Optimización de matrices origen-destino estimadas a partir de datos georeferenciados en redes sociales

Alejandro J. Perez[1], Leonardo D. Dominguez[1],
Aldo J. Rubiales[2], Pablo A. Lotito[2].

[1] PLADEMA, Universidad Nacional del Centro de la Provincia de Bueno Aires,
[2] CONICET

Resumen A la hora de aplicar cualquier política urbana de transporte es muy importante conocer el patrón de movilidad de la población y por lo tanto las matrices Origen Destino que la describen. Los grandes volúmenes de datos arrojados por las nuevas tecnologías permiten obtener información dinámica del comportamiento de sus usuarios. En contraposición con los métodos clásicos de obtención de las matrices O-D el uso de las redes sociales aporta una manera más económica de realizar el estudio y un mayor dinamismo. En este trabajo se desarrolla un método para la actualización de matrices Origen-Destino correspondientes a tráfico vehicular, a partir de la información obtenida de datos disponibles en la red social Twitter. Para actualizar una matriz OD anterior con nueva información se consideró un problema de maximización de entropía restringiendo la distancia a la matriz original. El enfoque presentado se aplicó a analizar la movilidad diaria de las personas de CABA obteniendo las matrices O-D que la caracterizan y los resultados obtenidos se compararon con estudios previos que se realizaron utilizando otras metodologías comprobando la viabilidad del nuevo enfoque propuesto.

Keywords: Distribución de viajes, Maximización de la entropía, Matrices Origen Destino, Redes Sociales

1. Introducción

El tratamiento de grandes volúmenes de datos arrojados por las nuevas tecnologías, le permite a quienes planifican las ciudades, saber cuáles son los comportamientos que tienen a diario sus habitantes. De esta manera, se pueden tomar decisiones apropiadas para generar nuevas alternativas de transporte o mejorar el flujo a través de las arterias. La suma de los desplazamientos individuales es lo que C. Miralles Guasch[1] denomina "Movilidad Cotidiana" y debe ser considerada como el punto de partida de cualquier política urbana de transporte. A la hora de planificar un cambio, la estimación de las matrices Origen-Destino (O-D) es de suma importancia para conocer cómo se generan y distribuyen los viajes de la población.

Hasta no hace mucho tiempo, contar con estos volúmenes de datos masivos no era una tarea sencilla, y hasta incluso hoy en día en Argentina, se siguen generando las Encuestas Origen Destino (EOD) y las Encuestas de Movilidad Domiciliaria (EMD) a cargo del Proyecto de Transporte Urbano para Áreas Metropolitanas (PTUMA) para recabar este tipo de información. Si bien estas tienen algún tipo de ventaja sobre los medios digitales por estar específicamente diseñadas para tal fin, se ven muy perjudicadas por el tiempo y la logística que requieren llevarlas a cabo, y por el escaso dinamismo de los datos. La encuesta ENMODO[2], parte del proyecto PTUMA llevado a cabo a fines del año 2009, permitió analizar la movilidad del Área Metropolitana de Buenos Aires (AMBA), encuestando 22.170 hogares, con un total de 70.321 habitantes. Como desventajas de esta metodología se deben mencionar los costos que involucra una encuesta de este tipo, que no solo requiere entre otras cosas capacitar al personal y pagar su movilidad, sino también realizar campañas de difusión para concientizar a la población.

En los últimos años, los teléfonos móviles se han convertido en una tecnología que acompaña a sus usuarios en casi todo momento. La movilidad de estos dispositivos ha transformado a los teléfonos celulares en uno de los principales sensores de la conducta humana. De hecho, cada vez que un suscriptor hace o recibe una llamada telefónica o un SMS, información sobre la interacción, así como la geolocalización del usuario se registra a efectos de facturación. En [15] se utilizan técnicas de optimización en combinación con reglas de asociación temporal sobre los datos obtenidos para la estimación de matrices O-D. A su vez, en [16] el mismo objetivo fue logrado utilizando técnicas de filtrado espacio-temporal. Un análisis similar del comportamiento de la población pero aplicado a una región de nuestro pais puede observarse en el trabajo de S. Anapolsky[3]. En el mismo se cuenta con un lapso de 5 meses consecutivos de llamadas anonimizadas de una empresa de telefonía celular, sumando casi un total de 50 millones de llamados. Luego de filtrar aquellas muestras que no aportan información suficiente, trabajan con un subconjunto de 2,1 millones de llamados, lo que les dá un total de 75 mil usuarios de la Ciudad Autónoma de Buenos Aires (CABA). Este trabajo permitió obtener resultados parciales comparables con los de ENMODO teniendo en cuenta que del área analizada por PTUMA, CABA representa un 23 % con un total aprox. de 3 millones de habitantes, lo que arroja una relacion de 1 usuario por cada 40 habitantes. También, se logró reducir drásticamente el costo de realización del análisis aunque se debió contar con los datos provistos por la empresa, los cuales generalmente no son de libre acceso.

La creciente popularidad de las nuevas herramientas de comunicación como lo son Facebook y Twitter, exponen nuevos recursos de acceso gratuito, los cuales sirven como fuente de datos para analizar el comportamiento espacio-temporal de las personas, y de esa manera contribuir en futuros proyectos urbanos. Además, la ventaja de las redes sociales es que los mensajes son generados por los propios usuarios en forma espontánea y voluntaria, aportando en este aspecto una calidad de datos superior al de las encuestas previamente mencionadas, en donde los voluntarios son interrogados y pueden surgir problemas como sesgos del mues-

treo, debido a malas interpretaciones, falta de memoria, respuestas socialmente mal o bien vistas, defectos en el cuestionario o en el encuestador (raza, clase social, personalidad), sustitución de individuos, etc.[4]

El objetivo de este trabajo es analizar la movilidad diaria de las personas de CABA obteniendo las matrices O-D que la caracterizan. El aporte novedoso del mismo reside en la utilización de la red social Twitter como fuente de información pública.

2. Redes sociales como fuente de datos

Dentro de las redes sociales más populares que presentan información georeferenciada de sus usuarios se puede mencionar a Facebook y Twitter. Para comenzar el análisis, primero se evaluó la posibilidad de utilizar Facebook por ser la más popular, sin embargo, requiere un vínculo directo con cada usuario que se quiera analizar, haciendo inviable la obtención de datos masivos. Por este motivo, se decidió utilizar Twitter (una plataforma online que permite publicar mensajes cortos de 140 caracteres llamados 'tweets', que pueden ser georeferenciados, siempre y cuando el usuario disponga de un dispositivo movil con GPS integrado, y tenga la voluntad propia de indicar en dónde se encuentra actualmente) que de America Latina, es en Argentina en donde se espera que experimente la mayor tasa de crecimiento hasta el 2018[5]. Teniendo en cuenta que Twitter solo permite obtener los mensajes de los últimos 7 días, fue necesario construir una herramienta capaz de almacenar los "tweets" georeferenciados día tras día, utilizando para esto, la biblioteca Twitter4J[6], apta para Java. Esta herramienta, permite obtener los tweets con un radio y un centro dado, que en el caso de estudio (CABA) fue de 11 km a partir de las coordenadas -34.6120185, -58.4338760. Cabe destacar también, que los datos almacenados son anonimizados desde el primer momento que se obtienen, asegurando así la privacidad de cada usuario.

3. Metodología

A la hora de decidir la planificación del transporte urbano, estimar las matrices O-D resulta de suma importancia para conocer cómo se generan y distribuyen los viajes de la población. La metodología que se aplicó en este trabajo, se basó en la utilizada por S. Anapolsky, en donde se implementa un modelo basado en viajes, con el objetivo de comparar ambos resultados y validar el uso de las redes sociales como fuente de información. Además, en nuestra propuesta se aplica una optimización de la matriz resultante, mediante la **Maximización de la Entropía Doblemente Acotada**, que permite determinar la distribución más probable dada alguna información conocida del entorno[8]. Otros enfoques posibles son los presentados en **Métodos de Inferencia Estadística**, que optimizan a partir de una función de probabilidad, en donde las diferencias entre la matriz OD objetivo y la matriz de partida es sólo el resultado de la varianza[9][10][11], y la **Aproximación del Gradiente**, en donde la

matriz OD de partida es modificada en cada iteración en la 'dirección' basada en el gradiente de la función objetivo[12].

3.1. Filtrado

Como paso previo a la aplicación de la metodología, se realiza un preprocesamiento de los datos obtenidos agrupando los tweets en comunas. Para esto, se utilizó un algoritmo geométrico nativo de JAVA, que utiliza la regla par-impar[13], asociando cada uno de los tweets con una comuna de CABA descartando aquellos que se encuentran fuera del área de análisis.

A continuación, se procedió a descartar los tweets de los usuarios con información insuficiente. Aquí, se consideró como usuario no válido aquel que no cumple con la siguiente condición:

- El usuario es válido si tiene al menos una aparición al mediodia (de 11am a 3pm) y una aparición por la noche (de 8pm a 5am) de Lunes a Jueves para alguna de las semanas analizadas. Se considera como una aparición, que el usuario haya realizado al menos un tweet en una franja horaria, para un periodo semanal, un número de semana y número de comuna dado.

Luego de aplicar el algoritmo geométrico se paso de 80.130 usuarios en solo 36 días, a tener 57.135 que tuvieron al menos una aparición en alguna comuna de CABA (un 71,3 % de la muestra). Si de eso, se descartan aquellos no validos, tenemos 10.077 usuarios que tienen al menos una aparición al mediodia y una aparición a la noche (un 12,57 % de la muestra), porcentaje muy por encima al logrado por S.Anapolsky (3,5 %).

En la **Figura 1** se visualiza el área en donde se capturaron los tweets tal como se explicó en la sección anterior y con un contorno negro se ven los tweets filtrados por el algoritmo geométrico.

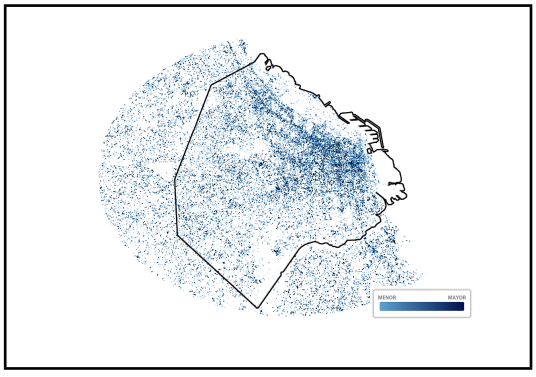

Figura 1: Mapa de densidad con tweets capturados en un radio de 11 km.

Presentamos también los mapas de calor que muestran la concentración de personas al mediodía **(Figura 2)** y la concentración por la noche **(Figura 3)**, introduciendo a simple vista, la variación de la concentración de personas en diferentes momentos del día, lo que motiva el estudio.

Figura 2: Tweets capturados al mediodía

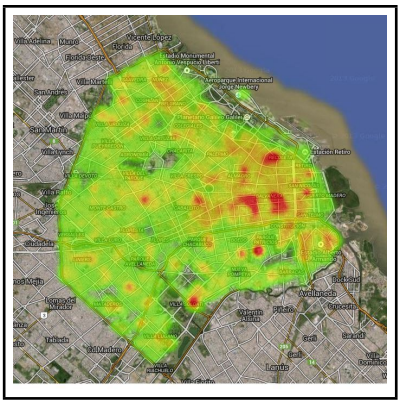

Figura 3: Tweets capturados a la noche

3.2. Metodología basada en el método de Anapolsky

Los modelos basados en viajes, analizan los trayectos de las personas para poder estimar los flujos futuros de pasajeros o vehículos, que habrá en una red de transporte. En el trabajo anteriormente citado, los autores decidieron separar su conjunto de datos en días y horas, basados en un análisis de la Encuesta de Movilidad Domiciliaria 2009[2]. Así, el mismo quedó clasificado en **cuatro períodos semanales**: días laborales clásicos (**de Lunes a Jueves**) y días con posibles variantes (**Viernes, Sábados y Domingos**). En cuanto a las **franjas horarias**, quedaron determinadas por: **Mañana** (de 5am a 11am), **Mediodía** (de 11am a 3pm), **Tarde** (de 3pm a 8pm) y **Noche** (de 8pm a 5am). Por una cuestión de poder comparar los resultados, en el presente trabajo se respetan las mismas divisiones. Posteriormente, se generó una tabla que presenta las comunas en las que apareció cada usuario en todo el intervalo analizado, discriminando de cada visita, semana, franja horaria y período semanal. Con esta tabla, se aplicó un procedimiento encargado de asignar a cada usuario una comuna considerada como su hogar, observando para tal fin, la franja nocturna e identificando la comuna en la que se encuentra el usuario un número mayor de veces. Una observación similar ocurre para determinar el lugar de trabajo, observando en ese caso la franja horaria del mediodía.

3.3. Primera matriz O-D

Una vez detectados los hogares y los lugares de trabajo de cada usuario, se armó una tabla de 15x15, en donde se observa en las columnas los origenes (donde viven) y en las filas los destinos (donde trabajan), completando cada celda bajo la siguiente pregunta: ¿Qué porcentaje de los usuarios que trabajan en la comuna Destino_1, viven en la Comuna_1?.

3.4. Optimización

La optimización que se aplico a la matriz se basó en el método propuesto por Ortúzar J. y Willumsen L[8]. Consiste en una maximización doblemente acotada de la entropia, que tiene por objetivo generar distribuciones de probabilidad con cierta información dada, más precisamente, entre todas las distribuciones de probabilidad que son factibles en un conjunto finito de estados bajo cierto tipo de información, se desea encontrar la que soporte mayor nivel de incertidumbre, que matemáticamente se describe como:

$$Max_{A_{ij}} \left[- \sum_{i=1}^{K} \sum_{j=1}^{L} A_{ij} \ln(A_{ij}) \right] \tag{1}$$

Sujeto a:

$$\sum_{i=1}^{K} A_{ij} = \sum_{i=1}^{K} S_{ij} = O_j \quad con\, j = 1, ..., L; \tag{2}$$

$$\sum_{j=1}^{L} A_{ij} = \sum_{j=1}^{L} S_{ij} = D_i \quad con\, i = 1, ..., K; \tag{3}$$

$$\left| A_{ij} - T_{ij} \right| \leq \varepsilon \tag{4}$$

$$siendo\ A_{ij} \geq 0,\ T_{ij} \geq 0\ y\ 1 \leq i \leq K, 1 \leq j \leq L$$

con K = L = cant. comunas

En donde las restricciones (2) y (3) establecen que O_j representa la cantidad de viajes generados por la comuna j, y D_i la cantidad de viajes atraídos por la comuna i, siendo ambos valores obtenidos en función de las encuestas realizadas por ENMODO [2]. Adicionalmente, proponemos una última restricción (4), siendo que T_{ij} es la matriz de distribución generada con los tweets georeferenciados

de los usuarios, A_{ij} es la matriz objetivo y S_{ij} es la matriz que se obtiene en la encuesta ENMODO [2], se busca minimizar la diferencia entre ambas matrices (A y T) con un error ε establecido.

Para resolver este problema de optimización utilizamos un algoritmo de Punto Interior provisto por la función $fmincon$ de Matlab con diferentes valores de ε. Las pruebas realizados y los valores de ε se detallan en la siguiente sección.

4. Resultados

Para resumir los resultados obtenidos, primero se presenta un gráfico que valida el uso de las redes sociales tal como se lo planteó en la sección 3. Se puede apreciar que los datos obtenidos de Twitter (derecha de la **Figura 4**), se asemejan a los obtenidos por S. Anapolsky[3] (izquierda de la **Figura 4**) con llamados telefónicos. Es importante tener en cuenta que la cantidad de días de nuestra muestra es de solo 36, mientras que S. Anapolsky cuenta con 5 meses consecutivos. Ambos gráficos aplican un factor de expansión que tiene en cuenta la población de cada comuna según el Censo 2010 [7].

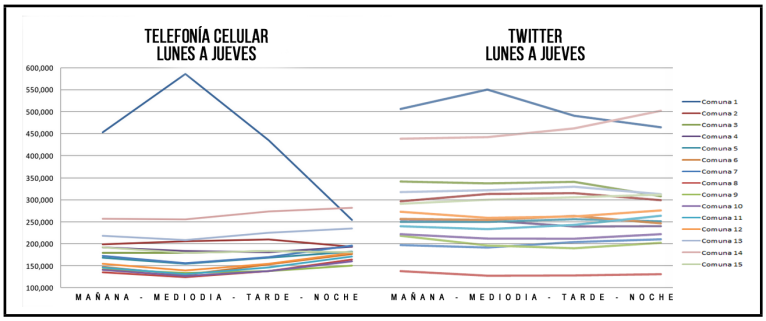

Figura 4: Población que utiliza el servicio en cada comuna.

Como siguiente paso, se muestra en la **Figura 5**, la matriz O-D obtenida por twitter para cada comuna. El orden de lectura es de Destino a Origen, por ejemplo: "el 8 % de las personas que trabajan en la comuna 1, viven en la comuna 2". Para llegar a estos resultados, fue necesario conocer primero los usuarios que se encuentran al mediodía en cada comuna, y de cada subconjunto, agruparlos según su hogar. En la diagonal, se encuentran los habitantes que trabajan en la misma comuna que viven.

En la **Figura 6** se muestran las matrices representadas con una escala de color que evidencia las zonas de mayor concentración.

		Origen														
	Comuna	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Destino	1	**40%**	8%	7%	7%	4%	4%	2%	1%	2%	2%	2%	3%	6%	9%	3%
	2	11%	**53%**	6%	2%	3%	1%	1%	1%	1%	1%	2%	2%	4%	12%	2%
	3	9%	5%	**45%**	5%	6%	4%	3%	2%	2%	3%	2%	3%	3%	5%	3%
	4	11%	2%	4%	**65%**	2%	2%	2%	3%	1%	1%	1%	1%	2%	2%	2%
	5	5%	2%	6%	4%	**56%**	6%	3%	1%	2%	1%	2%	2%	1%	4%	4%
	6	4%	1%	3%	3%	6%	**54%**	6%	1%	3%	3%	4%	2%	1%	2%	6%
	7	3%	1%	2%	4%	3%	6%	**56%**	5%	6%	6%	4%	1%	1%	1%	1%
	8	2%	0%	1%	4%	1%	1%	2%	**78%**	6%	1%	1%	0%	0%	1%	0%
	9	1%	1%	1%	1%	1%	1%	1%	7%	**74%**	8%	2%	1%	0%	0%	1%
	10	1%	1%	0%	1%	1%	1%	2%	2%	9%	**70%**	10%	1%	0%	1%	1%
	11	1%	1%	0%	1%	1%	2%	2%	0%	2%	8%	**70%**	4%	1%	1%	5%
	12	1%	1%	0%	1%	1%	1%	0%	0%	1%	2%	3%	**74%**	8%	2%	5%
	13	5%	3%	2%	0%	1%	2%	1%	0%	1%	1%	2%	10%	**56%**	10%	4%
	14	9%	7%	3%	1%	2%	2%	1%	1%	1%	1%	2%	2%	8%	**55%**	5%
	15	3%	2%	3%	1%	3%	4%	2%	1%	1%	2%	8%	6%	4%	5%	**56%**

Figura 5: Matriz Origen-Destino para cada comuna en la franja horaria mediodía

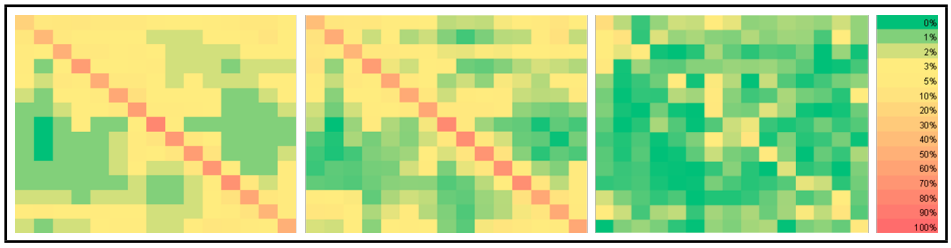

Figura 6: De izquierda a derecha: Matriz O-D de referencia, Matriz O-D estimada con $\varepsilon = 0.0101$, Diferencia entre las matrices y Escala de colores considerada

Presentamos también los valores más importantes que arroja el analisis de la Matriz Diferencia de la **Figura 6**. Se refleja de la comparación entre las matrices

Diferencia Máxima:	0.1472
Media:	0.0119

de la **Figura 6** que el método de optimización propuesto aplica una nivelación en los valores, logrando reducir las diferencias entre la matriz de referencia y la resultante a medida que ε se acerca a 0.0101, puesto que la condición (4) establece el módulo de la diferencia entre A_{ij} *y* T_{ij}.

Lo cierto es que el valor adecuado para ε dependerá de cuánto se quieran hacer corresponder los datos capturados con los de referencia. Si se conoce que la Matriz O-D de refencia es confiable, el valor de ε deberá ser pequeño, para que la diferencia entre ambas matrices sea mínima. En cambio, si la fuente es poco confiable o no se conoce su confiabilidad, dar valores más altos a ε nos evitará introducir sesgos a la matriz O-D estimada. En nuestro trabajo, arribamos a este valor luego de varias iteraciones, y considerando como muy confiable la matriz de referencia.

5. Conclusiones

Previo al análisis de los resultados obtenidos, es importante destacar que la metodología de optimización propuesta, permite desconocer información puntual de los viajes encontrando la distribución de probabilidad más factible en condiciones de máxima incertidumbre.

Luego de analizar la matriz y las imágenes, se puede apreciar que los datos que arroja Twitter en tan solo 36 días, conservan las tendencias de movilidad obtenidas por Anapolsky y por el Ministerio del Interior y Transporte, estableciendo que las redes sociales son aptas como fuentes de información para este tipo de estudios.

Otro aspecto a marcar es que a diferencia de los enfoques clásicos, este tipo de fuentes son públicas y gratuitas, lo que potencia aun más su utilidad, reduciendo los costos y aumentando el alcance. Además, por ser dinámicas y de origen digital, es posible procesarlas de forma ágil y rápida.

6. Trabajos Futuros

En el transcurso del trabajo, surgieron propuestas interesantes que estaban fuera del alcance del objetivo inicial, y que detallamos a continuación:

- Estimar los datos a partir de nuevas redes sociales.
- Incluir calculo del epsilon óptimo al problema a resolver.
- Probar la metodología en otras áreas urbanas.
- Generación automática de comunas utilizando métodos de clusterización, k-means, etc.
- Realizar comparaciónes con los otros dos métodos de estimación mencionados en la sección 3 (Inferencia Estadística y Aproximación del Gradiente)

Referencias

1. Miralles Guasch M.: Ciudad y transporte. El binomio imperfecto. Barcelona, Ariel (2002).

2. Secretaría de Transporte, Ministerio del Interior y Transporte: Encuesta de Movilidad Domiciliaria. PTUMA (2009).
3. Anapolsky S., Lang C., Ponieman N. y Sarraute C.: Exploración y análisis de datos de telefonía celular para estudiar comportamientos de movilidad en la Ciudad de Buenos Aires. XVIII CLATPU (2014).
4. Daly A. y Ortúzar J.: Forecasting and data aggregation: theory and practice. CEDEX (1990).
5. Emarketer: Emerging Markets Drive Twitter User Growth Worldwide "http://www.emarketer.com/Article.aspx?R=1010874"
6. Twitter4J - A Java library for the Twitter API. "http://twitter4j.org/en/index.html".
7. INDEC: Censo Nacional de Población, Hogares y Viviendas 2010. "http://www.censo2010.indec.gov.ar".
8. Ortúzar J. y Willumsen L.: Modelos de Transporte. Universidad de Cantabria (2008).
9. Cascetta E.: Estimation of Trip Matrices from Traffic Counts and Survey Data: A generalized least squares estimator. Transportation Research-B, Vol. 18B, No.4 (1984).
10. Cascetta E., Nguyen S.: A unified framework for estimating or updating origin/destination matrices from traffic counts. Transportation Research-B, Vol. 22B, No.6 (1988).
11. Carvalho, L., A Bayesian Statistical Approach for Inference on Static Origin-Destination Matrices in Transportation Studies, Technometrics, 56(2), 225–237 (2013).
12. Walpen J., Mancinellia E. M., Lotito P. A.: A heuristic for the OD matrix adjustment problem in a congested transport network. European Journal of Operational Research, Vol 242, No.3 (2014).
13. Foley J. D., Van Dam A.,Feiner S. K. y Hughes J. F.: Computer Graphics: Principles and Practice. The Systems Programming Series. Addison-Wesley (1990).
14. Colomer Ferrandiz J. y Ruiz Sanchez T.: Los modelos de cuatro etapas: Utilidad y limitaciones. CEDEX (1999).
15. Frias-Martinez, Vanessa and Soguero, Cristina and Frias-Martinez, Enrique: Estimation of urban commuting patterns using cellphone network data, Proceedings of the ACM SIGKDD International Workshop on Urban Computing. 9-16 (2012).
16. Bahoken, Françoise and Olteanu-Raimond, Ana-Maria: Designing Origin-Destination Flow Matrices from Individual Mobile Phone Paths: The effect of spatiotemporal filtering on flow measurement, Proceedings of 23rd International Cartography Conference (2013)