A, *Introduction to Urban Science: Evidence and Theory of Cities as Complex Systems*. Cambridge: MIT Press, 2021.

[4] M. Barthelemy, *The structure and dynamics of cities*. Cambridge: Cambridge University Press, 2016.

[5] A. J. Florczyk *et al.*, "Description of the GHS urban centre database 2015," *Publications Office of the European Union*, 2019. https://publications.jrc.ec.europa.eu/repository/handle/JRC115586.

[6] L. Dijkstra, H. Poelman, and P. Veneri, "The EU-OECD definition of a functional urban area," *OECD Regional Development Working Papers*, 2019. "https://doi.org/https://doi.org/10.1787/d58cb34d-en".

[7] A. Barabási and M. Pósfai, *Network Science*. Cambridge: Cambridge University Press, 2016.

[8] M. Newman, *Networks*. Oxford: Oxford University Press, 2018.

[9] L. Mumford, *The City in History: Its Origins, Its Transformations, and Its Prospects*. A Harbinger Book, Michigan: Harcourt, Brace & World, 1961.

[10] J. Lobo *et al.*, "Urban Science: Integrated Theory from the First Cities to Sustainable Metropolises," *Report submitted to the NSF on the Present State and Future of Urban Science*, 2020. http://dx.doi.org/10.2139/ssrn.3526940.

[11] UNDESA, *World Urbanization Prospects: The 2018 Revision*. Department of Economic and Social Affairs - Population Division - United Nations, New York: United Nations, 2019. https://www.un.org/en/desa/2018-revision-world-urbanization-prospects.

[12] F. L. Ribeiro and D. Rybski, "Mathematical models to explain the origin of urban scaling laws," *Physics Reports*, vol. 1012, pp. 1–39, 2023. https://doi.org/10.1016/j.physrep.2023.02.002.

[13] L. M. A. Bettencourt, J. Lobo, *et al.*, "Growth, innovation, scaling, and the pace of life in cities," *Proceedings of the National Academy of Sciences*, vol. 104, no. 17, pp. 7301–7306, 2007. https://doi.org/10.1073/pnas.0610172104.

[14] D. Rybski and M. C. González, "Cities as complex systems—Collection overview," *PloS One*, vol. 17, no. 2, p. e0262964, 2022. https://doi.org/10.1371/journal.pone.0262964.

[15] M. Acuto, S. Parnell, and K. C. Seto, "Building a global urban science," *Nature Sustainability*, vol. 1, no. 1, pp. 2–4, 2018. https://doi.org/10.1038/s41893-017-0013-9.

[16] M. Alberti, "Grand Challenges in Urban Science," *Frontiers in Built Environment*, vol. 3, 2017. https://doi.org/10.3389/fbuil.2017.00006.

[17] M. Batty, *Inventing Future Cities*. Cambridge: MIT Press, 2018.

[18] L. Bettencourt and G. West, "A unified theory of urban living," *Nature*, vol. 467, no. 7318, pp. 912–913, 2010. https://doi.org/10.1038/467912a.

[19] M. Mitchell, *Complexity: A guided tour*. Oxford: Oxford University Press, 2009.

[20] H. Simon, "The architecture of complexity," *Proceedings of the American Philosophical Society*, vol. 106, pp. 467–482, 1962.

[21] W. Alonso, *Location and land use: toward a general theory of land rent*. Harvard University Press, 1964.

[22] D. Pumain and F. Moriconi-Ebrard, "City size distributions and metropolisation," *Geojournal*, vol. 43, pp. 307–314, 1997.

[23] M. Batty, "Modelling cities as dynamic systems," *Nature*, vol. 231, no. 5303, pp. 425–428, 1971.

[24] J. Jacobs, *Muerte y vida de las grandes ciudades*. Madrid: Capitán Swing Libros, 2020.

[25] J. Hall *et al.*, "Engineering Cities: How can cities grow whilst reducing emissions and vulnerability?," *Tyndall Centre for Climate Change Research*, 10 2009. 10.13140/RG.2.1.3346.4809.

[26] M. Fujita, P. R. Krugman, and A. Venables, *The spatial economy: Cities, regions, and international trade*. Cambridge: MIT Press, 2001.

[27] G. Duranton and D. Puga, "Micro-foundations of urban agglomeration economies," in *Handbook of regional and urban economics*, vol. 4, pp. 2063–2117, Elsevier, 2004.

[28] M. Storper, *Keys to the city: How economics, institutions, social interaction, and politics shape development.* Princeton: Princeton University Press, 2013.

[29] E. L. Glaeser, *Cities, agglomeration, and spatial equilibrium.* Oxford: Oxford University Press, 2008.

[30] M. Storper and A. J. Venables, "Buzz: face-to-face contact and the urban economy," *Journal of economic geography*, vol. 4, no. 4, pp. 351–370, 2004.

[31] A. Losch *et al.*, *Economics of location.* New Haven: Yale University Press, 1954.

[32] P. Geddes, *Cities in evolution: an introduction to the town planning movement and to the study of civics.* London: Williams & Norgate, 1915.

[33] M. S. Granovetter, "The strength of weak ties," *American journal of sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[34] G. M. Feinman, "Size, complexity, and organizational variation: A comparative approach," *Cross-Cultural Research*, vol. 45, no. 1, pp. 37–58, 2011.

[35] M. Barthelemy, "The statistical physics of cities," *Nature Reviews Physics*, vol. 1, no. 6, pp. 406–415, 2019.

[36] E. Arcaute and J. J. Ramasco, "Recent advances in urban system science: Models and data," *PloS One*, vol. 17, no. 8, p. e0272863, 2022.

[37] G. Chadwick and F. Corazon, *A Systems View of Planning: Towards a Theory of the Urban and Regional Planning Process.* Oxford: Pergamon Press, 1971.

[38] L. Von Bertalanffy, *Teoría general de los sistemas.* México: Fondo de Cultura Económica, 1976.

[39] L. Skyttner, *General systems theory: ideas & applications.* Singapore: World Scientific, 2001.

[40] H. Sayama, *Introduction to the modeling and analysis of complex systems.* Geneseo: Open SUNY Textbooks, 2015.

[41] O. Miramontes, I. Lugo, L. B. Sosa, M. Mercado, J. Escandon, A. Rueda, G. de la Mora, O. DeSouza, and P. C. Souza, *Complejidad y Urbanismo:: Del organismo a la ciudad.* México: CopIt ArXives, 2017.

[42] D. J. Watts, "The "new" science of networks," *Annu. Rev. Sociol.*, vol. 30, pp. 243–270, 2004.

[43] E. L. Glaeser and J. E. Kohlhase, *Cities, regions and the decline of transport costs*, ch. Fifty Years of Regional Science, p. 197. Springer, 2004.

[44] T. C. Schelling, *Micromotives and macrobehavior*. New York: WW Norton & Company, 2006.

[45] M. Batty, *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*. MIT Press, 2007.

[46] J. Henrich, *The secret of our success: How learning from others drove human evolution, domesticated our species, and made us smart*. Princeton: Princeton University Press.[JH], 2016.

[47] R. L. Carneiro, "The transition from quantity to quality: A neglected causal mechanism in accounting for social evolution," *Proceedings of the National Academy of Sciences*, vol. 97, no. 23, pp. 12926–12931, 2000.

[48] A. W. Johnson and T. K. Earle, *The evolution of human societies: from foraging group to agrarian state*. Redwood City: Stanford University Press, 2000.

[49] S. Nordbeck, "Urban allometric growth," *Geografiska Annaler: Series B, Human Geography*, vol. 53, no. 1, pp. 54–67, 1971.

[50] B. J. Berry, "Cities as systems within systems of cities," *Papers in regional science*, vol. 13, no. 1, pp. 147–163, 1964.

[51] A. Bretagnolle, D. Pumain, and C. Vacchiani-Marcuzzo, "The organization of urban systems," *Complexity perspectives in innovation and social change*, pp. 197–220, 2009.

[52] I. Lugo, *Complejidad y Urbanismo: Del organismo a la ciudad*, ch. Red geoespacial de ciudades, pp. 29–43. México: CopIt-arXives, 2017.

[53] W. Weaver, "Science and complexity," *American scientist*, vol. 36, no. 4, pp. 536–544, 1948.

[54] C. Reynoso, *Complejidad y caos: una exploración antropológica*. Buenos Aires: Sb Editorial, 2006.

[55] J. Portugali, "Spatial cognitive dissonance and socio—spatial emergence in a self-organizing city," *Self-Organization and the City*, pp. 141–173, 2000.

[56] M. Gell-Mann, A. García, and R. Pastor, *El quark y el jaguar: aventuras en lo simple y lo complejo*. Colección Metatemas, Tusquets Editores, 1995.

[57] L. Sosa, *Complejidad y Urbanismo: Del organismo a la ciudad*, ch. Infraestructura urbana basada en sistemas complejos adaptativos, pp. 45–56. México: CopIt-arXives, 2017.

[58] M. Rosvall, A. Trusina, P. Minnhagen, and K. Sneppen, "Networks and cities: An information perspective," *Physical Review Letters*, vol. 94, no. 2, p. 028701, 2005.

[59] S. Yang, X. Yang, C. Zhang, and E. Spyrou, "Using social network theory for modeling human mobility," *IEEE network*, vol. 24, no. 5, pp. 6–13, 2010.

[60] T. Hossmann, T. Spyropoulos, and F. Legendre, "A complex network analysis of human mobility," in *2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS)*, pp. 876–881, IEEE, 2011.

[61] P. Krugman, *The self organizing economy*. Hoboken: John Wiley & Sons, 1996.

[62] M. Barthelemy, *Spatial Networks: A Complete Introduction: From Graph Theory and Statistical Physics to Real-World Applications*. Berlin: Springer Nature, 2022.

[63] F. Xie and D. Levinson, *Evolving transportation networks*. New York: Springer Science & Business Media, 2011.

[64] L. C. Freeman *et al.*, "Centrality in social networks: Conceptual clarification," *Social network: critical concepts in sociology. Londres: Routledge*, vol. 1, pp. 238–263, 2002.

[65] G. Bianconi, *Multilayer networks: structure and function*. Oxford: Oxford University Press, 2018.

[66] S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, and M. Zanin, "The structure and dynamics of multilayer networks," *Physics reports*, vol. 544, no. 1, pp. 1–122, 2014.

[67] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of complex networks*, vol. 2, no. 3, pp. 203–271, 2014.

[68] S. Sobolevsky, I. Sitko, R. Tachet des Combes, B. Hawelka, J. Murillo Arias, and C. Ratti, "Cities through the prism of people's spending behavior," *PloS One*, vol. 11, no. 2, p. e0146291, 2016.

[69] Alejandro P. Riascos, and José L. Mateos, "Networks and long-range mobility in cities: A study of more than one billion taxi trips in New York City," *Scientific Reports*, vol. 10, no. 1, pp. 1–14, 2020.

[70] Jaspe U. Martínez-González and Alejandro P. Riascos, "Activity of vehicles in the bus rapid transit system Metrobús in Mexico City," *Scientific Reports*, vol. 12, no. 1, pp. 1–11, 2022.

[71] D. Loaiza-Monsalve and Alejandro P. Riascos, "Human mobility in bike-sharing systems: Structure of local and non-local dynamics," *PloS One*, vol. 14, no. 3, p. e0213106, 2019.

[72] M. C. Gonzalez, C. A. Hidalgo, and A. L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[73] C. Song, T. Koren, P. Wang, and A. L. Barabási, "Modelling the scaling properties of human mobility," *Nature physics*, vol. 6, no. 10, pp. 818–823, 2010.

[74] T. Louail, M. Lenormand, O. G. Cantu Ros, M. Picornell, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy, "From mobile phone data to the spatial structure of cities," *Scientific Reports*, vol. 4, no. 1, pp. 1–12, 2014.

[75] T. Louail, M. Lenormand, M. Picornell, O. Garcia Cantu, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy, "Uncovering the spatial structure of mobility networks," *Nature communications*, vol. 6, no. 1, pp. 1–8, 2015.

[76] S. Çolak, A. Lima, and M. C. González, "Understanding congested travel in urban areas," *Nature communications*, vol. 7, no. 1, pp. 1–8, 2016.

[77] L. Alessandretti, P. Sapiezynski, V. Sekara, S. Lehmann, and A. Baronchelli, "Evidence for a conserved quantity in human mobility," *Nature human behaviour*, vol. 2, no. 7, pp. 485–491, 2018.

[78] C. Song, Z. Qu, N. Blumm, and A. L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[79] L. Alessandretti, U. Aslak, and S. Lehmann, "The scales of human mobility," *Nature*, vol. 587, no. 7834, pp. 402–407, 2020.

[80] K. Bhattacharya and K. Kaski, "Social physics: uncovering human behaviour from communication," *Adv. Phys.: X*, vol. 4, no. 1, p. 1527723, 2019.

[81] P. Melikov, J. A. Kho, V. Fighiera, F. Alhasoun, J. Audiffred, J. L. Mateos, and M. C. González, "Characterizing Urban Mobility Patterns: A Case Study of Mexico City," in *Urban Informatics* (W. Shi, M. Goodchild, M. Batty, M. Kwan, and A. Zhang, eds.), Springer The Urban Book Series, ch. 11, pp. 153–170, Springer Nature Singapore, 2021.

[82] Z. Chen, S. Kelty, A. G. Evsukoff, B. F. Welles, J. Bagrow, R. Menezes, and G. Ghoshal, "Contrasting social and non-social sources of predictability in human mobility," *Nature communications*, vol. 13, no. 1, pp. 1–9, 2022.

[83] M. E. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[84] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, p. 462, 2006.

[85] "Where's George? - Official Currency Tracking Project." `https://www.wheresgeorge.com/`.

[86] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in Foursquare," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, pp. 570–573, 2011.

[87] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, "A tale of many cities: universal patterns in human urban mobility," *PloS One*, vol. 7, no. 5, p. e37027, 2012.

[88] Alejandro P. Riascos, and José L. Mateos, "Emergence of encounter networks due to human mobility," *PloS One*, vol. 12, no. 10, p. e0184532, 2017.

[89] N. Serok and E. Blumenfeld-Lieberthal, "A simulation model for intra-urban movements," *PloS One*, vol. 10, no. 7, p. e0132576, 2015.

[90] X.-Y. Yan, X.-P. Han, B.-H. Wang, and T. Zhou, "Diversity of individual mobility patterns and emergence of aggregated scaling laws," *Scientific Reports*, vol. 3, no. 1, pp. 1–5, 2013.

[91] K. Zhao, M. Musolesi, P. Hui, W. Rao, and S. Tarkoma, "Explaining the power-law distribution of human mobility through transportation modality decomposition," *Scientific Reports*, vol. 5, no. 1, pp. 1–7, 2015.

[92] D. Monsivais, K. Bhattacharya, A. Ghosh, R. I. Dunbar, and K. Kaski, "Seasonal and geographical impact on human resting periods," *Scientific Reports*, vol. 7, no. 1, pp. 1–10, 2017.

[93] M. A. Canuto, F. Estrada-Belli, T. G. Garrison, S. D. Houston, M. J. Acuña, M. Kováč, D. Marken, P. Nondédéo, L. Auld-Thomas, C. Castanet, *et al.*, "Ancient lowland Maya complexity as revealed by airborne laser scanning of northern Guatemala," *Science*, vol. 361, no. 6409, p. eaau0137, 2018.

[94] S. G. Ortman, J. Lobo, and M. E. Smith, "Cities: Complexity, theory and history," *PloS One*, vol. 15, no. 12, p. e0243621, 2020. `https://doi.org/10.1371/journal.pone.0243621`.

[95] W. Christaller, *Central places in southern Germany*, vol. 10. Hoboken: Prentice-Hall, 1966.

[96] L. Mumford, "The neighborhood and the neighborhood unit," *The Town Planning Review*, vol. 24, no. 4, pp. 256–270, 1954.

[97] M. E. Smith, "The archaeological study of neighborhoods and districts in ancient cities," *Journal of anthropological archaeology*, vol. 29, no. 2, pp. 137–154, 2010.

[98] M. Batty and P. A. Longley, *Fractal cities: a geometry of form and function*. Cambridge, USA: Academic Press, 1994.

[99] T. A. Kohler and M. E. Smith, *Ten thousand years of inequality: the archaeology of wealth differences*. Tucson: University of Arizona Press, 2018.

[100] D. S. Massey and N. A. Denton, "Hypersegregation in us metropolitan areas: Black and hispanic segregation along five dimensions," *Demography*, vol. 26, pp. 373–391, 1989.

[101] T. Piketty, "Capital in the twenty-first century," in *Capital in the twenty-first century*, Harvard University Press, 2017.

[102] A. Marshall, *Principles of Economics*. London: Mc Millan, 1920.

[103] C. Grasland, "Spatial analysis of social facts," 2010.

[104] M. Lenormand, A. Bassolas, and J. J. Ramasco, "Systematic comparison of trip distribution laws and models," *Journal of Transport Geography*, vol. 51, pp. 158–169, 2016.

[105] G. Zipf, *Human Behaviour and the Principle of Least Effort*. Addison-Wesley Press, 1949.

[106] J. Q. Stewart, "Empirical mathematical rules concerning the distribution and equilibrium of population," *Geographical review*, vol. 37, no. 3, pp. 461–485, 1947.

[107] X. Gabaix, "Zipf's law and the growth of cities," *American Economic Review*, vol. 89, no. 2, pp. 129–132, 1999.

[108] L. M. Bettencourt, "The origins of scaling in cities," *Science*, vol. 340, no. 6139, pp. 1438–1441, 2013.

[109] J. Lobo, L. M. Bettencourt, M. E. Smith, and S. Ortman, "Settlement scaling theory: Bridging the study of ancient and contemporary urban systems," *Urban Studies*, vol. 57, no. 4, pp. 731–747, 2020.

[110] D. Rybski, "Commentary," *Environment and Planning A: Economy and Space*, vol. 45, no. 6, pp. 1266–1268, 2013.

[111] C. Cottineau, E. Hatna, E. Arcaute, and M. Batty, "Diverse cities or the systematic paradox of urban scaling laws," *Computers, environment and urban systems*, vol. 63, pp. 80–94, 2017.

[112] K. Jordahl *et al.*, "geopandas/geopandas: v0.8.1," July 2020. `https://doi.org/10.5281/zenodo.3946761`.

[113] INEGI, *Delimitación de las zonas metropolitanas de México*. Grupo Interinstitucional para la Delimitación de Zonas Metropolitanas, México: SEDESOL-CONAPO-INEGI, 2004.

[114] E. Pérez-Campuzano, L. Guzmán-Vargas, and F. Angulo-Brown, "Distributions of city sizes in Mexico during the 20th century," *Chaos, Solitons & Fractals*, vol. 73, pp. 64–70, 2015.

[115] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[116] R. A. Rossi and N. K. Ahmed, "The network data repository with interactive graph analytics and visualization," in *AAAI*, 2015. `https://networkrepository.com`.

[117] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[118] R. Albert, H. Jeong, and A. L. Barabási, "Diameter of the world-wide web," *Nature*, vol. 401, no. 6749, pp. 130–131, 1999.

[119] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[120] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[121] S. Fortunato, "Community detection in graphs/santo fortunato," *Physics Reports*, 2009.

[122] S. Porta, P. Crucitti, and V. Latora, "The network analysis of urban streets: a primal approach," *Environment and Planning B: planning and design*, vol. 33, no. 5, pp. 705–725, 2006.

[123] R. Prieto Curiel, J. E. Patino, J. C. Duque, and N. O'Clery, "The heartbeat of the city," *PloS One*, vol. 16, no. 2, p. e0246714, 2021.

[124] R. Louf and M. Barthelemy, "How congestion shapes cities: from mobility patterns to scaling," *Scientific Reports*, vol. 4, no. 1, pp. 1–9, 2014.

[125] W. Pan, G. Ghoshal, C. Krumme, M. Cebrian, and A. Pentland, "Urban characteristics attributable to density-driven tie formation," *Nature communications*, vol. 4, no. 1, pp. 1–7, 2013.

[126] S. Jiang, G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, and M. C. González, "A review of urban computing for mobile phone traces: current methods, challenges and opportunities," in *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*, pp. 1–9, 2013.

[127] S. Jiang, J. Ferreira, and M. C. Gonzalez, "Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore," *IEEE Transactions on Big Data*, vol. 3, no. 2, pp. 208–219, 2017.

[128] M. Gonzalez *et al.*, "Cell phone location data for travel behavior analysis NCHSP Research Report," *National Cooperative Highway Research Program*, 2018.

[129] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González, "Unravelling daily human mobility motifs," *Journal of The Royal Society Interface*, vol. 10, no. 84, p. 20130246, 2013.

[130] E. Toch, B. Lerner, E. Ben-Zion, and I. Ben-Gal, "Analyzing large-scale human mobility data: a survey of machine learning methods and applications," *Knowledge and Information Systems*, vol. 58, no. 3, pp. 501–523, 2019.

[131] J. L. Toole, Y.-A. d. Montjoye, M. C. González, and A. S. Pentland, "Modeling and understanding intrinsic characteristics of human mobility," in *Social phenomena*, pp. 15–35, Springer, 2015.

[132] L. Pappalardo and F. Simini, "Data-driven generation of spatio-temporal routines in human mobility," *Data Mining and Knowledge Discovery*, vol. 32, no. 3, pp. 787–829, 2018.

[133] X. Liu, Y. Liu, K. Aberer, and C. Miao, "Personalized point-of-interest recommendation by mining users' preference transition," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 733–738, 2013.

[134] Y. Liu, W. Wei, A. Sun, and C. Miao, "Exploiting geographical neighborhood characteristics for location recommendation," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 739–748, 2014.

[135] A. Cuttone, S. Lehmann, and M. C. González, "Understanding predictability and exploration in human mobility," *EPJ Data Science*, vol. 7, pp. 1–17, 2018.

[136] L. Yang, H. Jiang, S. Wang, L. Wang, and Y. Fang, "Characterizing pairwise contact patterns in human contact networks," *Ad Hoc Networks*, vol. 10, no. 3, pp. 524–535, 2012.

[137] S. Jiang, A. Alves, F. Rodrigues, J. Ferreira Jr, and F. C. Pereira, "Mining point-of-interest data from social networks for urban land use classification and disaggregation," *Computers, Environment and Urban Systems*, vol. 53, pp. 36–46, 2015.

[138] M. Zignani, S. Gaito, and G. Rossi, "Extracting human mobility and social behavior from location-aware traces," *Wireless Communications and Mobile Computing*, vol. 13, no. 3, pp. 313–327, 2013.

[139] N. Srnicek, *Platform capitalism.* Hoboken: John Wiley & Sons, 2017.

[140] M. Lenormand, B. Gonçalves, A. Tugores, and J. J. Ramasco, "Human diffusion and city influence," *Journal of The Royal Society Interface*, vol. 12, no. 109, p. 20150473, 2015.

[141] D. Yang, B. Fankhauser, P. Rosso, and P. Cudre-Mauroux, "Location prediction over sparse user mobility traces using RNNs: Flashback in hidden states!," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 2184–2190, 2020.

[142] G. Cornacchia and L. Pappalardo, "STS-EPR: Modelling individual mobility considering the spatial, temporal, and social dimensions together," *Procedia Computer Science*, vol. 184, pp. 258–265, 2021.

[143] D. Yang, B. Qu, J. Yang, and P. Cudre-Mauroux, "Revisiting user mobility and social relationships in LBSN: A hypergraph embedding approach," in *The world wide web conference*, pp. 2147–2157, 2019.

[144] Alejandro P. Riascos, and José L. Mateos , "Random walks on weighted networks: a survey of local and non-local dynamics," *Journal of Complex Networks*, vol. 9, 10 2021. `10.1093/comnet/cnab032`.

[145] L. Gauvin, A. Panisson, C. Cattuto, and A. Barrat, "Activity clocks: spreading dynamics on temporal networks of human contact," *Scientific Reports*, vol. 3, no. 1, pp. 1–6, 2013.

[146] L. Sun, K. W. Axhausen, D.-H. Lee, and X. Huang, "Understanding metropolitan patterns of daily encounters," *Proceedings of the National Academy of Sciences*, vol. 110, no. 34, pp. 13774–13779, 2013.

[147] A. Barrat and C. Cattuto, "Temporal networks of face-to-face human interactions," in *Temporal Networks*, pp. 191–216, Springer, 2013.

[148] D. Yang, D. Zhang, L. Chen, and B. Qu, "Nationtelescope: Monitoring and visualizing large-scale collective behavior in lbsns," *Journal of Network and Computer Applications*, vol. 55, pp. 170–180, 2015.

[149] "Foursquare City Guide." `https://foursquare.com/city-guide`.

[150] D. Yang, B. Qu, J. Yang, and P. Cudré-Mauroux, "Lbsn2vec++: Heterogeneous hypergraph embedding for location-based social networks," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[151] "Foursquare." `https://es.foursquare.com/`. Consultado: 2018-09-30.

[152] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu, "Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 1, pp. 129–142, 2015.

[153] D. Yang, B. Qu, and P. Cudre-Mauroux, "Location-centric social media analytics: Challenges and opportunities for smart cities," *IEEE Intelligent Systems*, vol. 36, no. 5, pp. 3–10, 2020.

[154] R. Gallotti, G. Bertagnolli, and M. De Domenico, "Unraveling the hidden organisation of urban systems and their mobility flows," *EPJ Data Science*, vol. 10, no. 1, p. 3, 2021.

[155] A. Noulas, B. Shaw, R. Lambiotte, and C. Mascolo, "Topological properties and temporal dynamics of place networks in urban environments," in *Proceedings of the 24th International Conference on World Wide Web*, pp. 431–441, 2015.

[156] K. D'Silva, A. Noulas, M. Musolesi, C. Mascolo, and M. Sklar, "Predicting the temporal activity patterns of new venues," *EPJ data science*, vol. 7, pp. 1–17, 2018.

[157] "44 Foursquare Statistics You Must Know: 2023 Market Share & Business Use - Financesonline.com." `https://financesonline.com/foursquare-statistics/`.

[158] "Foursquare.com Traffic Analytics & Market Share | Similarweb." `https://www.similarweb.com/website/foursquare.com/#demographics`.

[159] "Foursquare Categories and Core Attributes | Foursquare." `https://location.foursquare.com/places/docs/categories`.

[160] "Listing of Countries and Territories With POIs | Foursquare." `https://location.foursquare.com/places/docs/supported-countries`.

[161] "Making Foursquare Places Work For You | Foursquare." `https://location.foursquare.com/places/docs/how-does-places-work`.

[162] "Foursquare Global-scale Check-in Dataset." `https://sites.google.com/site/yangdingqi/home/foursquare-dataset`. Consultado: 2018-09-30.

[163] E. Cho, S. Myers, and J. Leskovec, "Friendship and mobility: Friendship and mobility: User movement in location-based social networks," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1082–1090, 2011.

[164] "SNAP: Network datasets: Brightkite." `http://snap.stanford.edu/data/loc-Brightkite.html`.

[165] "Yong Liu Datasets." `https://www.yongliu.org/datasets/index.html`.

[166] "PANDAS - Python Data Analysis Library." `https://pandas.pydata.org/`.

[167] "Python." `https://www.python.org/`.

[168] A. J. Florczyk *et al.*, "GHS Urban Centre Database 2015, multitemporal and multidimensional attributes, R2019A," *European Commission, Joint Research Centre (JRC)*, 2019. [Dataset] Available online: `https://data.jrc.ec.europa.eu/dataset/53473144-b88c-44bc-b4a3-4583ed1f547e` (accessed on May 2023).

[169] P. W. Anderson, "More is different: broken symmetry and the nature of the hierarchical structure of science.," *Science*, vol. 177, no. 4047, pp. 393–396, 1972.

[170] A. L. Barabási, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, no. 7039, pp. 207–211, 2005.

[171] K. Sparks, G. Thakur, A. Pasarkar, and M. Urban, "A global analysis of cities' geosocial temporal signatures for points of interest hours of operation," *International Journal of Geographical Information Science*, vol. 34, no. 4, pp. 759–776, 2020.

[172] M. Ye, K. Janowicz, C. Mülligann, and W.-C. Lee, "What you are is when you are: the temporal dimension of feature types in location-based social networks," in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 102–111, 2011.

[173] L. Li, M. F. Goodchild, and B. Xu, "Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr," *Cartography and geographic information science*, vol. 40, no. 2, pp. 61–77, 2013.

[174] J. Aschoff and R. Wever, "Human circadian rhythms: a multioscillatory system.," in *Federation proceedings*, vol. 35-12, pp. 236–232, 1976.

[175] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A. L. Barabási, "Modeling bursts and heavy tails in human dynamics," *Physical Review E*, vol. 73, no. 3, p. 036127, 2006.

[176] J. Saramäki and E. Moro, "From seconds to months: an overview of multi-scale dynamics of mobile telephone calls," *The European Physical Journal B*, vol. 88, no. 6, pp. 1–10, 2015.

[177] S. Jiang, J. Ferreira, and M. C. González, "Clustering daily patterns of human activities in the city," *Data Mining and Knowledge Discovery*, vol. 25, no. 3, pp. 478–510, 2012.

[178] L. Sun, K. W. Axhausen, D.-H. Lee, and M. Cebrian, "Efficient detection of contagious outbreaks in massive metropolitan encounter networks," *Scientific Reports*, vol. 4, no. 1, pp. 1–6, 2014.

[179] V. D. Blondel, A. Decuyper, and G. Krings, "A survey of results on mobile phone datasets analysis," *EPJ data science*, vol. 4, no. 1, p. 10, 2015.

[180] T. Aledavood, I. Kivimäki, S. Lehmann, and J. Saramäki, "Quantifying daily rhythms with non-negative matrix factorization applied to mobile phone data," *Scientific Reports*, vol. 12, no. 1, pp. 1–10, 2022.

[181] S. Sobolevsky, M. Szell, R. Campari, T. Couronné, Z. Smoreda, and C. Ratti, "Delineating geographical regions with networks of human interactions in an extensive set of countries," *PloS One*, vol. 8, no. 12, p. e81707, 2013.

[182] D. Yang, D. Zhang, and B. Qu, "Participatory cultural mapping based on collective behavior data in location-based social networks," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 3, pp. 1–23, 2016.

[183] L. Hedayatifar, A. J. Morales, and Y. Bar-Yam, "Geographical fragmentation of the global network of Twitter communications," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 30, no. 7, p. 073133, 2020.

[184] E. Ben Zion and B. Lerner, "Identifying and predicting social lifestyles in people's trajectories by neural networks," *EPJ Data Science*, vol. 7, no. 1, pp. 1–27, 2018.

[185] H. Wang, H. Zhang, G. Tang, L. Zhou, and S. Jiang, "Inter-city association pattern recognition by constructing cultural semantic similarity network," *Transactions in GIS*, 2022.

[186] G. McKenzie and D. Romm, "Measuring urban regional similarity through mobility signatures," *Computers, Environment and Urban Systems*, vol. 89, p. 101684, 2021.

[187] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79 – 86, 1951.

[188] T. Cover and J. Thomas, *Elements of Information Theory*. A Wiley-Interscience publication, Hoboken: John Wiley & Sons, 2006.

[189] F. Nielsen, "On the Jensen–Shannon symmetrization of distances relying on abstract means," *Entropy*, vol. 21, no. 5, p. 485, 2019.

[190] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

APÉNDICES

A Temporal visitation patterns of points of interest in cities on a planetary scale: a network science and machine learning approach

scientific reports

Check for updates

OPEN

Temporal visitation patterns of points of interest in cities on a planetary scale: a network science and machine learning approach

Francisco Betancourt[1,3]✉, Alejandro P. Riascos[1,3] & José L. Mateos[1,2,3]

We aim to study the temporal patterns of activity in points of interest of cities around the world. In order to do so, we use the data provided by the online location-based social network Foursquare, where users make check-ins that indicate points of interest in the city. The data set comprises more than 90 million check-ins in 632 cities of 87 countries in 5 continents. We analyzed more than 11 million points of interest including all sorts of places: airports, restaurants, parks, hospitals, and many others. With this information, we obtained spatial and temporal patterns of activities for each city. We quantify similarities and differences of these patterns for all the cities involved and construct a network connecting pairs of cities. The links of this network indicate the similarity of temporal visitation patterns of points of interest between cities and is quantified with the Kullback-Leibler divergence between two distributions. Then, we obtained the community structure of this network and the geographic distribution of these communities worldwide. For comparison, we also use a Machine Learning algorithm—unsupervised agglomerative clustering—to obtain clusters or communities of cities with similar patterns. The main result is that both approaches give the same classification of five communities belonging to five different continents worldwide. This suggests that temporal patterns of activity can be universal, with some geographical, historical, and cultural variations, on a planetary scale.

In recent years, cities have become a topic of considerable scientific interest[1–3]. In the last decade, the development of technologies applied to inform the activities of humans has made available an unprecedented amount of data associated with the digital trace of humans in cities[4,5]. For instance: the use of credit card transactions[6], the use of digital devices to access transportation services, like taxis[7], buses and subway[8], bicycle[9], and telecommunications networks like cell phones[10–16] and GPS devices[17,18].

A very important source of information for the study of human behavior in cities are Location-Based Social Networks (LBSN)[19]. In these, people share information about the places they visit that may be of interest to other people in this social network. These venues, also known as Points of Interest (POIs), correspond to places within the city that can be characterized by features, such as restaurants, bars, gyms, hospitals, museums, parks and so on. Besides the type of feature, LBSN identify, for each POI, the spatial location (coordinates or addresses) and the time of visit (date and hour); these visits are recorded as check-ins using, typically, a mobile-phone application. Therefore, we end up with a data set with detailed information of the type of place and the exact location, together with the day of the week, and time of the visit. LBSN provide valuable information on the interaction of people among themselves, as well as the physical places where these interactions occur. This interplay has been explored in studies of urban mobility[20], human behavior[21,22], social interactions[23] and encounter networks[24].

In this study, we use data from one of the more relevant location-based social networks: Foursquare[25]. This LBSN has been used previously in several studies[23,24,26–34], of human mobility and social relationships, encounter networks due to human mobility, spatial and temporal activity patterns, among others. A more detailed

[1]Instituto de Física, Universidad Nacional Autónoma de México, Ciudad Universitaria, 04510 Mexico City, Mexico. [2]Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, 04510 Mexico City, Mexico. [3]These authors contributed equally: Francisco Betancourt, Alejandro P. Riascos and José L. Mateos. ✉email: cfrancisco@ciencias.unam.mx

description of Foursquare will be given in the next sections. In this paper, we will focus mainly on the temporal patterns of human activities in cities throughout the world. This temporal signature of the vastness of human dynamics can be captured through time series, using different sources such as email records[35], mobile phone calls[36], data sets from the public sector[37] and the analysis of the frequency of check-ins during the visitation of POIs in cities[38]. In this work, we will study the temporal patterns that emerge due to the frequent visitation of points of interest in different cities around the world. The data set we analyzed contains more than 11 million POIs of 632 cities of 87 countries located in 5 continents. We will focus mainly on the temporal patterns of activities on a weekly basis, that distinguish weekdays and weekend. Using network science[39], we generate an undirected weighted network connecting pairs of cities. The links between each pair of cities have a weight that quantifies the similarity of the temporal patterns of visitation of POIs. To measure the similarity between the two distributions of temporal patterns, we used the relative entropy (Kullback-Leibler divergence)[40] between the corresponding distributions. Once we have our network of cities, we obtained the community structure of this network using a well-known method of community detection: the Louvain algorithm. Then we locate these communities of cities geographically and notice a strong correlation between geography and temporal patterns of activities. On the other hand, besides the network science approach, we use a Machine Learning approach to classify the clusters in the data set of temporal distributions of activity. In particular, we used the unsupervised agglomerative hierarchical clustering algorithm[41] to obtain clusters or communities of cities that have similar time distributions of check-ins in Foursquare. Both approaches, network science and machine learning, give a very similar classification of five communities that distribute geographically in five different continents throughout the world.

The dataset used comprises over 90 million records made in more than 11 million POIs. Each record contains information about the place and time at which a person visited a POI. Additionally, using another dataset constructed from public information on Foursquare, it is possible to know the precise coordinates of venues and the type of place based on different categories used by this social network to classify sites. Thus, this is a database that provides spatial and temporal information on 2,733,324 individuals around the world, as well as information on the type of activities they perform. In this manner, we obtained spatial and temporal patterns of activity in many cities. In this work, we will focus mainly on the temporal patterns of activities on a weekly basis.

Finally, it is worth mentioning part of the motivation that led to this research. The origins of the present paper can be traced back to our interest in the interplay between human mobility in cities and the encounter (or contact) networks that emerge. In a paper published in 2017[24] by two of the authors, we explore precisely this emergence by analyzing data from Foursquare in two cities: New York City and Tokyo. We had the spatial and temporal data corresponding to the check-ins of many users visiting POIs in both cities. Thus, we explore the co-occurrences of two users in the same place at the same time in order to obtain the encounter or contact network. The motivation was, among other things, to study the propagation of viruses during an epidemic. By the way, this study was done and published previously to the current COVID-19 pandemic, where these contact networks were important to track the onset of the pandemic at the beginning of 2020. Years later, we keep studying the rich data set of Foursquare, but, instead of two cities, we explore more than 600 cities around the globe. In particular, the points of interest in cities are precisely the places where many people congregate and therefore can lead to an explosion of ideas and innovation, but at the same time can lead to the origin of epidemics of infectious diseases. That is part of our motivation to study in more detail the mobility and the temporal patterns of activity of POIs and to make a quantitative comparison of these patterns using metrics such as the Kullback-Leibler divergence. With this metric, we constructed a network of cities on a planetary scale and obtained the community structure of this network.

The paper is structured as follows: After the Introduction, we show first the Results with all the details in separate sections, then a Discussion and finally the Methods we employ.

Results

Check-ins and points of interest in Foursquare. In this work, we use Foursquare check-ins as a proxy of human activity in cities. Check-ins are spatio-temporal interactions between users and points of interest (POIs). They provide vast information about people's interests, site characteristics, and behavioral patterns in cities, among many others. This is why Foursquare metadata and location-based social networks in general, can provide useful information for studies of mobility, infrastructure, human behavior, and public policy, just to mention a few examples.

We study a large-scale and long-term Foursquare data set collected by Yang et al.[23,28] that is publicly available[42]. The data set contains 90,048,627 check-ins made by 2,733,324 users from April 2012 to January 2014. The authors collected active check-ins from Twitter by searching a hashtag generated automatically for users that linked their Foursquare activity with that social network. Then, with the set of places where check-ins occurred, they collected the POIs description; this information is available to the public from the Foursquare platform. The records include detailed information about the POIs, their coordinates, and the exact time of each user check-in[23,28]. In this manner, the database provides the space-temporal activity obtained from the sequence of successive check-ins for each user. Furthermore, the registers include a short description that labels each POI with specific keywords (for example "restaurant", "metro", "university") and a social network of users defined via the mutual following between Twitter accounts. We did not incorporate this information in our study (see the "Methods" section for a detailed description of the data).

The complete dataset includes information of Foursquare users in every country in the world. From the 253 country codes, check-ins on countries with 5000 or more POIs represent 98.92% of data. Almost all of the data is concentrated in a third of the countries. To visualize the geographical distribution of the POIs worldwide, we grouped them according to their coordinates. We generate a two-dimensional histogram dividing the latitude

and longitude using square bins with side 0.1° and counting the number of POIs in each coordinate square. This allowed us to identify regions with non-null POIs and those with the highest concentration. In Fig. 1 we depict the spatial distribution obtained; it consists of 3600 × 1410 squares covering latitudes from −56.6° and 84.6°, and longitudes from −180° to 180°; latitudes excluded are due to the lack of POIs on those regions. The number of POIs in each region is codified in the colorbar, where a logarithmic scale was chosen to show the wide range of values. The great coverage of this dataset is clear, as well as the existence of regions with high concentrations of POIs reaching up to 70,128 POIs in a single square defined above. The plot is the result of projecting the coordinates of the globe on a rectangle which implies spatial distortions; still, the frontiers of continents are identifiable even though no borders were used or drawn. A high density of POIs can be appreciated in North America and Europe, but active regions are also present in South America, the Middle East, and East Asia. Although to a much lesser extent, there are some localized regions with high numbers of POIs in Africa and Oceania. The details of this information are discussed in the "Methods" section in Table 1, where the 15 countries with most of the POIs are listed; 80% of the venues belong to these countries.

Temporal activity of users in Foursquare. The classification of POIs and check-ins at the level of cities allows us to perform temporal, spatial and spatio-temporal analysis of features. For example, in Ref.[24] a detailed study of the activity in Foursquare, of users in New York and Tokyo, is presented. We focused our study on cities that hold the majority of the data. From the total number of points of interest in the database, we selected only those belonging to cities with more than 10,000 check-ins. This resulted in a set of 632 cities located in 87 countries, as illustrated in Fig. 2. These 632 cities, represented by red dots on the map in panel 2a, are located in countries across all continents, reflecting the diversity in social, economic, cultural, linguistic, and other terms, crucial for the study of human activity in cities. As we will see later, this diversity allows us to identify commonalities among all cities, but also and especially differences in activity patterns of cities. To illustrate these commonalities and contrasts, we selected eight cities that capture some of this diversity. These cities are Barcelona (Spain), New York (United States), Mexico City (Mexico), Sao Paulo (Brazil), Istanbul (Turkey), Jakarta (Indonesia), Tokyo (Japan), and Moscow (Russia). These cities are shown with different colors on the map in Fig. 2a. In Fig. 2b, we present the temporal distribution of check-ins made during the nearly 22 months of observation in these eight cities that were intentionally selected to represent the behaviors in different regions with a great variety of cultural and linguistic characteristics. Also, these cities were chosen because they have a greater number of check-ins. Temporal information of the check-ins was grouped by urban area and was generated using the local time, obtained through the Universal Coordinated Time and adding the minutes corresponding to the timezone correction (see "Methods" section for details). In this way, we can identify daily routines for many geographical areas around the world. For this study, we defined the time granularity as 168 hours (corresponding to a full week), so we have, for each region, a histogram with all the check-ins made in the period of observation gathered by hour. We considered this histogram as a characteristic print of human temporal activity within a region. In doing so, clear patterns emerge with slight differences between regions. Some regularities are evident: there is low activity of check-ins during the night and high activity during the day; this matches the behavior reported by Yang, et al.[23], for a subset of this data at a global scale. In this case, the pattern persists at a local scale, as has

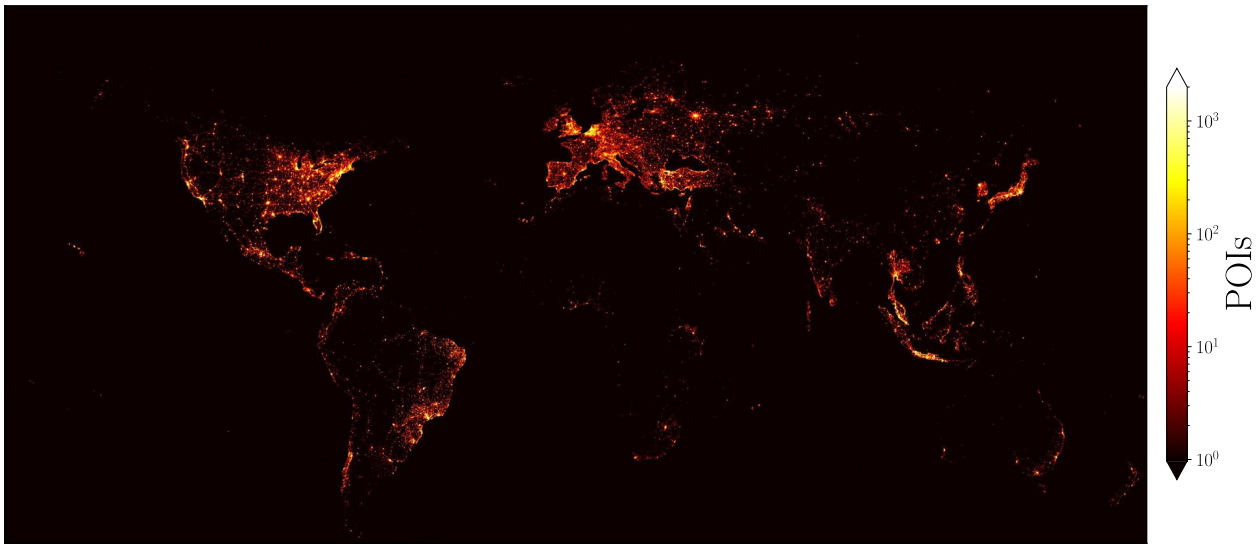


Figure 1. Points of interest worldwide from human activity in Foursquare. This analysis includes 11,179,790 points of interest (POIs). The counts codified in the colorbar show the number of POIs in each rectangle of a grid from −180° to 180° in the longitude and from −56° to 85° in the latitude (the dimensions of the grid are 3, 600 × 1, 410 squares defined by sides with 0.1°). A logarithmic scale was used to show the non-null frequencies of POIs found in the squares; the zones with null counts of POIs are depicted in black. This representation only considers geographical information of POIs, no map was used for this analysis. This figure was created using python 3.8 and the matplotlib (3.5.0) package (https://matplotlib.org).

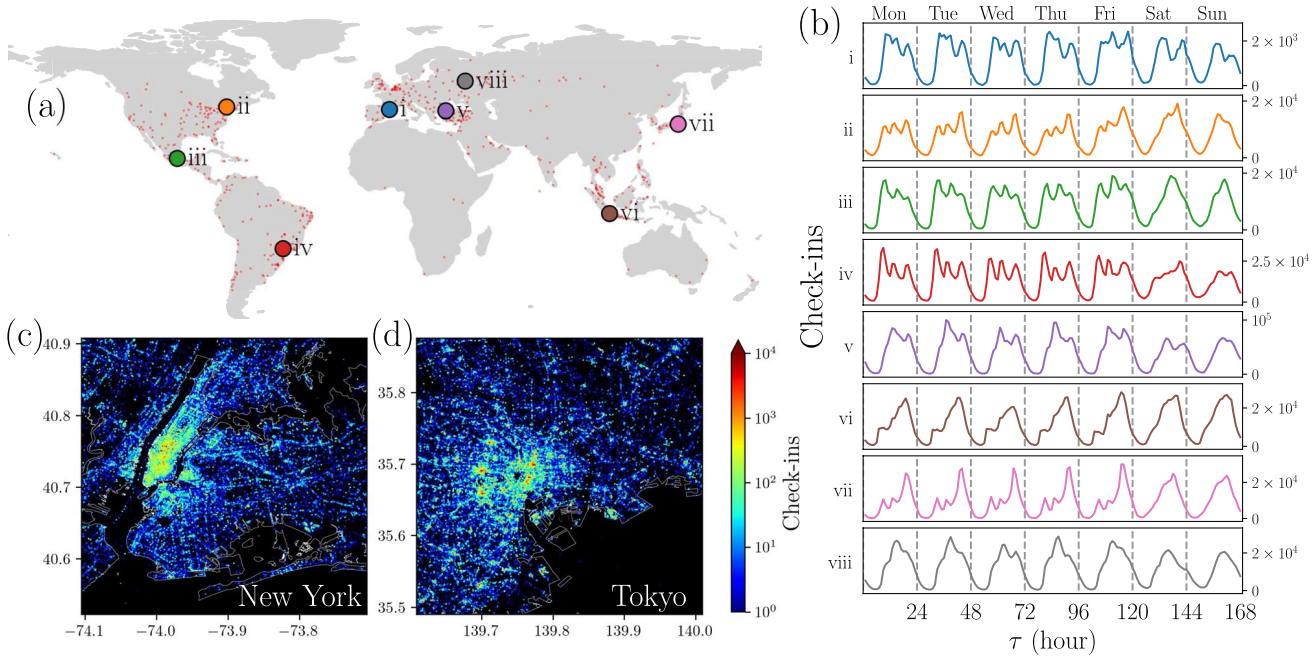


Figure 2. Spatial and temporal analysis of check-ins in different cities. (**a**) 632 cities worldwide analyzed, represented with small dots, and eight selected cities: (i) Barcelona, (ii) New York, (iii) Mexico City, (iv) Sao Paulo, (v) Istanbul (vi), Jakarta, (vii) Tokyo, and (viii) Moscow. (**b**) Frequency counts for all the check-ins in the cities (i) to (viii) presented in (**a**) gathered using the local time when they were made. The counts reported in the histograms correspond to the number of check-ins in each hour of the week, from the first hour on Monday to the last hour on Sunday. (**c**) Points of interest in the Foursquare dataset in New York and (**d**) Tokyo. In this representation, the number of check-ins in the respective POI is codified in the colorbar. All figures were created using python 3.8 and the matplotlib (3.5.0) package (https://matplotlib.org). The map in panel (**a**) was created using the geopandas (0.12.1) package (https://geopandas.org). Land borders in panels (**c**) and (**d**) were obtained from Open Street Maps Boundaries webpage (https://osm-boundaries.com/).

been shown for other phenomena[16,37,38]. But, at this level, differences in patterns become noticeable; the part of the day that concentrates most of the activity, the maximum number of check-ins in a day, and the change of the activity patterns for weekdays and weekends describes the collective behavior of city inhabitants that varies from one city to another.

In Fig. 2c,d, we show the spatial distribution of POIs for two cities: New York and Tokyo. When placing a point at the coordinates of each POI, different spatial patterns emerge directly linked to urban infrastructure in the first place. Street grids and blocks in urban areas can be identified. But some patterns emerge, which are very different in each case, related to the quantity and density of the POIs. For example, a large area of high density can be seen in the case of New York, in the Manhattan area, while in Tokyo the high-density areas form more compact and dispersed clusters in a larger region. In addition, the attractiveness (total number of check-ins in a POI) of the venues is depicted in the colorbar that codifies the number of check-ins. The statistical analysis of the number of check-ins at each POIs reveal characteristics of a complex urban environment with a power-law behavior in which most POIs register a lower number of check-ins while a few POIs have an attractiveness several orders of magnitude greater. Again, a similar result has been reported by Yang et al.[23], at a global scale for the same dataset. This power-law behavior in the attractiveness of POIs is also observed at different scales in other studies; for example, in the attractiveness of sites measured from the activity of taxis in New York City[7] or the importance of airports in the United States[43]. In this context, our findings in Fig. 2c,d show that the activity of users in Foursquare is associated with the complexity in the distribution of POIs. This feature is observed in all the cities in this study.

In Fig. 3 we extend the temporal analysis of check-ins to $N = 632$ cities in 87 countries. We consider activity counts as in panels in Fig. 2b, all the information of check-ins in a city i were gathered using the corresponding local time τ. The activity counts are normalized over the week to obtain the relative frequency or probability $\mathscr{P}_i(\tau)$. The time τ can take the values from 1 (associated to the first hour of Monday) to 168 (the last hour of Sunday). In Fig. 3a, we present the probability $\mathscr{P}_i(\tau)$ of check-in at times τ in cities denoted by $i = 1, 2, \ldots, 632$. The results reveal marked differences between low nocturnal activity and high daytime activity. Also, patterns emerge in all of them by grouping check-ins by their local time τ of occurrence. For all the cities, fluctuations in the values of $\mathscr{P}_i(\tau)$ are small at night, but during the day we see different behaviors of the curves describing each city with all kinds of deviations; for example, in some cities $\mathscr{P}_i(\tau)$ changes considerably in the early hours of the morning, others at night, whereas other cities differ considerably at weekends. To show these variations more clearly, in Fig. 3b we depict $\mathscr{P}_{\text{Total}}(\tau)$, obtained for all the check-ins in the $N = 632$ cities. This quantity is equivalent to calculate the average between all cities

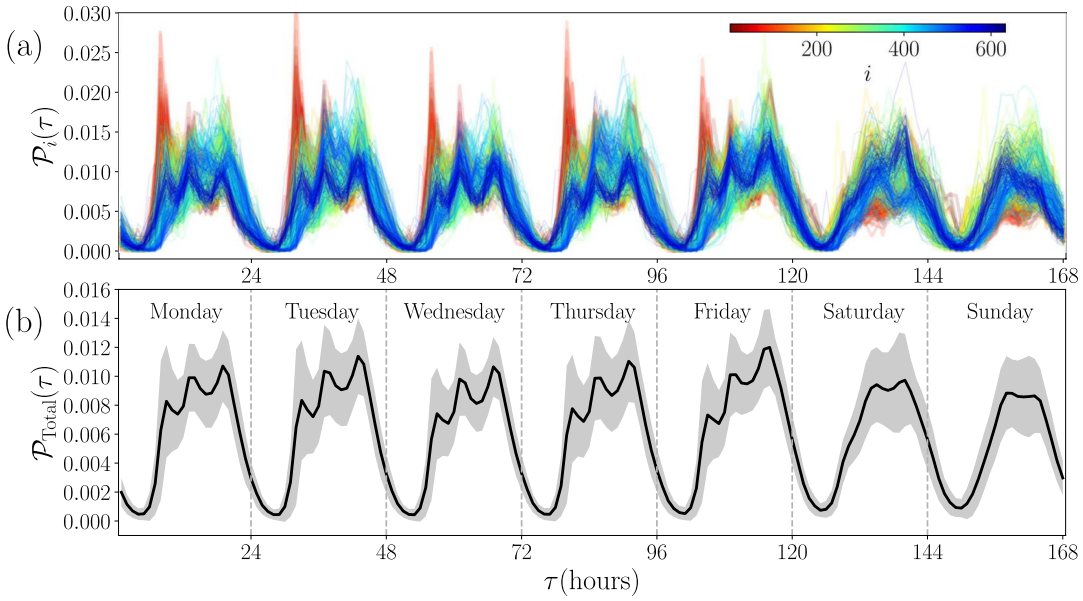


Figure 3. Temporal analysis of check-ins in 632 cities. (**a**) Probability $\mathscr{P}_i(\tau)$ of check-in at times τ in $N = 632$ cities denoted by $i = 1, 2, \ldots, N$. All the check-ins in each city were gathered by the local time when each register was made. Frequency counts are normalized over the week to obtain each $\mathscr{P}_i(\tau)$. The time τ can take the values from 1 (the first hour of Monday) to 168 (the last hour of Sunday), and the results for each city i (codified in the colorbar) are presented with continuous lines. (**b**) The probability $\mathscr{P}_{\text{Total}}(\tau)$ obtained for all the check-ins in the 632 cities is plotted with a continuous line. In this case, the standard deviation $\sigma(\tau)$ for all the check-ins and all the cities in (**a**) evaluated at time τ defines variations of this probability. The results are shown as a region defined by $\mathscr{P}_{\text{Total}}(\tau) \pm \sigma(\tau)$ presented with gray color. All figures were created using python 3.8 and the matplotlib (3.5.0) package (https://matplotlib.org).

$$\mathscr{P}_{\text{Total}}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \mathscr{P}_i(\tau). \tag{1}$$

On the other hand, the respective standard deviation $\sigma(\tau)$ of the values $\mathscr{P}_i(\tau)$ give us a measure of the differences observed in the cities considered. Our findings are shown in Fig. 3b as a shaded region defined by $\mathscr{P}_{\text{Total}}(\tau) \pm \sigma(\tau)$. The dispersion of the values in particular hours of the day can be seen, and noticeable variations between weekdays and weekends can be observed.

Comparison of temporal activity between cities using the Kullback–Leibler divergence. The variety of results observed for the probabilities $\mathscr{P}_i(\tau)$ in Fig. 3a motivates the exploration of a criterion to compare the temporal activity between two particular cities i, j. To this end, we use the Kullback-Leibler divergence, also known as relative entropy, defined by[40,44]

$$\mathscr{D}_{KL}(i,j) = \sum_{\tau} \mathscr{P}_i(\tau) \log\left[\frac{\mathscr{P}_i(\tau)}{\mathscr{P}_j(\tau)}\right], \tag{2}$$

where the sum in time τ ranges from 1 to 168 (all the hours in a week) and $i, j = 1, 2, \ldots, N$ (see the "Methods" section for details). The Kullback-Leibler divergence satisfies our interest to compare the temporal distributions of check-ins, but its definition does not generate a symmetric measure since $\mathscr{D}_{KL}(i,j)$ is different from $\mathscr{D}_{KL}(j,i)$. This would cause that even if a city i is determined to be similar to another city j, j is not necessarily similar to i. However, the average of the Kullback-Leibler divergence between pairs (i, j) and (j, i)

$$\mathscr{D}_{KLS}(i,j) \equiv \frac{\mathscr{D}_{KL}(i,j) + \mathscr{D}_{KL}(j,i)}{2} \tag{3}$$

is symmetric. In this manner, similar distributions produce a small value and dissimilar ones are associated with larger values. Then, this symmetric quantity is adequate to describe the similarity between the temporal distributions of cities.

In Fig. 4, we present the results obtained from the evaluation of $\mathscr{D}_{KLS}(i,j)$, in Eq. (3), to compare the temporal activity presented in Fig. 3a for all the cities $i, j = 1, 2, \ldots, 632$, considering the information of user's check-ins in Foursquare. In Fig. 4a we present the statistical analysis of $\mathscr{D}_{KLS}(i,j)$ for all pairs of cities. The results are depicted as a probability density $\rho(\mathscr{D}_{KLS})$ for the values $\mathscr{D}_{KLS}$, that is, we calculate the values of $\mathscr{D}_{KLS}(i,j)$ for all the pairs (i, j). With these values, we obtained the histogram shown. We show, as an inset, the $N \times N$ matrix with elements

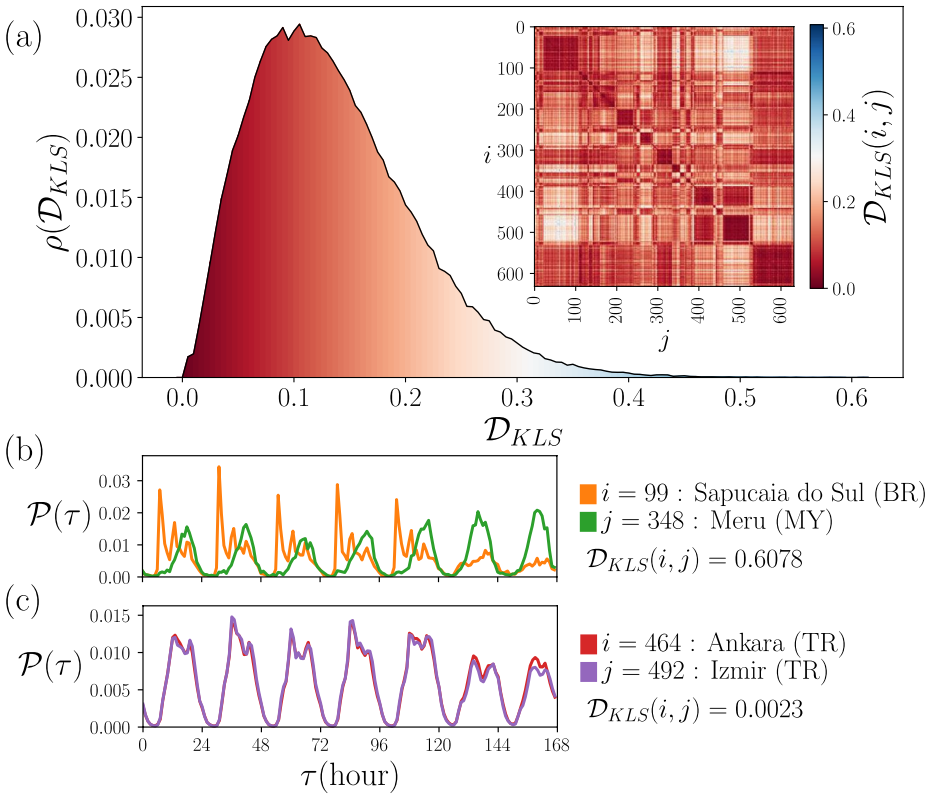


Figure 4. Comparison of temporal activity between cities. (**a**) Statistical analysis of the symmetric Kullback-Leibler distances $\mathscr{D}_{KLS}$ from the comparison of check-ins temporal distribution by pairs. The values are obtained from Eq. (3) for all the city pairs $i, j = 1, \ldots, 632$. The probability density $\rho(\mathscr{D}_{KLS})$ is obtained using bin counts in intervals with $\Delta\mathscr{D}_{KLS} = 0.005$. The matrix with all the elements $\mathscr{D}_{KLS}(i,j)$ is presented as an inset. (**b**) The two most different probability densities of the set are of the cities Sapucaia do Sul in Brazil and Meru in Malaysia, with $\mathscr{D}_{KLS} = 0.6078$; the maximum value found. (**c**) The two most similar cities of our sample, are Ankara and Izmir both in Turkey, with $\mathscr{D}_{KLS} = 0.0023$, the minimum non-null value. All figures were created using python 3.8 and the matplotlib (3.5.0) package (https://matplotlib.org).

$\mathscr{D}_{KLS}(i,j)$; the respective values are codified in the colorbar. The results for $\rho(\mathscr{D}_{KLS})$ show that a high fraction of the entries in the matrix have values $\mathscr{D}_{KLS}$ around 0.1. For the diagonal elements we have $\mathscr{D}_{KLS}(i,i) = 0$. On the other hand, the maximum value for two different cities is $\mathscr{D}_{KLS} = 0.6078$, and occurs between the Brazilian city of Sapucaia do Sul and Meru, in Malaysia, as shown in Fig. 4b. At the opposite extreme, the minimum non-null value reached is $\mathscr{D}_{KLS} = 0.0023$, found for the comparison between Ankara and Izmir, both in Turkey. The respective probabilities are shown in Fig. 4c. The results observed in Fig. 4b are reasonable since we are comparing the activity in two complete different urban areas. In contrast, for the cities considered in Fig. 4c, the resemblance between these two cities is remarkable considering that they are more than 520 km apart. Additionally, a review of the data shows that of the 82,285 active users in Ankara and the 90,923 active users in Izmir, only 11,141 made check-ins in both cities; this represents only 6.87% of the total users with activity in these regions. The similarity in their patterns is not explained by common users but by comparable urban behavior in the same country.

Networks and temporal patterns between cities. In this section, we apply methods of network science to analyze the similarities between cities. To this end, we define a network in which nodes represent cities and links the similarity relationship between cities. In this way, two nodes are connected if the respective cities have similar temporal activity. To generate this structure, it is necessary to define what is considered sufficiently similar. We use a threshold value H. If two cities have values $\mathscr{D}_{KLS}$ in Eq. (3) lower or equal than H then these cities are considered similar. All this information defines a similarity network for each value H. The respective $N \times N$ adjacency matrix is denoted as $\mathbf{A}(H)$, with elements i, j given by

$$A_{ij}(H) = \begin{cases} 1 & \mathscr{D}_{KLS}(i,j) \leq H, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Additionally, we require $A_{ii}(H) = 0$ for $i = 1, 2, \ldots, N$, to avoid self loops. From the symmetry of the distance $\mathscr{D}_{KLS}$, follows the symmetry of $\mathbf{A}(H)$, defining an undirected network.

In Fig. 5, we analyze similarity networks as a function of H. In Fig. 5a we depict the adjacency matrix $\mathbf{A}(H)$ for different values of H. In this manner, for each H all the information of $\mathscr{D}_{KLS}(i,j)$ in Fig. 4a is converted into

a binary matrix with entries 0 and 1. In Fig. 5b, we present the fraction of nodes that belong to the Largest Connected Component, $\nu_{LCC} = S_{LCC}/N$, where S_{LCC} is the size of the Largest Connected Component (LCC). The LCC obtained for the values of H explored in Fig. 5a are shown as insets.

In the results in Fig. 5b, it is worth noticing that for $H = 0.005$, the network is formed by disconnected subnetworks with a few nodes and the LCC is a linear graph with 6 nodes; for $H < 0.0023$ each node is disconnected. On the other hand, the network is fully connected for $H > 0.6078$; the maximum value of $\mathscr{D}_{KLS}$ found. The transition between these two limits gives rise to a convenient choice of H corresponding with a network that captures the nature of the similarity between the cities that we are analyzing. The results for ν_{LCC} reveal that the size of the LCC contains more than 90% of the nodes at $H = 0.031$, 99% at $H = 0.052$. On the other hand, $H = 0.0723$ is the minimal value of the similarity threshold that produces a connected network that includes all the $N = 632$ nodes (cities). Finally, we see that around $H = 0.02$, ν_{LCC} suffers an abrupt change with H that is analogous to a percolation threshold[39], separating two regimes: one with $\nu_{LCC} < 0.3$, defining small sub-networks with high similarity in the temporal information and a second one with $\nu_{LCC} > 0.6$ where the connected networks incorporate a high fraction of the cities.

On the other hand, in order to explore the relationship between the edges in each network generated with the value H and the geographical distance between cities, we define

$$\langle D\rangle_H \equiv \frac{\sum_{i,j=1}^{N} A_{ij}(H) d_{ij}}{\sum_{i,j=1}^{N} A_{ij}(H)} = \frac{\sum_{i,j=1}^{N} A_{ij}(H) d_{ij}}{2\,\mathscr{E}\ (H)}, \tag{5}$$

where d_{ij} denotes the geographical distance between cities i and j and $\mathscr{E}\ (H)$ is the number of edges of the graph associated with $\mathbf{A}(H)$ (see "Methods" section for details on the calculation of geographical distances between cities). In Fig. 5b, we present as an inset the value $\langle D\rangle_H$ as a function of H. The results show that for small H, cities having very similar temporal activity histograms are also geographically closer. These average distances remain relatively low up to the network with $H = 0.05$. However, after this value, the average distances grow to 8512 km which is the average distance between all the cities analyzed.

Once we established a criteria to build similarity networks using the temporal activity of users of Foursquare in Fig. 3a, we can apply community detection algorithms. The concept of community has emerged in network

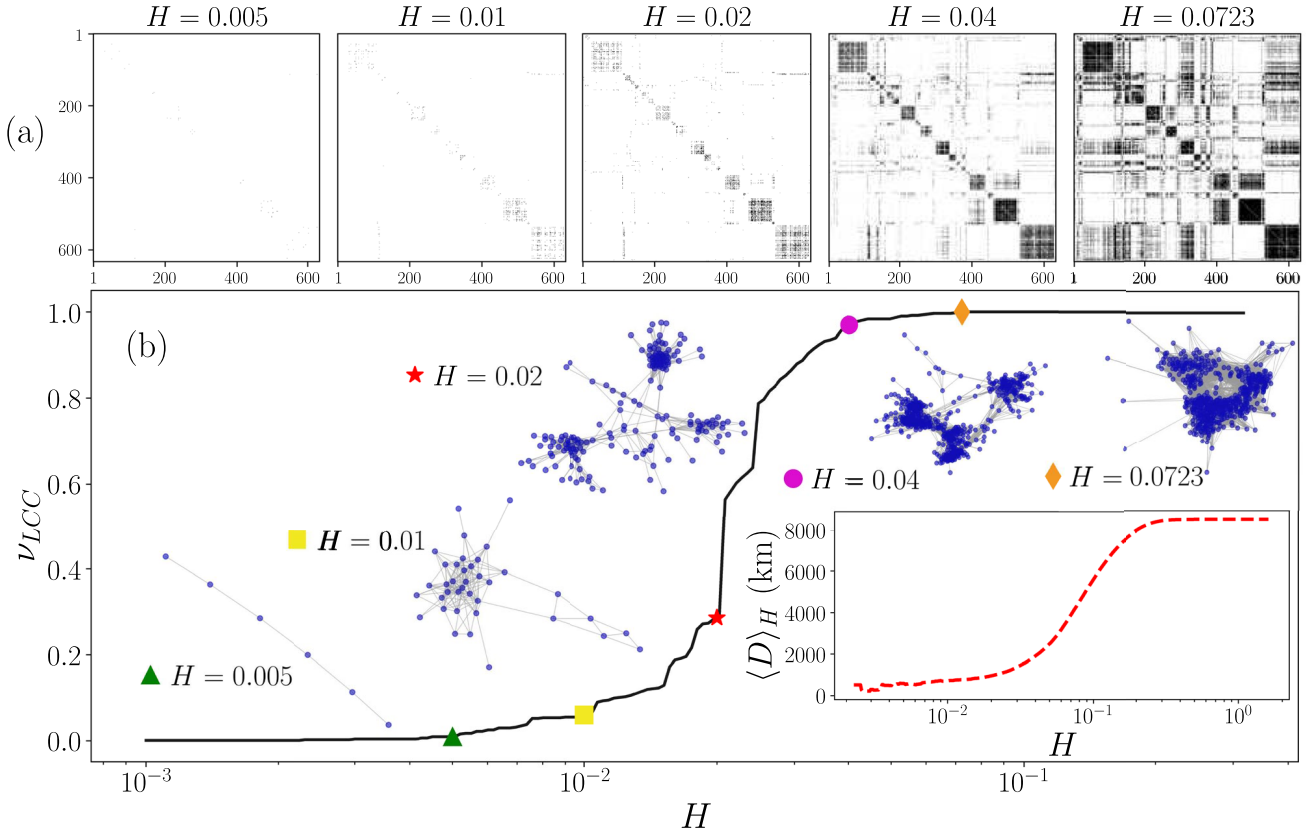


Figure 5. Similarity networks generated using different threshold values H. (**a**) Adjacency matrices $\mathbf{A}(H)$ with elements given by Eq. (4) using the values $H \in \{0.005, 0.01, 0.02, 0.04, 0.0723\}$, binary entries $A_{ij}(H)$ are depicted in white for 0 and black for 1. (**b**) Fraction of nodes in the largest connected component ν_{LCC} as a function of H in the interval $0.001 \leq H \leq 0.32$. The largest connected component of the networks generated with $\mathbf{A}(H)$ in panel (**a**) are presented and the inset shows $\langle D\rangle_H$ calculated using Eq. (5). All figures were created using python 3.8 and the matplotlib (3.5.0) package (https://matplotlib.org).

science as a method for finding groups within complex systems identifying sub-networks with statistically significantly more links between nodes in the same group than nodes in different groups[45–47]. In our similarity network, these communities represent groups of cities with comparable activity $\mathscr{P}_i(\tau)$. In Fig. 6 we present the results for a network with $N = 632$ nodes generated with $H = 0.0723$.

In Fig. 6a we depict five communities $\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_5$ detected using the Louvain's algorithm[48] implemented in the library NetworkX in Python[49], see "Methods" section for a description of this algorithm. The number of nodes on each community $\mathscr{C}_s$ (with $s = 1, 2 \ldots, 5$) is denoted by $\mathscr{N}_s$ and, from this analysis we obtain groups of cities with $\mathscr{N}_1 = 148$, $\mathscr{N}_2 = 136$, $\mathscr{N}_3 = 133$, $\mathscr{N}_4 = 113$, and $\mathscr{N}_5 = 102$. In Fig. 6b we plot with thin gray lines the probabilities $\mathscr{P}_i(\tau)$ showed in Fig. 3a in groups defined by each community $\mathscr{C}_s$; each panel contains $\mathscr{N}_s$ curves. In addition, we include the statistical analysis considering all the check-ins in each group; the results are shown with colored thick lines. When grouped in this way, the curves observed within each community $\mathscr{C}_s$ are similar, evidencing the fact that they have the same shape as the averages.

The average curves in Fig. 6b show that the $\mathscr{C}_1$ community, whose average is plotted in blue, has a behavior from Monday to Friday characterized by three peaks throughout the day (at 8, 12 and 18 hours). On weekends, this pattern is broken and a single maximum is observed around 20 hours on Saturday and at noon on Sunday. The $\mathscr{C}_2$ community, with the orange average line, is characterized by a pronounced maximum at 13 hours and a second relative maximum at 19 hours, from Monday to Friday. This behavior is maintained on weekends but with less check-ins. The community $\mathscr{C}_3$, with a curve in green, is the one with the least contrast between the shape from weekdays versus weekends. Every day the maximum is found at 18 hours. Still, from Monday through Friday there is a small local maximum at 8 hours that disappears on the weekend. $\mathscr{C}_4$ with the average presented in red, has the same features as $\mathscr{C}_1$ but with different relative sizes; in this case, the first daily maximum dominates over

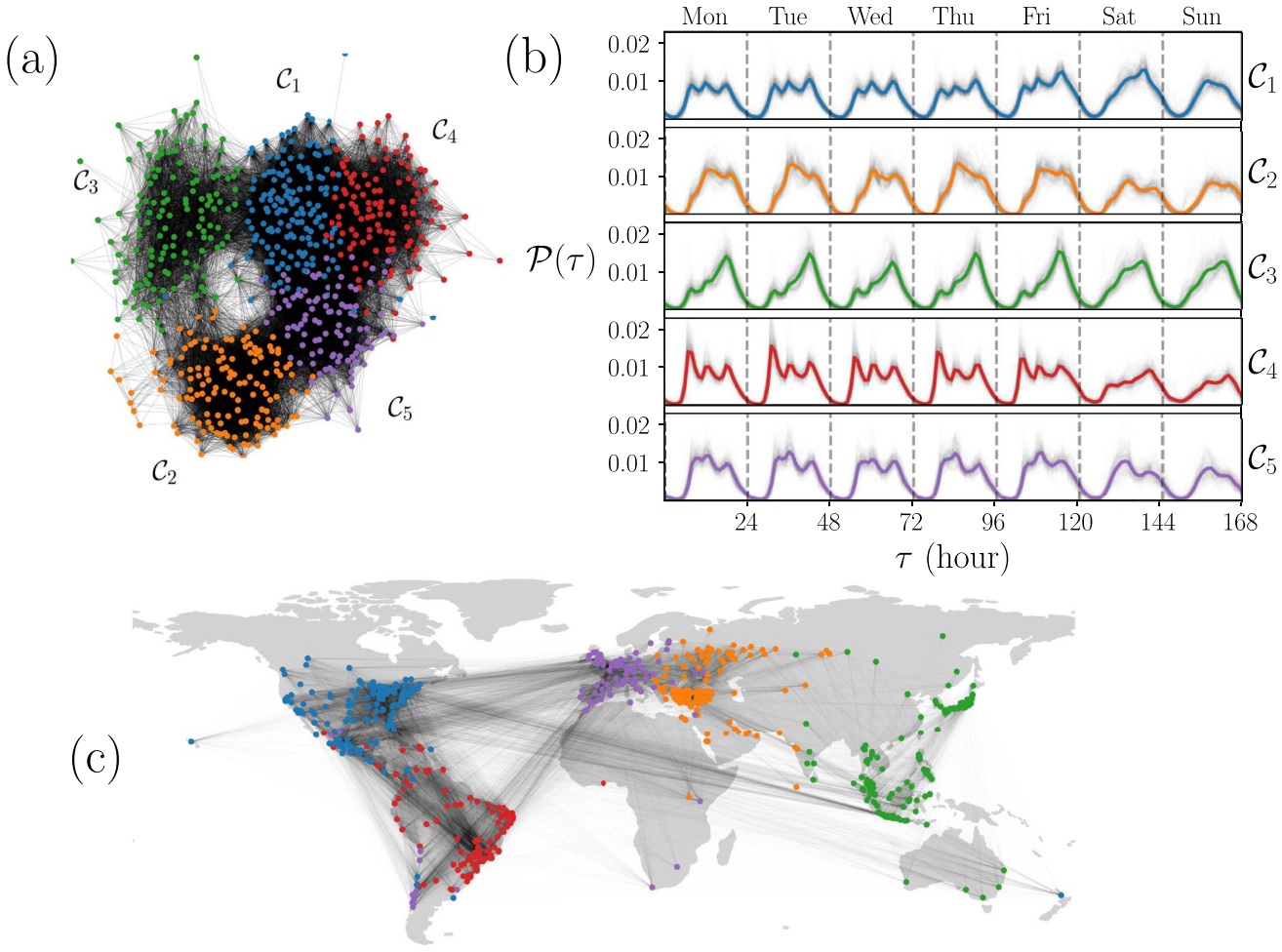


Figure 6. Human activity patterns identified using community detection in similarity networks. (**a**) Community structure of the similarity network generated with $H = 0.0723$, five communities $\mathscr{C}_1, \mathscr{C}_2, \mathscr{C}_3, \mathscr{C}_4, \mathscr{C}_5$ were detected using the Louvain's algorithm and are represented with different colors. (**b**) Probability $\mathscr{P}(\tau)$ for the temporal activity of the cities in each community. Thin gray lines present the curves $\mathscr{P}_i(\tau)$ depicted in Fig. 3a whereas the colored thick line represents the statistical analysis of all the check-ins in the cities of the community. (**c**) Geographical representation of the network in (**a**). In this case, each node is a city depicted on a world map. All figures were created using python 3.8 and the matplotlib (3.5.0) package (https://matplotlib.org). The map in panel (**c**) was created using the geopandas (0.12.1) package (https://geopandas.org).

the rest. The $\mathscr{C}_5$ community is the smallest and whose average behavior is depicted in purple. In this case the curve suffers different changes throughout the week. While there is a pattern with maxima at 9, 12 and 19 hours, and a valley at 15 hours, from Monday to Friday, the relative sizes are not always equal; on Monday the first and second peaks dominate, on Tuesday the second maximum is the largest, on Wednesday the three are practically the same size, on Thursday the first two peaks almost merge and dominate over the third, and on Friday the first almost disappears while the second peak dominates. Finally, on Saturday there is a change giving rise to two maximums at 13 and 20 hours, also present on Sunday although smaller.

In addition, considering that each probability curve $\mathscr{P}_i(\tau)$ corresponds to a node i, which in turn is a city with given coordinates, in Fig. 6c we plot the network with each node located in a world map. The node colors and the whole network connectivity are the same as in Fig. 6a. This network representation shows that the communities formed from the similarity of city behaviors correspond to well-defined geographic regions. Cities in North America belong predominantly to the $\mathscr{C}_1$ community; those in Eastern Europe and the Middle East belong to $\mathscr{C}_2$; the $\mathscr{C}_3$ community is composed of several cities in Eastern India, and cities in East Asia and Oceania; $\mathscr{C}_4$ contains most cities in South America, and most of the cities in Western Europe belong to the $\mathscr{C}_5$ community. This is a remarkable result that we will discuss further below.

Once we defined the communities of the network, we can compare the fraction of intra-community and inter-community links. Defining L_s as the number of links between nodes in community $\mathscr{C}_s$, and L_{st} as the number of links between a node in $\mathscr{C}_s$ and a node in $\mathscr{C}_t$, we calculate the fraction of inter-community links as $L_{st}/(\mathscr{N}_s\mathscr{N}_t)$ and intra-community links as $2L_s/(\mathscr{N}_s(\mathscr{N}_s-1))$. The values obtained allow us to compare the number of links with the total number of possible links. The values for the fraction of intra-community links are 0.79, 0.68, 0.51, 0.76 and 0.75 for $\mathscr{C}_1,\ldots,\mathscr{C}_5$, respectively. Regarding the fraction of inter-community links, the values are below 0.1 except between communities $\mathscr{C}_1$ and $\mathscr{C}_4$ (North America and South America) with 0.32 and, with 0.26, between $\mathscr{C}_1$ and $\mathscr{C}_5$ (North America and Europe); there are a fraction 0.15 of inter-community nodes between $\mathscr{C}_2$ (Middle East) and $\mathscr{C}_5$ (Europe), and 0.12 between $\mathscr{C}_5$ and $\mathscr{C}_4$ (South America). The results also show that $\mathscr{C}_3$ has little similarity with the other communities. Europe is a community with many elements in common with America and the Near East. East Asia, on the other hand, has few elements in common with the rest of the world in terms of temporal patterns of human behavior. These features are observed in the map in Fig. 6c evidencing cultural and historical features of each region.

Finally, it is worth mentioning that, unlike other cluster identification methods, e.g., agglomerative clustering (discussed in the next section), information on similarity between specific regions within the same communities is preserved in the individual links. In Fig. 6c, it is clear the high density of links between North America and the United Kingdom, as well as Brazil and Portugal, regions with cultural and linguistic relationships. The results also show that not all cities belong to the same community as would be expected based on their geographic region. Specifically, there are 8 cities in Chile, 3 in Mexico and 1 in Uruguay that are more similar to European cities than to American ones.

Identification of patterns using machine learning. Machine learning can be used as an alternative approach, with many algorithms involved, to classify objects or patterns in a very efficient way[41,50]. In particular, we used the unsupervised hierarchical agglomerative clustering to classify our temporal patterns of activity in cities. When we apply this algorithm to our data set, we obtained five clusters, as shown in Fig. 7. This number of clusters or communities is obtained by maximizing the modularity. The clusters of cities detected are presented in Fig. 7a. For comparison, these clusters are depicted using the same colors as those used in Fig. 6c. In this classification of the dataset of check-ins, each cluster obtained by this procedure is denoted as $\mathscr{C}_s{}'$ and contains $\mathscr{N}'_s$ elements (with $s=1,2\ldots,5$). We have: $\mathscr{N}'_1=178$, $\mathscr{N}'_2=128$, $\mathscr{N}'_3=140$, $\mathscr{N}'_4=90$, and $\mathscr{N}'_5=87$. A comparison of Fig. 7a and our findings in Fig. 6c shows that the differences between $\mathscr{N}_s$ and $\mathscr{N}'_s$ are mainly due to cities from South America and the Caribbean that were in $\mathscr{C}_1$ (blue) are now in $\mathscr{C}_4{}'$ (red), and between Russian

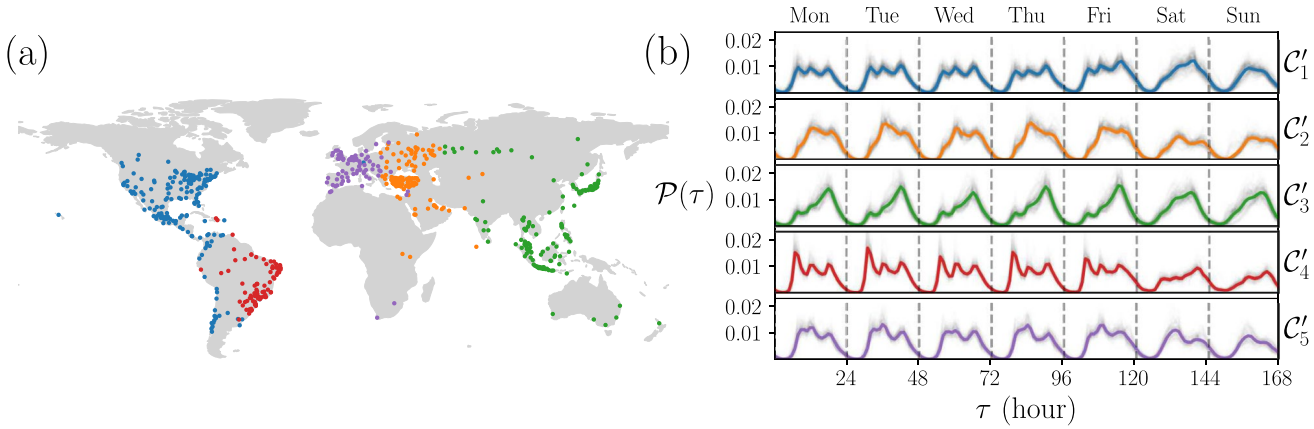


Figure 7. Pattern identification using agglomerative hierarchical clustering. (**a**) Geographical representation of 5 clusters detected. (**b**) Activity patterns in each group. All figures were created using python 3.8 and the matplotlib (3.5.0) package (https://matplotlib.org). The map in panel (**a**) was created using the geopandas (0.12.1) package (https://geopandas.org).

and Indian cities that were in $\mathscr{C}_2$ (orange) now belong to $\mathscr{C}_3{}'$ (green). In addition, Fig. 7b shows the probabilities $\mathscr{P}_i(\tau)$ grouped according to this classification. Again, in each panel, corresponding $\mathscr{C}_s{}'$, the $\mathscr{N}'_s$ curves of the cities are shown in thin gray, while the average behavior for each time τ is shown with a colored thick line. The resemblance to the behaviors shown in Fig. 7b is remarkable showing that the emergent patterns in the dataset can be detected by different algorithms.

Discussion

In this work, we use Foursquare check-ins as a proxy of human activity in cities. The data set explored provides vast information about people's interests, site characteristics, and behavioral patterns in cities, among many others. The information analyzed contains 90,048,627 check-ins made by 2,733,324 users from April 2012 to January 2014 with active check-ins obtained from registers on Twitter by searching a hashtag generated automatically for users that linked their Foursquare activity with that social network. Information from 632 cities from 87 countries, with more than 10,000 POIs, was used. We explore the spatial properties of POIs and check-ins and the temporal features of the check-ins which, as self-reported spatio-temporal interactions between people and places, provide valuable information about activities carried out in cities. In particular, we analyze statistically the check-ins per hour on weeks for each city.

From the probabilities describing the temporal activity of each city, we apply a symmetrized Kullback-Leibler divergence to compare these probabilities across all 632 cities. From this information, we define similarity networks in terms of a threshold value H to decide if two cities have a similar activity or not. We explore the size of the largest connected component as a function of H; in particular, we found that around $H = 0.02$ a critical percolation threshold exists, whereas for $H = 0.0723$ the largest connected component includes all the nodes of the network. Our findings reveal collective emergent behaviors that go beyond the spatial and mobility aspects that define metropolitan areas, megacities, or functional urban areas, and allows for the tracking of higher-order structures, as was proven in some cities that have great similarity, despite the fact that their geographical separation prevents them from being the same.

We apply Louvain's algorithm for community detection to the similarity network generated with $H = 0.0723$; the results define 5 groups of cities with similar activity of check-ins. By locating all the cities in a map, we notice a remarkable pattern: The five communities of cities belong to five different regions that correspond to different continents. In addition, we use a Machine Learning algorithm to classify the activity patterns obtained. Although this method is completely different from the Network Science approach, we nevertheless obtained basically the same classification. Specifically, the Machine Learning algorithm that we used was the Unsupervised Agglomerating Clustering[41,50]. Both approaches, although very different, lead to the same classification of five different communities worldwide. Even more surprising, these five communities clearly correspond to five different regions on a planetary scale: 1) North America, (2) South America, (3) West Europe, (4) East Europe and Middle East, and (5) East Asia.

It can be clearly noticed that practically in all cities there are these patterns of maxima and minima of human activity at the same hour during the week. At the same time, we notice a different pattern during the weekdays and another one slightly different during weekends. These patterns are very robust in the sense that the maxima and minima are clearly noticeable, even in the average curves involving the 632 cities in the analysis. Since these 632 cities are distributed in many parts of the world, spanning 5 continents, we think that the patterns are related to some universal aspects of human behavior. For instance, the fact that we humans are universally a diurnal species. This is one of the so-called human universals in the literature of human evolution. This fact can be seen in the minima of activity for all cities at deep night (around of 2 or 3 am). This universal periodicity is also related to circadian rhythms in humans and other internal clocks studied in chronobiology. However, the 3 peaks during the day, which appear early in the morning (between 6 and 7 am), at noon (around 12 or 13 hours), and in the afternoon (around 19 and 20 hours), can be related more with modern routines associated with working hours.

In summary, using a vast data set from location-based social networks, we analyze 632 cities of 87 countries around the world to obtain temporal patterns of human activity that characterize points of interest within the cities. Using network science and machine learning algorithms we unravel communities with different patterns of human behavior. This suggests that human activity patterns can be universal with some geographical, historical and cultural variations on a planetary scale.

Methods

Foursquare data and demography. In this section, we recall some recent demographics of Foursquare. The user demographics is the following[51]:

- urban communities: 64%, towns and rural communities: 36%.
- medium-sized town: 28%, large city: 26%

This means that we have users of Foursquare from both urban, town and rural communities. Even though the urban percentage is larger, as expected, the difference is not that high; the rural or town percentage is roughly 1 in 3. This means that the data set used in our study reflects not only an urban feature but reflects small town or rural features as well.

On the other hand, the gender and age distribution is as follows[52]:

- male: 52%, female: 48%.

Country		POIs		Check-ins		Users
Code	Name	Number	%	Number	%	Number
US	United States	1990327	17.80	12778097	14.19	426341
ID	Indonesia	1198611	10.72	7765315	8.62	361193
BR	Brazil	1159258	10.37	9991354	11.10	261079
TR	Turkey	1098373	9.82	17500113	19.43	592630
RU	Russia	546532	4.89	4291601	4.77	122268
JP	Japan	519409	4.65	4784080	5.31	81293
MY	Malaysia	493453	4.41	4926145	5.47	127390
MX	Mexico	408434	3.65	3981409	4.42	147563
TH	Thailand	353444	3.16	2633608	2.92	82765
PH	Philippines	219097	1.96	1998063	2.22	60197
ES	Spain	212161	1.90	1083153	1.20	67638
GB	United Kingdom	210777	1.89	1271622	1.41	77949
IT	Italy	197007	1.76	867931	0.96	52394
CL	Chile	195226	1.75	2209981	2.45	53714
DE	Germany	142347	1.27	623759	0.69	45574

Table 1. Foursquare data set by country. The first 15 countries sorted by the number of points of interest (POI) are shown, as well as the number of check-ins, users, and the percentage of the total dataset that represents.

- ages: 18-24 years old: 19%, 25-34 years old: 32%, 35-44 years old: 20%, 45-54 years old: 14%, 55-64 years old: 10%, 65+ years old: 6%.

Here we notice that the percentage of males and females is basically the same: half and half. Therefore, our data set and analysis can be applied to both females and males. Regarding age distribution, we notice that about half (50%) of users of Foursquare have ages between 25 and 44 years old. This is expected since people at that age: young adults, are the ones that show more mobility and activity. It is clear that the mobility, activity and interest in reporting through check-ins to the social network tends to diminish with age. However, the distribution is somewhat broad.

Regarding the type of places visited and checked as POIs, it is worth mentioning that the categories are very broad. There are in Foursquare more than 1100 venue categories, distributed in ten major Foursquare categorization groups of points of interest POIs[53]: Arts and Entertainment, Business and Professional Services, Community and Government, Dining and Drinking, Event, Health and Medicine, Landmarks and Outdoors, Retail, Sports and Recreation, Travel and Transportation. This points to a large diversity of interests that are captured in the data set that we used in our analysis.

Other relevant data about Foursquare are the following[54,55]:

- Foursquare Places has over 100 million POIs across 247 countries and territories, as of January 4, 2023.
- 100M+ global POI
- 200+ countries and territories
- 14 billion user verified check-ins
- 1100+ venue categories
- 1 billion + photos, tips, reviews

Regarding the numbers of the particular data set used in this research, let us point out that we used data of Foursquare from April 2012 to January 2014. All the details of the data set and the processing, data mining, and data analytics, can be found in the following sections: Discussion, Methods—Dataset description, and Methods—Grouping POIs by urban area. Many of the results can be found in Tables 1 and 2. Just to summarize the size of the data set studied, let us list the numbers involved:

1. User: 2,733,324
2. POIs: 11,180,160
3. Check-ins: 90,048,627
4. Countries: 253 (87 of which include 99% of the data)
5. Cities: 6,463 (632 with more than 10,000 check-ins)
6. Continents: 5
7. Categories of POIs: 519

Given all these demographics together with the numbers, diversity, and vastness of the data set analyzed, and the use of aggregate data, the results can be a good description of the temporal activity in cities worldwide.

City	Country	POIs	Check-ins	POIs / km^2	Check-ins PC
Istanbul	TR	334517	7343552	249.640	0.520404
Jakarta	ID	398154	2908026	79.488	0.080083
Kuala Lumpur	MY	200199	2558716	150.752	0.403605
Tokyo	JP	191162	2405876	35.946	0.072842
Mexico City	MX	137655	1804660	65.116	0.092265
Bangkok	TH	169220	1741638	65.896	0.118231
Moscow	RU	160868	1643199	85.477	0.116726
São Paulo	BR	137494	1478616	68.576	0.077356
Izmir	TR	74237	1443711	210.303	0.516692
Quezon City [Manila]	PH	123941	1343955	61.055	0.061959
New York	US	137841	1311179	25.602	0.082202
Santiago	CL	89115	1276748	123.771	0.201507
Ankara	TR	59610	1021287	158.537	0.340152
Bandung	ID	119357	933686	117.709	0.114121
Singapore	SG	72981	827715	83.027	0.119596
Surabaya	ID	110852	756628	63.308	0.090586
Los Angeles	US	91011	698605	16.157	0.048916
Osaka [Kyoto]	JP	66574	667631	21.081	0.042544
Yogyakarta	ID	83635	609749	53.785	0.119592
Rio de Janeiro	BR	55765	570441	40.794	0.058220
Belém	BR	39125	490162	143.842	0.234762
Chicago	US	55319	485198	14.444	0.071565
Saint Petersburg	RU	56800	482715	107.780	0.112237
Kuwait City	KW	43547	473904	91.485	0.149590
London	GB	48778	430524	26.168	0.044801
Lima	PE	34989	404408	39.942	0.043646
Denpasar	ID	54399	393148	130.453	0.209318
Bursa	TR	25214	392454	120.067	0.233577
Manaus	BR	34633	390084	132.693	0.193399
Medan	ID	46528	387255	62.961	0.097886
Seoul	KR	79306	382646	32.383	0.017714

Table 2. Foursquare data by city. The 31 cities with the highest number of check-ins are shown. This selection contains information on a wide variety of countries with different geographic, cultural, linguistic, economic, and religious characteristics.

Dataset description. Check-ins information was collected by Yang, et al.[22,23,28,29] and it is publicly available[42]. The datasets are divided into two sets. The first one is a list of check-ins obtained through an automated search on Twitter with the help of its API streaming service[29]. Foursquare allows linking users' accounts with other social media such as Twitter or Facebook, to share the check-ins on these platforms too. If this is the case, when one user registers their presence in a place, automatically, a tweet or a post is generated with the information of the check-in and some elements like hashtags and the URL of the venue's page at Foursquare. Although the users that link their accounts in this way are only a subset of all the Foursquare users, this dataset contains information on about 2,733,324 users made in almost 22 months (from April 4, 2012, to January 29, 2014). With the information of tweets, a check-in dataset D_{4S} was made with 90,048,627 rows, each one corresponding to a check-in, i.e. the interaction between a user, with an anonymized ID, and a POI, where the ID is an alphanumeric identifier of the site in the Foursquare platform. Each record includes the Coordinated Universal Time (UTC) when the check-in occurred and the fourth column is the correction of the UTC corresponding to the Time Zone where the check-in was made. The user ID data has 2,733,324 different values, which corresponds to the activity of the same number of users. The POI ID column has 11,180,160 different values, which is equivalent to the number of POIs in the dataset.

From the information in the tweets, it is possible to obtain the POI profile from Foursquare's site and with it the information contained in the second dataset, D_{POI}. Each POI ID is the same alphanumeric code that appears in D_{4S}; this column contains 11,180,160 different values. In addition, each place is described by its geographical coordinates: latitude, and longitude, described by floating numbers that obey the World Geodetic System 1984 standard (WGS84) with 6 decimal places, which is equivalent to a precision of less than one meter. Each POI is described by a category name; the dataset contains 519 different categories of which "Home (private)" is the most common with 1,310,012 sites, followed by "Residential Building (Apartment / Condo)" with 354,858; the third most popular is "Office" with 317,149; the fourth place is occupied by "Building" with 255,121 sites; the

following are "Café", "Restaurant", "Bar" and "Hotel" with 188436, 153027, 145878 and 138476 sites, respectively. This information also includes the country code of the POI according to the two-letter ISO 3166-1 standard; containing 253 different codes. Then, the dataset includes check-in on every country in the world. From the 253 country codes, 84 countries with 5000 POIs or more represent 98.92% of the data. Using the information in the country code, we group all the POIs by their code, obtaining 253 datasets, D_{POI}(code), each containing the POIs of a single country. In Table 1, 15 countries with most of the POIs are listed; 80% of the venues belong to these countries. From this, we can highlight the presence of countries with remarkable diversity in terms of culture and geography.

With the information of each set D_{POI}(code), D_{4S} can be filtered, grouping the check-ins by country code. In this manner, it is possible to know the number of check-ins per country and the number of users who had activity in each country, as shown in the Table 1. In addition to the number of POIs and check-ins, this table shows the percentage that this number represents of the total. The same is not done in the case of users since many users have activity in more than one country. Although the order in the ranking varies, the same countries that concentrate the majority of the POIs add up to the largest number of check-ins. Regarding check-ins, the records in the 85 countries that contain 25,000 check-ins or more, represent 99.54% of the data.

Grouping POIs by urban area. Our topic of interest is the behavior of people in cities, so classifying POIs and check-ins by country is not enough. The definition of what a city is and what its borders are is a complex topic and has been dealt with by different authors[3,56,57]. In this work, we opt for the definition of *Functional Urban Area* used by the European Commission, which integrates factors such as infrastructure, population, and economy, and with which the Joint Research Centre generated the Global Human Settlement—Urban Centre Database (GHS), a dataset with the borders of 13,135 urban areas worldwide. This information is contained in a shapefile publicly available[58] that includes the name of the city, its population, coordinates, area, the country, and region in which it is located, if it extends beyond the borders of a single region within the same country (New York, whose functional urban area is divided into counties belonging to the states of New York and New Jersey, in the United States) or even in more than one country (for example, Detroit, in the United States, whose functional urban area extends to Ontario, Canada, including the city of Windsor); among much more information. We use this dataset to group POIs by functional urban areas using Geopandas, a Python package for data analysis with geographic information[59]. Of all the urban areas contained in the GHS, 6,463 cities have, at least, one Foursquare POI. The POIs that are located within these cities represent 74 percent of the total; 82% of the total check-ins were carried out in them. We focused on the 632 urban areas that have more than 10,000 check-ins, which represent 63% of the total POIs (7,026,688), and 76% of the total check-ins (68,356,896). That is, more than three-quarters of the total check-ins in the database were made in urban areas with more than 10,000 check-ins. The 31 cities with more records are shown in Table 2. Again, we find great diversity in cultural, social, and geographic terms, giving the Foursquare dataset great relevance for the study of urban dynamics. Along with the number of check-ins, the number of POIs per square kilometer and the number of check-ins per city inhabitant were calculated to give us an idea of the urban environments reflected by the Foursquare activity.

Geographic analysis of distances. Each city, as well as each POI, has associated coordinates. To measure the distance between cities, the Haversine formula is required[60]. This formula calculates the physical distance between cities based on the great-circle distance between two points on a sphere, specifically the Earth's surface, as follows

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right), \tag{6}$$

where φ_i and λ_i are, respectively, the latitude and the longitude of the point i and r the radius of the sphere. To perform this calculation, we utilized the Haversine 2.8.0 library for Python[61].

Kulback–Liebler divergence. The Kullback-Leibler divergence, also known as relative entropy, is an important quantity to calculate the difference between two probability distributions $P(z)$ and $Q(z)$ describing a stochastic variable z. For discrete distributions, this divergence is given by[40,44]

$$\mathscr{D}_{\mathrm{KL}}(P||Q) = \sum_z P(z) \log\left[\frac{P(z)}{Q(z)}\right]. \tag{7}$$

Here Q acts as a reference distribution. Also, it is important to emphasize that $\mathscr{D}_{\mathrm{KL}}(P||Q)$ is not a distance in the sense of a metric since the distance between P and Q is not necessarily the same as between Q and P. Also, from the definition in Eq. (7), it is clear that $\mathscr{D}_{\mathrm{KL}}(P||Q) > 0$ and is null when $P = Q$.

Networks and community detection. Undirected networks with N nodes are described by an adjacency $N \times N$ matrix $\mathbf{A}$ with entries 1 if two different nodes are connected and 0 otherwise. An important quantity in the study of networks is the degree of node i given by $k_i = \sum_{l=1}^{N} A_{il}$, which gives the number of connections to that node.

In many real networks, it is common to have subsets of nodes called communities. A community is defined as a locally dense connected subgraph in a network. The Louvain method is an algorithm to detect communities from large networks. The method optimizes the modularity Q, given by[45,46,48]

$$Q = \frac{1}{2\mathscr{E}} \sum_{i,j=1}^{N} \left[A_{ij} - \frac{k_i k_j}{2\mathscr{E}} \right] \delta_{c_i,c_j}, \tag{8}$$

where c_i, c_j are the communities of the nodes i and j, $\delta_{x,y}$ denotes the Kronecker delta and, $\mathscr{E} = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij}$ is the total number of edges in the network.

In the implementation of the method, the main goal is to generate a partition of the set of nodes in communities with labels c_i for $i = 1, 2, \ldots, N$. The method works with the iteration of two steps: in the first step, for each node i, the change in modularity ΔQ is calculated for removing i from its community and moving it into the community of each neighbor. Then, once this value is calculated for all communities, i is placed into the community that resulted in the greatest modularity increase. If no increase is possible, the node i remains in its original community. This process is applied repeatedly and sequentially to all nodes until no modularity increase can occur. Once a local maximum of modularity is reached, the first step is ended. In the second step, all the nodes in each community are grouped building a new network where nodes are the communities from the previous step. Any links between nodes of the same community are now represented by self-loops on the new community node and links from multiple nodes in the same community to a node in a different community are represented by weighted edges between communities. Once the new network is created, the second step has ended and the first step is applied to the new network. The values of Q define a scale that measures the relative density of edges inside communities in comparison with the edges outside communities (see Ref.[47,48] for details).

Machine learning and agglomerative clustering. Machine learning (ML) is a branch of artificial intelligence that focuses on creating methods that can learn and improve their performance on tasks by leveraging data[62]. ML algorithms build models based on training data to make predictions or decisions. There are two main categories of Machine Learning algorithms: supervised and unsupervised. In supervised learning, the algorithm is provided with labeled data, which consists of a set of input-output pairs. The goal of the algorithm is to learn a general rule or function that maps inputs into outputs. This is done by building a mathematical model of the data. The algorithm iteratively optimizes an objective function to learn a function that can accurately predict the output associated with new inputs.

In unsupervised learning, the algorithm is provided with unlabeled data and must identify patterns or structure in the data on its own[62]. Unsupervised learning algorithms discover patterns in the data and adapt their behavior based on the presence or absence of these patterns in new data. As a part of unsupervised learning, cluster analysis is the process of dividing a group of observations into subsets or clusters, where each one contains similar observations[50]. The objective of clustering is to categorize untagged data into clusters based on a specific metric of similarity or distance. Essentially, a cluster is regarded as a collection of data points that exhibit some common pattern or structure. Clustering techniques vary in their assumptions on the data's structure and use different similarity metrics to evaluate the internal compactness and separation of clusters.

Hierarchical clustering is one of the most commonly used methods for grouping unlabeled data into clusters based on some similarity measure. This approach involves using either agglomerative or divisive algorithms to create nested clusters by either combining or separating previous clusters. Agglomerative clustering initially considers each data point as a separate cluster and merges them pairwise until all data is contained in a single cluster, or until a specific condition is met, such as grouping data into a specific number of clusters. To accomplish this, a distance metric is used to determine the distance between each pair of data points, and a criterion for determining which clusters to merge at each stage is employed. In this study, the Euclidean distance was used to measure the distance between check-in distributions, and the criterion for merging clusters was to combine the two whose union minimized the variance of distances within all clusters. To do this, the Scikit-learn[63] python library was used.

Data availability

The datasets analysed during the current study are available in the webpage: https://sites.google.com/site/yangdingqi/home/foursquare-dataset.

Received: 18 January 2023; Accepted: 22 March 2023
Published online: 25 March 2023

References

1. Batty, M. *The New Science of Cities* (MIT Press, 2013).
2. Barthelemy, M. *The Structure and Dynamics of Cities* (Cambridge University Press, 2016).
3. Bettencourt, L. M. *Introduction to Urban Science: Evidence and Theory of Cities as Complex Systems* (MIT Press, 2021).
4. Shi, W., Goodchild, M., Batty, M., Kwan, M. & Zhang, A. (eds.) *Urban Informatics*. The Urban Book Series (Springer Singapore, 2021).
5. Rybski, D. & González, M. C. Cities as complex systems-collection overview. *PloS One* **17**, e0262964. https://doi.org/10.1371/journal.pone.0262964 (2022).
6. Sobolevsky, S. *et al.* Cities through the prism of people's spending behavior. *PloS One* **11**, e0146291. https://doi.org/10.1371/journal.pone.0146291 (2016).
7. Riascos, A. P. & Mateos, J. L. Networks and long-range mobility in cities: A study of more than one billion taxi trips in New York City. *Sci. Rep.* **10**, 4022. https://doi.org/10.1038/s41598-020-60875-w (2020).
8. Melikov, P. *et al.* Characterizing Urban Mobility Patterns: A Case Study of Mexico City. In Shi, W., Goodchild, M., Batty, M., Kwan, M. & Zhang, A. (eds.) *Urban Informatics*, Springer The Urban Book Series, chap. 11, 153–170, https://doi.org/10.1007/978-981-15-8983-611 (Springer Nature Singapore, 2021).
9. Loaiza-Monsalve, D. & Riascos, A. P. Human mobility in bike-sharing systems: Structure of local and non-local dynamics. *PloS One* **14**, e0213106. https://doi.org/10.1371/journal.pone.0213106 (2019).

10. González, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779–782. https://doi.org/10.1038/nature06958 (2008).
11. Song, C., Koren, T., Wang, P. & Barabási, A.-L. Modelling the scaling properties of human mobility. *Nat. Phys.* **6**, 818–823. https://doi.org/10.1038/nphys1760 (2010).
12. Louail, T. *et al.* From mobile phone data to the spatial structure of cities. *Sci. Rep.* **4**, 5276. https://doi.org/10.1038/srep05276 (2014).
13. Louail, T. *et al.* Uncovering the spatial structure of mobility networks. *Nat. Commun.* **6**, 6007. https://doi.org/10.1038/ncomms7007 (2015).
14. Çolak, S., Lima, A. & González, M. C. Understanding congested travel in urban areas. *Nat. Commun.* **7**, 10793. https://doi.org/10.1038/ncomms10793 (2016).
15. Alessandretti, L., Sapiezynski, P., Sekara, V., Lehmann, S. & Baronchelli, A. Evidence for a conserved quantity in human mobility. *Nat. Hum. Behav.* **2**, 485–491. https://doi.org/10.1038/s41562-018-0364-x (2018).
16. Song, C., Qu, Z., Blumm, N. & Barabási, A.-L. Limits of predictability in human mobility. *Science* **327**, 1018–1021. https://doi.org/10.1126/science.117717 (2010).
17. Alessandretti, L., Aslak, U. & Lehmann, S. The scales of human mobility. *Nature* **587**, 402–407. https://doi.org/10.1038/s41586-020-2909-1 (2020).
18. Bhattacharya, K. & Kaski, K. Social physics: uncovering human behaviour from communication. *Adv. Phys.: X* **4**, 1527723, https://doi.org/10.1080/23746149.2018.1527723 (2019).
19. Wei, X., Qian, Y., Sun, C., Sun, J. & Liu, Y. A survey of location-based social networks: Problems, methods, and future research directions. *GeoInformatica* **26**, 159–199. https://doi.org/10.1007/s10707-021-00450-1 (2022).
20. Chen, Z. *et al.* Contrasting social and non-social sources of predictability in human mobility. *Nat. Commun.* **13**, 1–9. https://doi.org/10.1038/s41467-022-29592-y (2022).
21. Lenormand, M., Gonçalves, B., Tugores, A. & Ramasco, J. J. Human diffusion and city influence. *J. R. Soc. Interface* **12**, 20150473. https://doi.org/10.1098/rsif.2015.0473 (2015).
22. Yang, D., Fankhauser, B., Rosso, P. & Cudre-Mauroux, P. Location prediction over sparse user mobility traces using rnns: Flashback in hidden states! In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, 2184–2190, https://doi.org/10.24963/ijcai.2020/302 (2020).
23. Yang, D., Qu, B., Yang, J. & Cudre-Mauroux, P. Revisiting user mobility and social relationships in LBSN: A hypergraph embedding approach. In *The World Wide Web Conference*, 2147–2157, https://doi.org/10.1145/3308558.3313635 (2019).
24. Riascos, A. P. & Mateos, J. L. Emergence of encounter networks due to human mobility. *PloS One* **12**, e0184532. https://doi.org/10.1371/journal.pone.0184532 (2017).
25. Foursquare City Guide. https://foursquare.com/city-guide.
26. Noulas, A., Scellato, S., Lambiotte, R., Pontil, M. & Mascolo, C. A tale of many cities: Universal patterns in human urban mobility. *PloS One* **7**, e37027. https://doi.org/10.1371/journal.pone.0037027 (2012).
27. Yang, D., Zhang, D., Zheng, V. W. & Yu, Z. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Trans. Syst. Man Cybern.: Syst.* **45**, 129–142, https://doi.org/10.1109/TSMC.2014.2327053 (2015).
28. Yang, D., Qu, B., Yang, J. & Cudré-Mauroux, P. Lbsn2vec++: Heterogeneous hypergraph embedding for location-based social networks. *IEEE Trans. Knowl. Data Eng.*https://doi.org/10.1109/TKDE.2020.2997869 *(2020).*
29. Yang, D., Qu, B. & Cudre-Mauroux, P. Location-centric social media analytics: Challenges and opportunities for smart cities. *IEEE Intell. Syst.* **36**, 3–10. https://doi.org/10.1109/MIS.2020.3009438 (2020).
30. Gallotti, R., Bertagnolli, G. & De Domenico, M. Unraveling the hidden organisation of urban systems and their mobility flows. *EPJ Data Sci.* **10**, 3. https://doi.org/10.1140/epjds/s13688-020-00258-3 (2021).
31. Noulas, A., Shaw, B., Lambiotte, R. & Mascolo, C. Topological properties and temporal dynamics of place networks in urban environments. In *Proceedings of the 24th International Conference on World Wide Web*, 431–441, https://doi.org/10.1145/2740908.2745402 (2015).
32. Noulas, A., Scellato, S., Mascolo, C. & Pontil, M. An empirical study of geographic user activity patterns in foursquare. In *Proceedings of the International AAAI Conference on Web and Social Media* **5–1**, 570–573. https://doi.org/10.1609/icwsm.v5i1.14175 (2011).
33. Cornacchia, G. & Pappalardo, L. STS-EPR: Modelling individual mobility considering the spatial, temporal, and social dimensions together. *Procedia Comput. Sci.* **184**, 258–265. https://doi.org/10.1016/j.procs.2021.03.035 (2021).
34. D'Silva, K., Noulas, A., Musolesi, M., Mascolo, C. & Sklar, M. Predicting the temporal activity patterns of new venues. *EPJ Data Sci.* **7**, 1–17. https://doi.org/10.1140/epjds/s13688-018-0142-z (2018).
35. Barabási, A.-L. The origin of bursts and heavy tails in human dynamics. *Nature* **435**, 207–211. https://doi.org/10.1038/nature03459 (2005).
36. Saramäki, J. & Moro, E. From seconds to months: An overview of multi-scale dynamics of mobile telephone calls. *Eur. Phys. J. B* **88**, 1–10. https://doi.org/10.1140/epjb/e2015-60106-6 (2015).
37. Prieto Curiel, R., Patino, J. E., Duque, J. C. & O'Clery, N. The heartbeat of the city. *PloS One* **16**, e0246714. https://doi.org/10.1371/journal.pone.0246714 (2021).
38. Sparks, K., Thakur, G., Pasarkar, A. & Urban, M. A global analysis of cities' geosocial temporal signatures for points of interest hours of operation. *Int. J. Geogr. Inf. Sci.* **34**, 759–776. https://doi.org/10.1080/13658816.2019.1615069 (2020).
39. Barabási, A. & Pósfai, M. *Network Science* (Cambridge University Press, 2016).
40. Cover, T. & Thomas, J. *Elements of Information Theory*. A Wiley-Interscience publication (Wiley, 2006).
41. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics (Springer, New York, 2017).
42. Dingqi YANG's Homepage. https://sites.google.com/site/yangdingqi/home/foursquare-dataset.
43. Ruiz-Gayosso, J. A. & Riascos, A. P. Human mobility in the airport transportation network of the United States. *Int. J. Mod. Phys. C* 2350072, https://doi.org/10.1142/S0129183123500729 (2023).
44. Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86. https://doi.org/10.1214/aoms/1177729694 (1951).
45. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113. https://doi.org/10.1103/PhysRevE.69.026113 (2004).
46. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103**, 8577–8582. https://doi.org/10.1073/pnas.0601602103 (2006).
47. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174. https://doi.org/10.1016/j.physrep.2009.11.002 (2010).
48. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech: Theory Exp.* **2008**, P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008 (2008).
49. NetworkX - Network Analysis in Python. https://networkx.org/.
50. Müller, A. & Guido, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists* (O'Reilly Media, 2016).
51. 44 Foursquare Statistics You Must Know: 2023 Market Share & Business Use - Financesonline.com. https://financesonline.com/foursquare-statistics/.
52. foursquare.com Traffic Analytics & Market Share. Similarweb. https://www.similarweb.com/website/foursquare.com/demographics.

53. Foursquare Categories and Core Attributes. Foursquare. https://location.foursquare.com/places/docs/categories.
54. Listing of Countries and Territories With POIs. Foursquare. https://location.foursquare.com/places/docs/supported-countries.
55. Making Foursquare Places Work For You. Foursquare. https://location.foursquare.com/places/docs/how-does-places-work.
56. Rozenfeld, H. D. *et al.* Laws of population growth. *Proc. Natl. Acad. Sci.* **105**, 18702–18707. https://doi.org/10.1073/pnas.0807435105 (2008).
57. Ortman, S. G., Lobo, J. & Smith, M. E. Cities: Complexity, theory and history. *PloS One* **15**, e0243621. https://doi.org/10.1371/journal.pone.0243621 (2020).
58. Florczyk, A. J. *et al.* GHS Urban Centre Database 2015, multitemporal and multidimensional attributes, r2019a. *European Commission, Joint Research Centre (JRC)* (2019). [Dataset]. Available online https://data.jrc.ec.europa.eu/dataset/53473144-b88c-44bc-b4a3-4583ed1f547e (accessed on May 2022).
59. Jordahl, K. *et al.* geopandas/geopandas: v0.8.1, https://doi.org/10.5281/zenodo.3946761 (2020).
60. Korn, G. A. & Korn, T. M. *Mathematical handbook for scientists and engineers: definitions, theorems, and formulas for reference and review* (Courier Corporation, 2000).
61. Rouberol, B. haversine 2.8.0. https://pypi.org/project/haversine/ (2023).
62. Mehta, P. *et al.* A high-bias, low-variance introduction to machine learning for physicists. *Phys. Rep.* **810**, 1–124. https://doi.org/10.1016/j.physrep.2019.03.001 (2019).
63. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830. https://doi.org/10.5555/1953048.2078195 (2011).

Acknowledgements

F.B. acknowledges support from CONACYT México. This work was supported by PAPIIT-UNAM grant No. IN116220.

Author contributions

F.B., A.P.R. and J.L.M. designed the research, performed the research, and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

B Sistema de 632 ciudades

	Ciudad	CC	POIs	Ch-ins
1	Istanbul	TR	334517	7343552
2	Jakarta	ID	398154	2908026
3	Kuala Lumpur	MY	200199	2558716
4	Tokyo	JP	191162	2405876
5	Mexico City	MX	137655	1804660
6	Bangkok	TH	169220	1741638
7	Moscow	RU	160868	1643199
8	São Paulo	BR	137494	1478616
9	Izmir	TR	74237	1443711
10	Quezon City [Manila]	PH	123941	1343955
11	New York	US	137841	1311179
12	Santiago	CL	89115	1276748
13	Ankara	TR	59610	1021287
14	Bandung	ID	119357	933686
15	Singapore	SG	72981	827715
16	Surabaya	ID	110852	756628
17	Los Angeles	US	91011	698605
18	Osaka [Kyoto]	JP	66574	667631
19	Yogyakarta	ID	83635	609749
20	Rio de Janeiro	BR	55765	570441
21	Belém	BR	39125	490162
22	Chicago	US	55319	485198
23	Saint Petersburg	RU	56800	482715
24	Kuwait City	KW	43547	473904
25	London	GB	48778	430524
26	Lima	PE	34989	404408
27	Denpasar	ID	54399	393148
28	Bursa	TR	25214	392454
29	Manaus	BR	34633	390084
30	Medan	ID	46528	387255
31	Seoul	KR	79306	382646
32	Fortaleza	BR	35535	371596
33	San Jose	US	42630	366703
34	Porto Alegre	BR	28823	360906
35	Belo Horizonte	BR	36376	349709
36	Antalya	TR	25866	348245
37	Washington D.C.	US	35571	344754
38	Curitiba	BR	30060	319598
39	Recife	BR	31727	315185
40	San José	CR	21928	292786
41	Buenos Aires	AR	32123	283371
42	Miami	US	38703	277545
43	Nagoya	JP	30843	277212
44	Adana	TR	17879	274392
45	Eskisehir	TR	14636	264926
46	Dallas	US	34971	260906
47	Paris	FR	37467	255049
48	Bogota	CO	24660	247400
49	Macapá	BR	16811	246513
50	Asuncion	PY	22733	244594
51	İzmit	TR	13337	233816
52	Monterrey	MX	21084	230510
53	Madrid	ES	27799	219399
54	Trabzon	TR	9262	218830
55	Mersin	TR	14873	218475
56	Guadalajara	MX	23312	218232
57	Atlanta	US	21406	216031
58	Kyiv	UA	23993	214314
59	Philadelphia	US	26967	209209
60	Cebu City	PH	19155	201851
61	Boston	US	21550	199112
62	Sakarya [Adapazarı]	TR	11722	196129
63	Las Vegas	US	17271	189859
64	Houston	US	28421	188927
65	Viña del Mar [Valparaíso]	CL	15540	188670
66	Toronto	CA	27796	183437
67	Barcelona	ES	25270	177784
68	George Town	MY	15615	174440
69	Semarang	ID	29561	168119
70	Phoenix	US	23243	164197
71	Riyadh	SA	15834	163228
72	Seattle	US	22958	163171
73	Samsun	TR	10038	162442
74	Tijuana	US	22223	161661
75	Natal	BR	13770	159665
76	Minneapolis [Saint Paul]	US	17270	155982
77	Denizli	TR	12327	155312
78	Milwaukee	US	15497	155092
79	Detroit	US	23679	153695
80	Makassar	ID	21904	153217
81	Dubai	AE	17110	147096
82	Athens	GR	19853	146705
83	Yekaterinburg	RU	13842	143529
84	Milan	IT	20716	141181
85	Puebla	MX	13833	140126
86	Salvador	BR	18682	138477
87	Riga	LV	12820	133134
88	Tampa	US	18850	132009
89	Goiânia	BR	14883	131663
90	Sincan	TR	9637	129214
91	Austin	US	12536	127603
92	Mérida	MX	10618	127215
93	Konya	TR	11920	121719
94	Kota Kinabalu	MY	13393	119700
95	Kuching	MY	10738	119176
96	Minsk	BY	11822	115631
97	Concepción	CL	9446	114225
98	Gaziantep	TR	10789	113227
99	Melaka City	MY	11128	112466
100	Fukuoka	JP	12230	111892

	Ciudad	CC	POIs	Ch-ins
101	Kayseri	TR	9968	111277
102	Orlando	US	13427	110586
103	Jeddah	SA	11120	109793
104	Florianópolis	BR	10735	107873
105	Balıkesir	TR	6716	107710
106	Sydney	AU	18181	107048
107	Manado	ID	12438	106302
108	Chiang Mai	TH	15014	105688
109	Indianapolis	US	12262	102048
110	Panama City	PA	8159	101269
111	San Antonio	US	12656	99106
112	Baltimore	US	12664	98844
113	Berlin	DE	15726	97724
114	Campinas	BR	9844	96408
115	Amsterdam	NL	13593	95610
116	João Pessoa	BR	10703	93885
117	Manisa	TR	6275	93388
118	Denver	US	14507	91519
119	Santo Domingo	DO	9515	91097
120	Mumbai	IN	15450	90171
121	Ipoh	MY	10466	88642
122	Rostov-on-Don	RU	10039	88247
123	Brussels	BE	13028	86955
124	Cleveland	US	10898	86817
125	Villahermosa	MX	6344	86799
126	Perm	RU	8810	86613
127	Brasília	BR	7288	86418
128	Saratov	RU	7771	85470
129	Melbourne	AU	16246	85033
130	Budapest	HU	11894	83043
131	Antwerp	BE	11988	82176
132	Bukit Mertajam	MY	9085	82159
133	Kazan	RU	9019	81498
134	Campo Grande	BR	8950	79349
135	Rotterdam [The Hague]	NL	13771	78796
136	Columbus	US	10684	77906
137	Sapporo	JP	9117	77592
138	Palembang	ID	13430	77158
139	Kansas City	US	10781	77046
140	Krasnoyarsk	RU	8554	76513
141	Rome	IT	13293	76084
142	Cairo	EG	10567	75494
143	Vancouver	CA	11451	75476
144	Aydin	TR	5120	75039
145	Tekirdağ	TR	4518	70471
146	Maceió	BR	8499	70176
147	St. Louis	US	9760	69439
148	Dnipro	UA	7269	69394
149	Edirne	TR	4259	69019
150	Teresina	BR	9332	68417
151	Belgrade	RS	8166	67588
152	Santos	BR	8094	65852
153	Montreal	CA	11525	64794
154	Kuantan	MY	5955	64315
155	São Luís	BR	9216	64123
156	Beirut	LB	7762	63981
157	Vila Velha	BR	7169	62983
158	Ceilândia	BR	7039	62826
159	Pittsburgh	US	8271	62163
160	Sendai	JP	6479	61600
161	Yalova	TR	3627	60480
162	Novosibirsk	RU	8328	60387
163	Lisbon	PT	9056	59629
164	Querétaro	MX	6442	59505
165	Toluca	MX	6851	59335
166	Shanghai	CN	12516	58672
167	Batam City	ID	6821	57962
168	Medellín	CO	7584	57641
169	New Orleans	US	7685	57330
170	Davao City	PH	5928	57249
171	Nairobi	KE	8543	56329
172	Seremban	MY	7894	56076
173	Alor Setar	MY	5622	56025
174	Sao Goncalo	BR	6734	55964
175	Colombo	LK	7941	55392
176	Manchester	GB	7619	53969
177	Chelyabinsk	RU	6678	53372
178	Malé	MV	5714	53169
179	Pekanbaru	ID	10468	52916
180	Novo Hamburgo	BR	5349	52116
181	Ghent	BE	7563	51969
182	Samara	RU	6941	51831
183	Caxias do Sul	BR	5651	51388
184	San Salvador	SV	5702	51288
185	Charlotte	US	5076	51096
186	Odesa	UA	6725	50970
187	Joinville	BR	5296	50904
188	Veracruz	MX	5483	49413
189	Vienna	AT	8698	49148
190	Çanakkale	TR	3698	49109
191	İskenderun	TR	3784	48354
192	Honolulu	US	5973	48074
193	Kütahya	TR	3972	47991
194	Çorlu	TR	3717	47805
195	Stockholm	SE	8681	47771
196	Isparta	TR	4211	47685
197	Al Khobar	SA	4429	46919
198	Prague	CZ	8818	46729
199	Río Piedras [San Juan]	PR	6954	46714
200	Birmingham	GB	6678	46700

Ciudad	CC	POIs	Ch-ins
201 Xalapa	MX	4517	46403
202 Nizhny Novgorod	RU	5706	45487
203 Delhi [New Delhi]	IN	8968	45423
204 Bandar Lampung	ID	9454	45288
205 Cincinnati	US	6866	45239
206 Caracas	VE	7209	45142
207 Juazeiro do Norte	BR	4334	44804
208 Buffalo	US	6430	44785
209 Sao Jose dos Campos	BR	5409	44229
210 Virginia Beach	US	7328	44010
211 Valencia	ES	7899	43877
212 Banjarmasin	ID	8448	43588
213 Cheboksary	RU	5400	43431
214 Helsinki	FI	8019	43171
215 Málaga	ES	6356	43130
216 Kota Bharu	MY	5456	42622
217 Tampico	MX	4490	42192
218 Sacramento	US	8010	41939
219 Louisville	US	6167	41794
220 Grand Rapids	US	5143	41369
221 Temuco	CL	3758	40693
222 Balikpapan	ID	6939	39862
223 Khon Kaen	TH	5193	39592
224 Omsk	RU	4901	39344
225 Nashville	US	4593	39273
226 Uberlândia	BR	4983	39256
227 Guayaquil	EC	5451	38870
228 Munich	DE	6973	38657
229 Cancún	MX	3575	38544
230 Johannesburg	ZA	8099	38449
231 Silivri	TR	2538	38259
232 Giresun	TR	2881	37906
233 Chonburi	TH	6268	37480
234 León	MX	4705	37429
235 Hamburg	DE	7480	37403
236 Alanya	TR	4116	37261
237 La Serena	CL	3912	37132
238 Volgograd	RU	4212	36608
239 Amman	JO	5049	36495
240 Ribeirao Preto	BR	4570	36438
241 Thessaloniki	GR	4877	36392
242 Turin	IT	5918	36331
243 Antofagasta	CL	3464	36309
244 Doha	QA	4789	36292
245 Des Moines	US	4302	36222
246 Banda Aceh	ID	5294	36078
247 Aracaju	BR	4692	36068
248 Cuiabá	BR	4314	35763
249 Cali	CO	4388	35678
250 Seville	ES	5568	35646

Ciudad	CC	POIs	Ch-ins
251 Accra	GH	5729	35638
252 Beijing	CN	8966	35553
253 Canoas	BR	3706	35308
254 Blumenau	BR	3519	35280
255 Bengaluru	IN	7577	34918
256 Antakya	TR	3489	34562
257 Voronezh	RU	3993	34536
258 Bandırma	TR	2735	34364
259 Dublin	IE	5632	34361
260 Ponta Grossa	BR	3535	34303
261 Samarinda	ID	6168	34210
262 Kharkiv	UA	5104	34059
263 Mossoró	BR	3961	34002
264 Utrecht	NL	4828	33638
265 Dortmund	DE	7025	33226
266 Rize	TR	2647	32852
267 Krasnodar	RU	4830	32638
268 Ufa	RU	4671	32499
269 Pontianak	ID	6765	32422
270 Tolyatti	RU	3485	32362
271 Jundiaí	BR	3664	32235
272 Vladivostok	RU	3953	32195
273 Uberaba	BR	3357	32133
274 Ordu	TR	2543	31910
275 Montes Claros	BR	4418	31894
276 Nicosia	CY	4033	31534
277 New Taipei [Taipei]	TW	9957	31511
278 Baku	AZ	4377	31327
279 Aguascalientes	MX	3874	31202
280 Padang	ID	7116	31153
281 Izhevsk	RU	3663	30672
282 Raleigh	US	3720	30606
283 Quito	EC	4914	30549
284 Guará	BR	3569	30377
285 Zonguldak	TR	2371	30325
286 Tunis	TN	4600	30276
287 Tuxtla Gutiérrez	MX	3657	30147
288 Juiz de Fora	BR	3599	30102
289 Brisbane	AU	5805	30079
290 Iquique	CL	3263	30001
291 Iloilo City	PH	3526	29809
292 Bauru	BR	3487	29596
293 Bucharest	RO	5705	29426
294 Leeds	GB	4249	29391
295 Takasaki [Maebashi]	JP	4422	29385
296 San Luis Potosí	MX	3963	29384
297 Tasikmalaya	ID	6411	29212
298 Cuernavaca	MX	4090	28783
299 Salt Lake City	US	5413	28783
300 Tel Aviv	IL	6021	28757

	Ciudad	CC	POIs	Ch-ins
301	Cirebon	ID	5924	28732
302	Santarém	BR	3324	28325
303	Malatya	TR	3014	28232
304	Cologne	DE	5355	28227
305	Maringá	BR	3920	28208
306	Ho Chi Minh City	VN	8009	28175
307	Manama	BH	4829	28148
308	Talca	CL	2723	28121
309	Gölcük	TR	2740	28113
310	Barranquilla	CO	3215	28021
311	Londrina	BR	3953	28006
312	Pune	IN	5297	27810
313	Karabük	TR	2262	27780
314	Kuala Terengganu	MY	3551	27366
315	Madison	US	3573	27299
316	Hat Yai	TH	4758	27230
317	Hiroshima	JP	4110	27173
318	Pattaya	TH	4826	27063
319	Purwokerto	ID	6085	26975
320	Providence	US	4291	26968
321	Morelia	MX	3786	26914
322	Rochester	US	3782	26848
323	Jacksonville	US	4121	26703
324	Boa Vista	BR	3173	26637
325	Guangzhou	CN	7650	26619
326	Almaty	KZ	3980	26585
327	Khabarovsk	RU	3763	26583
328	Sorocaba	BR	3943	26453
329	Irkutsk	RU	3747	26397
330	Perth	AU	5442	26264
331	Şanlıurfa	TR	3382	26067
332	Rancagua	CL	2873	26015
333	Warsaw	PL	5789	26008
334	Montevideo	UY	4054	25991
335	Tomsk	RU	3747	25728
336	Bolu	TR	2333	25711
337	Erzurum	TR	2486	25669
338	Richmond	US	3262	25396
339	Adelaide	AU	4935	25350
340	Omaha	US	4492	25339
341	Naha	JP	4760	25332
342	Bologna	IT	3922	24560
343	Memphis	US	3515	24455
344	Sofia	BG	4322	24403
345	Mataram	ID	5344	24390
346	Frankfurt am Main	DE	4681	24353
347	Sivas	TR	2569	24216
348	Hamamatsu	JP	3377	23989
349	Kahramanmaraş	TR	3333	23923
350	Hermosillo	MX	3412	23911
351	Ulyanovsk	RU	2676	23691
352	Sungai Petani	MY	2973	23640
353	Bruges	BE	3395	23587
354	Cordoba	AR	3440	23502
355	Busan	KR	6000	23292
356	Yaroslavl	RU	2904	23278
357	Bacolod	PH	3323	23184
358	Kitakyushu	JP	3523	23163
359	Donetsk	UA	3329	23135
360	Salihli	TR	2227	23047
361	Cape Town	ZA	4503	23044
362	Santa Maria	BR	2484	22962
363	Düzce	TR	1938	22953
364	Cagayan de Oro	PH	3238	22855
365	Bridgeport	US	4093	22819
366	Guatemala City	GT	3983	22745
367	Pelotas	BR	2470	22706
368	Brick	US	3638	22689
369	Copenhagen	DK	5402	22533
370	Kanazawa	JP	3421	22488
371	Chennai	IN	5641	22417
372	Moriyama [Otsu]	JP	2300	22413
373	Kingston	JM	3342	22394
374	Allentown	US	3724	22363
375	Yakutsk	RU	2598	22357
376	Presidente Prudente	BR	2351	22269
377	Dusseldorf	DE	3732	22143
378	Feira de Santana	BR	4086	22110
379	Angeles [San Fernando]	PH	3228	21955
380	Diyarbakir	TR	2447	21866
381	Liverpool	GB	3556	21811
382	Syracuse	US	2899	21759
383	Campina Grande	BR	2848	21661
384	Florence	IT	3816	21526
385	Oklahoma City	US	4044	21516
386	Bradenton	US	4112	21463
387	Calgary	CA	4271	21406
388	Cartago	CR	2123	21361
389	Oaxaca	MX	2973	21273
390	Nakhon Ratchasima	TH	3546	21199
391	Balneário Camboriú	BR	2591	21139
392	Acapulco	MX	2705	21115
393	Crato	BR	2443	20938
394	Meru	MY	1536	20274
395	Okayama	JP	3081	20233
396	Santana	BR	2259	20187
397	Concord	US	4045	20117
398	Kaliningrad	RU	2824	20115
399	Campeche	MX	2348	20100
400	Edinburgh	GB	3176	20078

	Ciudad	CC	POIs	Ch-ins
401	Ottawa	CA	3611	20070
402	Smolensk	RU	2237	19983
403	São José dos Pinhais	BR	1584	19895
404	Uşak	TR	2148	19798
405	Chiang Rai	TH	2902	19781
406	Playa del Carmen	MX	1819	19769
407	Americana	BR	2567	19679
408	Encarnacion	PY	1848	19612
409	Abu Dhabi	AE	3395	19579
410	Glasgow	GB	3287	19478
411	Pachuca	MX	2381	19213
412	Fethiye	TR	2008	19144
413	Cabanatuan	PH	1923	19142
414	Manavgat	TR	2360	19099
415	Auckland	NZ	4441	19036
416	Managua	NI	2839	18988
417	Casablanca	MA	3015	18933
418	Nazilli	TR	1721	18808
419	Nakhon Pathom	TH	2832	18800
420	Ciudad del Este	PY	2293	18795
421	Zurich	CH	3632	18783
422	Tucson	US	3621	18767
423	Newcastle upon Tyne	GB	2966	18648
424	Araraquara	BR	2084	18446
425	Afyonkarahisar	TR	2112	18382
426	Naples	IT	4162	18373
427	Kirov	RU	2061	18264
428	Shizuoka	JP	2989	18124
429	Porto Velho	BR	2737	18123
430	Itajaí	BR	2118	18116
431	Eindhoven	NL	2830	18095
432	Lansing	US	2473	18079
433	Porto	PT	3509	18047
434	Elazig	TR	2139	18040
435	Gravataí	BR	2446	18023
436	Tver	RU	2271	17941
437	Tulsa	US	3301	17847
438	Albuquerque	US	3379	17806
439	Birmingham	US	2538	17723
440	Tyumen	RU	2771	17688
441	Hyderabad	IN	3668	17324
442	Akçaabat	TR	1285	17223
443	Jambi	ID	3542	16927
444	Udon Thani	TH	2623	16913
445	Rio Branco	BR	2481	16801
446	Sheffield	GB	2475	16796
447	Pereira	CO	1906	16727
448	Edmonton	CA	3789	16616
449	Santiago de los Caballeros	DO	2189	16615
450	Ereğli	TR	1371	16607
451	São José do Rio Preto	BR	2559	16591
452	Maracaibo	VE	2902	16292
453	Ivanovo	RU	1828	16269
454	Baguio	PH	2220	16245
455	Bristol	GB	2549	16218
456	Al Ghubra	OM	1794	16198
457	Iguatu	BR	1636	16123
458	Nice	FR	3261	16119
459	Dayton	US	2467	16040
460	Caruaru	BR	2312	15988
461	Zaragoza	ES	3161	15984
462	Lampang	TH	2320	15961
463	Medina	SA	2053	15941
464	Chillán	CL	1719	15925
465	Çorum	TR	1847	15896
466	Potengi	BR	1929	15869
467	Patos	BR	1716	15822
468	Harrisburg	US	2394	15797
469	Banjarbaru	ID	3345	15715
470	Cascavel	BR	2374	15705
471	Tallinn	EE	2900	15649
472	Tawau	MY	1990	15547
473	Morioka	JP	1939	15515
474	Stuttgart	DE	3128	15506
475	İnegöl	TR	1787	15476
476	Karachi	PK	2681	15444
477	Kumamoto	JP	2671	15431
478	Greensboro	US	2226	15275
479	Charleston	US	1903	15273
480	Puerto Vallarta	MX	1914	15262
481	Alacant / Alicante	ES	2942	15189
482	Roswell	US	2222	15164
483	Magelang	ID	3209	15133
484	Cilegon	ID	3124	15089
485	Taubaté	BR	2024	15061
486	Lajeado	BR	1347	15055
487	Niigata	JP	2260	14986
488	Culiacán	MX	2233	14984
489	Dammam	SA	3318	14849
490	Ōita	JP	2181	14793
491	Green Bay	US	1913	14761
492	Lyon	FR	3169	14684
493	Sapucaia do Sul	BR	1911	14552
494	Valdivia	CL	1617	14536
495	Utsunomiya	JP	1638	14487
496	Mecca	SA	2192	14444
497	Puerto Montt	CL	1551	14421
498	Petrolina	BR	2299	14406
499	Granada	ES	2703	14350
500	Zagreb	HR	3382	14323

	Ciudad	CC	POIs	Ch-ins
501	Durango	MX	2195	14319
502	Erie	US	1975	14287
503	Mithatpaşa Mah.	TR	1208	14242
504	Gorontalo	ID	2510	14175
505	Turgutlu	TR	1573	14096
506	Toyohashi	JP	1798	14057
507	Palma de Mallorca	ES	2863	14046
508	Lipetsk	RU	1804	14042
509	Saltillo	MX	1985	14009
510	Colima	MX	1902	13986
511	La Paz	MX	1874	13953
512	Lviv	UA	1881	13942
513	Arnhem	NL	2178	13905
514	Kemerovo	RU	1952	13802
515	Cardiff	GB	2044	13792
516	Taiping	MY	2094	13630
517	Jember	ID	3554	13550
518	Brighton	GB	2325	13532
519	Cartagena	CO	1869	13522
520	Bilbao	ES	2936	13517
521	Hartford	US	2358	13507
522	Atlantic City	US	1465	13497
523	Vitória da Conquista	BR	2312	13460
524	Phra Nakhon Si Ayutthaya	TH	1976	13442
525	Akron	US	1806	13417
526	Divinópolis	BR	1832	13394
527	Kortrijk	BE	2368	13294
528	Sobral	BR	1714	13291
529	Skopje	MK	2652	13289
530	Arapiraca	BR	2033	13284
531	Nottingham	GB	2107	13271
532	Penza	RU	1946	13209
533	Batu Pahat	MY	1934	13032
534	Mar del Plata	AR	1909	12946
535	Toyama	JP	1800	12944
536	Novi Sad	RS	1800	12873
537	Alexandria	EG	2180	12818
538	Valencia	VE	2078	12810
539	Indio	US	2137	12795
540	Ubon Ratchathani	TH	2258	12788
541	Mendoza	AR	1949	12763
542	Leuven	BE	1953	12739
543	Volta Redonda	BR	1694	12722
544	Zelenograd	RU	1788	12716
545	Lille	FR	2729	12701
546	Colorado Springs	US	2416	12684
547	Tashkent	UZ	2550	12676
548	Tarsus	TR	1400	12563
549	Phuket	TH	2959	12504
550	Arica	CL	1886	12465
551	Port of Spain	TT	2031	12414
552	Fort Wayne	US	2090	12392
553	Kupang	ID	1977	12336
554	Yamagata	JP	1902	12308
555	Torreon	MX	2229	12290
556	Nuremberg	DE	2417	12275
557	Trujillo	PE	1915	12215
558	Mechelen	BE	1627	12203
559	Ann Arbor	US	1656	12156
560	Marseille	FR	2297	12149
561	Boise	US	2168	12136
562	Newport News	US	2063	11936
563	Guanajuato	MX	1255	11908
564	Poza Rica	MX	1503	11890
565	Purwakarta	ID	2416	11849
566	Muar	MY	1848	11845
567	Colina de Laranjeiras	BR	1600	11818
568	Governador Valadares	BR	1596	11790
569	Tomohon	ID	1638	11769
570	Palm Bay	US	2298	11761
571	Mito	JP	1474	11751
572	Wichita	US	1985	11749
573	Leipzig	DE	1867	11737
574	Takamatsu	JP	2073	11732
575	Lexington	US	2086	11690
576	Murcia	ES	2083	11689
577	Parintins	BR	1416	11676
578	Winnipeg	CA	2465	11667
579	Toledo	US	1815	11602
580	Arkhangelsk	RU	1754	11582
581	Baton Rouge	US	2202	11515
582	Celaya	MX	1719	11251
583	Vladimir	RU	1610	11199
584	Ryazan	RU	1556	11195
585	Savannah	US	1945	11154
586	Bryansk	RU	1508	11129
587	Kampala	UG	2201	11101
588	Dumaguete	PH	1524	11099
589	Mexicali	MX	2086	11052
590	La Plata	AR	1712	11022
591	Chihuahua	MX	1975	10982
592	Dordrecht	NL	1819	10947
593	Belfast	GB	1897	10883
594	Eagan	US	1665	10858
595	Rabat	MA	1731	10810
596	Pati	ID	3523	10753
597	Pematang Siantar	ID	1933	10751
598	Groningen	NL	1900	10708
599	Salatiga	ID	2479	10654
600	Coatzacoalcos	MX	1430	10627

Ciudad	CC	POIs	Ch-ins
601 Tula	RU	1856	10617
602 Madiun	ID	3200	10594
603 Santa Cruz do Sul	BR	1086	10589
604 Tegal	ID	2931	10588
605 Koriyama	JP	1389	10560
606 Leiden	NL	2118	10557
607 Gemlik	TR	1091	10539
608 Bucaramanga	CO	1882	10509
609 Limeira	BR	1454	10504
610 Podgorica	ME	1156	10480
611 Venice	IT	1804	10479
612 Kediri	ID	2798	10472
613 Bordeaux	FR	2148	10466
614 Champaign	US	1396	10434
615 Fresno	US	2381	10429
616 Praia Grande	BR	2097	10402

Ciudad	CC	POIs	Ch-ins
617 Haarlem	NL	1896	10302
618 Appleton	US	1814	10289
619 Kalamazoo	US	1552	10273
620 Reno	US	1862	10267
621 Cluj-Napoca	RO	1687	10260
622 Akhisar	TR	1085	10257
623 Portland	US	1498	10237
624 Liège	BE	2067	10237
625 Toulouse	FR	2438	10222
626 Norristown	US	1168	10211
627 Palu	ID	2726	10196
628 Albany	US	1620	10195
629 Pensacola	US	2017	10176
630 Daejeon	KR	2784	10089
631 Bintulu	MY	1643	10065
632 Amasya	TR	1281	10013