UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

Matriz origen destino utilizando datos de telefonía móvil

TESIS

Que para obtener el título de

Ingeniero en Computación

P R E S E N T A

Ludwing Van Christopher Chávez Jiménez

DIRECTOR DE TESIS

Dr. Carlos Gershenson García

Ciudad Universitaria, Cd. Mx., 2018

Gracias ...
A mis padres, Carmen y Marco, por tanta dedicación, esfuerzo, regaños, apoyo y amor que siempre me han brindado.
A Carlos, Pamela, Sandra y Osvaldo, por hacerme ver todo lo bueno y malo que hay en mi, las constantes risas y enojos.
A mis Abuelos, Don Guadalupe y Doña Antonia, que han sido como padres para mi, por brindarme su amor, techo, consejos y vivencias.
A mis tíos y tías por su compromiso constante con la educación de sus sobrinos.
Al Dr. Carlos Gershenson, por ser más que un tutor y ejemplo, un mentor de vida y valores. Gracias por su paciencia, apoyo y consejos.
A mis amigos que no sé qué hice pero siguen aquí.
A Anga por aprender juntos.

Este trabajo se realizó gracias a la beca otorgada por el CONACYT con clave CB-2013/221341 y al apoyo del proyecto *Ciudades grises, infraestructura verde: innovación colaborativa y diseño para infraestructuras urbanas saludables*, apoyado bajo el marco de la Convocatoria Conjunta British Council-Newton Fund – CONACYT, Institutional Links 2014.

Contenido

1. Introducción

1.1. Planteamiento del problema

Hoy en día las ciudades crecen a un ritmo exponencial, en algunos casos particulares podemos ver el aumento de autos al doble, como el caso de la Ciudad de México donde el parque vehicular creció a casi el doble en 10 años. En 2004 el censo vehicular registró una cantidad de 2,556,032, para 2014 la cantidad era de 4,737,749, el doble en tan solo 10 años. Beijing es un caso extremo en el cual en el año 2005 contaba con 2.6 millones de automóviles y al finalizar el año 2010 contaban con 4.8 millones de automóviles, el doble en un tiempo muy corto, con un crecimiento anual del 13 %. Esto obligó a las autoridades de Beijing a partir de 2011 a realizar un sorteo para aquellos ciudadanos que pretendían adquirir un automóvil. Con esta política se logró reducir la congestión vehicular, aunque este esfuerzo no es suficiente. Otras ciudades no crecieron a este ritmo exponencial, el caso de Nueva York parece ser diferente pues para el año 2004 el número de automóviles registrados era de aproximadamente de 1.92 millones mientras que para 2014 este número incrementó a 2.05 millones. Este incremento tan poco representativo puede deberse a las altas cuotas para registrar un automóvil[1], la ineficiencia de un automóvil en una ciudad con poco espacio o la basta cantidad de transporte público.

Una matriz de origen-destino es una herramienta que se suele utilizar para el análisis de movilidad urbana y poder desarrollar iniciativas de transporte público o planeación urbana. La construcción de una matriz origen-destino, tradicionalmente, se ha hecho con base en encuestas a usuarios o contando el tránsito de automóviles en vialidades. Estos análisis tienen un espacio muestral muy pequeño puesto que los conteos o encuestas levantadas no son suficientes para cubrir la toda la población de una ciudad, pues una ciudad medianamente grande está en el orden de millones, además de ser análisis muy costosos en relación a la cantidad de datos recolectados, por lo que los resultados pueden ser sesgados o insignificantes y actualizar estos datos de manera periódica tomaría mucho tiempo.

[1] http://www.crainsnewyork.com/article/20170206/TRANSPORTATION/170209944

En el caso de conteo de automóviles los sensores suelen ser costosos, por lo que implementar sensores en una cantidad suficiente en el que estudio sea considerado correcto, esto es que cuente con una cantidad de información suficiente, llegaría a ser altamente costoso, además de todas las implicaciones de logística que habrían de hacerse pues dichos sensores tendrían que colocarse en distintos puntos a lo largo de la avenida, recibir mantenimiento, etc. Una alternativa prometedora serían los servicios tales como Waze, Google Maps, etc. en el que el usuario comparte sus datos de geolocalización a cambio de encontrar la ruta más óptima a su destino. Sin embargo se suele denegar acceso a estos datos, por obvias razones de privacidad. Por lo que primero habría que establecer condiciones y reglas para garantizar la privacidad de los datos recolectados.

Actualmente las compañías telefónicas recolectan datos de los dispositivos móviles con fines técnicos o para generar estados de cuenta. Los CDR por sus siglas en inglés (Call Detail Records) contienen información del dispositivo cada que interactúa con la red, es decir cada que se recibe o manda un mensaje o cada que se recibe o hace una llamada. Así mismo registra el tiempo en que se hizo la interacción y con que antena tuvo interacción. Las compañías telefónicas no suelen abrir estos datos, ya que se podría malinterpretar su uso, pero si se les trata de manera correcta podemos obtener patrones de movilidad. En el caso de México la telefonía celular representa un gran sector pues tienen una penetración de 89 suscriptores por cada 100 habitantes [2]. Mientras que a nivel mundial en países desarrollados la penetración de la telefonía móvil es de 128 % (cada habitante tiene uno o más teléfonos móviles) mientras que en países en vías de desarrollo es de 89 %(International Telecommunication Union (2015)). Por lo que de hacer un análisis de estos datos móviles permitirá caracterizar la movilidad a un nivel muy específico (granulado). Es por ello que generar una matriz origen-destino derivada de datos de telefonía móvil resultaría ser más eficiente que los métodos tradicionales. La recolección de datos será fácil, la actualización de la matriz podría ser en tiempo real y el espacio muestral sería más grande. La generación de una matriz origen-destino derivada de datos móviles da pie a otros análisis, que se pueden derivar directamente de esta matriz o que puede complementar. No solo podremos conocer qué cantidad de viajes existen entre dos puntos, sino también cuál es la ruta utilizada, que no es el caso de estudio de esta tesis.
La matriz origen-destino puede ayudar a determinar sitios donde la población reside y labora, estimar tiempos de viaje, horas pico, demanda de viajes, modo en que se transportan los usuarios. Es decir, podemos hacer un análisis más detallado utilizando este tipo de datos.

[2] http://www.ift.org.mx/sites/default/files/contenidogeneral/estadisticas/anuario-estadistico-2015-acc_1.pdf

En el caso particular de la Zona Metropolitana del Valle de México no se ha podido hacer un experimento con estas características. La última encuesta origen-destino data de 2007, cuyo costo fue $53,525,000.00 de pesos y la cantidad de viviendas encuestadas fue de 46,500 (Instituto Nacional de Estadística y Geografía (2007)), dado que la cantidad de habitantes en la zona era de aproximadamente 29 millones (Instituto Nacional de Estadística y Geografía (2011)), el tiempo total entre levantamiento de encuestas y resultados fue de 6 meses.
Con estos datos es fácil apreciar que un experimento de este tipo realizado de la manera convencional es muy costoso en tiempo, dinero y eficacia. De hacer este experimento para la Zona Metropolitana se podría dar cuenta del flujo y comportamiento diario de la ciudad, para poder dar solución a los problemas de movilidad con el fin de reducir los tiempos de viaje.

1.2. Objetivos

1.2.1. Objetivo principal

Desarrollar una matriz origen-destino mediante un algoritmo heurístico, haciendo uso de los viajes y lugares más comunes para cada usuario en el conjunto de datos.

1.2.2. Objetivos secundarios

- Estimar la cantidad de viajes en días laborables y no laborables. De ser los datos muy refinados, estimar el método de transporte y transbordo.

- Determinar las horas con mayor concentración de viajes así como el tiempo promedio de los viajes.

- Determinar zonas habitacionales y no habitacionales.

1.3. Hipótesis

Es posible desarrollar un algoritmo capaz de identificar viajes y sitios comunes para cada usuario además de generar una matriz de origen-destino, usando un conjunto de datos conformado por CDRs.

1.4. Justificación

La justificación de este trabajo se centra en la posibilidad de mejorar las condiciones de movilidad en las grandes urbes, que actualmente presentan un gran problema, puesto que muchas ciudades no fueron planeadas para albergar tal cantidad de población lo que resulta fatal para trasladarse dentro y fuera de las ciudades. Las vialidades presentan problemas de congestión que nunca antes se habían visto, de igual manera el transporte público es insuficiente para la cantidad de usuarios. Por lo tanto es un problema al que se deben de encontrar soluciones rápidas y eficaces, dado el crecimiento exponencial de las ciudades. En resumen este trabajo parte de la motivación de entender, conocer y mejorar las condiciones viales o de movilidad que cada ciudad afronta.

1.5. Estructura del documento

En el capítulo segundo se presentan los antecedentes teóricos que sirven como base para el presente trabajo. En este capítulo se explica de forma general el funcionamiento de una red celular, las categorías de células, los rangos de los teléfonos móviles dentro de la red, las variables para calcular este rango y una breve descripción del uso de teléfonos móviles a nivel mundial. Posteriormente se define lo qué es una matriz de origen-destino así como investigaciones anteriores que han generado resultados utilizando CDRs.

En el capítulo tercero se describe el conjunto de datos a utilizar, el área geográfica donde toma lugar el estudio. Posteriormente se explica a detalle el algoritmo propuesto, además de las consideraciones que se hacen para un mejor desempeño en la obtención de resultados. También se explica la complejidad del algoritmo. Se muestra la posible implementación para su ejecución en paralelo. Finalmente se muestran las especificaciones técnicas de una plataforma web en la que se visualizarán los resultados de manera gráfica.

En el capítulo cuarto se describe a detalle los resultados. En la primer sección se concentran los resultados de la frecuencia de actividad en la red para un muestreo aleatorio. En la segunda sección se muestra la matriz origen-destino resultante y tres análisis derivados de la matriz. El primer análisis es la distribución de los viajes, para obtener la dinámica de los viajes. El segundo análisis muestra resultados temporales, para identificar la dinámica de los viajes en días no ordinarios. Finalmente se muestra la plataforma de visualización gráfica, la cual facilita la conceptualización de los resultados.

El capítulo quinto presenta las conclusiones, objetivos cumplidos así como las contribuciones de este trabajo y el posible trabajo a futuro.

2. Antecedentes

2.1. Datos de telefonía celular

Los CDR(call detail record) son datos históricos que las compañías celulares almacenan con fines de administración y contabilidad, estos datos contienen detalles acerca de la interacción del teléfono móvil con la red, es decir recaba información de mensajes o llamadas entrantes y salientes, a que antena se conecto, tiempo de conexión, número marcado, tiempo en que se realizó y terminó conexión etc. (Horak, 2007)

Existen otro tipo de datos asociados a los teléfonos móviles como los IPDR(Internet Protocol Detail Record) que almacena registros del tráfico de datos en la red, similar al CDR pero con mayor granularidad[1]. El Signalling es la comunicación constante entre el teléfono y antena con la finalidad de mantener, liberar la conexión, generando datos de manera pasiva[2].
Aunque se ha comprobado que al tener un conjunto de datos más robustos, una combinación entre IPDRs, CDRs y Signalling, se genera una mayor granularidad en los datos en espacio y tiempo, por lo que existe mayor exactitud en la determinación de comportamientos de movilidad y se generan trayectorias más precisas ((Pinelli & Calabrese, 2015)). Aunque en este trabajo de tesis no se cuente con los dos conjuntos de datos antes mencionados, IPDRs y Signalling, sí podemos determinar con precisión las ubicaciones más concurridas por personas (Pinelli & Calabrese, 2015). Que en particular es uno de los casos que sí aborda este trabajo.

2.2. Redes celulares

En 1947 Bell Telephone se desarrolló formalmente el concepto de células de radio, el cual estaba compuesto de antenas transmisoras y receptoras de baja potencia.
Las antenas están distribuidas en un área tal que está dividida en células, pequeñas áreas de cobertura, de manera que esta arquitectura sirve para incrementar la capacidad

[1] https://www.incognito.com/tips-and-tutorials/faq-bandwidth-monitoring-with-ipdr/
[2] http://ace.ucv.ro/sintes11/Volume2/5\%20ELECTRONICS/E06\%20-\%20CONSTANTINACHE\%20Pompiliu.pdf

de suscriptores (Horak, 2007). Por lo que los canales pueden ser rehusados, el rehusó de canales está basado en la premisa de que si múltiples sesiones de comunicación están lo suficientemente lejanas pueden evitar la interferencia, cada frecuencia puede ser rehusada en células no adyacentes. Además, cuando el tráfico de conexiones o sesiones aumenta, estás células pueden volverse a subdividir y aplicar el mismo principio de rehusó de canales. Por lo que el sistema se vuelve altamente escalable.

El diseño de la red está pensado como una malla de células hexagonales (Figura **2-1**). El uso del hexágono es razonable ya que se pueden representar grandes áreas de cobertura además debido a que aproxima a un círculo. El la practica las células no son hexagonales, no tienen orden y muchas de ellas están empalmadas.

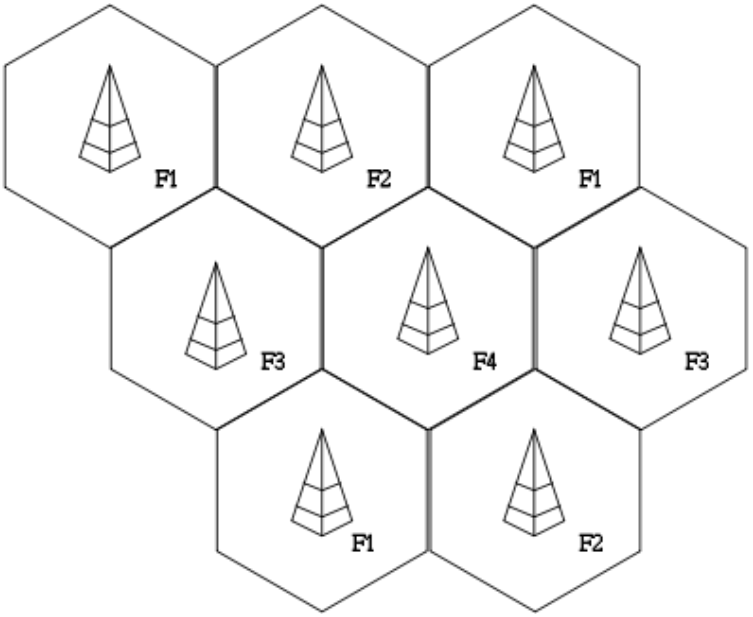

Figura 2-1.: Diseño de red

2.2.1. Categorías de Células

Las células están categorizadas en tres grandes grupos: macrocélulas, microcélulas y pico células. Como se mencionó con anterioridad, mientras las células se van haciendo más pequeñas la reusabilidad de canales aumenta (Figura **2-2**).

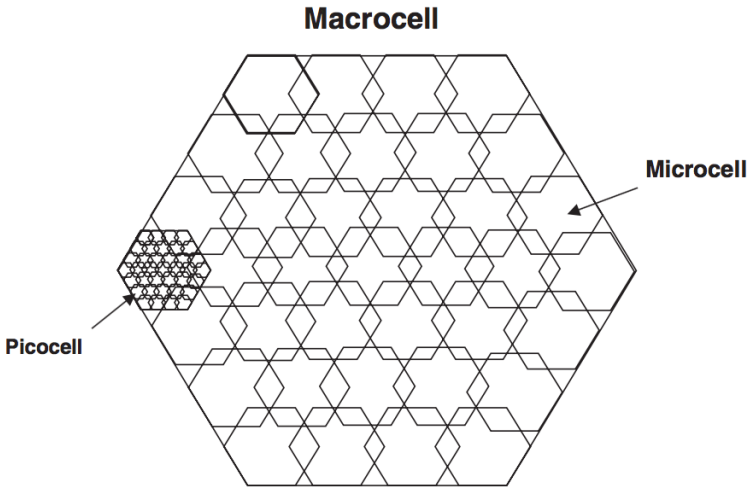

Figura 2-2.: Tipos de célula

Como se puede apreciar en la Figura **2-3** la configuración de la Macrocélula está compuesta de microcélula y por tanto reutilizar canales no contiguos. Sin embargo la red se vuelve más compleja y costosa. Aunque una macrocélula pueda cubrir una gran extensión no permite múltiples sesiones, por tanto si se divide la macrocélula en picocélulas se puede dar soporte a más sesiones.

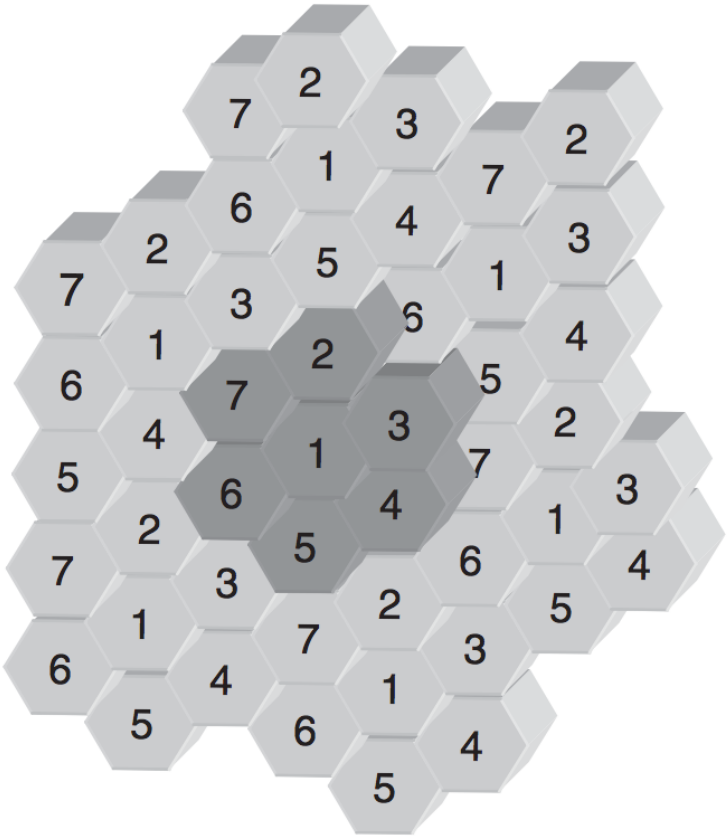

Figura 2-3.: Macrocélula

2.2.2. Rango Celular

Para determinar el rango es necesario calcular todas las pérdidas y ganancias que hay entre el transmisor y receptor, a través del medio, puesto que existe ruido externo al sistema. Es por ello que se necesita un parámetro fiable para poder descartar todos los ruidos y adicionar las ganancias del sistema. Este parámetro es denominado *link budget*, el cual está en función de distintas variables, el número de variables dependerá de la principal funcionalidad que se le quiera dar a la red como: cobertura, capacidad, calidad etc. Algunas de las variables fundamentales a considerar son:

- Transmitted carrier power
- Transmitter antenna gain
- Path loss

- Receiver antenna gain
- Received signal power
- Receiver sensitivity
- Receiver noise power
- Receiver noise density
- Receiver noise bandwidth
- Thermal noise density
- Interference margin
- Handoff gain
- Location probability
- Blocking probability
- Area type information
- Propagation condition
- Spectrum available
- Traffic density information
- Coverage regions

De manera general se puede definir al *link budget* como una forma de cuantificar el desempeño del enlace de comunicación, del transmisor en el medio. El *link budget* en general es una función de parámetros de enlaces de comunicación tales como: la energía transmitida, pérdidas y ganancias en el transmisor y receptor.
Los cálculos del *link budget* estiman el valor máximo de atenuación de señal, llamado *path loss*, entre el dispositivo móvil y la antena base. El valor máximo del *path loss* permite determinar el valor máximo del rango de la célula, el cual puede determinarse con un modelo de propagación adecuado. El rango de la célula estima el valor del número de antenas que se necesitan para cubrir un área geográfica determinada (Holma, 2009).

Para calcular el valor máximo del *path loss* se requiere de los siguientes parámetros:

Base station transmission power	P_b
Base station antenna gain	G_b
Cable loss in the base station	L_c
Equivalent Isotropic Radiated Power (EIRP)	$EIRP = P_b + G_b - L_c$
Noise figure	N_f
Boltzmann constant	k
Termal noise	t
Noise bandwidth	N_b
Terminal noise or total input noise	$Nt = (k)(t)(N_b)$
Energy-to-noise ratio	$E_b/N0$
Interference margin	I_m
Control channel overhead	O_c
User equipment antenna gain	G_u
Body loss	L_b

Tabla 2-1.: Parámetros para el cálculo del *path loss*

El *path loss* se define como:

$$Max(L_p) = EIRP - (N_f + N_t + E_b/N0) - I_m - O_c + G_u - L_b$$

Como se mencionó antes se debe escoger un modelo de propagación adecuado tales como Walfish-Ikegami, Lee-Young y Okumura-Hata. El modelo más comúnmente usado es el de Okumura-Hata descrito de la siguiente manera. En el modelo de Okumura-Hata la definición del path loss es la siguiente:

$$P_l = A + Blog(d) + C$$

Donde A ,B y C son factores que dependen de la frecuencia y altura de la antena, mientras que d es el rango en km. Sea

- f_c: frecuencia de transmisión
- h_b: Altura de la antena
- h_m: Altura de antena del dispositivo móvil

De tal manera que.

$$A = 69{,}55 + 26{,}16log(f_c) - 13{,}82log(h_b) - a(h_m)$$

$B = 44{,}9 - 6{,}55log(h_b)$

La ecuación C y la función $a(h_m)$, dependen del medio de transmisión. Por lo que para mediana y pequeñas ciudades se tiene que:

$a(h_m) = (1{,}1log(f_c) - 0{,}7)h_m - (1{,}56log(f_c) - 0{,}8)$

$C = 0$

Para áreas metropolitanas

$$a(h_m) \begin{cases} 8{,}29(log(1{,}54h_m))^2 - 1{,}1, f \leq 200Mhz \\ 3{,}2(log(11{,}75h_m))^2 - 4{,}97, f \geq 400Mhz \end{cases}$$

$C = 0$

Para áreas suburbanas

$C = -2(log(\frac{f_c}{28}))^2 - 5{,}4$

Para zonas rurales

$C = -4{,}78(log(f_c))^2 + 18{,}33log(f_c) - 40{,}98$

La función $a(h_m)$ en las áreas suburbanas y rurales es la misma qué expresión para medianas y pequeñas ciudades. Dadas todas estas funciones podemos determinar el valor del rango(d) y finalmente obtener la distancia máxima que tendrá la célula. Para una banda de 1900MHz para un red WCDMA(3G) los valores del rango promedio de la célula son los siguientes:

Tipo	Rango(km)
Urbano	0.447
Suburbano	1.178
Rural	9.387

Tabla 2-2.: Rangos promedio por áreas

El tipo de tecnología GSM, 3G, 4G no demuestra un incremento del rango de manera significativa, por lo que las variaciones son muy pequeñas, el cambio más notorio es el incremento de un intercambio de datos mayor, de igual manera el cambio de banda no representa una alteración importante en el rango de la célula (Holma, 2009) en la Figura **2-4** se puede observar la distribución del rango en kilómetros comparado con el tipo de banda que se utiliza.

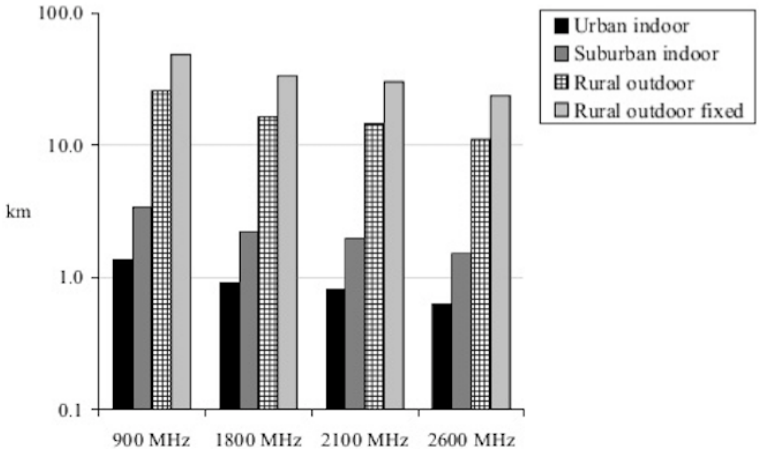

Figura 2-4.: Rango de frécuencias (Horak (2007))

2.2.3. Funcionamiento general

La red debe localizar cada dispositivo móvil para poder proveer del servicio de manera rápida y con poco retraso. Por lo que la red es capaz de localizar la célula en la cual se encuentra registrado el dispositivo móvil. Las dos operaciones básicas de la red son: la actualización de la ubicación y paginación.

La actualización de la ubicación consiste en enviar una señal al dispositivo móvil para poder actualizar su ubicación. Después el registro es actualizado en la base de datos del sistema con el mensaje regresado por el dispositivo móvil con su ubicación. Este funcionamiento es activado en ciertas circunstancias como: el dispositivo móvil se conecta a la red, es decir es encendido o dado de alta en la red, el dispositivo móvil se mueve a una nueva célula, el temporizador de la red expira y automáticamente tiene que registrar la localización de este siga en la misma célula.

La paginación es una operación realizada por la red celular. Esta se da cuando una llamada entra a un dispositivo móvil, entonces la red busca al dispositivo móvil por la red, en todas las posibles células, enviando un mensaje de paginación. Este procedimiento intenta encontrar en cuál célula está el dispositivo móvil para que la llamada entrante pueda ser transmitida a la antena correspondiente.

Un teléfono celular dentro de esta red celular tiene dos funciones básicas; debe de localizar estaciones(antenas) activas e inactivas, una vez ubicada debe conectarse a la que esté disponible. Las siguiente tarea requiere de monitoreo constantemente para asegurar que la conexión sea de calidad, además toma en adición factores como la topología y la cantidad de tráfico en la red, en caso contrario se elegirá una nueva base para conectarse.

2.3. Matrices Origen-Destino

Las Matrices Origen-Destino son aquellas representaciones matriciales de la demanda de viajes entre dos puntos. Son utilizadas como información necesaria en la toma de decisiones en construcción, planeación de transporte y vías de comunicación. Así mismo para la optimización y futuro re escalamiento de estas. El resultado final de toda esta planeación se vería efectiva en los tiempos de viaje más cortos y cubrir la demanda de manera eficaz en todos los puntos.

Actualmente las matrices Origen-Destino son construidas a base de encuestas a hogares o censos y encuestas de tráfico. Los principales problemas que enfrentan estas metodologías se presentan a continuación.

Las encuestas solo representan una muestra muy pequeña de la gran cantidad de viajes realizados. Por ejemplo la zona metropolitana de la Ciudad de México la última encuesta origen-destino realizada en 2007 encuesto a un total de 46,500(Instituto Nacional de Estadística y Geografía, 2007) participantes mientras que la población total de la zona metropolitana era de 29 millones según datos del último censo de población (Instituto Nacional de Estadística y Geografía, 2011).

Limitación en espacio y tiempo en el conjunto de datos. Los viajes no son representados en su totalidad, puede que tengan escalas, por lo que el nivel de granularidad es poca. Dado que las encuestas solo incluyen viajes cotidianos no se incluyen aquellos realizados en fines de semana, temporada vacacional o aquellos realizados en condiciones extraordinarias. Que en términos de movilidad podrían resultar interesantes para determinar factores de caos vial.

La frecuencia de actualización es difícil. Dado que las encuestas llevan mucho tiempo en realizar y obtener resultados, así como el costo de realizarlas es alto. Se vuelve difícil la actualización de las matrices de manera continua. Por lo que no representan el rápido crecimiento de las zonas urbanas y la transformación de los comportamientos de movilidad.

Actualmente los dispositivos móviles como GPS y teléfonos móviles, brindan una manera alternativa para mostrar donde está y no está la gente. Con las que podemos crear una nueva tendencia del estudio de sociedades, de manera actualizada, granular y de bajo costo, comparado con el método tradicional de encuesta.

2.4. Penetración de telefonía celular

El crecimiento de la telefonía celular a nivel mundial ha ido en aumento de manera acelerada durante los últimos 10 años(Figura **2-5**) la penetración mundial es del 97 % y el 95 % de la población mundial vive en una área con cobertura de telefonía celular International Telecommunication Union (2015). En los países desarrollados el índice de penetración es de 128 % mientras que en los países en vías de desarrollo es del 89 %.

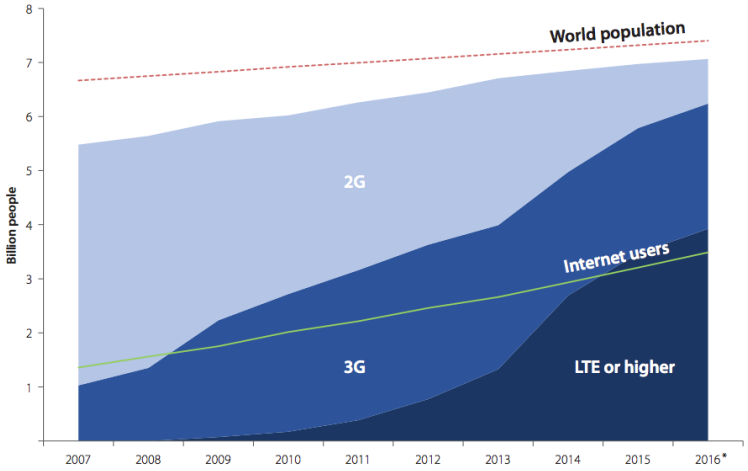

Figura 2-5.: Penetración celular (International Telecommunication Union, 2015)

Es decir que casi toda la población mundial tiene acceso a un teléfono móvil de segunda generación(2G) y en el caso de los países desarrollados, los habitantes tienen uno o más teléfonos móviles por persona.
La penetración de otras tecnologías móviles de datos, tales como 3G o LTE, aún no tiene una penetración tan grande. En países desarrollados es del 90 % mientras que en países en vías de desarrollo es del 40 %. Este tipo de tecnología arrojaría huellas digitales más granuladas, en adición a los CDR como demuestra (Pinelli & Calabrese, 2015).

2.5. Aplicaciones anteriores

Anteriormente se ha abordado este problema desde diferentes perspectivas, soluciones, conjuntos de datos y ciudades.

Uno de los primeros esfuerzos realizados en la determinación de viajes y matrices origen-destino utilizando datos de telefonía celular es abordado por White (2002) hace diferencia entre los tipos de datos que pueden ser usados para la determinación de viajes. En su estudio hacen uso de los registros telefónicos CDR para el análisis del área conurbada

de Kent, UK. Su matriz resultante es comparada y actualizada con otra anteriormente construida la cual se originó de un conteo de tráfico realizado en 1992. En este primer esfuerzo se detalla el esfuerzo de privatizar los datos de los usuarios, la diferencia entre los tipos de datos telefónicos y cuales son ideales para este tipo de experimento, así mismo recalca las posibles implementaciones del análisis de estos datos.

Gonzáles (2008) utiliza los CDRs para caracterizar la movilidad humana, la cual muestra que las trayectorias humanas tienen un alto grado de regularidad en espacio y tiempo, en el que cada individuo está caracterizado de manera independiente por una distancia de viaje y una significativa probabilidad de regresar a algunas locaciones altamente visitadas. Y de manera casi obvia, mientras más se observa al individuo más alta la probabilidad de que visite un lugar nuevo.

Calabrese (2010) utiliza un conjunto de datos más amplio, además de los CDRs recolecta datos de cuando alguna aplicación del teléfono hace uso de la red. Por lo que su conjunto de datos será más robusto y granulado como se demuestra en Pinelli & Calabrese (2015). Realizó un estudio detallado de la generación de una matriz origen-destino para la la ciudad de Boston, con el conjunto de datos mencionado. Caracterizan el tiempo promedio de viaje así como una distribución de los viajes. Así mismo hacen énfasis en que este método puede caracterizar algo que no pueden las encuestas, como días festivos o anormalidades como eventos deportivos, en los que la dinámica de la ciudad se comporta diferente.

Sus resultados son comparados con dos fuentes el US censo 2000 y el CTPP (Census Transportation Planning Package). Su primera comparación es basada en el censo US 2000 dado que logran caracterizar zonas residenciales con el supuesto de que los individuos están en su hogar en un rango de 6 p.m. a 8 p.m. de esta manera establecen los hogares de los individuos. Usando el censo y la estimación de población en conjunto con los datos móviles, logra crear un factor de escalamiento lineal.
El CTPP mapea el flujo de los trabajadores entre su lugar de trabajo y hogar. Por lo que al emplear los factores de escala para poder comparar resultados, resulta en una alta correlación con el CTPP, por lo que la matriz OD estimada es muy parecida a la matriz OD generada con otro tipo de información. También hacen una fuerte diferenciación entre los viajes semanales y aquellos en fin de semana, asimismo compara días festivos, un lunes festivo contra lunes regulares, para mostrar cómo la actividad reduce o aumenta. Lo mismo hace para eventos extraordinarios como un concierto o evento deportivo, por lo que la actividad en ese específico lugar varía significativamente a la actividad usual.

Iqbal (2014) hace un análisis para Dhaka, Bangladesh. El estudio no cuenta con información de censos o estudios respecto a la movilidad, por lo que no puede comparar resultados, algo común en países en vías de desarrollo. Sin embargo emplea un conteo de tráfico utilizando cámaras que detectan la cantidad de autos en cierta vialidad. La matriz OD es generada mediante CDRs y factores de escalamiento determinados por una simulación que tiene como entrada los resultados del conteo de tráfico, para obtener una matriz más realista. Así mismo en su algoritmo toman en cuenta un error de desplazamiento causado por la red celular cuando tiene mucho tráfico tiende a balancear el tráfico y por tanto crear un desplazamiento de antena, que no necesariamente significa un desplazamiento físico. Este estudio hace consideraciones técnicas para rectificar los errores de desplazamiento, por lo que sus resultados finales podrán ser más precisos a comparación de otros estudios del mismo tópico.
No solo se pueden usar los datos móviles para determinar orígenes y destinos. Como el experimento llevado a cabo por Ratti (2007) en el cual, con datos proporcionados por una telefónica Austriaca, desarrolló un mapa en tiempo real de la movilidad de la ciudad Graz, Austria, en la que se se siguen miles de individuos por la ciudad que tienen su celular prendido. Los tres tipos de mapas que fueron desarrollados fueron: densidad del tráfico de dispositivos móviles, origen y destino de las llamadas y el movimiento de los individuos a través de la ciudad. Esto desde una una perspectiva de planeación urbana resulta interesante pues se tiene el pulso de la ciudad en tiempo real. Podríamos detallar de manera precisa problemas viales, poder dar predicciones etc.

Uno de los trabajos más actuales y que utiliza puramente CDRs es el de Dong (2015) en el cual desarrolla todo un *framework* para detectar actividades inusuales en multitudes, pueden diferenciar entre una actividad cotidiana en la multitud o una conglomeración inusual como una manifestación. El experimento se llevaba a cabo con datos extraídos durante 5 meses en 2012 en Costa de Marfil, cabe resaltar que no se usan más datos móviles extra como en trabajos anteriores, se basan totalmente en CDRs, lo que hace interesante el trabajo pues logran detectar con cierta granularidad la localización de los eventos extraordinarios. Con esto existe un gran potencial de detectar y monitorear a las multitudes en tiempo real. Que en términos de seguridad resultaría interesante.

3. Metodología

3.1. Conjunto de datos

Para demostrar el funcionamiento de lo propuesto se realizará el experimento con una base de datos real. La base de datos se obtuvo del concurso *Data for Development Senegal* (D4D Senegal) 2014[1], este concurso libero millones de CDRs de la compañía telefónica Orange, en la edición de 2014 se liberaron datos de Senegal. El objetivo de este concurso es ayudar al desarrollo socio-económico de la región encontrando soluciones o aplicaciones usando los datos.
El conjunto de datos se compone de CDRs detallados de 300,000 usuarios, elegidos aleatoriamente, durante dos semanas. Y cada dos semanas se renovaban los usuarios. Estos datos se recolectaron entre 1 de enero de 2013 y 31 de diciembre de 2013. Por lo tanto se tienen 25 subconjuntos de dos semanas cada uno con CDRs de 300,000 usuarios. Así mismo se proporcionó datos complementarios como: tiempo de llamadas, tipo de actividad y coordenadas geográficas de cada antena de la red.

3.1.1. Descripción de los datos

Para cada usuario el registro contiene un identificador de usuario, hora-fecha y antena a la que se conectó como se muestra en la Tabla **3-1**. Esta información se registra cada 10 minutos para cada usuario, si es que interactúa con la red.

Tabla 3-1.: Datos iniciales

Usuario_id	Timestamp	Site_id
1	2013-03-18 21:30:00	716
1	2013-03-18 21:40:00	718
1	2013-03-19 20:40:00	716
1	2013-03-19 20:40:00	716
1	2013-03-19 21:00:00	716
1	2013-03-19 21:30:00	718
1	2013-03-20 09:10:00	705
1	2013-03-21 13:00:00	705

[1] http://www.d4d.orange.com/en/Accueil

La descripción de las antenas contiene el sector geográfico al que pertenecen, longitud y latitud, la Tabla **3-2** muestra este conjunto. Cada antena pertenece a un sector asignado por la telefónica. Está sección se apega a la distribución geográfica de Senegal, es decir, va acorde a las divisiones departamentales de Senegal.

Tabla 3-2.: Datos antenas

site_id	arr_id	lon	lat
1	2	-17.525142	14.746832
2	2	-17.524360	14.747434
3	2	-17.522576	14.745198
4	2	-17.516398	14.746730

Se puede observar en la Tabla **3-2** que la antena con identificador(site_id) 1 pertenece al sector(arr_id) 2 y sus correspondientes longitud(lon) y latitud(lat).

La última tabla de importancia es aquella en la que cada sector es asignado a un área geográfica de Senegal, esto nos permitirá elegir el área a analizar. Esta contiene el identificador del sector, la región a la que pertenece, departamento y barrio, como lo muestra la Tabla **3-3**. El sector(arr_id) número 1 pertenece a la región(reg) de Dakar, al departamento(dept) de Dakar y al barrio(arr) Parcelles Assainies.

Tabla 3-3.: Datos departamentos

arr_id	reg	dept	arr
1	DAKAR	DAKAR	PARCELLES ASSAINIES
2	DAKAR	DAKAR	ALMADIES
3	DAKAR	DAKAR	GRAND DAKAR
4	DAKAR	DAKAR	DAKAR PLATEAU
5	DAKAR	GUEDIAWAYE	GUEDIAWAYE
6	DAKAR	PIKINE	PIKINE DAGOUDANE

Estas tres tablas, aquí detalladas, serán suficientes y necesarias para realizar el análisis de la matriz origen-destino propuesta.

3.2. Descripción del área de estudio

Senegal se encuentra en el oeste del continente Africano con una extensión territorial de 196,712 km^2 su población es de 13.5 millones de habitantes (Agence Nationale de la Statistique et de la Démographie, 2013) la penetración de la telefonía móvil para Senegal en el año 2013 era del 96.8 % (Autorité de Régulation des Télécommunications et des Postes, 2013) por lo que podríamos suponer que al menos cada habitante tendría un teléfono móvil. En términos de movilidad el número de automóviles en 2013 era de 401,910. Donde el 72.8 % se concentraba en la ciudad capital de Dakar. Menos del 30 % eran vehículos de modelo reciente (Agence Nationale de la Statistique et de la Démographie, 2013), por lo que la mayoría son vehículos usados.

Para este caso de estudio nos centraremos en la ciudad capital de Senegal, Dakar. Dakar tiene una superficie de 550 km2 y es una ciudad costera. La ciudad está dividida en cuatro departamentos: Dakar, Pikine, Guédiawaye y Rufisque. La población en la ciudad de Dakar hasta 2013 era de 3 millones un 23.3 % de la población de Senegal, en su mayoría jóvenes. Dakar es la ciudad con mayor densidad poblacional de Senegal con 5704 personas por kilómetro cuadrado. El 96.5 % es población urbana mientras que el restante es población rural (Service Régional de la Statistique et de la Démographie de Dakar, 2013).

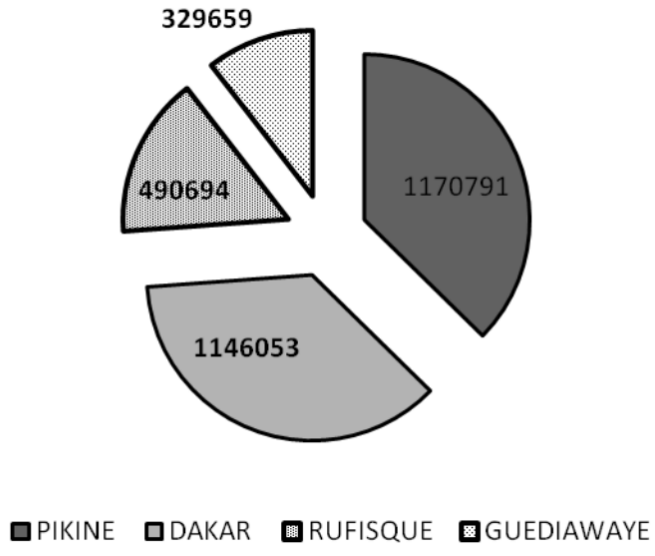

Figura 3-1.: Población por departamento

La distribución de la población en los cuatro departamentos de la ciudad de Dakar se representan en la Figura **3-1**

Tabla 3-4.: Encuesta de movilidad (Service Régional de la Statistique et de la Démographie de Dakar, 2013)

	Encuesta 2011	Encuesta 2015
Desplazamiento por día (de todas las maneras)	3440000	6545000
Motivo de desplazamiento por día		
Trabajo, estudio	35.3 %	25.9 %
Domestico, personal	23.3 %	16.8 %
Social, recreación	19.3 %	13.7 %
Regreso a casa	S/D	42.9 %
Otros	22.1 %	0.7 %
Horas pico		
Mañana		
Caminando	7-14: 46 %	7-11: 29 %
Motorizado	7-11: 30 %	7-10: 26 %
Tarde		
Caminando	17 - 20: 21 %	17 - 20: 23 %
Motorizado	16 - 20: 20 %	16 - 19: 24 %
Duración de los desplazamientos por día de la semana		
	<= 15 min : 69 %	<= 14 min : 53 %
	16 - 30 min : 16 %	15 - 29 min : 23 %
	31 - 60 min : 11 %	30 - 59 min : 17 %
	>60 min : 4 %	>= 60 min : 7 %

[1] Todos los encuestados mayores de 14 años

El transporte público en Dakar se compone principalmente por autobuses, microbuses, camionetas y taxis. Además de contar con un transporte ferroviario PTB (Petit Train de Banlieue) que ofrece un servicio entre los departamentos de Dakar y Rufisque, con un número promedio de pasajeros anual de 2.5 millones (Service Régional de la Statistique et de la Démographie de Dakar, 2013). Uno de los pocos datos de tiempo de traslado que se tienen es de la construcción de la carretera Dakar-Diamniadio, carretera que conecta el departamento de Dakar con el de Rufisque, el cual redujo un tiempo de trayecto dos horas a menos de 30 minutos. Datos de una encuesta de movilidad de 2015 nos pueden dar una perspectiva más amplia del transporte en Dakar (Tabla **3-4**).

Se cuenta con 256 vehículos motorizados por cada 1000 casas habitación y 40 vehículos motorizados por cada 1000 habitantes. Existen 169 vehículos particulares por cada 1000 casas habitación y 25 vehículos particulares por cada 1000 habitantes. Hay 1.7 millones de viajes en transporte público por día en Dakar, de los cuales el 40.5 % pertenecen

al departamento de Dakar. El 80 % de los desplazamientos en vehículos motorizados se hace en transporte público. De manera más general la movilidad en Dakar se podrá resumir, para individuos de más de 11 años, de tal manera que de Lunes a Viernes se registran un total de 7.2 millones de desplazamientos diarios, de todas las formas. Se registran 3.36 viajes por persona por día y 70 % de los desplazamientos son peatonales, es decir el medio de transporte más usado es caminar y el tiempo estimado de transporte es de 61 min (Conseil Exécutif des Transports Urbains de Dakar, 2015).

3.3. Depuración de Datos

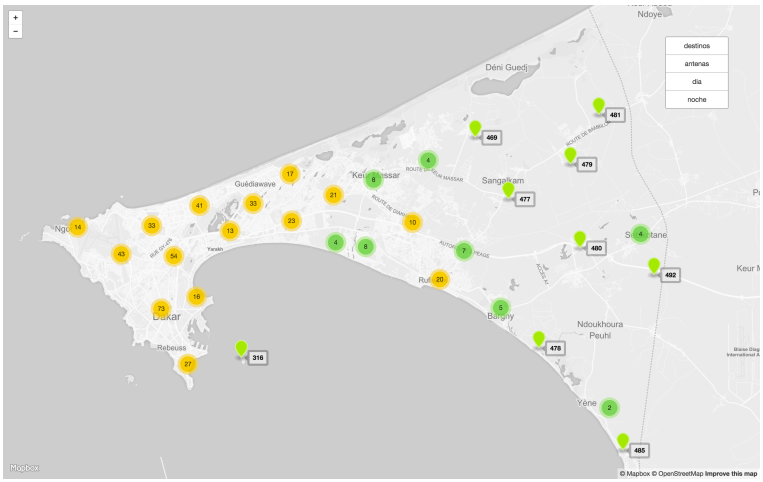

Figura 3-2.: Antenas en la ciudad de Dakar

El conjunto de datos está depurado al momento de su recolección, con el fin de proteger la privacidad de cada usuario, es decir, para cada conjunto de dos semanas el identificador de usuario puede ser el mismo pero no implica que sea el mismo usuario. Por lo tanto es poco probable rastrear al mismo usuario.

El conjunto de datos en su totalidad contiene CDRs de todo Senegal, en este presente trabajo trabajaremos con la ciudad de mayor densidad en Senegal, Dakar, con el propósito de hacer una mejor caracterización de una gran urbe. La siguiente etapa de depuración será excluir todos aquellos CDRs que no pertenezcan a la región de Dakar. En la región de Dakar al momento de recolección de datos se contaban con un total de 489 antenas para una región de 550km^2 es decir casi un antena por kilómetro cuadrado (Figura **3-2**). Para extraer los datos que solo pertenecen a Dakar, se excluirán todas las antenas que no pertenezcan a la región. Utilizando las tablas **3-1**, **3-2**, **3-3**. Se filtraran primero todas las antenas que pertenezcan a la región de Dakar y posteriormente todos los CDRs que cumplen la condición de las antenas.

La tabla preliminar de la intersección de los datos pertenecientes a las tablas antes mencionadas, será la siguiente (Tabla **3-5**).

Tabla 3-5.: Datos por usuario

id	uid	timestamp	siteid	lat	lon
1	6000201358	2013-03-18 09:20:00	10	14.747767	-17.508766
2	6000201358	2013-03-18 09:40:00	10	14.747767	-17.508766
3	6000201358	2013-03-18 09:40:00	11	14.751808	-17.505067
4	6000201358	2013-03-18 09:50:00	6	14.748411	-17.512103
5	6000201358	2013-03-18 10:20:00	10	14.747767	-17.508766
6	6000201358	2013-03-18 11:10:00	10	14.747767	-17.508766
7	6000201358	2013-03-18 11:20:00	10	14.747767	-17.508766
8	6000201358	2013-03-18 11:30:00	11	14.751808	-17.508766

Como se puede observar el identificador único del registro se denomina por *id*, al identificador de usuario se le denomina por *uid*, el registro de fecha y hora por *timestamp*, al identificador de antena por *siteid*, finalmente longitud y latitud.
De esta última tabla (Tabla **3-5**) se filtra de manera final, que tiene como objetivo quitar los CDRs repetidos, existen CDRs que están en la misma hora y fecha pero en diferentes antenas como el registros de la Tabla 4.4 con id: 2,3. Esto sucede debido a un falso desplazamiento, es decir el teléfono móvil cambia de antena sin necesariamente haber cambiado de posición, esto debido al funcionamiento de la red, ya que al existir alta demanda en la red esta necesita balancear el tráfico de la red y desplazar el flujo de aquellas antenas que tienen alta demanda a aquellas que no, de tal manera que la carga del tráfico de manera distribuida. Es por ello que esto ocasiona un falso desplazamiento y registra un mismo usuario a la misma hora pero en distinto lugar.

Por lo tanto con el fin de evitar errores se eliminaron los valores duplicados, sólo se mantendrá el primer registro y los demás incidentes, repetidos, serán eliminados. Una vez eliminados estos datos la tabla se redujo en número de registros a procesar por el algoritmo así como reducción de errores.
Finalmente se creó una tabla con datos de cada mes, es decir 12 tablas o subconjuntos de datos. Con ello podemos procesar en etapas por cada conjunto de datos y no tener que procesar todo en un solo conjunto.

3.4. Algoritmo

3.4.1. Definición de viaje

Para poder hacer un análisis de movilidad y generar la matriz de origen-destino se necesita analizar el movimiento individual de cada usuario del conjunto de datos, se definirán dos tipos de movimientos viajes y segmentos de viaje. Para cada usuario se toma de manera consecutiva los registros de los CDR que serán definidos como segmentos de viaje. Los viajes están compuestos de un origen y destino, mientras que los segmentos de viajes conforman la totalidad de un viaje. Es decir serán transbordos. El origen del viaje se definen como el primer registro (CDR) del usuario y los destinos son aquellos en los que el usuario permanece, de manera consecutiva, estático por más de cierto tiempo (Δt)
Este trabajo no pretende generar las rutas más concurridas por las que un usuario se desplaza, solo se almacenan los puntos iniciales y finales, es decir orígenes y destinos.
Los orígenes y destinos, referentes a la matriz no son los mismos que los de viajes, más adelante determinaremos el proceso para extraer esta información.

3.4.2. Algoritmo

Antes de la creación del algoritmo, se hicieron algunas consideraciones para poder determinar la demanda de viajes lo mejor posible de los CDRs, que como se mencionó en el capítulo anterior carecen de una buena resolución en espacio y tiempo.
Se considera que el usuario avanzó, si este se desplaza entre dos sitios(antenas) más de 3 km o si el origen y sitio(antena) actual se encuentran a una distancia mayor a 3 km. Se considera un viaje completo si el usuario de manera consecutiva está estacionado o estático por más de 20 minutos. Después de esto se crea un nuevo viaje, con un nuevo origen. Si el viaje se completa en un plazo mayor a 3 horas entonces se descarta y empezara uno nuevo.

El Algoritmo 1 servirá para extraer los viajes de cada usuario. Debido a la gran cantidad de datos, se analiza información mensualmente, por lo que el conjunto se dividirá en doce subconjuntos, respectivos a cada mes. De cada mes se extraerán los identificadores de usuarios y entonces se seguirá el siguiente análisis.
Para justificar las constantes nos basaremos en algunas presunciones y factores técnicos. Al ser un área conurbada es de esperar que los viajes no sean prolongados entonces basados en tiempos promedios de grandes urbes como Beijing 52 min[2], New York 48 min[3] y D.F. 1 hora 21 min ((Instituto Nacional de Estadística y Geografía, 2007)). Por lo

[2] http://www.bbc.com/capital/story/20170221-the-gruelling-six-hour-commute-of-beijings-workers
[3] http://pfnyc.org/news/

tanto delimitamos este parámetro a máximo 3 horas. El parámetro de 3 km es debido a que la falta definición espacial y temporal escogeremos un parámetro mayor al del rango celular urbano y suburbano (.5 km - 1.1 km) y menor al rural (9.1 km). Con esto podemos caracterizar movimientos reales y no falsos desplazamientos.

3.4.3. Análisis de algoritmo

Para determinar la complejidad del algoritmo (Algoritmo 2), su máximo tiempo de ejecución y si puede determinarse una mejor implementación utilizaremos la notación big O (O(f(n))). Para cada parte del algoritmo se analiza su complejidad y posteriormente se suman todas las variables y/o constantes que nos resultan en el tiempo de ejecución total. Por lo que la suma de las constantes y variables

$$T(n) = n(m + m - 1 + m - 1 + c_1 + c_2 + c_3 + c_4) = O(n(m)), m < n$$

El tiempo máximo de ejecución será polinomial, de los datos que se tienen sabemos que por cada conjunto de datos existen 300 mil usuarios y que se registra la actividad del usuario, si se conecta a la red, cada 10 minutos. Por lo que en el peor escenario todos los usuarios tendrán actividad cada 10 minutos de manera continua es decir para cada usuario del conjunto se tendrían 2160 CDRs, por lo que el tiempo de ejecución por conjunto sería de:

$$T(n) = n(m) = 300000(2160)$$

Que en un supuesto teórico tardaría 10800 minutos en determinar los viajes de cada conjunto de datos. Claro el supuesto teórico nos puede dar una estimación pero dependiendo de la tecnología usada, lenguaje, memoria, etc. este parámetro podría incrementarse.

3.5. Paralelización

Dado que la cantidad de datos resulta inmensa, intentar ejecutar el algoritmo de manera secuencial podría resultar tardado dado que su complejidad es del orden de $O(n^m)$ y en el supuesto de que m tiende a n podríamos decir que $O(n^2)$. Por tanto para poder incrementar su velocidad de respuesta es necesario poder ejecutar el algoritmo en paralelo. Python cuenta con bibliotecas para hacer uso de paralización por hilos o procesos, al usar hilos no hacemos uso total del CPU dado que Python cuenta con un Global Interpreter Lock, es decir un candado global de intérprete, esto es principalmente porque el manejo de memoria no es segura de usar con hilos (thread-safe)[4]. Por lo que los hilos se ejecutan en una sola instancia de GIL y de Python. Al usar paralelización por procesos, se crean

[4] `https://wiki.python.org/moin/GlobalInterpreterLock`

Algoritmo 1: Extracción de viajes por usuario

Datos:
S={conjunto de sitios}
U={conjunto de usuarios}
$\Delta t^- = 20min$
$\Delta t^+ = 180min$
$\Delta d = 3km$
Resultado: Viajes individuales por usuario

```
para todo u ∈ U hacer
    para cada s_i ∈ S hacer
        si origen nuevo entonces                    /* Origen nuevo por defecto */
            origen = s_i
        si s_i ≠ s_{i+1} entonces
            si d(s_i, s_{i+1}) > Δd entonces
                continua viaje
            en otro caso
                si d(origen, s_i) > Δd y Δt(s_i, s_{i+1}) > Δt⁻ y Δt(origen, s_i) < Δt⁺
                 entonces
                    destino = s_i
                    viaje_j=(origen, destino)
                    Establece origen nuevo
        en otro caso
            si d(origen, s_i) > Δd entonces
                si Δt(s_i, s_{i+1}) > Δt⁻ y Δt(origen, s_i) < Δt⁺ entonces
                    destino = s_i
                    viaje_j=(origen, destino)
                    Establece origen nuevo
                si no, si s_i = s_{i+1} = s_{i+2} y Δt(origen, s_i) < Δt⁺ entonces  /* Evita
                 pérdida de viajes */
                    destino = s_i
                    viaje_j=(origen, destino)
                    Establece origen nuevo
                en otro caso
                    continua viaje
            en otro caso
                Establece origen nuevo
```

Algoritmo 2: Complejidad de Algoritmo

```
1  para todo u ∈ U hacer                                              /* n */
2      para cada s_i ∈ S hacer                                        /* m */
3          si origen nuevo entonces
4              origen = s_i
5          si s_i ≠ s_{i+1} entonces                                  /* m-1 */
6              si d(s_i, s_{i+1}) > Δd entonces                       /* c_1 */
7                  continua viaje
8              en otro caso                                           /* c_2 */
9                  si d(origen, s_i) > Δd y Δt(s_i, s_{i+1}) > Δt⁻ y Δt(origen, s_i) < Δt⁺
                    entonces
10                     destino = s_i
11                     viaje_j=(origen, destino)
12                     Establece origen nuevo
13         en otro caso                                               /* m-1 */
14             si d(origen, s_i) > Δd entonces                        /* c_3 */
15                 si Δt(s_i, s_{i+1}) > Δt⁻ y Δt(origen, s_i) < Δt⁺ entonces
16                     destino = s_i
17                     viaje_j=(origen, destino)
18                     Establece origen nuevo
19                 si no, si s_i = s_{i+1} = s_{i+2} y Δt(origen, s_i) < Δt⁺ entonces
20                     destino = s_i
21                     viaje_j=(origen, destino)
22                     Establece origen nuevo
23                 en otro caso
24                     continua viaje
25             en otro caso                                           /* c_4 */
26                 Establece origen nuevo
```

diferentes instancias de Python, que no comparten memoria. Entonces se puede ejecutar simultáneamente el algoritmo, aprovechando al máximo la capacidad del CPU.

Para su funcionamiento se hicieron modificaciones al Algoritmo 1, que no afectan la parte central del cálculo de rutas, solo se asignan usuarios a cada proceso, es decir si se tienen 32 instancias que ejecutan el algoritmo entonces los n usuarios se repartirán entre esas 32 instancias. Con eso se asegura que los datos no se corrompan. Por tanto se particiona de manera balanceada los usuarios entre n instancias.

3.6. Matriz resultante

3.6.1. Matriz transitoria

La matriz resultante de la determinación de viajes (Algoritmo 1), es un matriz que contiene todos los viajes realizados por el usuario, de manera que se denomina matriz transitoria, dado que no es la matriz final. Esta matriz transitoria sirve para obtener una perspectiva de la dinámica general de la ciudad. Pues toma en cuenta viajes no habituales: viajes de fin de semana o eventos especiales.

3.6.2. Matriz origen-destino

La matriz origen-destino se deriva de la matriz transitoria, para obtener los lugares donde los usuarios residen y donde realizan sus actividades diarias: trabajo, escuela, etc. Para ello utilizaremos el rango de las horas pico para Dakar, como se muestra en la tabla **3-4**. Por la mañana se tendrá un horario de 7 am a 11 pm y por la tarde de 5 pm a 8 pm de esta manera podremos obtener los lugares habituales por la mañana y tarde. Por lo que para todo el conjunto de usuarios no se pueden determinar lugares cotidianos, dado que su actividad no fue la suficiente para determinar más viajes y por tanto su lugar de residencia o trabajo, es por ello que estos usuarios se discriminaran, por lo que la matriz origen-destino resultante está compuesta únicamente con usuarios que cumplan la condición de cotidianidad en el rango establecido. El Algoritmo 3 explica el procedimiento para obtener la matriz origen-destino resultante.

Algoritmo 3: Extracción lugares más comunes

Datos:
T={conjunto de viajes}
U={conjunto de usuarios}
X={conjunto de orígenes}
Y={conjunto de destinos}
P={conjunto de pares ordenados $(x, y) \in$ T | x $\in$ X & y $\in$ Y }
Resultado: Viajes más comunes por usuario

```
1  para todo u ∈ U hacer
2      para cada p_i ∈ T hacer
3          si viaje nuevo entonces        /* (origen y destino de cada viaje) */
4              px_i = 1
5              py_i = 1
6          en otro caso
7              px_i += 1
8              py_i += 1
9          orígenes[x_i] = px_i
10         destinos[y_i] = py_i
11     origen frecuente = máximo(orígenes)
12     destino frecuente = máximo(orígenes)
```

3.7. Plataforma de visualización

Para poder visualizar los flujos de los viajes de una manera geográfica y gráfica, se desarrolló una plataforma interactiva, que explica el flujo de la ciudad de Dakar. Esta aplicación web permite observar detalladamente cuales son las áreas residenciales o laborales, así como los puntos de mayor demanda en la ciudad de Dakar, se puede acceder a las rutas o viajes con mayor demanda.

Los aspectos técnicos de la aplicación Web se presentan a continuación mientras que el desarrollo y modelado de esta aplicación no es la intención de este trabajo, podrá encontrase el repositorio público de BitBucket[5].
La aplicación Web está construida en Django, un *framework* Web para Python, los mapas qué se utilizan son *open-source* (OpenStreetMap). La arquitectura de la plataforma es un modelo básico de MVC (modelo-vista-controlador). Se alimenta de una base de datos (Postgres) con los valores finales de la matriz transitoria y la matriz de origen-destino. La interfaz de usuario está desarrollada en JavaScript y HTML.

[5] `https://bitbucket.org/ludwing_van/dakar`

4. Resultados

4.1. Frecuencia de actividad

Para poder tener más claro un panorama general de los datos analizados, se muestra la frecuencia de actividad o interacciones con la red por usuario, en un periodo de tiempo mensual o diario, la cual se eligió de manera aleatoria.

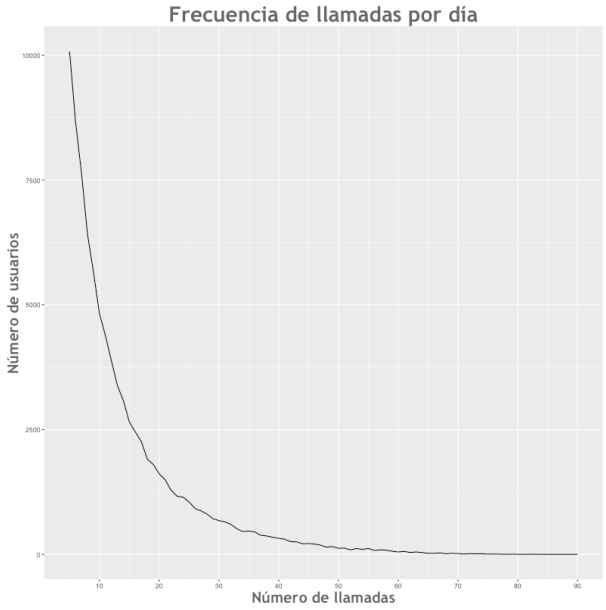

Figura 4-1.: Llamadas por día

La frecuencia de actividad diaria (Figura **4-1**) y mensual (Figura **4-2**) muestra que existe una gran cantidad de usuarios que interactúan con la red de manera muy esporádica, la gráfica nos muestra que existe una gran cantidad de usuarios con menos de 10 llamadas, lo que nos da la impresión de que no se podrán determinar tantos viajes para analizar. Sin embargo uniendo todos los meses analizados (Figura **4-3**), se observa que prevalece una cantidad enorme de usuarios con poca actividad, sin embargo se tiene una cantidad suficiente de usuarios en el rango de 10 a 100 llamadas, por lo que es suficiente el número de viajes.

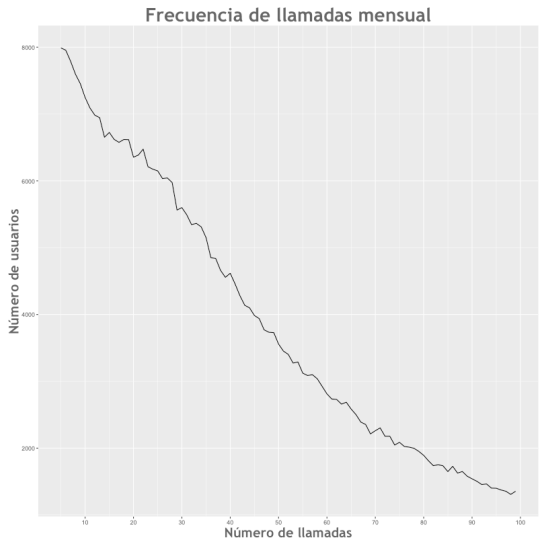

Figura 4-2.: Llamadas mensuales

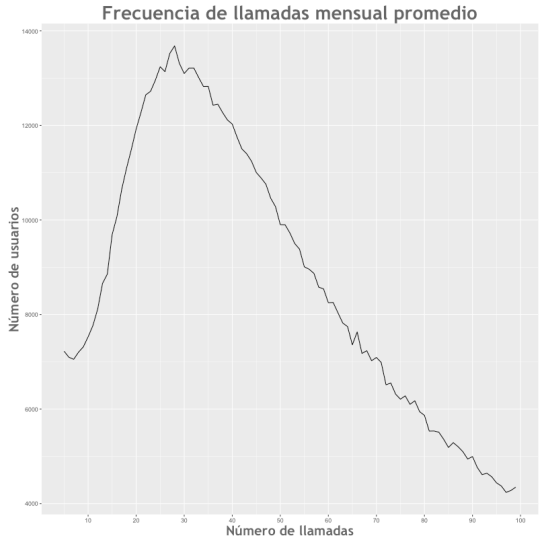

Figura 4-3.: Llamadas mensual promedio

4.2. Matriz Origen-Destino

La tabla de viajes resultante del análisis de los viajes es la que se presenta en la Tabla **4-1**. Esta tabla contiene un identificador único(id), el identificador de usuario (uid), el timestamp de inicio del viaje(t_i), identificador de la antena de inicio(site_i), latitud(lat_i) y longitud(lon_i) de la antena de inicio y de manera recíproca para los valores subsecuentes que pertenecen a la antena de destino, donde el viaje termina. Está tabla es la que se denomina matriz transitoria, pues contiene todos los orígenes y destinos de todos los viajes realizados

Tabla 4-1.: Datos por usuario (Matriz transitoria)

id	uid	t_i	site_i	lat_i	lon_i	t_d	site_d	lat_d	lon_t
1687755	400023928	2013-02-19 10:40	23	14.74	-17.49	2013-02-19 11:40	65	14.69	-17.46
687757	400023928	2013-02-20 08:40	23	14.74	-17.49	2013-02-20 10:10	47	14.69	-17.47
689070	40005031	2013-02-21 09:00	430	14.76	-17.30	2013-02-21 10:00	248	14.72	-17.43
689071	40005031	2013-02-22 11:00	39	14.71	-17.47	2013-02-22 11:10	216	14.66	-17.44

id	site_id	lon	lat	site_id_d_fk	count
1	456	-17.268864	14.783741	443	109
2	420	-17.316665	14.776986	447	107
3	481	-17.16592	14.82036	480	105
4	479	-17.184373	14.789131	171	85
5	486	-17.150519	14.743846	141	77
6	447	-17.284472	14.761459	296	72
7	412	-17.327504	14.776569	104	69
8	425	-17.310757	14.782069	247	62
9	488	-17.136879	14.745807	56	59
10	420	-17.316665	14.776986	272	56

Tabla 4-2.: Matriz origen destino

La matriz transitoria es la resultante de todos los viajes encontrados, posteriormente se deriva la matriz de origen-destino. La determinación de esta matriz se hace con el análisis de los lugares más frecuentados por cada usuario, definiendo dos locaciones más comunes de viaje.

Es de importancia poder analizar propiedades cuantitativas de los resultados, por ello para poder verificar la veracidad de este trabajo se procederá a comparar con resultados anteriores de experimentos similares. Dado que no existe un parámetro que nos permita verificar o comparar los resultados de la matriz resultante sí podemos identificar las zonas residenciales y aquellas donde se encuentran concentradas los centros de trabajo esto podremos verificarlo con el trabajo de Borderon (2010). Además verificaremos las horas pico y la duración de los viajes con los de la Tabla **3-1**.

La duración de los viajes como se muestra en la figura **4-4** se puede observar que el 14.7 % son viajes con duración menor a 30 minutos, el 25 % de los viajes se realiza entre 30 minutos y 1 hora, entre 1 hora - 2 horas corresponde el 39.4 % de los viajes mientras que el 20.9 de los viajes restantes duran más de 2 horas. Es decir que casi 40 % se realiza en menos de 1 hora, esta ligera variación respecto a la Tabla **3-4** se debe a la restricción del algoritmo, dado que los CDRs son poco precisos en espacio tiempo, necesitábamos un rango de tiempo de espera mayor para poderlo clasificar como un viaje, como la tabla **3-1**. supone que el 53 % de los viajes se realiza en menos de 14 minutos, no podemos determinar este tipo de viajes pues tenemos que tener una mayor granularidad en espacio y tiempo. Aunque si pudimos caracterizar algunos viajes con duración corta, estos son poco significativos.

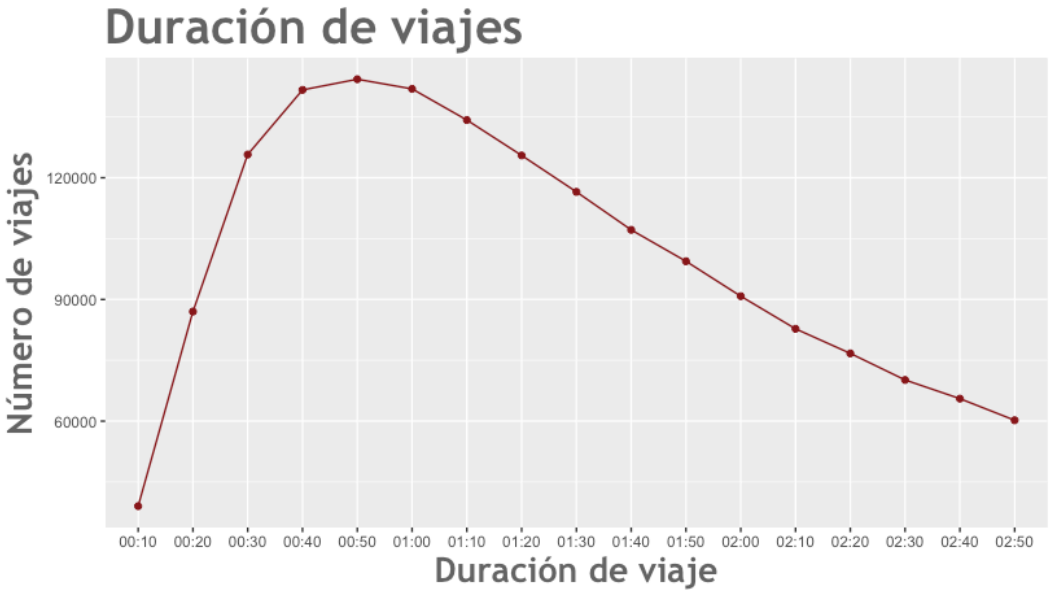

Figura 4-4.: Duración de viajes

Las horas pico (Figura **4-5**) se pueden observar en la mañana en un rango de 7am a 12pm, con un 27.5 % de todos los viajes, por la tarde la hora pico va desde las 4pm hasta las 8pm con un 39.7 % de viajes. Comparado con la tabla **3-1** las horas pico coinciden con los rangos aquí determinados, si bien la cantidad de viajes puede variar en la mañana, sí podemos determinar cuales son las horas de mayor concentración de viajes. También se puede observar que en el rango entre 8 pm y 6 am la cantidad de viajes es muy baja, esto debido a que la cantidad de CDRs registrados durante la noche/madrugada son muy pocos para poder determinar una mayor cantidad de viajes.

De ambas comparaciones podemos determinar que la extracción de viajes mediante CDRs tiene algunas limitaciones, sin embargo se pudieron determinar valores parecidos a los de la encuesta(**Tabla 3-4**), con lo que se muestra los posibles potenciales de trabajar con CDRs.

La detección de zonas residenciales y laborales, se obtiene en base a los sitios que cada usuario visita de manera frecuente, estableciendo rangos de hora, podemos establecer cuál es la residencia y cuál es el sitio donde laboral. Los rangos de hora para determinar una zona residencial es de 8 pm a 7 am y para las zonas laborales en un rango de 8 am a 7 pm. Es decir si un usuario está más de una vez en el mismo sitio durante el rango de horas definida, podemos clasificar por residencia o actividad laboral. En el trabajo presentado por Borderon (2010) identifica zonas residenciales de manera geográfica, utilizando fotografías satélites determinando el uso del suelo, comparándolo con los censos disponibles. Con eso logran determinar ciertas áreas y clasificarlas por su uso de suelo. Este trabajo nos sirve para comparar las áreas detectadas como residenciales y de actividad diurna. Además en el último censo para Dakar, no muestra cuales son los departamentos

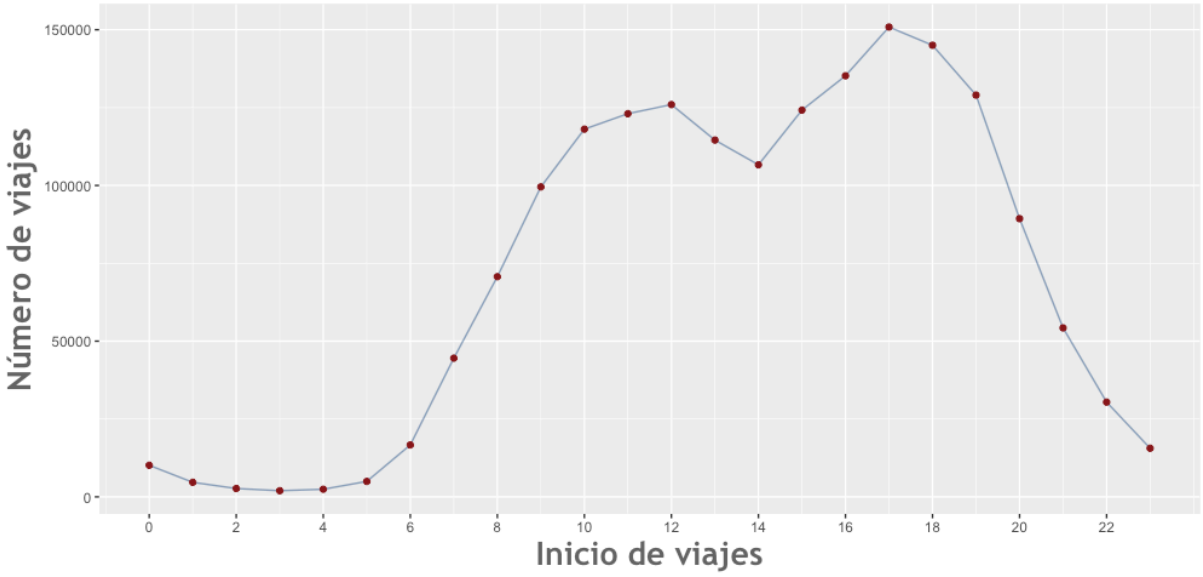

Figura 4-5.: Horas pico

más poblados, lo que es un inconveniente para generar esta comparación.

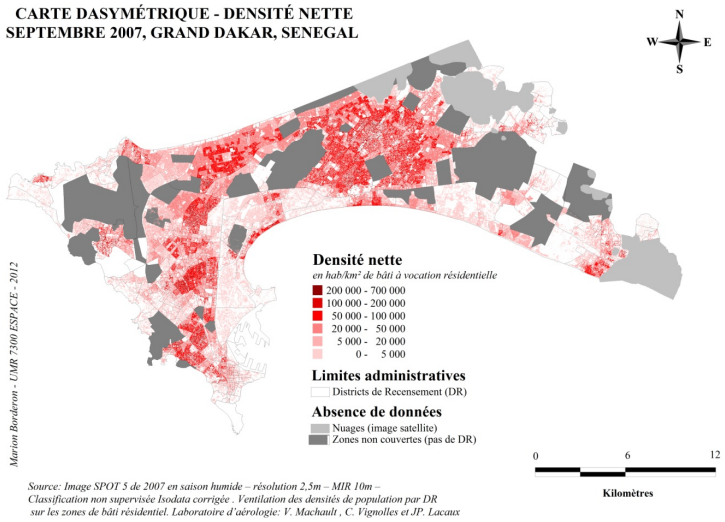

Figura 4-6.: Población en la región de Dakar

En los resultados presentados por Borderon (2010) hemos de identificar dos figuras importantes. En la figura **4-6** podemos apreciar la densidad de población en una gama de blanco a rojo, de menor a mayor poblado respectivamente. Se puede observar con precisión las áreas que cuentan con mayor concentración residencial, una gran parte de la población se encuentra concentrada en los departamentos de Dakar y Pikine. En la Figura **4-7** se muestra como se encuentran divididas la zonas urbanas determinadas por

Borderon (2010), nos debe de importar aquella representada por el color amarillo pues es una área no residencial, es decir se encuentran principalmente ocupada por industria, oficinas o comercios. Las demás zonas, representadas por otros colores, muestran áreas residenciales dependiendo del estatus de dichas viviendas, bien establecidas, irregulares o zonas rurales. Estas no son estrictamente residencial, pero cuentan con una gran cantidad de casas habitación por lo que se podrán encontrar comercios, oficinas etc.

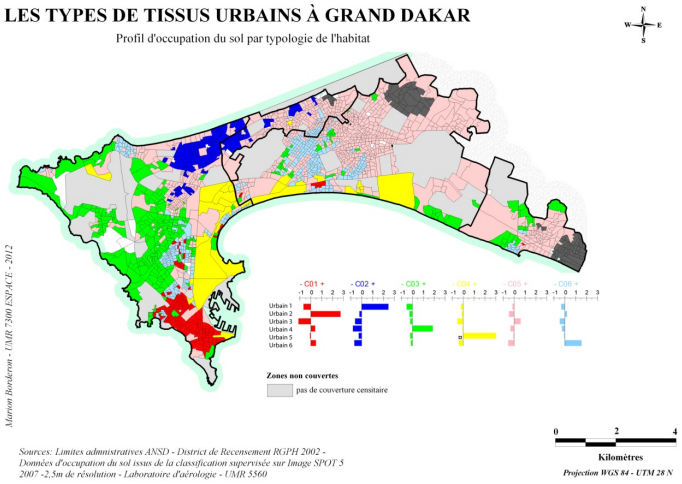

Figura 4-7.: Uso de suelo de Dakar

Los resultados obtenidos de la matriz origen destino para cada usuario nos muestran que para la zona delimitada como mayormente industrial en Borderon (2010) es bastante parecida al resultado de la Figura **4-8** resultante de este trabajo. Representado con un mapa de calor toda la costera y parte del departamento de Dakar, se muestra con una fuerte actividad diurna, que nos puede indicar que son lugares de trabajo. Además nos da una estimación de la vida diurna de Dakar.
Para las zonas habitacionales (Figura **4-9**), estas zonas antes mencionadas aparecen ahora sin actividad o con poca actividad, mostrando aquella regiones donde podemos establecer casas habitación, nos muestra otra vez con una gran distribución en la región de Dakar y Pikine. Sin embargo no podemos apreciar la densidad de la población. Esto debido a la falta de resolución en los CDRs en espacio tiempo. Por lo que al obtener esta figura final los datos de usuarios con sitios frecuentes son muy pocos.
A pesar de este inconveniente si podemos distinguir entre zonas potencialmente residenciales e industriales. En la figura **4-10** se aprecian de un manera más clara la demanda de viajes en la región de Dakar. Podemos notar que existe una gran demanda de la zona conurbada hacia el departamento de Dakar, es decir gran parte de la actividad diurna se concentra en este departamento. Por lo que la mayoría de los traslados se hacen desde las afueras de la ciudad. Podemos comparar estos resultados con los presentados en

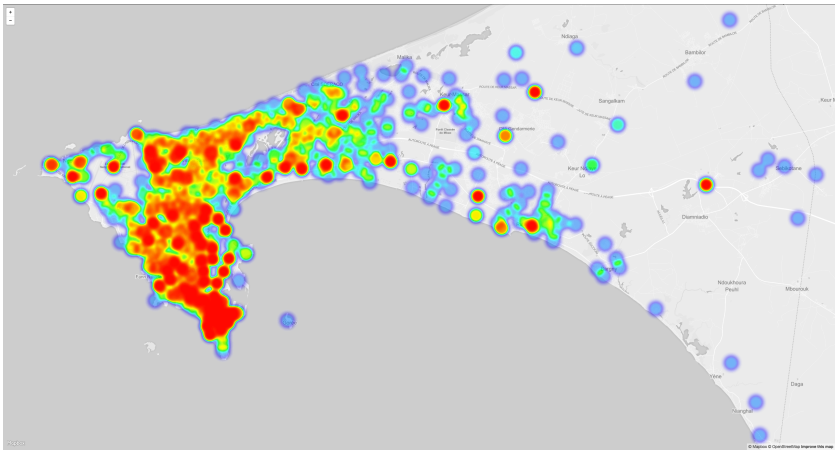

Figura 4-8.: Zonas industriales y comerciales

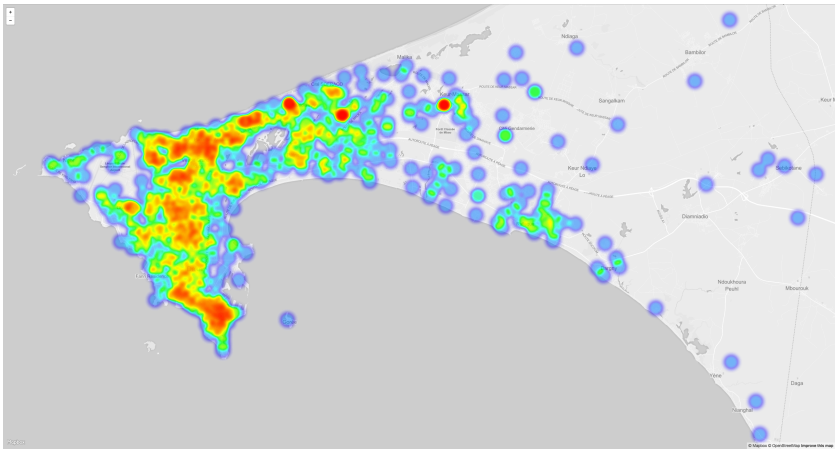

Figura 4-9.: Zonas residenciales

Gundlegard (2015) donde realiza una demanda de viajes a nivel estatal pero también se hace un análisis de Dakar. En los resultados que él presenta se aprecia un gran parecido a la figura aquí presentado. Con una gran actividad desde las afueras de la ciudad de Dakar, por lo que se puede notar un comportamiento similar con el resultado aquí presentado

4.2.1. Distribución de viajes

Una característica de importancia es determinar la dinámica de las trayectorias humanas que impactaría directamente al planeamiento urbano, prevención de epidemias, detección de eventos masivos imprevistos, es decir el entendimiento de la movilidad humana. Como se menciona en Gonzáles (2008) las trayectorias humanas tienen una distribución de probabilidad, independientemente de las trayectorias individuales, la cual sigue sim-

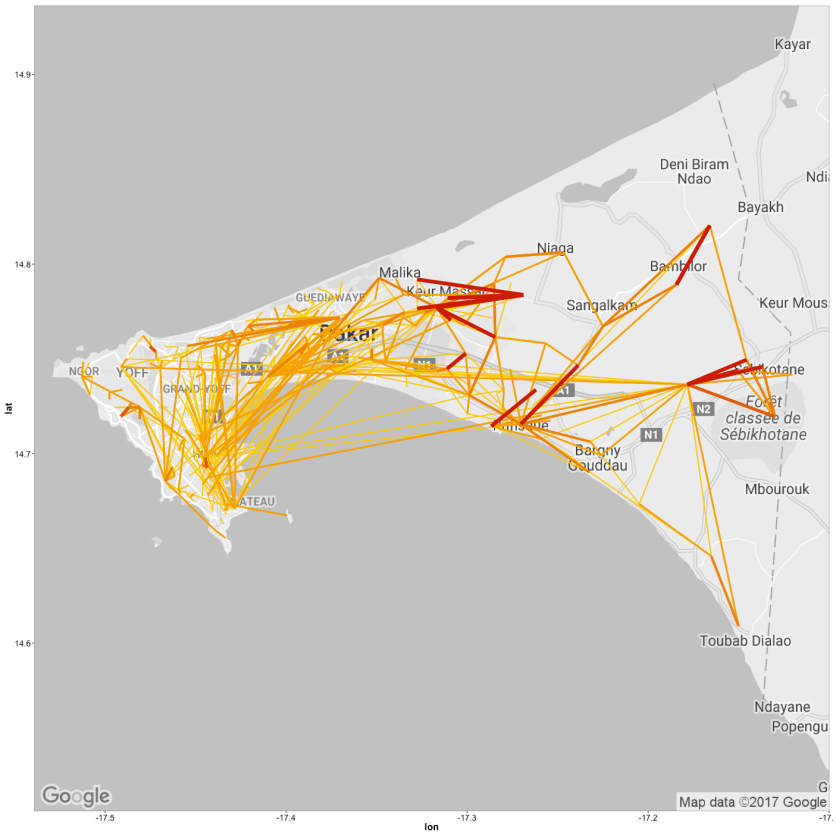

Figura 4-10.: Zonas industriales y comerciales

ples patrones reproducibles. En dicho trabajo la distribución de las trayectorias está bien aproximada por una ley de potencia (distribución de Lévy truncada).

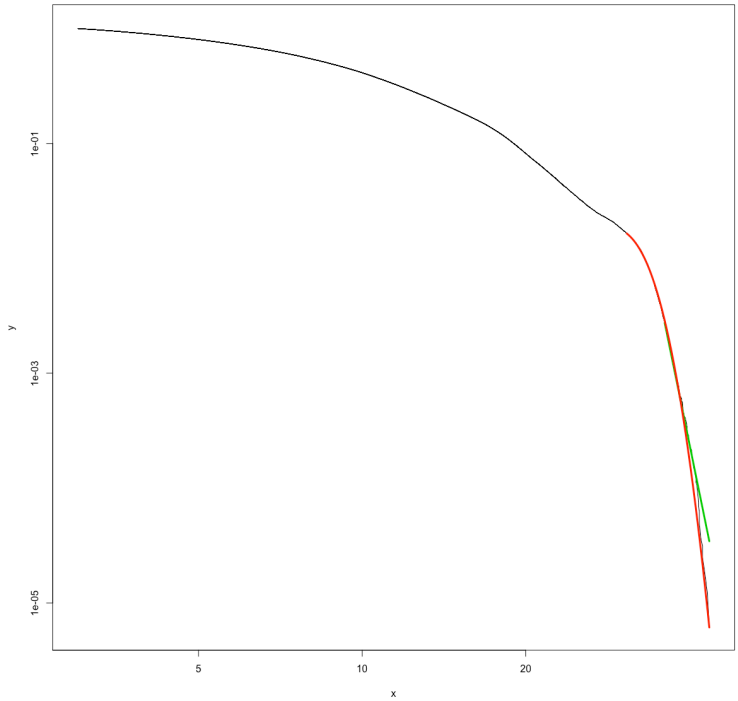

Figura 4-11.: Distribución de distancias de los viajes

Sin embargo en distintos trabajos como se demuestra en Alessandretti (2016) la distribución de trayectorias humanas están mejor descritas por dos distribuciones, dependiendo de la distancia de las trayectorias la distribución mejor aproximada para distancias cortas es log-normal mientras que para trayectorias lo suficientemente largas la distribución que mejor describe el comportamiento de movilidad es una ley de potencia. Además se muestra que la distribución log-normal es la que mejor describe trayectorias nuevas, es decir nuevas rutas que un humano pueda descubrir o tomar, por lo que es una característica del comportamiento de movilidad humano. Dado que este trabajo cuenta con distancias tanto cortas como muy largas podemos verificar los valores en ambas distribuciones.

En la figura **4-11** se muestra la distribución actual para los viajes, como se vuelve a mostrar una vez más aquí, se tiene una gran cantidad de viajes de gran distancia, que como se mencionó se debe a la poca granularidad en espacio y tiempo de los CDRs, por lo que como se propone en Alessandretti (2016) la distribución que mejor se ajusta a las trayectorias humanas es una distribución log-normal. En este trabajo se hizo un ajuste a los datos para verificar a cuál distribución se ajustaba de mejor manera.

Para determinar los distribuciones se tomaron muestras del conjunto de datos, cada muestra con cardinalidad n = 1000, el muestreo de los datos se realizó de manera aleatoria. Para ciertos intervalos de distancia tenemos distintas distribuciones que se ajustan a los datos, en un intervalo menor a 15 km se tiene una distribución log-normal con pará-

metros $\mu = 2{,}14$ y $\alpha = 0{,}58$. Para distancias menores a 22 km la distribución que mejor se adapta es una ley de potencia de parámetros $\alpha = 2{,}05$. Como se mencionó, no es tema principal de este proyecto el estudio de las distribuciones de las trayectorias humanas, sin embargo nos ayuda a sustentar nuestros resultados puesto que los valores son similares a los encontrados en (Gonzáles, 2008) o (Alessandretti, 2016) .

4.2.2. Análisis temporal

Para una análisis temporal se examinarán eventos extraordinarios, algo que los CDR nos permiten, como el día del trabajo o el día de la independencia, que son días feriados en Senegal, por lo que se debería notar alguna diferencia respecto a los viajes. Además se analizó días que muestran actividad diferente a la habitual, es decir se analizaron los fines de semana contra la semana laboral habitual(lunes-viernes). Como se puede observar en **4-12** la actividad semanal y de fin de semana sigue una comportamiento similar pero en cantidad de viajes mucho menor. Esto muestra lo que es común en un fin de semana pues la actividad decrece de manera considerable.

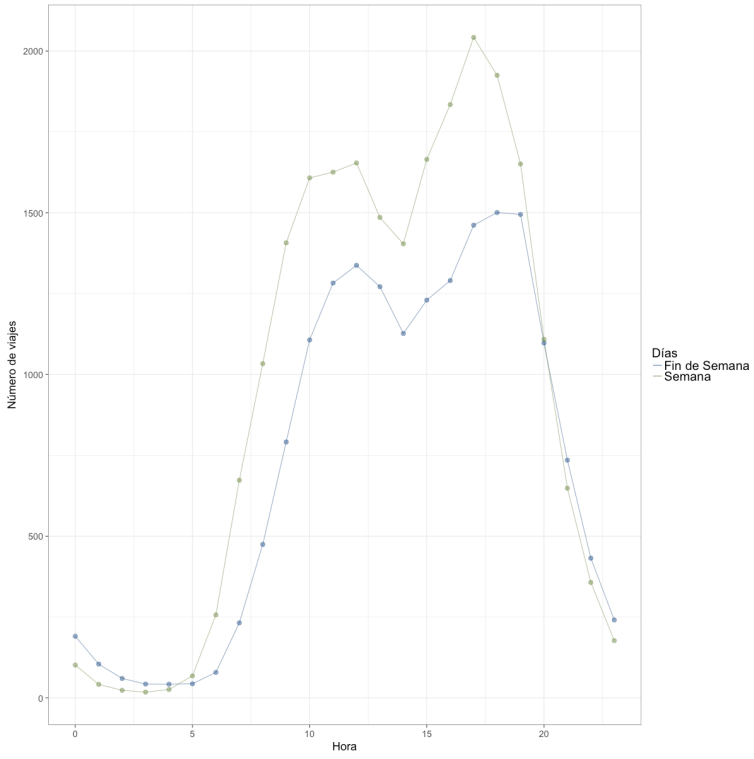

Figura 4-12.: Actividad semanal y fin de semana

Como el comportamiento es similar en la actividad en horas, se analizó a nivel más detallado para poder determinar las variaciones en las actividades por hora. En la figura **4-13** se puede determinar que la actividad del domingo difiere completamente del sábado o lunes, aunque el sábado parece tener una actividad parecida al lunes, se puede detectar claras variaciones respecto a la cantidad de viajes y a las horas en que estos se realizan. En el caso del día de la independencia celebrado el 4 abril, día Jueves en el año 2013, podemos notar una clara afluencia por la mañana, comparable con cualquier otro día de jueves, tomado aleatoriamente, sin embargo por la tarde, después de medio día, la actividad disminuye de manera significativa, que no se asemeja a cualquier otro jueves, esto puede indicar que ese día las actividades laborales fueron casi nulas, por ser un día feriado (Figura **4-14**).

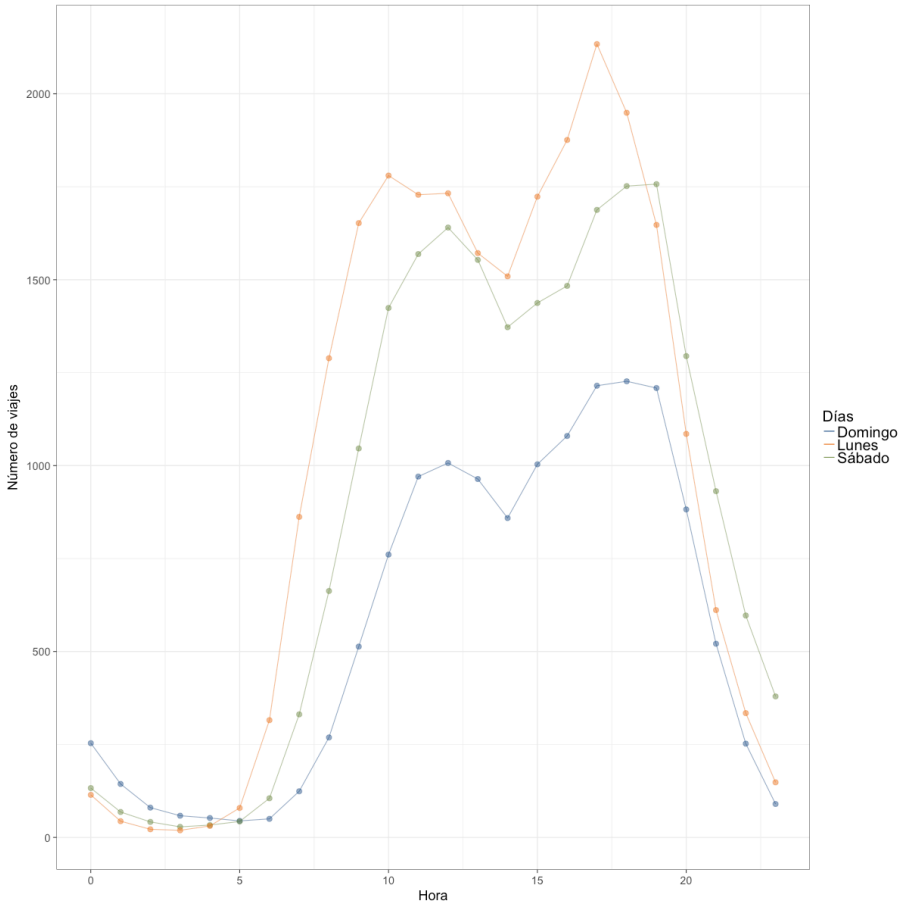

Figura 4-13.: Actividad en fin de semana y lunes

Respecto al día del trabajo, se tiene una actividad similar aunque por la mañana se registra un pico no visto en otros días Miércoles, esto pudiera indicar dos escenarios, que el conjunto de datos analizado resultó mayor, sin embargo no se registra un pico tan alto por la tarde, por lo que podemos descartar esta opción. El segundo escenario pudiera parecer que las actividades del día del trabajo se centran todas en la mañana, es decir la mayoría trabajó media jornada esto pudo concentrar toda la actividad diurna durante gran parte de la mañana y por la tarde como se puede observar, la actividad decae después de las 3pm (Figura **4-15**).

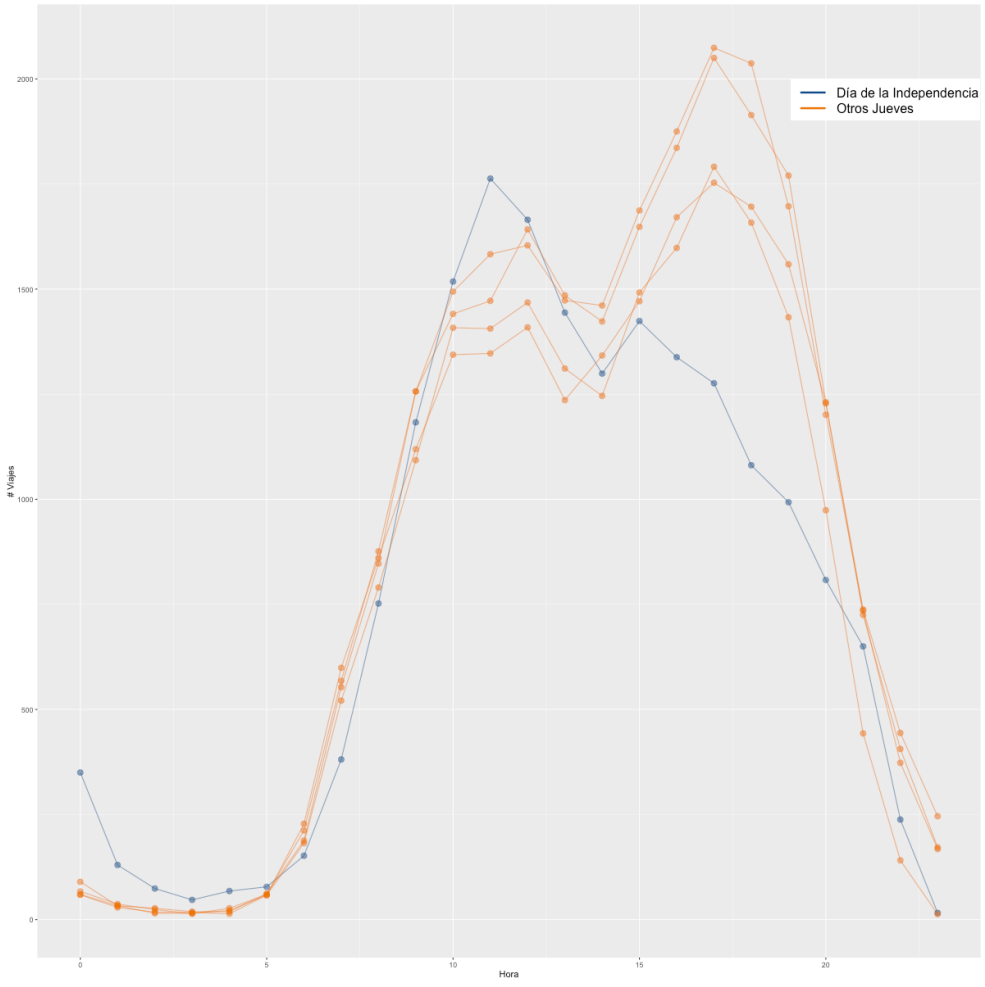

Figura 4-14.: Actividad en día de la independencia de Senegal

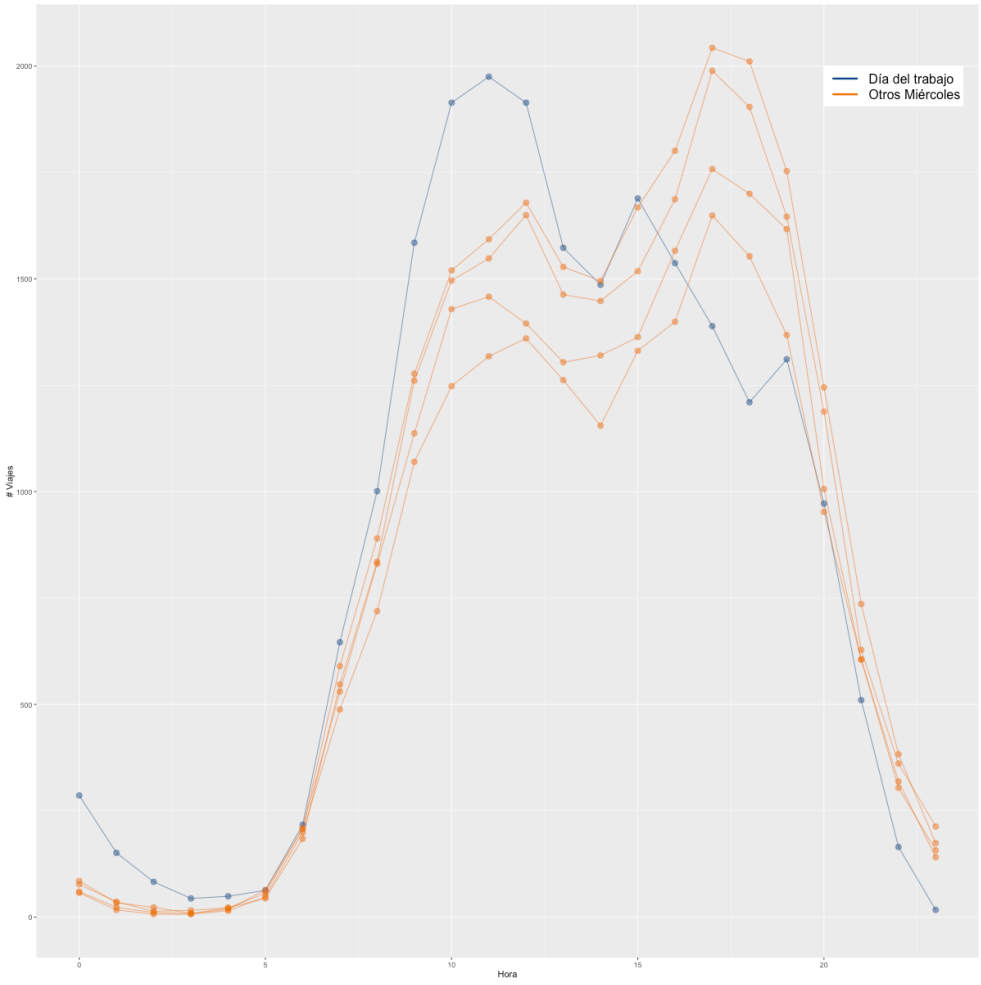

Figura 4-15.: Actividad en día del trabajo

4.2.3. Plataforma de visualización

La plataforma de visualización muestra el número de viajes, demanda entre dos puntos de cualquier región de Dakar, esto ayuda a determinar en qué zonas de la ciudad se realizan la mayor cantidad de desplazamientos humanos y con eso poder analizar si existen los medios para que se puedan realizar estos desplazamientos o poder distribuir la carga de los desplazamientos.
Como se muestra en la figura **4-16** la plataforma en un inicio muestra todas las antenas, se puede elegir alguna para obtener más información acerca de ese punto geográfico.

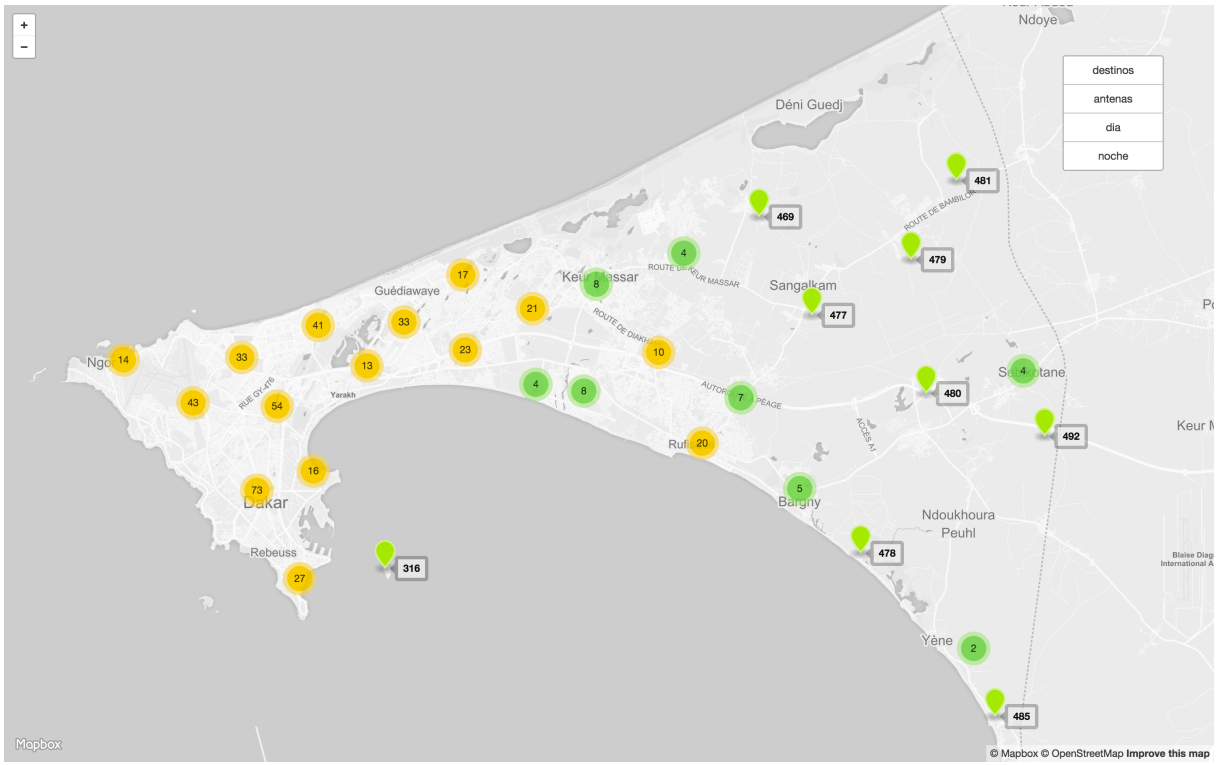

Figura 4-16.: Plaforma web interactiva

Después de seleccionar la antena correspondiente, se muestra el conjunto de destinos a los que está relacionado. En una escala de colores de azul a rojo (menor a mayor intensidad de viajes) para distinguir los puntos con más viajes entre sí, como se muestra en la figura **4-17**.

Al seleccionar un destino particular se podrá observar la cantidad de viajes que existe entre los dos puntos seleccionados, figura **4-18**. Con flechas que indican la dirección de los viajes y la cantidad de viajes realizados.
La plataforma muestra todos los datos de la matriz temporal, es decir todos los viajes registrados por todos los usuarios, por lo que podemos tener una radiografía de la movilidad en la ciudad.
Se puede filtrar por rangos de tiempo: día, noche o total de los viajes. Con esto poder observar las diferencias entre los viajes diurnos y nocturnos.

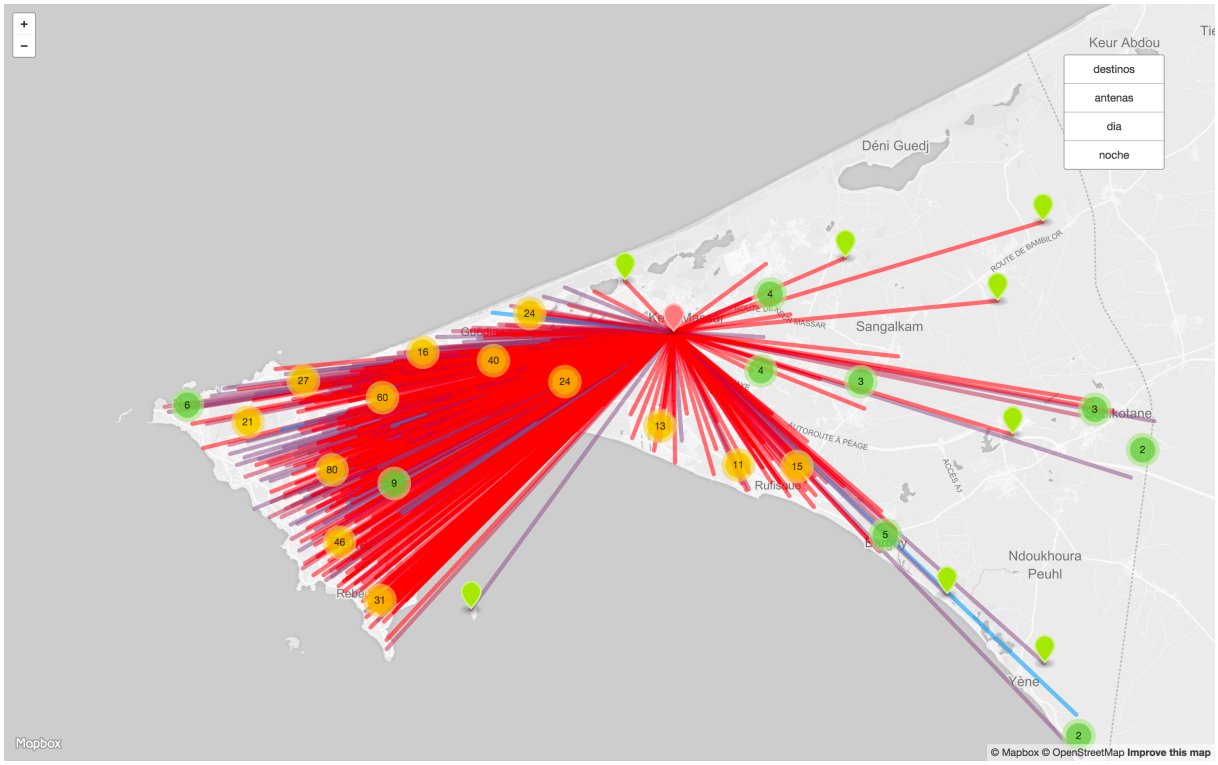

Figura 4-17.: Plataforma web interactiva

Figura 4-18.: Plataforma web interactiva

5. Conclusiones

5.1. Conclusión

Se ha podido dar una alternativa para el análisis de movilidad humana, en particular el estudio y extracción de matrices origen-destino. Esta alternativa de extraer con mayor facilidad, relativa a la manera usual de entrevistar o hacer conteo de tráfico, resulta más económica y representa un muestreo más grande de la que una entrevista puede aportar, puesto que los CDRs se almacenan de manera periódica con fines de facturación. Además nos permite tener información actualizada y analizar en un rango de tiempo más corto que las entrevistas. En un caso muy favorable, se podría analizar los datos en tiempo real, esto daría la posibilidad de poder predecir acontecimientos irregulares, poder manipular el flujo del tráfico, transporte, prevenir la propagación de enfermedades, etc.

Una de las grandes limitantes de los CDRs, dependiendo de la operadora telefónica, es la granularidad de los datos en espacio y tiempo. Agregado a este problema de granularidad está el hecho de que no se puede tener un análisis completo utilizando exclusivamente estos datos, para este trabajo en particular, debido a las pocas fuentes de información no se puede comparar o complementar con encuestas o conteos de tráfico ya existentes. De haber sido así se pudo complementar para poder obtener una mejor aproximación de la movilidad como bien se plantea Iqbal (2014). Donde utiliza una combinación de CDRs con un conteo de tráfico en las principales vías de comunicación para después poder hacer una simulación a mayor escala y poder utilizar un factor de escalamiento, que posteriormente le ayuda a construir la matriz de origen-destino final.

Debido a esta falta de datos extra en este presente trabajo, se tuvieron que hacer un número mayor de consideraciones así como un análisis más grande respecto al conjunto de datos por analizar. Una de la grandes consideraciones particulares es la especificación técnica del rango de las antenas, en muchos trabajos anteriores este valor es omitido, pero juega un papel importante pues el rango de las antenas varía en función de las zonas donde la señal es transmitida además con esto se puede evitar el problema de ubicar a un usuario al mismo tiempo en distintas localidades, es decir un falso desplazamiento. Con este parámetro podemos controlar la mínima distancia para considerar un verdadero desplazamiento dependiendo de la zona donde se realice el estudio, en una zona urbana o rural.

Se pudo determinar que a pesar de la ausencia de datos adicionales, para poder extrapolar, los CDRs en su sola naturaleza nos pueden ayudar a determinar zonas potencialmente residenciales o zonas potencialmente laborales. Con lo que se puede determinar una matriz de origen-destino y distinguir aquellas regiones en las que se tiene una demanda de movilidad importante. Con esto poder aportar una solución a la demanda entre esos dos puntos de conflicto.

Cabe mencionar que respecto a las distribuciones de distancia en comparación con los demás estudios como Gonzáles (2008) o Alessandretti (2016) los parámetros tienden a variar en gran medida pues los datos analizados en este conjunto cuentan con una gran diferencia en espacio y tiempo. Además el parámetro mínimo de desplazamiento es de 3 kilómetros para evitar falsos desplazamientos por lo que los viajes finales tienden a ser más largos. Esto provoca que los factores en las distribuciones log-normal y potencial difieran en cierta medida con lo parámetros encontrados en Gonzáles (2008).

Otro gran impedimento es que el conjunto de datos se encuentra parcialmente completo, por lo que la actividad de un usuario podría estar interrumpida lo que deriva en un análisis incompleto. Además los datos, en algunos subconjuntos de datos mensuales, cuentan con un mayor número de registros repetidos, debido a los falsos desplazamientos o errores en la recolección de los datos. Esto ocasionó que se detectarán más viajes en unas semanas y otras tuvieran una representación significativamente menor.

La plataforma de visualización ayudó a tener un mejor panorama de la actividad en la ciudad de Dakar, con esto podemos tener a mayor detalle los puntos de desplazamiento, el nivel y dirección del flujo de los viajes. Podemos determinar áreas de mayor intensidad respecto a la actividad, poder generar un análisis más detallado. Además es una manera visual e interactiva de poder interpretar los resultados, que ayudan a entender el contexto y resultado final de este trabajo.

5.2. Trabajo futuro

El algoritmo **1** fue desarrollado de tal manera que funcione con cualquier conjunto que se componga de puntos geográficos, puntos con latitud y longitud, por lo que se puede reutilizar o adaptar para poder analizar cualquier conjunto de datos que cumpla estas condiciones, un caso particular son los CDRs, pues cuentan con la localización de un usuario con puntos granulados en espacio y tiempo. Uno de los trabajos propuestos, es poder utilizar el Algoritmo **1** para datos similares en la CDMX y poder comparar con la encuesta origen-destino 2017.

A. Anexo:

```
"""
 Algoritmo para determinar viajes por usuario, multiproceso.
"""
import psycopg2
import psycopg2.pool
import geopy
import pdb
import multiprocessing
import math
from geopy.distance import vincenty

conn_string = "host='localhost' dbname='senegal' user='chris' password=''"
conn = psycopg2.connect(conn_string)
cur = conn.cursor()


class multithre(multiprocessing.Process):
    """
        crea multiples procesos para analizar en paralelo
    """
    def __init__(self, target, args):
        self.target = target
        self.args = args
        multiprocessing.Process.__init__(self)

    def run(self):
        self.target(*self.args)

def idslist(n, month):
    """
        Funcion para dividir la cantidad de usuarios en vectores de magnitud homogenea
        :param n: Numero de procesos en los que se dividiran la carga de los usuarios a analizar
        :param month: String mes del cual se extraera la información
        :return: listas de tamaño n con los usuarios
    """
    cur.execute('SELECT DISTINCT(uid) from {};'.format(month))
    ids = cur.fetchall()
    print ids, "\n"
    nu = len(ids)
    div = int(math.ceil(nu/n))
```

```
41      tids = []
42      val = nu % n
43      for x in range(0, n):
44          if val != 0 and x == (n-1):
45              tids.append(ids[x*div:])
46          else:
47              tids.append(ids[x*div:(x+1)*div])
48      if val != 0:  # Si la lista esta dividida en partes desiugales
49          aux = tids[n-1][:div]  # Valores normalizados ultimo espacio
50          aux2 = tids[n-1][div:]  # Sobrantes
51          tids[n-1] = aux
52          for x in range(0, val):
53              tids[x].append(aux2[x])
54      return tids
55
56  def createconnection():
57      """
58          crea conexion con la base de datos
59          :return: conexion a base de datos
60      """
61      conn_string1 = "host='localhost' dbname='senegal' user='chris' password=''"
62      conna = psycopg2.connect(conn_string1)
63      return conna
64
65  def closeconnection(connection, cursor):
66      """
67          Cierra conexion
68      """
69      cursor.close()
70      connection.close()
71
72  def insertondb(connection, cursor, site1, site2):
73      """
74          Funcion para inserta los viajes determinados en la tabla matrizodaux(matriz transitoria)
75          :param connection: conexion de a la base de datos
76          :param cursor: cursor de la base de datos
77          :param site1: array con los datos de sitio de origen
78          :param site2: array con los datos de sitio de destino
79      """
80      cursor.execute("INSERT INTO matrizodaux(uid,forigen, oid, olat, olon, fdestino, did, dlat, dlon) VAI
81                     (site1[0], site1[1], site1[2], site1[3], site1[4],
82                      site2[1], site2[2], site2[3], site2[4]))
83      connection.commit()
84
85  def trip(cursor, i, month):
86      """
87          Funcion regresa CDRs de usuario i
88          :param cursor: cursor de la base de datos
```

```
89          :param i: id del usuario a analizar
90          :param month: mes a analizar
91          :return: todos los CDR del usuario i
92      """
93      cursor.execute('SELECT uid,fechahora,siteid,lat,lon FROM {} WHERE uid = {} '.format(month, i))
94      totaltrip = cursor.fetchall()
95      return totaltrip
96
97  def tripi(connection, cursor, listids, month):
98      """
99          Funcion itera sobre todos los usuarios de listids, con su analisis respectivo
100         :param connection: conexion a la base de datos
101         :param cursor: cursor a la base de datos
102         :param listids: array con ids de usuarios
103         :param month: string mes a analizar
104     """
105     for i in listids:
106         totaltrip = trip(cursor, i, month)
107         analyzetrip(connection, cursor, totaltrip)
108
109 def setdest(connection, cursor, oraux, site):
110     """
111         Crea viaje y envia datos a la base
112         :param connection: conexion a la base de datos
113         :param cursor: cursor a la base de datos
114         :param oaraux: array origen del viaje
115         :param site: array destino del viaje
116         :return: boolean estado del nuevo origen
117     """
118     odaux = [oraux.pop(), site]
119     insertondb(connection, cursor, odaux[0], odaux[1])
120     orig = True
121     return orig
122
123
124 def time_in_place(connection, cursor, oraux, site1, site2, est):
125     """
126         Funcion analiza si usuario permanecio estatio en una delta de tiempo
127         :param connection: conexion a la base de datos
128         :param cursor: cursor a la base de datos
129         :param oaraux: array con origen del viaje actual
130         :param site1: array con el punto transitorio
131         :param site2: array con el punto destino potencial
132         :param est: boolean si se mantuvo estatico, para analisis si se encuentra lejano del orige
133         :return: boolean si se agrego o desecho el destino potencial
134     """
135     deltat = 1200  # segundos 20min
136     deltau = 10800  # un viaje no puede durar mas de un 3 horas
```

```
137     timedelta = site2[1] - site1[1]
138     deltai = timedelta.total_seconds()
139     timedelta = site1[1] - oraux[0][1]
140     deltaj = timedelta.total_seconds()
141 
142     if deltai >= deltat and deltaj < deltau:
143         orig = setdest(connection, cursor, oraux, site1)
144     elif est and deltaj < deltau:
145         orig = setdest(connection, cursor, oraux, site1)
146     elif deltaj > deltau:  # Nuevo origen pues rebaso el delta de 3 horas
147         orig = True
148     else:  # No se completo viaje o sigue viaje
149         orig = False
150     return orig
151 
152 def analyzetrip(connection, cursor, trip):
153     """
154         Funcion analiza CDRs del usuario para extraer los viajes
155         :param connection: conexion a la base de datos
156         :param cursor: cursor a la base de datos
157         :param trip: array con todos los CDRs del usuario
158     """
159     oraux = []
160     rdistance = 3000
161     nran = len(trip)
162     orig = True
163 
164     for i in range(nran-1):
165         if orig:  # origen con el primer elemento del array o si se determinar un nuevo origen
166             oraux.append(trip[i])
167             orig = False
168         if trip[i][2] != trip[i + 1][2]:  # Verifica si se movio o permanecio mismo lugar
169             site1 = trip[i][3:]  # Orden de la tabla id/fechahora/siteid/lat/ lon
170             site2 = trip[i+1][3:]
171             deltadistance = vincenty(site1, site2).meters  # Mide la distancia que debera ser mayor a 3(
172             if deltadistance > rdistance:  # El usuario se movio siguiente punto analizar
173                 pass
174             else:  # El usuario se quedo estatico
175                 if vincenty(oraux[0][3:], site1).meters > rdistance:  # verifica si la distancia del ori
176                     estatico = False
177                     orig = time_in_place(connection, cursor, oraux, trip[i], trip[i + 1], estatico)  # I
178                     if orig: # Si origen nuevo se vacia el origen actual
179                         oraux = []
180                 else:
181                     pass
182         else: # verifica si del origen al destino se movio una distancia verdadera
183             if vincenty(oraux[0][3:], trip[i][3:]).meters > rdistance:
184             ^^I# verifica si la distancia del origen al punto actual es mayor a 3km
```

```
185                 if i < (nran-2):
186                     estatico = trip[i][2] == trip[i + 1][2] and trip[i + 1][2] == trip[i + 2][2]
187                 else:
188                     estatico = False
189                 orig = time_in_place(connection, cursor, oraux, trip[i], trip[i + 1], estatico) #
190                 if orig:  # Si origen nuevo se vacia el origen actual
191                     oraux = []  # agregado por threading
192             else:  # Permanecio estatio
193                 orig = True
194                 oraux = []
195     oraux = []
196
197
198 def genpooldb(poolc):
199     """
200         Crea un pool de conexiones para los multiples procesos
201         :param poolc: ThreadedConnectionPool para generar conexiones
202         :return: pool de conexiones
203     """
204     connecn = poolc.getconn()
205     connecn.set_isolation_level(0)
206     return connecn
207
208
209 def detstate(theadlist):
210     """
211         Determinar el estado de cada proceso, si es que finalizo
212         :param theadlist: array de procesos
213         :return: boolean si todos los procesos ya terminaron
214     """
215     aux = True
216     for x in theadlist:
217         aux = x.is_alive() or aux
218     return aux
219
220
221 def main():
222     """
223         Funcion main que genera los procesos y
224         la cantidad de segmentaciones a crear
225     """
226     connection_list = []
227     processlist = []
228     n = multiprocessing.cpu_count()  # Numero de cores a utilizar dadas las capacidades del CPU
229     month = 'november'  # Mes a analizar
230     tids = idslist(n, month)
231     poolc = psycopg2.pool.ThreadedConnectionPool(1, n+1, host='localhost',
232                                                  dbname='senegal', user='chris',
```

```
233                                                 password='')
234
235     for x in range(0, n):
236         connection_list.append(genpooldb(poolc))
237     for j in range(0, n):
238         processlist.append(multithre(target=tripi, args=(connection_list[j],
239                            connection_list[j].cursor(), tids[j]), month))
240
241     for i in range(0, n):  # Inicializa los procesos
242         processlist[i].start()
243
244     for h in xrange(0, n):  # Espera a todos los procesos a terminar
245         processlist[h].join()
246
247     if not(detstate(processlist)):  # Cierra conexiones
248         poolc.closeall()
249         closeconnection(conn, cur)
250
251 if __name__ == "__main__":
252     main()
```

Bibliografía

Agence Nationale de la Statistique et de la Démographie, ANSD. 2013. *Situation Economique et Sociale du Senegal 2013*.

Alessandretti, L, Sapiezynski P. Lehmann S. Baronchelli A. 2016. Multi-scale spatio-temporal analysis of human mobility.

Autorité de Régulation des Télécommunications et des Postes, ARTP. 2013. *Observatoire de la Téléphonie Mobile 2013*.

Borderon, M., Oleveau S. Machault V. 2010. Qualifier les espaces urbains à Dakar. *Cybergeo: European Journal of Geography.*

Calabrese, F., Lorenzo G. Ratti C. Liang L. 2010. Estimating Origin-Destination Flows Using Mobile Location Data. *Cell*, **10**, 36–44.

Conseil Exécutif des Transports Urbains de Dakar, CETUD. 2015. *Enquête Ménage sur la Mobilité, le Transport et L'accès aux Services Urbains dans L'agglomération de Dakar 2015*.

Dong, Yuxiao, Pinelli Fabio Gkoufas Yiannis et al. 2015. Inferring Unusual Crowd Events From Mobile Phone Call Detail Records.

Gonzáles, M., Hidalgo C. Barabási A.. 2008. Understanding individual human mobility patterns. *Nature*, **453**, 779–782.

Gundlegard, David, Rydergren Clas Barcelo Jaume. 2015. Travel demand analysis with differentially private releases. *Data for Development Challenge Senegal: Book of Abstracts*, 413–426.

Holma, Harri, Toskala Antti. 2009. *LTE for UMTS- OFDMA and SC-FDMA Based Radio Access*. Chippenham, United Kingdom: John Wiley & Sons, Inc.

Horak, Ray. 2007. *Telecommunications and data communications handbook*. New Jersey, USA.: John Wiley & Sons, Inc.

Instituto Nacional de Estadística y Geografía, INEGI. 2007. *Encuesta Origen-Destino 2007*.

Instituto Nacional de Estadística y Geografía, INEGI. 2011. *Perspectiva estadística México*.

International Telecommunication Union, ITC. 2015. *International Telecommunication Union: Facts and Figures 2015*.

Iqbal, Md Shahadat, Choudhury Charisma F. Wang Pu González Marta C. 2014. Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, **40**, 63–74.

Pinelli, F., Lorenzo G., & Calabrese, F. 2015. Evaluating urban sensing applications using actively and passively-generated mobile phone location data. *NetMob 2015: Book of Abstracs*, 6–8.

Ratti, Carlo, Sevtsuk Andres Huang Sonya Pailer Rudolf. 2007. Mobile Landscapes: Graz in Real Time. *Location Based Services and TeleCartography*, 433–444.

Service Régional de la Statistique et de la Démographie de Dakar, SRSDD. 2013. *Situation Economique et Social Régionale 2013*.

White, Joanna, Wells I. 2002. Extracting Origin Destination Information from Mobile Phone Data. *Road Transportation and Control*, 30–34.