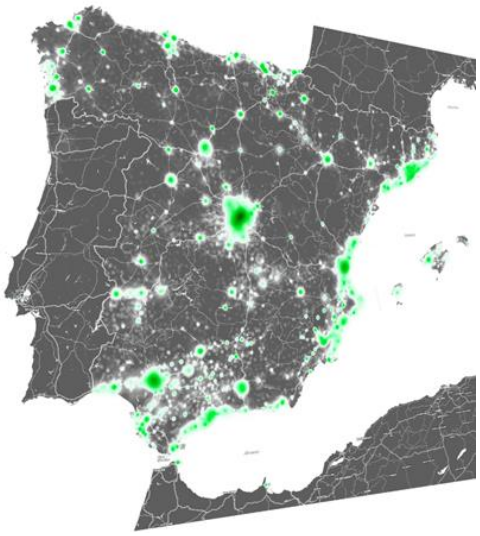
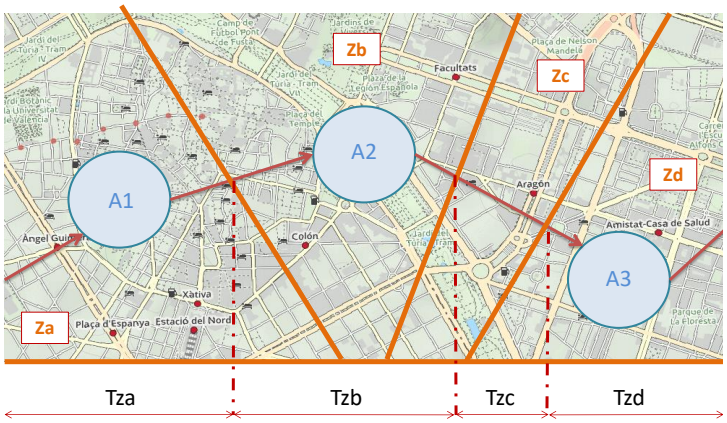
UNIVERSITAT POLITÈCNICA DE VALÈNCIA





METODOLOGÍA PARA LA EXTRACCIÓN DE PATRONES DE MOVILIDAD URBANA MEDIANTE EL ANÁLISIS DE REGISTROS DE ACTIVIDAD TELEFÓNICA (CALL DETAIL RECORD)

Tesis Doctoral

Realizada por:

Miguel Picornell Tronch

Dirigida por

Dr. Tomás Ruiz Sánchez

Universitat Politècnica de València

Escuela de Doctorado

Programa de Doctorado en Ingeniería Civil y Urbanismo

Valencia, 2 de julio de 2017

"A mis padres Miguel y Pilar por su apoyo incondicional en esta y en cada una de las etapas de mi vida"

Resum

En l'últim segle, Europa ha viscut una forta migració de l'àmbit rural a l'urbà. S'estima que al voltant del 70% de la població europea (aproximadament 350 milions de persones) viu en aglomeracions urbanes de més de 5.000 habitants (DG REGIO, 2010). La mobilitat urbana és fonamental per al desenvolupament econòmic i social de les ciutats però al mateix temps comporta a una sèrie d'importants efectes negatius, com ara la congestió o la contaminació de l'aire. L'entesa dels patrons de mobilitat urbana dels ciutadans és essencial perquè els gestors puguin avaluar quines són les polítiques i mesures més adequades per aconseguir un desenvolupament urbà sostenible. La majoria dels estudis empírics sobre mobilitat urbana es recolzen en enquestes, ja que aquestes proporcionen informació detallada sobre els patrons de mobilitat de la població aportant al mateix temps una gran quantitat d'informació sociodemogràfica. No obstant això, les enquestes presenten una sèrie de limitacions pràctiques importants (Ortúzar & Willumsen, 2011) com ara els seus elevats costos econòmics o els seus llargs terminis d'execució. L'ús generalitzat de dispositius mòbils per part de la població obre la possibilitat de recollir de manera anònima i passiva una gran quantitat d'informació espai-temporal d'una gran mostra d'usuaris, superant algunes de les limitacions dels actuals mètodes de recollida d'informació. En concret, les dades de la xarxa de telefonia mòbil presenten una sèrie d'avantatges que els posicionen com una de les millors fonts de dades per a l'estudi de la mobilitat general de grans nuclis de població (baixos costos d'extracció de les dades, grans dimensions de mostra, amplia cobertura espacial, etc.). En els últims anys, s'han dut a terme diferents estudis amb dades de telefonia mòbil per a l'anàlisi dels patrons d'activitat de la població a les ciutats (Reades et al. 2007), l'estudi dels patrons de mobilitat de les persones (Gonzalez et al. 2008), l'estimació de volums de trànsit (Càceres et al. 2012) o l'anàlisi del comportament dels turistes (Ahas et al., 2007). No obstant, encara hi ha un ampli marge de millora en aquesta disciplina pel que fa a aspectes metodològics sobre el tractament de les dades i un gran nombre d'aplicacions pràctiques per explorar.

L'objectiu principal d'aquesta investigació és contribuir als recents avenços en el camp de l'anàlisi de les dades de telefonia mòbil mitjançant el desenvolupament i validació d'una metodologia que permeti extreure informació de patrons d'activitat i mobilitat de la població en àmbits urbans. La metodologia desenvolupada presenta una sèrie de millores rellevants pel que fa a estudis previs, com l'estimació de localitzacions freqüents diferents de casa i treball, la millora en l'estimació de l'hora del viatge, procediments per a la selecció i expansió de la mostra o la millora en l'estimació del nombre de persones en una àrea específica a partir dels patrons d'activitat i mobilitat de les mateixes. Aquesta metodologia ha estat aplicada en tres casos d'ús per a: (1) l'obtenció d'estadístiques bàsiques de mobilitat i matrius origen-destinació en àmbits urbans, (2) l'anàlisi de la influència de la xarxa social en la mobilitat i (3) l'estudi de l'exposició de la població a la contaminació.

Els resultats mostren el gran potencial de les dades de telefonia mòbil per estimar estadístiques bàsiques de mobilitat (nombre de viatges per persona, distribució de distància dels viatges, etc.) i matrius origen-destinació superant algunes de les limitacions dels mètodes convencionals de recollida d'informació, com la reduïda grandària de mostra, els alts costos econòmics o els llargs terminis d'execució. Els resultats obtinguts amb aquesta nova metodologia han estat comparats amb estadístiques procedents d'enquestes. Les grans similituds trobades en comparar les dues metodologies posen en evidència, d'una banda, el potencial de les dades de telefonia mòbil per capturar patrons de mobilitat de la població i, d'altra banda, la validesa de la metodologia proposada.

Addicionalment, s'ha mostrat el potencial de les dades de telefonia mòbil per a analitzar de manera conjunta la xarxa social i la mobilitat. Els resultats reforcen la hipòtesi que 'altres' localitzacions freqüents diferents de casa i treball poden ser considerades com a llocs on potencialment es produeixen interaccions entre les persones d'una mateixa xarxa social. A més, s'ha observat que la majoria dels esdeveniments de co-ubicació tenen lloc a 'altres' localitzacions freqüents i en localitzacions no freqüentment visitades per les persones. La informació de xarxa social i mobilitat extreta de les dades de telefonia mòbil

pot ajudar a millorar la definició dels actuals models de transport basats en activitats, aportant noves variables a considerar a l'hora de simular on i quan es realitzen les diferents activitats. Aquestes millores són d'interessant aplicació per a l'anàlisi de serveis de mobilitat compartida, com ara el transport sota demanda o el carpooling.

En aquesta investigació també s'ha analitzat el potencial de les dades de telefonia mòbil per millorar les estimacions d'exposició de la població a la contaminació. Els resultats mostren com la informació de presència de població obtinguda a partir dels patrons d'activitat i mobilitat de la població millora dels resultats dels indicadors d'exposició obtinguts mitjançant metodologies convencionals. Per a estudis amb un cert nivell d'agregació (diverses desenes de quilòmetres), les metodologies tradicionals basades en dades censals aporten resultats satisfactoris, però, per a estudis a una escala més detallada és fonamental considerar els patrons d'activitat i mobilitat de la població.

Tot i les evidents avantatges que proporcionen les dades de telefonia mòbil, també s'han observat limitacions rellevants en els diferents estudis realitzats. Les limitacions vénen derivades principalment de les característiques espai-temporals de les dades i de la manca d'informació sociodemogràfica associada als mateixos. Per al cas de l'obtenció d'estadístiques sobre mobilitat en àmbits urbans s'ha identificat la impossibilitat, a dia d'avui, d'obtenir informació detallada sobre el motiu precís dels viatges, el mitjà de transport i la ruta dels mateixos, informació sociodemogràfica detallada de les persones o opinions de valor de la població. Per al cas de l'estudi de la xarxa social i la mobilitat, s'han identificat limitacions associades a la possibilitat que hi hagi relacions socials que es produeixin per canals de comunicació diferents al telèfon mòbil. A més, la resolució espai-temporal de les dades pot influir en la qualitat dels resultats sobre localitzacions compartides pels membres d'una mateixa xarxa social. Per al cas de l'exposició a la contaminació, la principal limitació identificada ha estat la possible manca de qualitat de la informació sociodemogràfica disponible, que pot afectar els processos d'elevació mostral.

Aquesta recerca obre la porta a un gran nombre de futures línies d'investigació. S'han identificat línies de millora metodològiques així com un gran nombre de futures aplicacions pràctiques. La informació de mobilitat extreta de les dades de telefonia mòbil ja aporta a dia d'avui una informació molt valuosa per a estudis de mobilitat. És previsible que la millora en la qualitat de les dades de telefonia mòbil així com noves millores metodològiques facin que aquesta font de dades es posicioni com una de les fonts de dades fonamental i imprescindible per a estudis de mobilitat en el curt termini.

Resumen

En el último siglo, Europa ha vivido una fuerte migración del ámbito rural al urbano. Se estima que alrededor del 70% de la población europea (aproximadamente 350 millones de personas) vive en aglomeraciones urbanas de más de 5.000 habitantes (DG REGIO, 2010). La movilidad urbana es fundamental para el desarrollo económico y social de las ciudades pero al mismo tiempo conlleva a una serie de importantes efectos negativos, tales como la congestión o la contaminación del aire. El entendimiento de los patrones de movilidad urbana de los ciudadanos es esencial para que los gestores puedan evaluar cuáles son las políticas y medidas más adecuadas para conseguir un desarrollo urbano sostenible. La mayoría de los estudios empíricos sobre movilidad urbana se apoyan en encuestas, ya que estas proporcionan información detallada sobre los patrones de movilidad de la población aportando al mismo tiempo una gran cantidad de información socio-demográfica. Sin embargo, las encuestas presentan una serie de limitaciones prácticas importantes (Ortúzar & Willumsen, 2011) tales como sus elevados costes económicos o sus largos plazos de ejecución. El uso generalizado de dispositivos móviles por parte de la población abre la posibilidad de recoger de manera anónima y pasiva una gran cantidad de información espacio-temporal de una gran muestra de usuarios, superando algunas de las limitaciones de los actuales métodos de recogida de información. En concreto, los datos de la red de telefonía móvil presentan una serie de ventajas que los posicionan como una de las mejores fuentes de datos para el estudio de la movilidad general de grandes núcleos de población (bajos costes de extracción de los datos, gran tamaño de muestra, amplía cobertura espacial, etc.). En los últimos años, se han llevado a cabo diferentes estudios con datos de telefonía móvil para el análisis de los patrones de actividad de la población en las ciudades (Reades et al. 2007), el estudio de los patrones de movilidad de las personas (Gonzalez et al. 2008), la estimación de volúmenes de tráfico (Cáceres et al. 2012) o el análisis del comportamiento de los turistas (Ahas et al., 2007). No obstante, aún existe un amplio margen de mejora en esta disciplina en lo referente especialmente a aspectos metodológicos sobre el tratamiento de los datos y un gran número de aplicaciones prácticas por explorar.

El objetivo principal de esta investigación es contribuir a los recientes avances en el campo del análisis de los datos de telefonía móvil mediante el desarrollo y validación de una metodología que permita extraer información de patrones de actividad y movilidad de la población en ámbitos urbanos. La metodología desarrollada presenta una serie de mejoras relevantes con respecto a estudios previos, como la estimación de localizaciones frecuentes distintas de casa y trabajo, la mejora en la estimación de la hora del viaje, procedimientos para la selección y expansión de la muestra o la mejora en la estimación del número de personas en un área específica a partir de los patrones de actividad y movilidad de las mismas. Esta metodología ha sido aplicada en tres casos de uso para: (1) la obtención de estadísticas básicas de movilidad y matrices origen-destino en ámbitos urbanos, (2) el análisis de la influencia de la red social en la movilidad y (3) el estudio de la exposición de la población a la contaminación.

Los resultados muestran el gran potencial de los datos de telefonía móvil para estimar estadísticas básicas de movilidad (número de viajes por persona, distribución de distancia de los viajes, etc.) y matrices origen-destino superando algunas de las limitaciones de los métodos convencionales de recogida de información, como el reducido tamaño de muestra, los altos costes económicos o los largos plazos de ejecución. Los resultados obtenidos con esta nueva metodología han sido comparados con estadísticas procedentes de encuestas. Las grandes similitudes encontradas al comparar ambas metodologías ponen en evidencia, por un lado, el potencial de los datos de telefonía móvil para capturar patrones de movilidad de la población y, por otro lado, la validez de la metodología propuesta.

Adicionalmente, se ha mostrado el potencial de los datos de telefonía móvil para analizar de manera conjunta la red social y la movilidad. Los resultados refuerzan la hipótesis de que ‘otras’ localizaciones frecuentes distintas de casa y trabajo pueden ser consideradas como lugares donde potencialmente se producen interacciones entre las personas de una misma red social. Además, se ha observado que la mayoría de los eventos de co-ubicación tienen lugar en ‘otras’ localizaciones frecuentes y en localizaciones no frecuentemente

visitadas por las personas. La información de red social y movilidad extraída de los datos de telefonía móvil puede ayudar a mejorar la definición de los actuales modelos de transporte basados en actividades, aportando nuevas variables a considerar a la hora de simular dónde y cuándo se realizan las distintas actividades. Estas mejoras son de interesante aplicación para el análisis de servicios de movilidad compartida, como por ejemplo el transporte bajo demanda o el carpooling.

En esta investigación también se ha analizado el potencial de los datos de telefonía móvil para mejorar las estimaciones de exposición de la población a la contaminación. Los resultados muestran como la información de presencia de población obtenida a partir de los patrones de actividad y movilidad de la población mejora los resultados de los indicadores de exposición obtenidos mediante metodologías convencionales. Para estudios con un cierto nivel de agregación (varias decenas de kilómetros), las metodologías tradicionales basadas en datos censales aportan resultados satisfactorios, sin embargo, para estudios a una escala más detallada es fundamental considerar los patrones de actividad y movilidad de la población.

A pesar de las evidentes ventajas que proporcionan los datos de telefonía móvil, también se han observado limitaciones relevantes en los distintos estudios realizados. Las limitaciones vienen derivadas principalmente de las características espacio-temporales de los datos y de la falta de información socio-demográfica asociada a los mismos. Para el caso de la obtención de estadísticas sobre movilidad en ámbitos urbanos se ha identificado la imposibilidad, a día de hoy, de obtener información detallada sobre el motivo preciso de los viajes, el modo y la ruta de los mismos, información socio-demográfica detallada de las personas u opiniones de valor de la población. Para el caso del estudio de la red social y la movilidad, se han identificado limitaciones asociadas a la posibilidad de que existan relaciones sociales que se produzcan por canales de comunicación distintos al teléfono móvil. Además, la resolución espacio-temporal de los datos puede influir en la calidad de los resultados sobre localizaciones compartidas por los miembros de una misma red social. Para el caso de la exposición a la contaminación, la

principal limitación identificada ha sido la posible falta de calidad de la información socio-demográfica disponible, que puede afectar a los procesos de elevación muestral.

Esta investigación abre la puerta a un gran número de futuras líneas de investigación. Se han identificado líneas de mejora metodológicas así como un gran número de futuras aplicaciones prácticas. La información de movilidad extraída de los datos de telefonía móvil ya aporta a día de hoy una información muy valiosa para estudios de movilidad. Es previsible que la mejora en la calidad de los datos de telefonía móvil así como nuevas mejoras metodológicas hagan que esta fuente de datos se posicione como una de las fuentes de datos fundamental e imprescindible para estudios de movilidad en el corto plazo.

Abstract

In the last century, Europe has seen a strong migration from rural to urban areas. It is estimated that around 70% of the EU population (approximately 350 million people) live in urban agglomerations of more than 5,000 inhabitants (DG REGIO, 2010). Urban mobility is key to the economic and social development of cities, but at the same time it generates a significant number of negative effects such as congestion or air pollution. Understanding urban mobility patterns is essential to evaluate which are the most appropriate policies and measures to achieve sustainable urban development. Most of the empirical studies on urban mobility are based on surveys, since they provide detailed information about population mobility patterns and a large amount of socio-demographic information. However, surveys have several practical limitations (Ortúzar & Willumsen, 2011) such as their high costs and long lead times. The pervasive used of mobile devices opens the opportunity of gather large amounts of anonymised, passively-collected geolocation data overcoming some of the limitations of traditional surveys. Mobile phone data are probably one of the best data sources from which extract population mobility patterns at city scale because of their advantages (large samples, wide spatial coverage, low data collection costs, etc.). In recent years, different studies have used mobile phone data to analyse population activity patterns in cities (Reades et al., 2007), population mobility patterns (Gonzalez et al., 2008), traffic volumes (Cáceres et al., 2012) and tourist behaviour (Ahas et al., 2007). However, there is still room for improvement regarding methodological aspects and a large number of new practical applications to explore.

The main objective of this research is to contribute to the recent advances in the analysis of mobile phone data by developing and validating a new methodology to extract population activity and mobility patterns in urban areas. The methodology developed present several improvements with respect to previous studies, such as the identification of frequent locations different from home and work, better trip time estimations, sample selection and expansion procedures and improvements on population density estimations. The methodology developed has been tested in three different case studies:

(1) estimation of mobility statistics and origin-destination matrices, (2) analysis of the relationship between social network and travel behaviour and (3) evaluation of population exposure to air pollution taking into account population activity and mobility patterns.

Results show the great potential of mobile phone data to estimate mobility statistics (trips per person, travel distance distribution, etc.) and trip matrices, overcoming some of the limitations of conventional data collection methods. The results obtained from mobile phone data have been compared with those coming from surveys. The similarity between the statistics obtained from both methodologies demonstrates the validity of the new methodology proposed.

The potential of mobile phone data to characterize the relationship between social network and travel behaviour has also been analysed. Results show that 'other' locations different from home and work are frequently associated to social interaction. Additionally, the importance of non-frequent locations in social network co-location events has been shown. Information about the relationship between social network and travel behaviour can help to improve current activity-based models by providing new variables and rules to consider when simulating population behaviour. These modelling improvements may lead to a better evaluation of policies where social interaction is relevant, such as transport on demand or carpooling.

This research has also analysed the potential of mobile phone data to improve population exposure assessments to air pollution. Results show that, for spatially aggregated analysis, the conventional static methodology (census based) and the dynamic methodology (mobile phone based) provide similar results. However, relevant discrepancies have been found when analysing results at more disaggregated levels; being the consideration of mobility patterns essential to determine the actual population exposed to air pollution.

Although it has been shown that mobile phone data have the potential to provide rich information about population activity and mobility patterns, they are not free of drawbacks and limitations. Main limitations are related to the spatio-temporal resolution

of the data and the lack of socio-demographic information. Limitations related to the travel purpose identification, mode and route estimations, socio-demographic population classification and population subjective perceptions have been shown. With respect to the study about social network and travel behaviour, limitations related to the fact that other social interactions may be conducted through other communication channels different from mobile phone have been pointed out. Additionally, it has also been remarked that the spatio-temporal resolution of the data may influence on the quality of the results related to the identification of locations shared by social contacts. Finally, the limited socio-demographic information available has also been pointed out as a limitation when analysing the population exposure to air pollution.

This research has thrown up many questions in need of further investigation. Future methodological improvements and new possible applications have been proposed. Activity and mobility information extracted from mobile phone data already provides valuable information for mobility studies. It is expected that improvements on data quality and new methodological advances will position mobile phone data as one of the most relevant and essential data sources for urban mobility studies in the short term.

Índice

Listado de Tablas

Listado de Figuras

INTRODUCCIÓN

Antecedentes

En el último siglo, Europa ha vivido una fuerte migración del ámbito rural al urbano. Se estima que alrededor del 70% de la población europea (aproximadamente 350 millones de personas) vive en aglomeraciones urbanas de más de 5.000 habitantes (DG REGIO, 2010). La movilidad urbana es fundamental para el desarrollo económico y social de las ciudades, pero al mismo tiempo, conlleva a una serie de importantes efectos negativos. La congestión, la contaminación del aire, la exposición al ruido o la seguridad vial, son algunos de los principales retos a los que se enfrentan las ciudades a día de hoy (European Commission 2011). El entendimiento de los patrones de movilidad urbana de los ciudadanos es esencial para que los gestores puedan evaluar cuáles son las políticas y medidas más adecuadas para conseguir un desarrollo urbano sostenible. La mayoría de los estudios empíricos sobre movilidad urbana se apoyan en encuestas, ya que éstas proporcionan información detallada sobre los patrones de movilidad de la población aportando al mismo tiempo una gran cantidad de información socio-demográfica. Sin embargo, las encuestas presentan una serie de limitaciones prácticas importantes (Ortúzar & Willumsen, 2011) tales como sus elevados costes económicos o sus largos plazos de ejecución. Estas limitaciones prácticas hacen que las encuestas no se realicen con el alcance y frecuencia deseados, lo que conlleva que, en muchas ocasiones, la información sobre movilidad de la cual se dispone no sea de la calidad deseada y/o se encuentre desactualizada. Durante las últimas décadas, se ha producido un gran despliegue tecnológico dentro del sector de las Tecnologías de la Información y la Comunicación (TIC), llevando al mercado una gran variedad de sensores y dispositivos móviles tales como redes de telefonía móvil, smartphones, receptores GPS o redes WiFi. A través de los registros de actividad de estas tecnologías se puede recoger una gran cantidad de datos espacio-temporales de las personas, de los que es posible extraer información de movilidad que permite superar muchas de las limitaciones prácticas que

presentan las encuestas. Los datos de la red de telefonía móvil en concreto presentan una serie de ventajas que los posicionan como una de las mejores fuentes de datos para el estudio de la movilidad general en grandes núcleos de población (bajos costes de extracción de los datos, gran tamaño de muestra, amplía cobertura espacial, etc.). Además, los datos de telefonía móvil, a diferencia de otras fuentes de datos, también proporcionan información muy relevante sobre la red social de las personas (a través de la información de SMS y llamadas que éstas intercambian) lo que permite tener un mejor conocimiento sobre las interacciones sociales que se producen en la ciudad. En los últimos años, se han llevado a cabo diferentes estudios con datos de telefonía móvil para el análisis de los patrones de actividad de la población en las ciudades (Reades et al. 2007), el estudio de los patrones de movilidad de las personas (Gonzalez et al. 2008), la estimación de volúmenes de tráfico (Cáceres et al. 2012) o el análisis del comportamiento de los turistas (Ahas et al., 2007). No obstante, aún existe un amplio margen de mejora en esta disciplina, especialmente en lo referente a aspectos metodológicos sobre el tratamiento de los datos, y un gran número de aplicaciones prácticas por explorar.

Objetivos Principales e Hipótesis de la Investigación

El objetivo principal de esta investigación es contribuir a los recientes avances en el campo del análisis de los datos de telefonía móvil mediante el desarrollo y validación de una metodología que permita extraer información de patrones de actividad y movilidad de la población en ámbitos urbanos. Al inicio de este trabajo de investigación, a nuestro leal saber y entender, no existía ninguna metodología detallada y validada que cubriera todo el proceso de análisis de los datos (pre-procesado, depuración, extracción de patrones, elevación muestral y cálculo de indicadores) para la obtención de estadísticas de actividad y movilidad de las personas en ámbitos urbanos. Estudios previos relevantes, como Calabrese et al. 2011a, realizaban algunas comparaciones de estadísticas básicas de movilidad obtenidas mediante telefonía móvil con encuestas (en concreto, número de viajes por persona) y generaban modelos de estimación de viajes ajustados mediante información de viajes casa-trabajo procedentes de datos censales, obteniendo unos

resultados esperanzadores pero no plenamente satisfactorios (R^2= 0,76 a nivel de condado y R^2= 0,36 a nivel de sección censal para el área de Boston). Aspectos como la identificación precisa de la hora del viaje, los propósitos de los viajes, la selección de la muestra, los procedimientos de expansión/ajuste no basados en datos procedentes de las propias encuestas o ejercicios de validación de la información expandida al total de la población, son aspectos muy relevantes no tratados en estudios previos.

Otro de los objetivos de esta investigación es aplicar la metodología desarrollada para la extracción de patrones de actividad y movilidad de la población en distintos casos prácticos. Entre las posibles aplicaciones prácticas relevantes, esta investigación se ha centrado en estudiar: (1) cómo los datos de telefonía móvil pueden proporcionar estadísticas básicas de movilidad y matrices origen-destino (matrices OD) en ámbitos urbanos; (2) cómo pueden ayudar a entender mejor las relaciones entre la red social y la movilidad; y (3) cómo el conocer los patrones de movilidad de la población puede ayudar a mejorar los estudios de exposición de la población a la contaminación. Varios estudios previos han utilizado datos de telefonía móvil para obtener matrices OD (White and Wells 2002; Cáceres et al. 2007; Sohn and Kim 2008), siendo menos habituales aquellos que han analizado matrices OD de movilidad general en ámbitos urbanos (p.ej. Calabrese et al. 2011a). Por otro lado, son escasos los estudios que han analizado de manera conjunta las relaciones entre la red social y la movilidad de las personas a partir de datos de telefonía móvil (p.ej. Phithakkitnukoon et al. 2011 y 2012; Calabrese et al. 2011b). Tanto los estudios para la extracción de matrices OD en ámbitos urbanos como los estudios que analizan conjuntamente la red social y la movilidad, presentan aún gran margen de mejora, especialmente en lo referente a aspectos metodológicos y de validación, como se ha señalado anteriormente. Por otro lado, a nuestro leal saber y entender, al inicio de esta investigación no existía ningún estudio que hubiera analizado la exposición de la población a la contaminación a partir de datos de telefonía móvil.

La principal hipótesis que se plantea en esta investigación es que la información que proporcionan los registros de actividad telefónica es una alternativa real a las actuales

encuestas de movilidad, permitiendo la obtención de información de una manera más eficaz y eficiente.

Metodología de la Investigación

La metodología que se ha empleado para llevar a cabo esta investigación consta principalmente de las siguientes etapas:

- **Revisión bibliográfica**: revisión del estado del arte sobre las características de los datos de telefonía móvil, revisión de las técnicas de análisis para la extracción de patrones de actividad y movilidad de la población a partir de datos de telefonía móvil, revisión de estudios sobre el análisis conjunto de red social y la movilidad, y revisión de estudios sobre el análisis de la exposición de la población a la contaminación mediante telefonía móvil.

- **Definición de la metodología para la extracción de patrones de actividad y movilidad**: definición detallada de una metodología para la extracción de patrones de actividad y movilidad de la población a partir de datos de telefonía móvil tomando como punto de partida el estado del arte actual e incorporando un conjunto de mejoras.

- **Desarrollo software de la solución**: desarrollo de las especificaciones, diseño técnico y funcional, implementación del código y ejecución de tests de verificación para el desarrollo software de la metodología propuesta.

- **Validación de la solución**: comparación de los resultados que proporciona la solución desarrollada con estadísticas procedentes de encuestas para su validación.

- **Aplicación práctica en entornos reales:** aplicación de la metodología desarrollada en distintos casos prácticos, para evaluar sus potencialidades y sus limitaciones para distintos casos de uso.

Para su desarrollo, la investigación ha requerido principalmente de: datos de telefonía móvil facilitados por operadores de la red de telefonía móvil, herramientas de software libre (Python, QGIS, etc.) y de una infraestructura hardware capaz de almacenar y procesar los datos de telefonía móvil de manera eficaz y eficiente.

Visión General del Proceso de Investigación

En esta sección se presenta brevemente una visión general de todo el proceso de investigación llevado a cabo, desde sus inicios hasta la actualidad. En la Figura 1 se muestra cada una de las principales etapas del proceso así como los hitos más importantes referentes a publicaciones científicas derivadas de la presente investigación. Dado que esta disciplina ha experimentado un avance muy importante en los últimos años, es importante ubicar temporalmente los estudios llevados a cabo dentro de esta investigación para poder entender mejor los avances y contribuciones de esta Tesis Doctoral. El estudio empieza en el año 2013 con la revisión del estado del arte sobre el uso de datos de telefonía móvil para estudios de movilidad y con la identificación y planteamiento del problema de investigación. Durante ese año, se empieza a definir y desarrollar la metodología para la extracción de patrones de actividad y movilidad de la población a partir de datos de telefonía móvil. Entre finales de 2013 y principios de 2014, se valida el primer prototipo de la solución mediante la comparación de estadísticas de movilidad obtenidas mediante telefonía móvil con estadísticas procedentes de encuestas. Los resultados de esta validación se publican a principios del año 2015 como parte del informe técnico para la Comisión Europea "*D6.3 Case Study 3: Barcelona*" (enlace) dentro del marco del proyecto EUNOIA. Durante el año 2014, se lleva a cabo el estudio sobre el análisis de la red social y la movilidad a partir de datos de telefonía móvil. Los resultados de este estudio son publicados a principios de 2015 en la revista Transportation (Ranking: Q1), con el título "*Exploring the potential of phone call data to characterize the relationship between social network and travel behavior*" (enlace). Durante el año 2015, se desarrolla la metodología para estimar la ubicación de la población en la ciudad a diferentes horas del día (lo que denominaremos en esta Tesis Doctoral presencia de

población) a partir de los patrones de actividad y movilidad extraídos de los datos de telefonía móvil. A principios de 2016 se aplica dicha metodología en un caso práctico real para la mejora de las estimaciones de exposición de la población a la contaminación. En mayo de 2016 se presentan los primeros resultados en la conferencia HARMO'17 (enlace) y, actualmente, una publicación científica derivada de dicho estudio con el título "*Population dynamics based on mobile phone data to improve air pollution exposure assessments*" ha sido enviada para su revisión a una revista de alto impacto (Ranking: Q1). Una copia del artículo está disponible en el Anexo II de este documento.

Agradecimientos

En primer lugar, quiero agradecer a mi Director de Tesis Tomás Ruiz su apoyo, guía y consejo durante todo el proceso de planteamiento y ejecución de la Tesis así como de las publicaciones científicas que se derivan de esta. También quiero agradecer a Ricardo Herranz y a Manuel Álvarez la oportunidad que me dieron en su día de formar parte de un proyecto innovador que me abrió las puertas al mundo del análisis de las nuevas fuentes de datos procedentes del Big Data, entre dichas fuentes de datos, los datos de telefonía móvil. Igualmente, agradecer a todos mis compañeros de trabajo y colegas de investigación sus contribuciones y enseñanzas que me han aportado durante todo este proceso. En último lugar, y no por ello menos importante, quiero agradecer a todas las entidades y programas de investigación que han contribuido en mayor o menor medida a la financiación de los trabajos presentes en esta Tesis Doctoral; destacando entre ellos al Séptimo Programa Marco de la Unión Europea FP7/2007-2013 a través del acuerdo de subvención nº 318367.

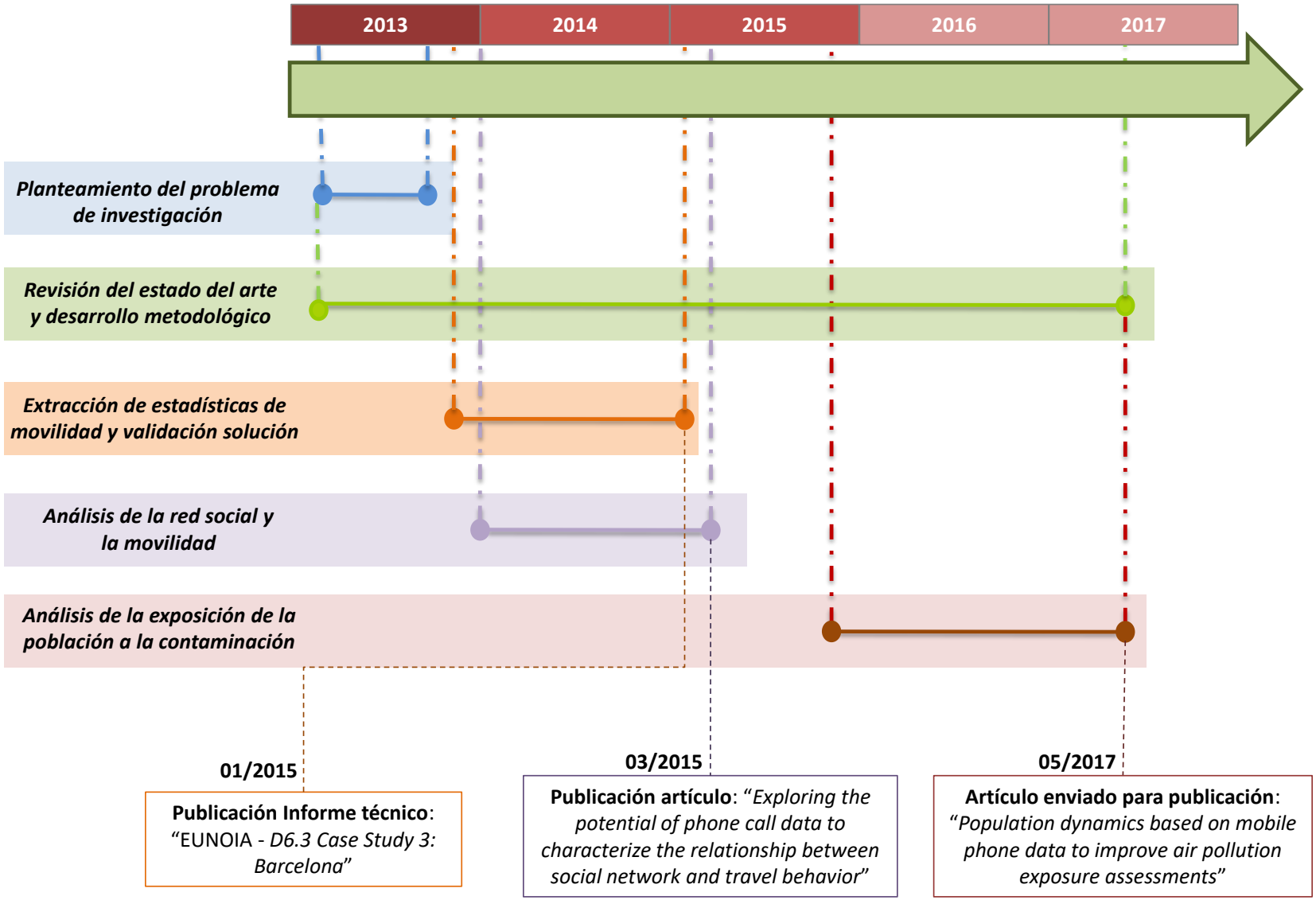


Figura 1. Cronograma del proceso de investigación con los hitos de publicación más importantes

Estructura del Documento

El presente documento se compone de cinco capítulos principales. En el **Capítulo I** se presenta el problema de investigación, se identifican las principales limitaciones asociadas a los actuales métodos de recogida de información sobre movilidad, se analizan las oportunidades que las nuevas fuentes de datos pueden aportan, y se presentan los objetivos y alcance del estudio. En el **Capítulo II** se presenta una revisión del estado del arte sobre la tipología de datos de telefonía móvil disponibles, una revisión de las técnicas de análisis para la extracción de patrones de actividad y movilidad a partir de datos de telefonía móvil, y una revisión sobre las aplicaciones prácticas de dicha información para el estudio de la influencia de la red social en la movilidad y para el estudio de la exposición de la población a la contaminación. En el **Capítulo III** se presentan las metodologías desarrolladas para la obtención de patrones de actividad y movilidad de la población, la obtención de la red social de las personas y la determinación de indicadores de presencia de población a partir de datos de telefonía móvil. En el **Capítulo IV** se presentan tres aplicaciones prácticas sobre el uso de los datos de telefonía móvil para la obtención de indicadores de actividad y movilidad. Finalmente, en el **Capítulo V**, se presentan las conclusiones del estudio así como las futuras líneas de investigación motivadas por los análisis y resultados del mismo.

CAPÍTULO I: EL PROBLEMA DE INVESTIGACIÓN

En esta sección se presenta, en primer lugar, el contexto en el cual se enmarca la investigación. En segundo lugar, se identifican los principales problemas y limitaciones que presentan las actuales metodologías de recogida de información sobre movilidad urbana basadas en encuestas. En tercer lugar, se exploran las oportunidades que las nuevas fuentes de datos como la telefonía móvil pueden ofrecer para estudios de movilidad. Del mismo modo, también se analizan en detalle las limitaciones prácticas y las oportunidades que las nuevas fuentes de datos pueden aportar para las aplicaciones prácticas concretas asociadas al análisis de la red social y la movilidad y al análisis de la exposición de la población a la contaminación. Posteriormente, se identifican aquellos aspectos no resueltos o pendientes de mejora que justifican la realización de esta investigación. Por último, se presentan los objetivos del estudio y se detalla el alcance del mismo.

1.1 La Movilidad Urbana

En el último siglo, Europa ha vivido una fuerte migración del ámbito rural al urbano. Se estima que alrededor del 70% de la población europea (aproximadamente 350 millones de personas) vive en aglomeraciones urbanas de más de 5.000 habitantes (DG REGIO, 2010). Aunque se espera un máximo de población europea alrededor del periodo 2015-2020, se estima un incremento del 10% de la población urbana para el 2050 (United Nations, Department of Economic and Social Affairs, Population Division 2010). Las ciudades son claves para el crecimiento y el desarrollo, y en ellas se genera el 85% del PIB europeo (European Commission 2009). Para sustentar la economía y bienestar de los ciudadanos, las ciudades necesitan de infraestructuras y servicios. Uno de los aspectos clave de esas infraestructuras y servicios es el transporte urbano, que da respuesta a la necesidad de movilidad urbana por parte de las personas. La movilidad urbana puede definirse como la totalidad de los viajes generados diariamente por las personas dentro de áreas urbanas y las características asociadas a dichos viajes, tales como el propósito del viaje, el modo de transporte, la distancia recorrida, la duración del viaje, etc. (Foth 2008).

El transporte urbano es esencial para que los ciudadanos realicen sus actividades diarias, pero al mismo tiempo conlleva un considerable número de efectos negativos, destacando entre ellos la congestión, la contaminación del aire, la exposición al ruido y la seguridad vial (European Commission 2011).

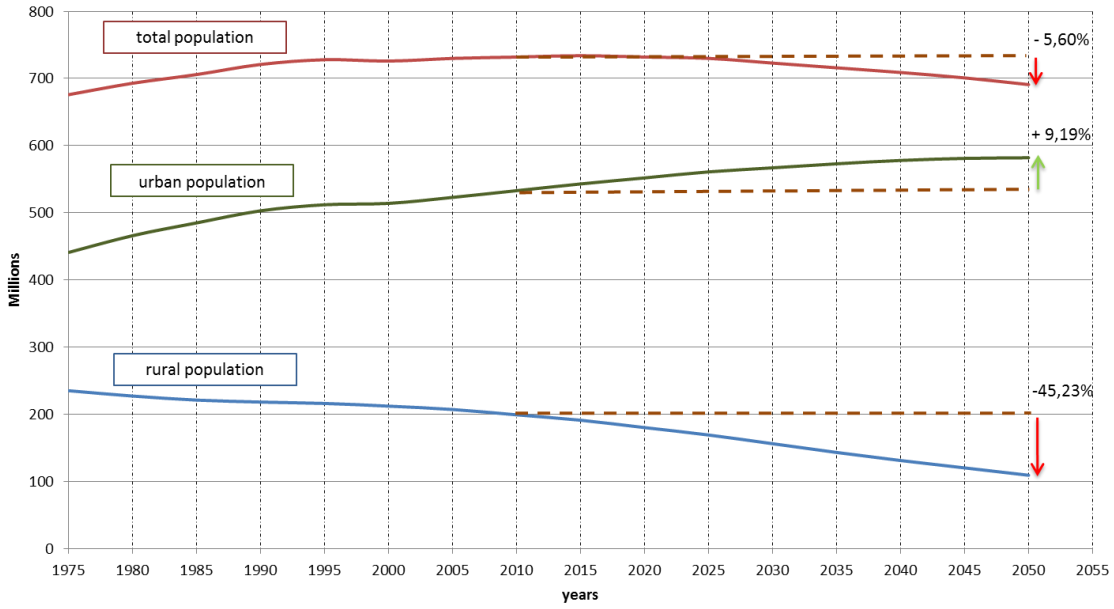


Figura 2. Crecimiento de la población en Europa desde 1975 a 2050. (Elaboración propia. Información extraída de: UN, Department of Economic and Social Affairs, 2010)

El transporte urbano constituye una de las mayores fuentes de polución (gases de efecto invernadero, calidad del aire, ruido, etc.) que afecta directamente a la salud y bienestar de los ciudadanos. Además, el transporte urbano es responsable del 25% de las emisiones de CO_2 derivadas del transporte y tiene una gran responsabilidad en la seguridad vial, produciéndose el 69% de los accidentes de circulación en entornos urbanos (European Commission 2011).

Si la tendencia actual continúa, las zonas urbanas se convertirán en entornos más congestionados, más contaminados y menos seguros para los ciudadanos. Existe la necesidad de tomar medidas urgentes en busca de una **movilidad más sostenible** para el **bienestar de los ciudadanos** y al mismo tiempo, asegurar el **desarrollo económico** y la **protección medioambiental** de los entornos urbanos.

1.2 Recogida de Información sobre Patrones de Movilidad Urbana

El entendimiento de los patrones de movilidad urbana es esencial para que los gestores puedan evaluar cuál es la estrategia más conveniente a adoptar de cara a conseguir un desarrollo urbano sostenible. Información sobre el perfil socioeconómico de los viajeros, el número de viajes que realizan al día, el origen y destino de los viajes, el propósito de los mismos, los modos de transporte y/o las rutas escogidas, son datos esenciales para el estudio de la movilidad. La mayor parte de esta información se ha recogido tradicionalmente a través de diferentes tipos de encuestas. Las encuestas permiten obtener información muy detallada sobre los patrones de movilidad de las personas e información sobre sus características socio-demográficas. Sin embargo, presentan también una serie de limitaciones prácticas importantes. A continuación se presentan las tipologías de encuestas comúnmente empleadas en estudios de movilidad, se detallan las principales limitaciones que presentan y se exploran las oportunidades que las nuevas fuentes de datos pueden ofrecer.

1.2.1 Principales encuestas para estudios de movilidad urbana

La gran mayoría de los estudios de movilidad se apoyan principalmente en la realización de encuestas como método para la recogida de información. Existen diferentes tipologías de encuestas para la recogida de información sobre movilidad. Las tipologías típicamente empleadas en estudios de movilidad son las denominadas encuestas domiciliarias y las encuestas de interceptación, que reciben su nombre en base al lugar donde se recoge la información (en el domicilio y en puntos específicos de la red de transporte donde se "intercepta" al usuario, respectivamente). A continuación se describe brevemente cada una de ellas (Ortúzar & Willumsen, 2011):

- **Encuestas domiciliarias**: las encuestas domiciliarias son un tipo de encuestas que recogen información sobre la movilidad (número de viajes, origen y destino de los viajes, modo de transporte, etc.) de todos los miembros de la familia, para todos sus viajes durante un periodo temporal de referencia. Del mismo modo, estas

encuestas suelen recoger información socio-demográfica detallada sobre el hogar (edad, género, disponibilidad de vehículo, renta, etc.). Dentro de las encuestas domiciliarias se incluyen las encuestas de diario de viajes, que aportan información similar a las encuestas domiciliarias tradicionales pero que buscan conseguir un mayor nivel de detalle a nivel individual.

- **Encuestas de interceptación**: estas encuestas se realizan en puntos específicos del área de estudio, como carreteras de acceso/salida del área de estudio, en medios de transporte público o en puntos de intercambio modal (estación de tren, aeropuertos, etc.). Este tipo de encuestas se realizan con el objetivo principal de recoger información sobre la movilidad de los no residentes, que no es capturada por las encuestas domiciliarias. Del mismo modo, la información recogida se utiliza para verificar y ampliar la información procedente de la encuesta domiciliaria. Este tipo de encuestas suelen ser muy breves, y recogen información reducida sobre el viaje (típicamente, al menos, información sobre el origen, destino y propósito del viaje). Generalmente se suelen realizar encuestas de cordón externo, para capturar los viajes con origen y/o destino fuera del área de estudio, y encuestas de cordón interno o líneas pantalla para analizar la movilidad de los no residentes dentro del área de estudio.

1.2.2 Limitaciones prácticas de las encuestas

Las encuestas aportan una descripción detallada de los viajes de las personas a nivel individual o por hogar, aportando al mismo tiempo gran cantidad de información socio-demográfica. No obstante, las encuestas también presentan una serie de limitaciones prácticas importantes (Ortúzar & Willumsen, 2011), destacando las siguientes:

- **Recursos económicos disponibles**: para la determinación de ciertas variables críticas para los estudios de movilidad (por ejemplo, información sobre la cantidad de viajes entre los diferentes orígenes y destinos de una ciudad) se requiere de información de una gran muestra de personas. En la mayoría de los casos, el

tamaño de muestra viene condicionado por las restricciones presupuestarias del estudio, lo que implica que en muchas ocasiones el alcance del estudio sea menor al que idealmente se necesitaría para evaluar de manera adecuada la movilidad.

- **Plazo de ejecución del estudio**: determina el tiempo y el esfuerzo que es posible dedicar a las tareas de recogida y análisis de la información. En función de la tipología de la encuesta, tamaño de la muestra y periodo temporal a considerar, los tiempos necesarios para la recogida y análisis de la información pueden variar entre unas pocas semanas/meses (p. ej. encuestas de interceptación) hasta incluso varios años (p. ej. encuestas domiciliarias). En muchas ocasiones, los plazos de ejecución de los proyectos de análisis de demanda son incompatibles con los tiempos necesarios para poder llevar a cabo una recogida de datos acorde a las necesidades del proyecto. En estos casos, lo habitual es utilizar información de movilidad disponible de estudios previos y, si es posible, se complementa dicha información con trabajos de campo puntuales. Estudios que requieren información de demanda en un corto plazo de tiempo son, por ejemplo, las licitaciones asociadas a concesiones de servicios e infraestructuras (autobuses, carreteras, etc.), en las que el estudio debe realizarse en un plazo de pocos meses o incluso semanas.

- **Precisión y/o veracidad de las respuestas:** con carácter general, las personas son reacias a responder a las preguntas de los cuestionarios, siendo esta tendencia cada día más marcada. La contestación de los cuestionarios consume mucho tiempo y la gente en muchas ocasiones rehúsa contestar u ofrece respuestas muy simplificadas, con el objetivo de preservar su privacidad o reducir el tiempo de contestación del cuestionario. En otras ocasiones, las respuestas pueden ser imprecisas por el simple hecho de que el entrevistado no recuerda con precisión la información (por ejemplo, cuando se le pregunta sobre viajes realizados en el día y no recuerda aquellos que son esporádicos, distintos a los viajes habituales basados en el hogar o en el trabajo). La no respuesta o falta de veracidad de la

información recogida puede generar importantes sesgos y errores en los resultados.

- Otras limitaciones prácticas relevantes están relacionadas con la definición de los **límites del área de estudio** (a priori es complicado definir cuál es el área adecuada a considerar y dónde colocar los límites del estudio) o las **barreras físicas** (p.ej. zonas donde no se puede realizar una encuesta de interceptación por la imposibilidad de detener a las personas).

Estas limitaciones prácticas tienen como consecuencia que en muchas ocasiones sea **inviable la actualización de la información de movilidad** con la periodicidad deseada. Uno de los ejemplos más destacados es la falta de información actualizada y detallada sobre movilidad en las grandes áreas metropolitanas, debido al alto coste asociado a la recogida de la información y a los posteriores trabajos de modelado. Por ejemplo, en el caso de España, en el momento de redactar esta Tesis, la última encuesta domiciliaria del área metropolitana de Madrid data del año 2004 y la del área metropolitana de Barcelona del año 2006.

Por tanto, para disponer de información detallada y actualizada sobre movilidad, **es necesario encontrar nuevas fuentes de datos y metodologías que permitan obtener dicha información superando las limitaciones de los actuales métodos** de recogida de información.

1.2.3 Nuevas fuentes de datos: retos y oportunidades

En los últimos años se ha experimentado un importante aumento del número y tipo de dispositivos móviles y sensores instalados en nuestro entorno. En el marco de este estudio, se entiende por dispositivos móviles aquellos dispositivos de carácter personal con alguna de las siguientes capacidades: procesamiento interno, conexión a internet y/o memoria. Dentro de esta definición de dispositivos móviles se encuentran por ejemplo los teléfonos móviles, las tarjetas inteligentes o los navegadores de conducción. Por otro

lado, los sensores hacen referencia a dispositivos capaces de detectar una determinada acción externa y transmitirla adecuadamente. Ejemplos de sensores son los sistemas de CCTV, las redes WiFi o los sensores de Bluetooth. Las personas, en su actividad diaria, utilizan o interaccionan con distintos dispositivos móviles y sensores, dejando una huella digital de sus actividades. Por ejemplo, cada vez que una persona paga con su tarjeta de crédito en un establecimiento, los datos sobre la hora, la localización del establecimiento y el importe de la compra (entre otros muchos datos) quedan registrados. Del mismo modo, cuando activamos las opciones de WiFi o Bluetooth en nuestro teléfono móvil, las redes de WiFi o Bluetooth presentes en nuestro entorno detectan que nuestro dispositivo se encuentra en los alrededores. El uso generalizado de dispositivos móviles, así como la generalización de las tecnologías de comunicación inalámbrica (Bluetooth, WiFi, etc.), abren la puerta a la recogida masiva de datos geolocalizados que, tras un adecuado análisis, permiten la obtención de información relevante sobre los patrones de actividad y movilidad de la población de manera anónima y agregada.

1.2.3.1 Ventajas y retos generales de las nuevas fuentes de datos

En función del dispositivo móvil o sensor (de ahora en adelante utilizaremos el término 'dispositivos' para referirnos a ambos conceptos), el tipo de datos que puede recogerse es muy variado. Con carácter general, todos los dispositivos suelen recoger información sobre la localización del usuario en determinados momentos del día y, dependiendo del tipo de dispositivo, recogen adicionalmente información contextual. Por ejemplo, las tarjetas inteligentes de transporte público aportan no solamente información sobre la localización y la hora a la cual el usuario accede al servicio de transporte, sino que también nos informa de que el usuario está utilizando el transporte público, la línea o estación a la cual accede o el modo de transporte elegido. Independientemente de la tipología de los datos recogidos, los nuevos dispositivos aportan una serie de ventajas con respecto a los métodos tradicionales:

- **Recogida pasiva**: el dato se recoge de manera pasiva, sin que haya necesidad de intervención por parte del usuario. Este aspecto resuelve uno de los problemas

señalados anteriormente con respecto a la predisposición a responder por parte de los encuestados.

- **Objetividad**: el dato se recoge sin que el usuario realice ningún juicio de valor. Nuevamente este aspecto puede ayudar a resolver algunos de los problemas señalados anteriormente, como la imprecisión de los usuarios a la hora de responder a ciertas preguntas (por ejemplo, la hora o la duración del viaje).

- **Datos actualizados e históricos**: el dato puede recogerse de manera continua y puede almacenarse con el objetivo de disponer de información histórica. Esto permite poder disponer de información para distintos años y distintos periodos del año de manera inmediata, reduciendo drásticamente los tiempos de recogida de la información. Del mismo modo, al recogerse la información de manera continua, pueden realizarse estudios a posteriori, algo que no resulta posible mediante los métodos tradicionales, ya que previamente se debe de haber planificado los trabajos de campo. Un ejemplo de estudios a posteriori podría ser el análisis de la movilidad en un día en el que se produjeron importantes cortes de circulación (por ejemplo por nevadas, accidentes, etc.) con el objetivo de analizar cómo respondió la población a dichos acontecimientos.

- **Coste de recogida**: generalmente, el coste de recogida del dato es sustancialmente inferior al coste de recogida de esa misma información por métodos tradicionales.

Sin embargo, las nuevas fuentes de datos no están exentas de limitaciones y retos importantes a superar:

- **Almacenamiento y procesamiento de los datos**: normalmente, la cantidad de datos que se recogen a través de estos dispositivos es de un tamaño muy elevado. En algunos casos puede tratarse de miles o millones de registros al día (por ejemplo, datos de tarjeta inteligente de transporte público), y en otros casos de miles de millones al día (por ejemplo, datos de la red de telefonía móvil). Es necesario almacenar y procesar estos datos de manera eficiente, con el objetivo

de que los análisis a realizar sobre los mismos puedan ejecutarse en un plazo de tiempo razonable.

- **Limpieza de los datos**: en muchos casos, los datos presentan un porcentaje significativo de errores de partida. En algunos casos, los errores pueden proceder de problemas en la configuración de los dispositivos (por ejemplo, la configuración de la hora del dispositivo) y en otros casos de registros administrativos asociados al dato (por ejemplo, el código postal de residencia del usuario).

- **Análisis de los datos**: a día de hoy, para muchas de estas fuentes de datos, no existe una metodología generalmente aceptada sobre cómo obtener información de movilidad a partir dichos datos. En la literatura se presentan diversos métodos, los cuales pueden conducir a resultados muy dispares entre sí. Un aspecto fundamental es, por lo tanto, realizar una validación rigurosa de la metodología a utilizar.

- **Adaptación a los modelos de transporte actuales**: la información recogida de estas fuentes de datos es, en muchos aspectos, distinta a la información recogida por fuentes tradicionales. Es necesario adaptar la información de tal forma que pueda ser utilizada en los actuales modelos de transporte y, quizá, en un futuro, adaptar también los actuales modelos de transporte para sacar el máximo partido de las nuevas fuentes de información.

1.2.3.2 Análisis comparativo de distintas fuentes de datos

La información sobre movilidad urbana que puede extraerse de cada una de las nuevas fuentes de datos (telefonía móvil, WiFi, Bluetooth, etc.) depende de las características concretas de cada dato. De cara a evaluar la calidad y la viabilidad de utilizar las distintas fuentes de datos es necesario tener en cuenta las siguientes características:

- **Tamaño de muestra**: hace referencia al número de personas de las cuales se dispone de información frente al total de la población de estudio. Es una de las

características clave. Algunas de las variables principales de los estudios de movilidad requieren de grandes muestras para proporcionar información con niveles de error aceptable. En particular, la calidad de las matrices origen-destino (la distribución de los viajes) está muy relacionada con el tamaño de muestra.

- **Alcance espacial**: el ámbito de estudio que se quiere analizar es normalmente muy extenso. Por lo tanto, es necesario que la información cubra todo el ámbito de estudio e, idealmente, todo el espacio definido por los orígenes y destinos que tienen influencia en el área de estudio.

- **Precisión espacial**: define el rango de incertidumbre espacial del dato. En función de la precisión espacial del dato se podrá definir mejor la localización concreta del usuario durante la realización de actividades (p.ej. la localización de su lugar de residencia) o durante sus desplazamientos (origen, destino, ruta, etc.).

- **Granularidad temporal**: se puede definir como el número de registros asociados a un mismo usuario por unidad de tiempo. Es necesario una granularidad suficiente para poder registrar todos los desplazamientos que el usuario puede realizar en el día. Del mismo modo, una granularidad temporal elevada puede mejorar la estimación de la ruta o el modo de transporte empleado por el usuario.

- **Disponibilidad de datos socio-demográficos**: es un aspecto relevante a la hora de caracterizar la movilidad en función de las características del usuario (edad, género, renta, etc.).También es un aspecto importante a la hora de llevar a cabo los procesos de elevación muestral.

- **Limitaciones prácticas:** hay otros aspectos como la necesidad del despliegue de infraestructura o la necesidad de instalación de software por parte del usuario que deben tenerse en cuenta a la hora de evaluar la idoneidad de cada fuente de datos.

En la *Tabla 1* se muestra un análisis comparativo de distintas fuentes de datos para las variables antes mencionadas. Los valores que se presentan están relacionados con las características habituales asociadas a cada tipo de dato. No obstante, es importante señalar que, dependiendo del caso concreto, algunas de las características pueden variar sensiblemente (por ejemplo, la precisión espacial de los datos de telefonía móvil podría considerarse alta si se dispusiera de información de posicionamiento obtenida mediante triangulación).

Del análisis comparativo de las distintas fuentes de datos se desprende que los **datos de la red telefonía móvil** presentan una serie de ventajas con respecto a otras fuentes de datos que los posicionan como una de las mejores alternativas a la hora de estudiar la **movilidad general de grandes volúmenes de población**:

- **Tamaño de la muestra**: el tamaño de muestra potencial procedente de los usuarios de la red de telefonía móvil suele ser muy elevado. Normalmente, los operadores con red propia suelen tener una cuota de mercado elevada. Por ejemplo, en el caso de España, existen cuatro operadores con red propia: Movistar, Orange, Vodafone y Yoigo; con una cuota de mercado móvil del 29,9%, 27,6%, 25,8% y 8,3% respectivamente, siendo la tasa de penetración en España de 110,2 dispositivos móviles por cada 100 habitantes (CNMC, enero 2017). Las muestras potenciales son por tanto muy superiores a las que podrían obtenerse con métodos tradicionales o mediante otras nuevas fuentes de datos.

- **Alcance espacial**: la red de telefonía móvil está desplegada por todo el territorio a nivel nacional. Esto permite analizar la movilidad de los usuarios tanto dentro como fuera del área de estudio definida, pudiendo caracterizar con detalle el origen y destino final de los viajes.

- **Precisión espacial**: la precisión espacial en zonas urbanas consolidadas suele ser media-alta, en torno a unas decenas o cientos de metros. Según la tecnología de la cual se disponga, se puede obtener una precisión espacial asociada al área de

cobertura de las antenas de telefonía móvil o información triangulada del dispositivo.

- **Granularidad temporal**: los datos de telefonía móvil suelen presentar una granularidad temporal elevada que permite disponer de información del usuario a lo largo de gran parte del día. Al igual que sucede con la precisión espacial, la granularidad temporal depende de la tecnología de extracción de datos de que se disponga. Se puede recoger eventos de manera activa, sólo cuando el usuario hace uso del móvil, o eventos de carácter pasivo (p.ej. cambio de zona de cobertura), estos últimos con una granularidad temporal mucho mayor.

- **Información socio-demográfica:** en algunos casos se puede disponer de información socio-demográfica básica del usuario, como la edad o el género. La información socio-demográfica disponible suele estar limitada por motivos de privacidad.

- **Despliegue de la solución:** la recogida de información de la red no necesita del despliegue de sensores adicionales a los ya instalados para la gestión de la red. Del mismo modo, tampoco es necesario que el usuario instale ningún software en su dispositivo móvil o que active ciertas configuraciones de su dispositivo para la recogida de información. Estos aspectos son una ventaja importante respecto a los sensores o aplicaciones móviles.

Procedencia del dato	Tamaño de muestra[1]	Alcance espacial	Precisión espacial	Granularidad temporal	Datos socio-demográficos	Limitaciones prácticas
Sensores (Bluetooth, WiFi)	Bajo (usuarios con sensores activos y zonas donde estén instalados los sensores)	Reducido (solo en la zona donde se instalen los dispositivos)	Muy Alta (metros)	Muy Baja (solo durante la permanencia en la zona de despliegue de los sensores)	Generalmente No	• Necesidad de activación por parte del usuario • Necesidad de despliegue
Red de telefonía móvil	Muy Alto (usuarios con teléfono móvil conectados a la red del operador)	Extenso (cobertura de red)	Media-Alta (decenas o cientos de metros)	Media-Alta (gran variabilidad, depende de los registros que se almacenen)	Algunas veces Sí (cartera de clientes del operador – edad, género, etc.)	• Necesidad de que el dispositivo esté conectado a la red de telefonía móvil
Aplicaciones móviles, GPS	Bajo-Medio (usuarios con la App instalada y opciones de recogida de datos habilitadas)	Extenso (cobertura de red)	Alta-Muy Alta (decenas o pocos cientos de metros)	Baja-Alta (gran variabilidad, depende del sistema de recogida de los datos)	Generalmente No (depende de los permisos e información introducida por el usuario)	• Instalación por parte del usuario • Consumo de batería para la recogida de información

[1]Referido al total de la población presente en un área de estudio (considerando ámbitos urbanos).

Procedencia del dato	Tamaño de muestra[2]	Alcance espacial	Precisión espacial	Granularidad temporal	Datos socio-demográficos	Limitaciones prácticas
Tarjeta inteligente de transporte público	Bajo-Alto (gran variabilidad. Personas con tarjeta de transporte público)	Reducido (solo ámbito dentro del transporte público)	Alta-Muy Alta (nivel de parada o estación)	Muy Baja (solo cuando se accede al transporte público)	Algunas veces Sí (información sobre el tipo de tarjeta)	• Información solo asociada al transporte público
Navegadores de conducción	Bajo-Medio	Extenso (cobertura de red)	Alta-Muy Alta (decenas o pocos cientos de metros)	Baja-Media (solo cuando se circula en vehículo y depende del sistema de recogida de datos)	Generalmente No	• Información solo asociada al transporte en vehículo

Tabla 1. Análisis comparativo de distintas fuentes de datos para estudios de movilidad urbana.

[2]Referido al total de la población presente en un área de estudio (considerando ámbitos urbanos).

1.3 La influencia de la Red Social en la Movilidad Urbana

La movilidad no obligada (aquella asociada a actividades distintas del trabajo o estudio) tiene un peso muy relevante en la movilidad urbana. Un gran número de viajes vienen derivados de la realización de actividades de tipo social, como por ejemplo ir al teatro, salir a cenar, o ir a visitar a un familiar. Sin embargo, los viajes relacionados con actividades sociales han recibido tradicionalmente menos atención que los viajes asociados a otro tipo de propósitos (Carrasco 2008a). Existen evidencias de que las características de la red social influyen de manera determinante en la decisión sobre las actividades sociales a realizar (Axhausen 2005; Arentze & Timmermans 2006; Carrasco & Miller 2006). El lugar donde se realiza la actividad, la frecuencia, la elección del modo de transporte y otras muchas características sobre las actividades y los viajes derivados de las mismas vienen influenciados por la red social. Las encuestas típicamente realizadas para estudios de movilidad (encuestas domiciliarias y encuestas de interceptación, ver sección 1.2.1) no suelen recoger información detallada sobre la red social de las personas, lo que imposibilita la integración de esta información en los procesos de modelado posteriores. Los patrones de movilidad de las personas casi siempre se modelan como un conjunto de decisiones independientes de su red social. Este planteamiento proporciona resultados satisfactorios para las actividades regulares o muy inelásticas como los viajes asociados al trabajo, pero ignora el hecho de que las relaciones con el entorno familiar o las amistades juegan un papel fundamental en otros muchos viajes y actividades de carácter social.

1.3.1 Limitaciones de los estudios de red social y movilidad

Aunque se han realizado avances teóricos significativos para mejorar el entendimiento sobre cómo la red social influye en los patrones de movilidad, la **disponibilidad de datos e información** es, a día de hoy, una de las **mayores limitaciones** de este tipo de estudios. Normalmente, la información sobre red social y movilidad se recoge a través de encuestas diseñadas específicamente para este tipo de estudios. Se suele recoger información personal de carácter socio-demográfico (edad, género, número de miembros del hogar,

profesión, nivel educativo, situación laboral, etc.) e información relacionada con las interacciones con la red social, como por ejemplo información referente a las personas con las cuales se relaciona el encuestado en su tiempo libre, el tipo de actividad (excursiones, ocio nocturno, reuniones de una asociación, etc.) y la frecuencia con la que la realiza, el idioma en el que se comunica con los miembros de la red social, dónde y cómo conoció a sus contactos, etc. Como señalan Van den Berg et al. (2013), son pocos los recursos y esfuerzos que se han dedicado a incorporar la red social en modelos de demanda de transporte. Además, los datos recogidos a través de encuestas presentan limitaciones importantes con respecto al tamaño de muestra (información normalmente solo disponible para unos pocos cientos de personas) y al periodo temporal de análisis (normalmente solo unos pocos días). Por mencionar algunos ejemplos, Carrasco et al. (2008a) llevaron a cabo una encuesta de carácter general sobre un total de 350 personas y, posteriormente, una encuesta de detalle para una sub-muestra de 84 personas. Por otro lado, Van den Berg et al. (2013) diseñaron una encuesta para el análisis de la red social y la movilidad que combinaba un cuestionario general con un diario para anotar la interacción social para cada uno de los días de estudio. Se obtuvieron solamente un total de 747 respuestas, con un ratio de respuesta del 20%.

A la vista de lo comentado anteriormente, parece evidente la **necesidad de buscar nuevas fuentes de datos y desarrollar nuevas metodologías que permitan analizar de manera conjunta la red social y la movilidad**, superando las limitaciones de tamaño de muestra y periodo temporal reducidos de las actuales encuestas.

1.3.2 Nuevas fuentes de datos: retos y oportunidades

Las nuevas fuentes de datos procedentes de Twitter, Facebook o la telefonía móvil, que proporcionan al mismo tiempo información relevante sobre las relaciones sociales y la localización espacial de las personas, brindan la oportunidad de hacer frente a las limitaciones actuales asociadas a la escasa cantidad de datos disponibles. A diferencia de las encuestas, estas nuevas fuentes de datos proporcionan información sobre localización e interacción social de millones de usuarios durante largos períodos de tiempo. En

términos de movilidad, como se ha comentado anteriormente (ver sección 1.2.3.2) y como también señalan otros investigadores (Lane et al. 2010), los datos de telefonía móvil son una de las mejores fuentes de datos para obtener información espacio-temporal durante un largo período de tiempo de un gran número de personas. Adicionalmente, cuando se analizan las redes sociales, los datos de telefonía móvil tienen la ventaja de proporcionar información más relevante sobre las interacciones "cara a cara" que otras fuentes de datos como Twitter o Facebook (Phithakkitnukoon et al. 2012). Por lo tanto, **la telefonía móvil parece ser una de las fuentes de datos más apropiadas para el análisis conjunto de la red social y la movilidad**. No obstante, al igual que sucede con las encuestas, los datos de telefonía móvil también presentan sus propias limitaciones y desventajas (por ejemplo, información socio-demográfica limitada en comparación con las encuestas) que se deberán tener en cuenta a la hora de analizar este tipo de datos.

1.4 El impacto de la Movilidad Urbana en la Salud de las Personas

La contaminación del aire es, actualmente, uno de los principales desafíos a los que se enfrentan las ciudades. Más del 80% de la población que vive en zonas urbanas está expuesta a niveles de contaminación que exceden los límites fijados por la Organización Mundial de la Salud (WHO 2016). Un considerable número de estudios epidemiológicos han demostrado que la exposición a la contaminación atmosférica aumenta el riesgo de sufrir graves enfermedades como el cáncer de pulmón o afecciones respiratorias de carácter crónico (Le Tertre et al. 2002; Omori et al. 2003; Schwartz 2004, etc.). Los contaminantes se generan a partir de una amplia gama de fuentes, incluyendo la industria, el transporte, la agricultura, la gestión de residuos y los hogares. El transporte por carretera desempeña un papel importante en las ciudades, siendo uno de los principales contribuyentes a la generación de óxidos de nitrógeno (NOx) y partículas (PM). Es necesario llevar a cabo políticas efectivas que reduzcan los niveles de contaminación en las ciudades y que protejan a los ciudadanos de la exposición a los mismos.

1.4.1 Limitaciones en la estimación de la exposición a contaminantes

La estimación de la exposición de la población a la contaminación depende tanto de información de concentraciones de contaminantes como de información sobre la presencia de las personas en el territorio. Las estimaciones de concentración de contaminantes suelen basarse en medidas puntuales extraídas de los datos de estaciones de control de calidad del aire o en estimaciones realizadas con modelos más sofisticados que, basándose en los datos de las estaciones de control y otra información (por ejemplo, datos climatológicos), proporcionan una información con un alto nivel de precisión espacial y temporal (Nyhan et al. 2016). Por otra parte, la información sobre presencia de la población suele recopilarse a través de encuestas, datos censales o registros administrativos que, en muchos casos, proporcionan información poco fiable, escasa y/o desactualizada (Briggs et al. 2007). El enfoque más común es utilizar la ubicación del lugar de residencia como una aproximación de la localización de las personas a lo largo del día (Huynh et al. 2006; Boldo et al. 2011; Cesaroni el al. 2013; Brunekreef et al. 2015). El

principal motivo de utilizar esta aproximación es el fácil acceso a este tipo de datos, a través de información censal o registros sobre la residencia de las personas (por ejemplo, datos del padrón de habitantes). Sin embargo, existen varios estudios que han encontrado discrepancias importantes entre los valores de exposición personal y la exposición en el lugar de residencia (Avery et al. 2010). Se ha demostrado que los patrones de movilidad influyen de manera significativa en la exposición de las personas a la contaminación (Beckx et al. 2009, Dons et al. 2011). De cara a introducir los patrones de movilidad en los análisis de exposición a la contaminación, algunos autores proponen la utilización de modelos de simulación basados en actividades (Burke et al. 2001; Hatzopoulou et al.2010; Panis et al. 2010). La principal limitación de estos modelos está asociada a la calidad de la información de la que disponen para realizar sus estimaciones. Esta información suele proceder de encuestas, y como se ha comentado en la sección 1.2.2, las limitaciones prácticas de las encuestas hacen que, en muchos casos, no se disponga de información fiable y actualizada.

Mientras que las técnicas de medición y modelado para la estimación de concentración de contaminantes han experimentado una mejora significativa en las últimas décadas, los métodos para la estimación de presencia de las personas siguen basándose principalmente en información estática, limitada, costosa de recopilar y desactualizada. Una revisión crítica llevada a cabo por el panel de expertos del Health Effects Institute sobre estudios epidemiológicos concluyó que la mayoría de los estudios carecen de información fiable sobre la población expuesta a la contaminación (Health Effects Institute 2009). Para poder **definir medidas efectivas que ayuden a reducir el problema de la exposición a la contaminación**, es necesario disponer de **información sobre los patrones de movilidad de la población**.

1.4.2 Nuevas fuentes de datos: retos y oportunidades

El uso generalizado de dispositivos móviles y la disponibilidad de sensores de detección de bajo coste (por ejemplo, Bluetooth, detectores WiFi) abre nuevas oportunidades para la recogida de información sobre la presencia de la población en las ciudades. Existen varios

estudios que han analizado la presencia de la población y sus patrones de movilidad a partir de datos de sensores de Bluetooth o WiFi (Van Londersele et al. 2009; Versichele et al. 2012; Naini et al.2012). Recientemente Kontokosta & Johson (2017) utilizaron datos procedentes de sensores WiFi para estimar la presencia de población en la zona del bajo Manhattan (Nueva York, USA). Obtienen resultados para distintas horas del día y clasifican a las personas como residentes, trabajadores o visitantes en función de los patrones de actividad registrados en la zona de estudio. Destacan como limitaciones del estudio los problemas de representatividad de la muestra (al no disponer de datos socio-demográficos de los usuarios y no poder asumir una distribución uniforme del uso del WiFi entre los diferentes segmentos de la población), los errores de conexión con los sensores y la no disponibilidad de un ID único del dispositivo (MAC[3]) que imposibilita el análisis longitudinal de los patrones de actividad y movilidad de los usuarios. Otro de los aspectos limitantes de este estudio y de todos los estudios que utilizan sensores es el alcance espacial limitado de la solución y la necesidad de instalación y mantenimiento de dichos sensores en el tiempo. En el caso del estudio de Kontokosta & Johson (2017), se utilizaron 53 puntos de control para toda el área de estudio. Por otra parte, datos procedentes de aplicaciones móviles y servicios web también han sido utilizadas para realizar estudios de presencia de personas, como por ejemplo Twitter (Lenormand et al. 2014; Bassolas et al. 2016) o Panoramio (García-Palomares et al. 2015). Las principales limitaciones asociadas a estos datos son el posible sesgo en los resultados debido a que ciertos segmentos de la población no utilicen ciertas aplicaciones (o que la gente que las utilice presente unas características muy concretas) y la baja granularidad temporal de los datos, que depende del uso y configuración del servicio (Yin et al. 2016). Del mismo modo, en la última década se han llevado a cabo un número considerable de estudios de presencia a partir de datos de telefonía móvil (Ratti et al. 2006; Reades et al. 2007; Terada et al. 2013; Deville et al. 2014). Se han realizado distintos estudios de validación con datos censales que han mostrado la validez de esta fuente de datos para capturar la presencia

[3] Los dispositivos móviles están empezando a instalar software que genera códigos MAC aleatorios por motivos de privacidad y seguridad

de personas (Oyabu et al. 2013; Douglass et al. 2015). Las principales limitaciones asociadas este tipo de datos vienen derivadas de la falta de información socio-demográfica (Dewulf et al. 2016) y la baja resolución espacial de los datos (Oyabu etl al. 2013) que limita el grado de desagregación espacial de los resultados.

En comparación con otras fuentes de datos, los datos de telefonía móvil, a pesar de sus limitaciones, presentan una serie de ventajas importantes como el tamaño de muestra, la granularidad temporal o el alcance espacial que la posicionan como **una de las mejores fuentes de datos para realizar análisis de presencia de población**. No obstante, es necesario tener en cuenta las limitaciones mencionadas anteriormente a la hora de llevar a cabo los análisis e interpretar los resultados.

1.5 Justificación del Estudio de Investigación

El conocimiento de la movilidad urbana es esencial para la gestión y planificación eficientes de las ciudades. Actualmente, la mayoría de las ciudades no dispone de información fiable y actualizada sobre los patrones de movilidad de sus ciudadanos, principalmente por las limitaciones prácticas asociadas a los actuales métodos de recogida de información (elevados costes económicos, largos plazos de ejecución, etc.). Las nuevas tecnologías de recogida pasiva de información geolocalizada brindan la oportunidad de obtener información sobre los patrones de actividad y movilidad de la población superando gran parte de las limitaciones de los métodos tradicionales. En las secciones anteriores de este capítulo se han identificado los datos de telefonía móvil como una de las fuentes de datos más apropiadas para obtener estadísticas de movilidad de un gran número de personas, para analizar las interacciones entre la red social y la movilidad, y para determinar la presencia de la población en las ciudades. Existen estudios previos que han utilizado datos de telefonía móvil para este tipo de aplicaciones (Reades et al. 2007, Gonzalez et al. 2008, Phithakkitnukoon et al. 2012, etc.). Sin embargo, estos estudios aún presentan un gran margen de mejora, especialmente en lo referente a aspectos metodológicos. En los siguientes sub-apartados de esta sección se detallan las principales limitaciones que presentan los estudios previos para cada una de las aplicaciones prácticas mencionadas. Estas limitaciones justifican la necesidad de seguir investigando sobre estas cuestiones.

1.5.1 Extracción de patrones de movilidad a partir de datos de telefonía móvil

Muchas de las aplicaciones de los datos de telefonía móvil se han centrado en el análisis de los patrones de movilidad de la población (González et al. 2008; Song et al. 2010; Calabrese et al. 2013 etc.). A pesar del avance observado en los últimos años, no se dispone a día de hoy de una metodología general y detallada que cubra todo el proceso de análisis de los datos, desde su extracción hasta el cálculo de los indicadores de movilidad expandidos al total de la población. Algunos aspectos metodológicos relevantes que presentan margen de mejora son los siguientes:

- Muchos estudios han obtenido información del lugar de residencia o lugar de trabajo de las personas a partir del análisis longitudinal de los datos de telefonía móvil (Isaacman et al. 2011; Phithakkitnukoon et al. 2012; Calabrese et al. 2013). Sin embargo, ningún estudio ha profundizado en el análisis de **otras localizaciones frecuentes** visitadas por las personas, como por ejemplo actividades deportivas de cierta periodicidad semanal.

- Son pocos los estudios que proporcionan soluciones para el **filtrado de los usuarios de telefonía móvil** que presentan una granularidad temporal baja en sus datos. La consideración de estos usuarios en la muestra final puede llevar a errores importantes en los resultados. Algunos estudios utilizan una muestra de usuarios aleatoria (González et al 2008), otros comprueban que la granularidad temporal media de los datos en la muestra sea suficientemente elevada para el objeto de su estudio (Calabrese et al. 2013) y otros eliminan aquellos usuarios con pocos registros a largo del día (Song et al. 2010a; Schenider et al. 2013). La principal limitación de estas aproximaciones es que no tienen en consideración la variabilidad de la granularidad temporal de los registros a lo largo del día.

- La mayoría de los estudios no inciden sobre la **determinación de la hora del viaje**.

- Muchos estudios no proporcionan metodologías para la **elevación de la muestra** y su posible corrección con datos socio-demográficos.

Por otro lado, son escasos los ejercicios de **comparación con estadísticas oficiales** de movilidad que se han realizado para comprobar el potencial real de los datos de telefonía y, al mismo tiempo, validar las metodologías de análisis propuestas.

1.5.2 Análisis conjunto de la red social y la movilidad

El número de estudios que ha analizado de forma conjunta la red social y la movilidad a partir de datos de telefonía móvil es reducido. Respecto al análisis de los **lugares característicos de la red social**, la mayoría de los estudios se han centrado en el análisis de

la distribución de distancias entre los lugares de residencia de los miembros de la red social y algunos han analizado también la distribución con respecto al lugar de trabajo (por ejemplo, Phithakkitnukoon et al. 2012). Sin embargo, la caracterización de los lugares comúnmente visitados por los miembros de la red social así como la naturaleza de los mismos es un aspecto no explorado aún. Otros estudios también han analizado situaciones de co-ubicación (personas que se encuentran en un mismo lugar al mismo tiempo) de los miembros de la red social. Algunos estudios identifican **co-ubicación** cuando se produce una llamada entre dos personas que se encuentran en la misma zona (siendo esta diferente de sus respectivos lugares de casa y trabajo) como un indicador de coordinación para realizar la actividad social (Calabrese et al. 2011b). Por lo tanto, sólo se detecta co-ubicación si se realiza una llamada y esta se produce en un sitio cercano al del supuesto lugar de encuentro, lo que puede no detectar situaciones en las que no se produce coordinación o la coordinación se produce fuera del lugar de encuentro. Por otro lado, sólo se detecta la co-ubicación entre los miembros que realizan la llamada, y no entre otros posibles acompañantes que acudan al evento. Chen & Mei 2014 consideran que se produce co-ubicación cuando dos personas se encuentran en la misma zona durante un mismo periodo, dividiendo el día en dos periodos, mañana (8 a.m. – 8 p.m) y tarde (8:01 p.m – 7:59 a.m.). La amplia extensión de los periodos analizados hace posible que dos personas que aparecen en dichos periodos en la misma zona realmente no coincidan en el tiempo (por ejemplo, dos hermanos que acuden a casa de sus padres pero uno lo hace de 10 a.m. a 11a.m. mientras que otro lo hace a las 5 p.m. a 6 p.m.). Una mejora significativa respecto al estudio de co-ubicación sería proporcionar información más detallada sobre la ubicación de cada miembro de la red social a lo largo del día, con el objetivo de detectar co-ubicación con una precisión temporal mayor y con independencia de si se producen o no llamadas entre los miembros de la red social.

1.5.3 Análisis de la exposición de la población a la contaminación

Al inicio de esta investigación, a nuestro leal saber y entender, no existía ningún estudio que analizase la exposición de la población a la contaminación a partir de datos de

telefonía móvil. Actualmente, los estudios que abordan este asunto siguen siendo escasos (Dewulf et al. 2016; Nyhan et al. 2016 y Gariazzo et al. 2016). La principal contribución al estudio de la exposición a la contaminación que aportan los datos de telefonía móvil es la mejora en cuanto a la determinación de la presencia de las personas. La mayoría de los estudios de estimación de presencia de población en general (Sterly et al. 2013; Deville et al. 2015), y aquellos aplicados al estudio de exposición a la contaminación en particular, suelen tomar en consideración una o más de las siguientes hipótesis: (1) si no se dispone de información de localización del usuario en un periodo concreto del día, se supone que este no se ha desplazado y se encuentra ubicado en la última localización detectada y (2) la distribución espacio-temporal de los usuarios de telefonía móvil activos es representativa de la distribución real de la población. La validez de estas hipótesis depende en gran medida del alcance del estudio y de las características de los datos de telefonía móvil. La primera hipótesis solamente se podrá considerar valida si la granularidad temporal de los datos es suficientemente elevada o si la zonificación de estudio presenta un alto grado de agregación (cuando la mayoría de las actividades de las personas tienen lugar dentro de cada una las zonas definidas). Respecto a la segunda hipótesis, su validez dependerá de la penetración del operador (u operadores) de red en el mercado de telefonía móvil y de cómo sus clientes se distribuyen entre los diferentes segmentos de la población. Por otro lado, la mayoría de los estudios de presencia suelen plantear un enfoque basado en la torre o celda de telefonía, aforando todos los dispositivos móviles que se conectan a ella. Este planteamiento, al no considerar los patrones de movilidad de las personas, no permite detectar saltos en la red de telefonía (cambio en la conexión del dispositivo móvil de una antena de telefonía a otra sin que el dispositivo se haya desplazado) ni tampoco estimar con precisión el tiempo de estancia del dispositivo en la celda (si el dispositivo está realizando un desplazamiento, permanecerá menos tiempo en dicha área).

1.6 Objetivos y Alcance del Estudio

1.6.1 Objetivos

El objetivo principal de esta investigación es contribuir a los recientes avances en el campo del análisis de los datos de telefonía móvil mediante el desarrollo y validación de una metodología que permita extraer información de patrones de actividad y movilidad de la población en ámbitos urbanos. Los objetivos específicos asociados a dicho desarrollo son los siguientes:

- Definir una metodología que cubra todo el proceso de análisis de los datos (desde su pre-procesado hasta el cálculo de indicadores) y que sea adaptable a datos de telefonía móvil con distintas características espacio-temporales.
- Definir metodologías para la identificación de localizaciones frecuentes distintas del lugar de residencia y del lugar de trabajo.
- Definir un procedimiento para la estimación de la hora del viaje.
- Definir procedimientos para el filtrado y elevación de la muestra.
- Definir indicadores básicos de movilidad a partir de la información extraída.
- Validar la metodología mediante la comparación con encuestas.

Por otro lado, otro de los objetivos principales de esta investigación es aplicar la metodología desarrollada en distintos casos de uso relevantes. En concreto, esta investigación se centra en: (1) la obtención de estadísticas básicas de movilidad y matrices origen-destino en ámbitos urbanos, (2) el análisis de la influencia de la red social en la movilidad y (3) el estudio de la exposición de la población a la contaminación. En la *Tabla 2* se presentan los objetivos específicos asociados a cada uno de estos tres casos prácticos.

Aplicaciones prácticas	Objetivos
Caso práctico 1: obtención de estadísticas básicas de movilidad y matrices origen-destino en ámbitos urbanos	• Calcular indicadores básicos de movilidad y matrices OD a partir de los patrones de actividad y movilidad extraídos de los datos de telefonía móvil • Comparar los resultados obtenidos con información procedente de encuestas • Identificar las ventajas y las limitaciones frente a metodologías tradicionales
Caso práctico 2: análisis de la influencia de la red social en la movilidad	• Desarrollar una metodología para la determinación de la red social de las personas a partir de datos de telefonía móvil • Definir una metodología para identificar los lugares característicos visitados por las personas de una misma red social • Definir una metodología para estimar eventos de co-ubicación de los miembros de una misma red social basados en los patrones de actividad y movilidad obtenidos mediante telefonía móvil • Identificar las ventajas y las limitaciones frente a metodologías tradicionales
Caso práctico 3: estudio de la exposición de la población a la contaminación.	• Definir una metodología para determinar presencia de población basada en los patrones de actividad y movilidad extraídos de los datos de telefonía móvil • Comparar los resultados obtenidos mediante esta metodología con resultados procedentes de métodos convencionales. • Identificar las ventajas y las limitaciones frente a metodologías tradicionales

Tabla 2 Objetivos específicos de las aplicaciones prácticas

1.6.2 Alcance del estudio

El estudio aborda el análisis de datos de la red de telefonía móvil para extraer información sobre los patrones de actividad y movilidad de la población y su aplicación a tres casos de uso concretos: estadísticas de movilidad, red social y movilidad, y exposición a la contaminación. El estudio se centra principalmente en la obtención de estadísticas básicas de movilidad (número de viajes por persona, distancia de los viajes, estimación de los propósitos del viaje, etc.) y en la obtención de matrices origen-destino. Queda fuera del alcance de este estudio la estimación de la información de movilidad asociada al modo y a la ruta de los viajes. Por lo tanto, el foco del estudio es la obtención de información asociada a las dos primeras etapas (generación y distribución) del clásico modelo de transporte de cuatro etapas. Del mismo modo, los planteamientos e hipótesis que aquí se presentan están enfocados al estudio de la movilidad urbana. Metodologías y consideraciones adicionales con respecto al análisis de los viajes de media y larga distancia no están contempladas en este estudio. Por último, también queda fuera del alcance del estudio la definición y discusión sobre las tecnologías hardware y software de Big Data empleadas para la gestión y procesamiento eficiente de los datos de telefonía móvil.

CAPÍTULO II: ESTADO DEL ARTE

En este capítulo se presenta una revisión del estado del arte de los aspectos más relevantes relacionados con el objeto de la presente investigación. En primer lugar, se realiza una revisión de las características de los datos de la red de telefonía móvil disponibles para llevar a cabo estudios de movilidad. En segundo lugar, se realiza una revisión de las metodologías existentes para la extracción de patrones de actividad y movilidad a partir de los datos de telefonía móvil. En tercer lugar, se realiza una revisión de los estudios relacionados con el análisis de la influencia de la red social en los patrones de movilidad, con especial interés en aquellos estudios que utilizan datos de telefonía móvil. Por último, se realiza una revisión del estado del arte sobre la determinación de presencia de población a partir de datos de telefonía móvil, con especial atención a su aplicación en estudios de exposición de la población a la contaminación.

2.1 Descripción de los Datos de la Red de Telefonía Móvil

En este apartado se presenta, en primer lugar, una descripción general de la estructura de las redes de telefonía móvil. Posteriormente, se describen los datos producidos por las interacciones entre los dispositivos móviles y la red de telefonía y se presentan otro tipo de datos, disponibles por el operador de red, relevantes para los estudios de movilidad. En este estudio clasificamos los datos de telefonía móvil en tres categorías principales:

- **Datos de eventos**: datos asociados a los registros de comunicación del dispositivo con la red de telefonía móvil. Estos eventos pueden ser activos (p. ej. llamadas o SMS) o pasivos (actualización periódica de la posición del dispositivo).

- **Datos de la red de telefonía móvil**: datos sobre la infraestructura de la red, tales como la localización de las torres de comunicación o tecnología de las antenas.

- **Datos socio-demográficos**: información socio-demográfica de los clientes del operador de telefonía móvil, como por ejemplo datos de edad o género.

2.1.1 La estructura de las redes de telefonía móvil

Una red de telefonía móvil es aquella que permite la comunicación entre los distintos dispositivos móviles conectados a la misma. En las redes de telefonía móvil la cobertura geográfica viene proporcionada por un conjunto de torres distribuidas a lo largo del territorio. Cada torre dispone de un conjunto de antenas de una a varias tecnologías distintas. El área de cobertura de cada torre se divide generalmente en tres sectores de 120 grados cada uno. Cada sector puede contener a su vez múltiples celdas, que proporcionan cobertura a una misma área geográfica pero con distintas frecuencias. No obstante, en muchos casos, la distinción entre celdas y sectores no se realizada y ambos términos se utilizan indistintamente. La densidad de las celdas viene determinada por requisitos de capacidad de la red en un área determinada. En áreas densamente pobladas, las celdas de telefonía móvil son más pequeñas, estando las torres de comunicación separadas unos pocos cientos de metros. Por otro lado, en áreas con menor densidad de población (por ejemplo, zonas rurales), las celdas son de mayor tamaño debido a la menor necesidad de capacidad.

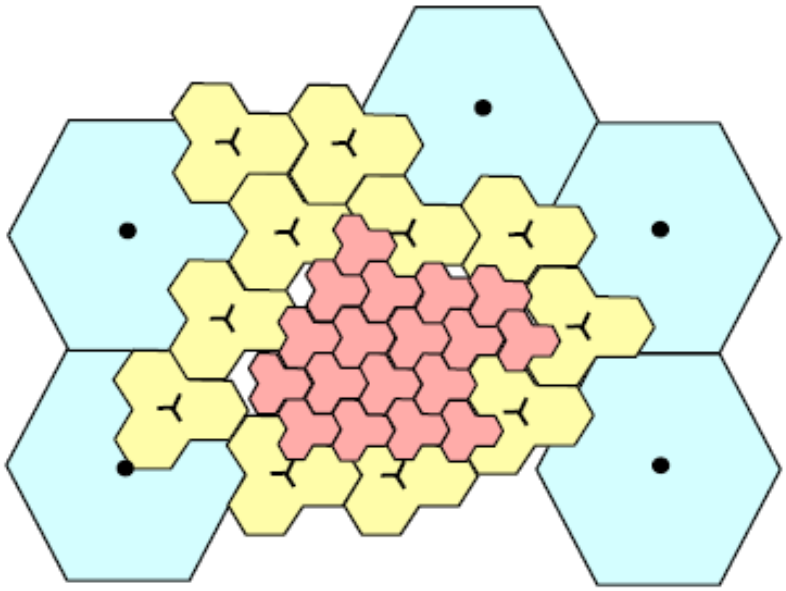


Figura 3. Ejemplo de distribución de densidad de celdas en una zona urbana, con mayor densidad en el centro de la ciudad y menor densidad en los alrededores (Ricciato et al. 2015)

2.1.2 Descripción de los datos asociados a eventos de red

Los registros asociados a la comunicación con la red (de ahora en adelante, eventos) son datos correspondientes a la interacción del dispositivo móvil con la red de telefonía. El dispositivo está continuamente comunicándose con la red de telefonía móvil, intercambiando una gran cantidad de información. Sin embargo, no toda la información es almacenada. Por motivos asociados a la facturación del servicio, siempre se almacenan los denominados "*Call Detail Records*" (CDRs). Estos registros se almacenan cada vez que el usuario realiza o recibe una llamada, cada vez que recibe o envía un SMS o cada vez que comienza una sesión de datos. En función de la tecnología que se emplee para la extracción de los datos, los CDRs pueden presentar un formato muy distinto e información muy variada, pero incluyen siempre una serie de elementos comunes. Por ejemplo, para el caso de una llamada telefónica, suelen contener, al menos, información sobre el número de teléfono que realiza la llamada, el número de teléfono que recibe la llamada, la fecha y hora de inicio de la llamada, la duración, el ID de celda a la cual se conecta el dispositivo, el tipo de llamada (p.ej. entrante, saliente), identificador del registro, etc.

2.1.2.1 Resolución espacial de los eventos

Los CDRs proporcionan información sobre el ID de la celda donde se produce la comunicación del dispositivo móvil con la red de telefonía. Este dato se puede cruzar con la información sobre las características de la red de torres y antenas del operador (ver apartado 3.1.3) para estimar la localización del usuario. Esta aproximación suele proporcionar una resolución espacial de decenas o cientos de metros en entornos urbanos (alta densidad de población) y de varios kilómetros en entornos rurales (baja densidad de población).

En algunas ocasiones, los operadores utilizan tecnologías de triangulación de señales para proporcionar información sobre la latitud y longitud aproximada del dispositivo móvil, presentando esta información una incertidumbre espacial variable en función de la tecnología empleada. Calabrese et al. (2013), por ejemplo, utilizan datos triangulados

generados por la empresa Airsage (www.airsage.com) reportando una precisión espacial media de 320 metros.

2.1.2.2 Granularidad temporal de los eventos

En función de la tipología de datos disponibles, se dispondrá de una mayor o menor granularidad temporal. Los CDRs asociados a llamadas o SMS presentan una granularidad temporal mucho más baja que los CDRs que contienen información de sesiones de datos. No obstante, los CDRs de llamadas y SMS aportan una información muy relevante con respecto a las interacciones personales entre los usuarios (red social), que no proporcionan otro tipo de datos. La granularidad temporal de los CDRs que contienen sesiones de datos suele ser bastante elevada, ya que los propios dispositivos móviles interactúan con la red automáticamente, sin que haya intervención directa por parte del usuario (por ejemplo, mediante las actualizaciones periódicas del software del dispositivo). La granularidad de los datos depende de múltiples factores, como el sistema de recogida de información de los CDRs, el tipo y configuración del dispositivo móvil o el uso del dispositivo por parte del usuario, siendo esta granularidad variable dentro de la muestra de usuarios de telefonía móvil. Algunos estudios han analizado la granularidad temporal de los datos de CDRs, obteniendo un tiempo medio entre registros de 500 minutos para el caso de CDRs de llamadas y SMS (Gonzalez et al. 2008) y alrededor de 260-340 minutos para el caso de CDRs con información de sesiones de datos (Calabrese et al. 2011a; Holleczek et al. 2014). No obstante, como se ha comentado anteriormente, la granularidad temporal de los datos varía significativamente de unos usuarios a otros; presentando, normalmente, valores muy inferiores para un número significativo de usuarios. En la *Figura 4* se muestra un ejemplo de la distribución del tiempo entre registros para una muestra de usuarios de telefonía móvil.

En los últimos años, debido al potencial que están mostrando los datos de telefonía móvil para su aplicación práctica en distintos sectores (transporte, turismo, geomarketing, etc.), algunos operadores de red están empezando a almacenar datos de eventos no necesariamente relacionados con eventos de facturación, como por ejemplo, información

sobre la localización del dispositivo móvil cada cierto tiempo, con el objetivo principal de mejorar la granularidad temporal de los datos.

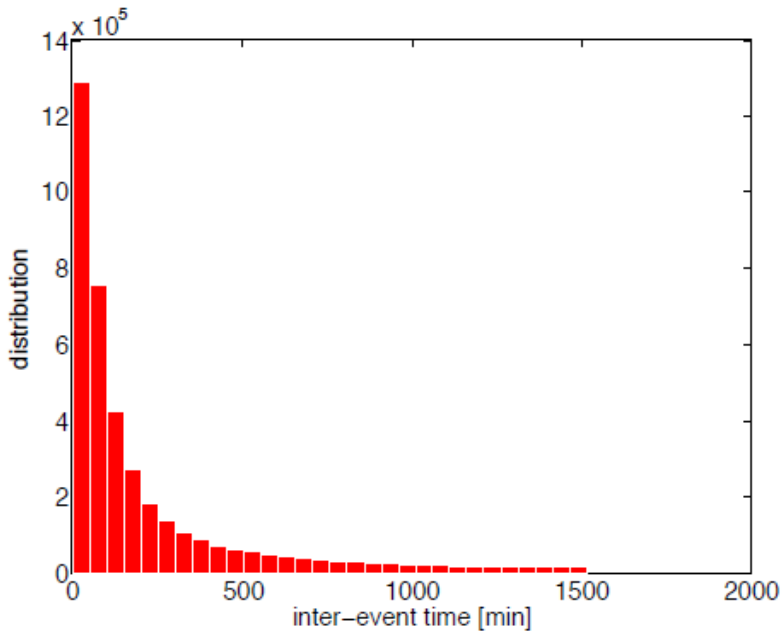


Figura 4. Distribución del tiempo entre eventos consecutivos para CDRs con sesiones de datos. (media = 320 min., 1er cuartil = 41 min., mediana = 114 min., 3er cuartil = 406 min.) (Holleczek et al. 2014)

2.1.3 Datos de la red de telefonía móvil

Los datos de la red de telefonía móvil proporcionan información sobre las características y localización de las torres de comunicación y las antenas. Al igual que sucede con los CDRs, el formato e información que contienen estos registros es muy variable en función del operador que gestione la red. Normalmente, contienen información sobre los distintos identificadores de celda, la dirección y coordenadas del emplazamiento de las antenas, el tipo de tecnología (2G, 3G, etc.), el azimut de la antena, la altura del emplazamiento con respecto al nivel del mar, etc.

Esta información se utiliza para estimar la localización del dispositivo móvil en el caso (el más habitual, al menos actualmente) de que no se disponga de información triangulada asociada a los eventos de los CDRs.

2.1.4 Datos socio-demográficos

Los datos socio-demográficos son datos que aportan información sobre los usuarios de los dispositivos móviles. Normalmente, sólo se dispone de información socio-demográfica en el caso de que el usuario tenga un contrato de post-pago con el operador de la red. La información que se proporciona generalmente está asociada al titular del contrato, que en algunos casos puede ser distinto del usuario del teléfono móvil. Es importante tener en cuenta este aspecto al utilizar esta información en análisis posteriores. La información asociada a estos registros puede contener información sobre la edad, género, código postal de residencia, profesión, tipo de contrato, facturación, etc. En muchos casos, por motivos de privacidad, esta información se facilita de manera categorizada (p.ej. rango de edad), sólo se facilita información muy restringida o directamente no se facilita dicha información.

2.2 Patrones de Actividad y Movilidad a partir de Datos de Telefonía

Estudios recientes en el ámbito de la investigación humana y social han demostrado la utilidad de los datos de telefonía móvil para el estudio de los patrones de actividad y movilidad de las personas. González et al. (2008), Song et al. (2010a, 2010b) y Bagrow & Lin (2012) han demostrado que la movilidad humana es altamente estructurada y está gobernada por patrones certeros. Silm & Ahas (2010) utilizaron información geolocalizada procedente de la telefonía móvil para identificar la residencia de las personas en Estonia. De forma similar, Isaacman et al. (2011) utilizaron los datos de telefonía móvil para identificar los lugares donde las personas permanecían más tiempo. Validaron los algoritmos utilizados contrastando los resultados obtenidos mediante la información veraz proporcionada por voluntarios. Los algoritmos identificaban casa y trabajo con un error medio menor a una milla. Becker et al. (2011) identificaron los lugares de residencia de los trabajadores en la ciudad de Morristown (New Yersey, USA) con el objetivo de analizar los flujos de entrada y salida de la ciudad. Ahas et al. (2010) también monitorizaron los desplazamientos por motivo laboral de los residentes del área periférica de la ciudad de Tallin. Song et al. (2010b) estudiaron la predictibilidad de la movilidad de la población a través de datos procedentes de GSM. Do & Gatica-Pérez (2012) desarrollaron algoritmos para predecir la movilidad de las personas usando diferentes tipos de datos procedentes de dispositivos móviles (GPS, WiFi APs, registros de llamadas, etc.) recogidos de un grupo de 153 voluntarios durante 17 meses. Por otro lado, los datos de posicionamiento procedentes de la telefonía móvil también se han utilizado para estudiar cómo se mueve la gente durante eventos sociales (Calabrese et al. 2010). En el sector del transporte, el interés en relación al uso de los datos de telefonía móvil se ha centrado en la estimación de tiempos de viaje o velocidades de recorrido (Bar-Gea 2007), reparto modal (Wang et al.2010; Doyle et al. 2011), matrices origen-destino (White and Wells 2002; Cáceres et al. 2007; Sohn and Kim 2008; Calabrese et al. 2011a) y estudios de intensidad de tráfico (Cáceres et al. 2012) entre otros aspectos. Revisiones sobre el estado del arte en el uso de datos de telefonía móvil pueden encontrarse en Yim (2003), Rose (2006), Cáceres et al. (2008) y Steenbruggen et al. (2011).

2.2.1 Métodos para la extracción de patrones de actividad y movilidad

En este apartado se presenta una revisión de las metodologías empleadas para la extracción de patrones de actividad y movilidad a partir de datos de telefonía móvil[4]. Tras una revisión de diferentes estudios (Gonzalez et al. 2008; Bayir et al. 2010; Calabrese et al. 2011a; Isaacman et al. 2011; Calabrese et al. 2013; Lenormand et al. 2014; Çolak et al. 2015, etc.), se han identificado las principales aproximaciones al problema. Los pasos más relevantes del proceso de extracción de patrones de movilidad a partir de datos de telefonía móvil pueden clasificarse en los siguientes grupos:

- Estimación de la localización del dispositivo móvil
- Depuración de errores en el posicionamiento
- Identificación de localizaciones frecuentes
- Determinación de estancias y viajes
- Procesos de depuración y elevación muestral

2.2.2 Estimación de la localización del dispositivo móvil

Existen principalmente dos tipologías de datos que aportan información sobre el posicionamiento del dispositivo móvil: (1) datos de la celda a la cual se conecta el dispositivo y (2) datos de posicionamiento estimado (latitud, longitud) del dispositivo, obtenidos mediante métodos de triangulación de señales. En función de la tecnología de extracción de datos empleada por el operador de la red, se dispondrá de una u otra información. En el segundo caso, la obtención del posicionamiento estimado del dispositivo es directa, y no requiere ningún procesamiento posterior. Sin embargo, en el primer caso (datos de celdas), existen diferentes aproximaciones al problema.

2.2.2.1 Estimación del posicionamiento a partir de datos de la red de telefonía

En los casos en los que no está disponible la información triangulada de los dispositivos móviles, se dispone de información sobre el ID de la celda a la cual se conecta el

[4] Este apartado ha sido revisado y actualizado a lo largo de la investigación con el objetivo de recoger los últimos avances metodológicos

dispositivo móvil. Adicionalmente, se suele disponer de información sobre la ubicación de la torre (Base Transceiver Station – BTS) donde se encuentran las antenas que dan cobertura a los diferentes sectores o celdas. En algunos casos, también se dispone de cierta información complementaria como el azimut asociado las antenas.

Algunos estudios consideran como estimación de la localización del dispositivo móvil la propia ubicación de la BTS (p.ej. Iovan et al. 2013); sin embargo, otros estudios (p. ej. Dewulf et al. 2016) estiman el área de cobertura asociada a la BTS. La aproximación más empleada para estimar el área de cobertura es utilizar áreas de Voronoi. El área de Voronoi asociada a una torre está definida por los puntos del plano que están más cerca de dicha torre que de cualquier otra. Esta aproximación está basada en la hipótesis de que el dispositivo móvil tiende a conectarse a la torre más cercana. En la *Figura 5* se muestra un ejemplo comparativo entre la cobertura real y la cobertura estimada mediante áreas de Voronoi. Una vez definida el área de Voronoi, se puede utilizar como localización aproximada del dispositivo móvil cualquier punto dentro de dicha área, como puede ser su centroide (habitualmente empleado) o cualquier otro punto asignado con algún criterio adicional (p. ej. datos de la red de transporte).

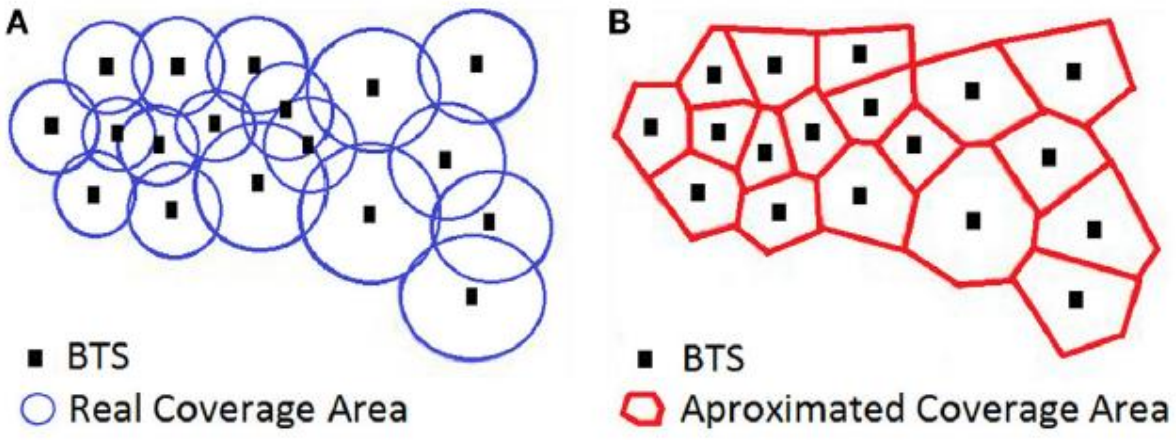


Figura 5. Ejemplo comparativo entre la cobertura real asociada a una BTS (caso A) y el área de cobertura estimada mediante la aproximación de áreas de Voronoi (Fuente: Oliver et al. 2015)

2.2.3 Depuración de errores de posicionamiento

Una zona del territorio, dentro de la red de telefonía móvil del operador, puede estar cubierta por varias celdas de telefonía móvil. Por lo tanto, un dispositivo móvil puede recibir al mismo tiempo señales de distintas antenas. El dispositivo móvil tenderá a conectarse a la antena que le proporcione una mejor cobertura en cada momento, por lo que variaciones del estado de la red (potencia, capacidad, etc.) pueden hacer que el dispositivo se conecte indistintamente a una u otra celda sin necesidad de que el dispositivo haya modificado su ubicación. Este hecho puede llevar a la confusión de considerar desplazamientos cuando en realidad éstos no se producen. Para evitar este problema, se utilizan filtros para eliminar datos referentes a las denominadas oscilaciones o saltos de señales. Existen distintos métodos al respecto, que se pueden dividir en dos grupos: aquellos basados en el análisis de las interacciones entre las señales (p.ej. Bayir et al. 2010, Iovan et al. 2013) y aquellos basados en métodos de clusterización (p.ej. Jiang et al. 2013, Alexander et al. 2015). Los primeros identifican saltos en la red si la velocidad de cambio entre antenas es mayor a un cierto umbral o si el número de cambios entre antenas es muy alto. El otro planteamiento (generación de clusters) se basa en agrupar señales basándose en criterios de distancia espacial y diferencia temporal.

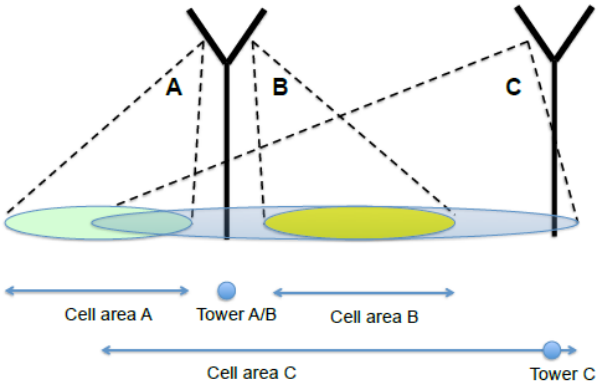


Figura 6. Ejemplo de solapamiento entre celdas (Ricciato et al. 2015)

2.2.4 Identificación de localizaciones frecuentes

De cara a caracterizar los lugares frecuentemente visitados por las personas, se realiza un análisis longitudinal de los datos de telefonía durante varios días. Este tipo de estudios se centra principalmente en la identificación del lugar de residencia y el lugar de trabajo de las personas. Existen numerosos estudios en los cuales se realizan este tipo de análisis (p.ej. Isaacman et al. 2011, Calabrese et al. 2013, Lenormand et al. 2014, Çolak et al. 2015) y en prácticamente la totalidad de los casos, la aproximación es la misma con algunas pequeñas modificaciones. Se suele considerar como el lugar de residencia la localización más visitada entre las 'A' p.m. y las 'B' a.m. y como lugar de trabajo la localización más visitada entre las 'C' a.m. y las 'D' p.m. Normalmente, para la estimación del lugar de residencia el análisis se realiza sobre días laborables y/o fines de semana, mientras que para el caso de identificación de trabajo, siempre se utilizan días laborables. Los valores de A, B, C y D son escogidos en función de los patrones generales de comportamiento de la población en el área de estudio. Por ejemplo, Calabrese et. al (2013) toman como valores de 'A' y 'B', 6 y 8 respectivamente; Lenormand et al. (2014) toman valores de A=8, B=7, C=9 y D=5 y Çolak et. al (2015) toman valores de A=7, B=8, C=9, D=6. Adicionalmente, algunos estudios también incluyen la variable de distancia entre el lugar de trabajo y el lugar de residencia para determinar el lugar de trabajo (Alexander et al. 2015). Estos estudios consideran que el lugar de trabajo es la actividad más visitada, dentro del periodo temporal de trabajo, que se encuentra a mayor distancia del lugar de residencia.

2.2.5 Determinación de estancias, actividades y viajes

El proceso de determinación de la cadena de viajes suele dividirse en dos fases: identificación de estancias e identificación de viajes. La principal hipótesis es que los lugares en los que las personas permanecen un cierto tiempo (estancia) son candidatos a lugares donde se están realizando actividades. Una vez identificadas las actividades, los viajes se definen como el desplazamiento entre actividades consecutivas.

Para identificar una estancia, se debe medir el tiempo que una persona permanece en una misma localización. Si ese tiempo es mayor a un cierto umbral, entonces se considera que la persona está realizando una actividad en la zona. No existe un acuerdo sobre qué umbral de tiempo es más conveniente utilizar. Algunos autores (Jiang. et al. 2013, Çolak et. al 2015) proponen utilizar un tiempo de 10 minutos mientras que otros (Holleczek et al. 2014) proponen 20 minutos. Una vez se han identificado las actividades, el viaje se determina como el desplazamiento entre dos actividades consecutivas. La principal incertidumbre en la determinación del viaje es la estimación de la hora de inicio y final del viaje. A partir de los CDRs se dispone de la última hora detectada en origen y la primera detectada en destino, lo que determina la mínima hora de inicio del viaje y la máxima hora de llegada del viaje respectivamente. Para determinar la hora de inicio del viaje, algunos estudios utilizan una función de probabilidad basada en estadísticas oficiales sobre distribución horaria de viajes (Alexander et al. 2015). Otros estudios, en cambio, proponen realizar estimaciones de tiempos de viaje para reducir el intervalo de incertidumbre y después asignar como hora del viaje la hora intermedia dentro del nuevo intervalo de horas posibles (Widhalm et al. 2015).

2.2.6 Procesos de depuración y elevación de la muestra

Dado que la granularidad temporal de los datos de telefonía móvil es muy variable de unos usuarios a otros, no todos los usuarios aportan información con el mismo nivel de detalle. Algunos estudios proponen utilizar solamente una muestra de usuarios con un alto número de registros (Song et al. 2010a; Onnela et al. 2011). Schneider et al. (2013) proponen dividir el día en intervalos de 30 minutos y descartar aquellos usuarios que no tengan información en al menos 8 intervalos. Otros estudios también proponen descartar usuarios con poca información asociada a sus lugares de casa y/o trabajo (Çolak et al. 2015, Alexander et al. 2015).

Respecto al proceso de elevación muestral, estudios previos proponen aplicar un factor de elevación basado en la relación entre la muestra de usuarios de telefonía y la población censal asociada a dicha muestra (Alexander et al. 2015). Otros estudios también proponen

utilizar un factor basado en población pero segmentando según las características socio-demográficas de la población (Terada et al. 2013).

2.3 Análisis de la Influencia de la Red Social en la Movilidad

2.3.1 La red social y los patrones de movilidad

Estudios previos han demostrado que las características de la red social influyen en las actividades sociales y en los viajes derivados de las mismas (Axhausen 2005; Arentze & Timmermans 2006; Carrasco & Miller 2006). Hay un número creciente de estudios que están empezando a incluir las redes sociales como un factor importante para mejorar los modelos de demanda de transporte. Las aplicaciones de la red social en estudios de planificación y patrones de movilidad datan de principios del presente milenio. Dugundji & Walker (2005) obtuvieron un modelo de elección de modo basado en datos de red social. Paez & Scott (2007) presentaron un enfoque similar para estimar la influencia de la red social en el porcentaje de teletrabajo de las empresas. Carrasco & Miller (2006) incluyeron de manera explícita información de la red social en un modelo conceptual basado en actividades para analizar los patrones de viaje. Marchal & Nagel (2006) presentan un modelo de simulación basado en agentes en el que se permite la compartición de información sobre localizaciones de actividades y sobre otros agentes, con el objetivo de optimizar la cadena de viajes. Arentze & Timmermans (2006) presentan un modelo de microsimulación basado en agentes que produce una red social dinámica que evoluciona en función de los patrones de actividad y movilidad de los agentes. Hackney et al. (2006) también estudiaron interdependencias entre la red social y los patrones de movilidad. Silvis et al. (2006) encontraron relaciones entre el número de viajes y los lugares visitados, y el tamaño de la red social y el número de contactos repetidos. Molin et al. (2007) analizan la influencia del tamaño y la composición de la red social sobre la demanda de transporte. Arentze & Timmermans (2008) analizan los efectos directos de la red social en los patrones de actividad y movilidad. Carrasco et al. (2008b) estudian la distribución espacial de las actividades sociales centrándose en el análisis de la distancia de dichas

actividades al lugar de residencia de las personas. Carrasco et al. (2008c) exploran la relación entre los patrones de movilidad, el uso de las TIC y las redes sociales. Carrasco & Miller (2009) estudian los efectos de las características de los individuos de la red social y las interacciones entre ellos en la frecuencia de determinadas actividades. Hackney & Marchal (2009) desarrollaron un modelo de microsimulación que incorporaba la red social dentro del diario de actividades. Más recientemente, Hackney & Marchal (2011) y Ronald et al. (2012a, 2012b) han tenido en cuenta el papel de la red social en los patrones de movilidad utilizando un modelo basado en agentes. Habib and Carrasco (2011) analizan los efectos de la red social con respecto a la hora en la cual suceden las actividades y la duración de las mismas. Van den Berg et al. (2013) estudian los efectos de la red social y las telecomunicaciones en los patrones de actividad y movilidad de las personas. Moore et al. (2013) estudian las relaciones entre la red social, el tiempo empleado y la localización geográfica de las personas. Sharmeen et al. (2013, 2014) analizan las interacciones sociales cara a cara junto con la accesibilidad geográfica.

2.3.2 Análisis conjunto de la red social y la movilidad a través de telefonía móvil

Un número considerable de estudios han utilizado datos de telefonía móvil para analizar la red social o los patrones de movilidad de manera aislada. Sin embargo, los estudios que han utilizado datos de telefonía móvil para analizar conjuntamente las redes sociales y los patrones de movilidad son escasos. Phithakkitnukoon et al. (2011) identificaron los lugares de residencia de los individuos con datos de telefonía móvil y cuantificaron la fuerza de los lazos sociales basándose en la duración de las llamadas. Encontraron que el cambio de residencia puede afectar a los lazos sociales a medida que pasa el tiempo: los lazos fuertes persisten después del cambio de residencia, mientras que los lazos débiles tienden a desaparecer. En un estudio posterior (Phithakkitnukoon et al., 2012), los autores encontraron que el 80% de las señales de telefonía móvil de los individuos se encontraban a menos de 20 km del lugar de residencia de sus contactos sociales más relevantes. Calabrese et al. (2011b) utilizaron un subconjunto de datos de telefonía móvil procedentes de un dataset de un millón de usuarios de telefonía móvil de Portugal para

estudiar la relación entre sus patrones de comunicación y sus posiciones geográficas. Encontraron que había una correlación positiva fuerte entre la frecuencia de las llamadas entre dos individuos y la frecuencia de situaciones de co-ubicación. Cho et al. (2011) estudiaron los viajes asociados a actividades sociales utilizando datos de telefonía móvil y datos de localización extraídos de dos redes sociales de Internet. Ythier et al. (2013) utilizaron datos de llamadas telefónicas, registros de SMS y GPS de 111 personas para investigar la influencia de la comunicación y los contactos sociales en los patrones de viajes. Encontraron que las personas socialmente conectadas tienden a viajar de una manera similar, resultado consistente con estudios previos sobre red social y movilidad. Chen y Mei (2014) identificaron los lazos sociales y caracterizaron los patrones básicos de movilidad utilizando un conjunto de datos de telefonía móvil de alrededor de 425.000 usuarios con información de ubicación y llamadas para una gran ciudad urbanizada en China.

2.4 Análisis de la Exposición de la Población a la Contaminación

2.4.1 Presencia de población mediante telefonía móvil

El estudio de presencia de la población fue una de las primeras aplicaciones prácticas de los datos de telefonía móvil. Ratti et al. (2006) generaron mapas de intensidad de la actividad humana y su evolución a lo largo del día para la ciudad de Milan basándose en datos de tráfico de la red de telefonía móvil (Erlangs). Reades et al. (2007) utilizan información de Erlangs para caracterizar los patrones de actividad de distintos puntos de interés en la ciudad de Roma. Calabrese et al. (2011c) utilizan datos de tráfico de telefonía móvil junto con otras fuentes de datos, como datos de posicionamiento de autobuses y taxis, para llevar a cabo una monitorización de la actividad de la población en Roma. Las limitaciones asociadas a los primeros estudios estaban relacionadas con el tipo de datos de telefonía móvil disponible, ya que normalmente se disponía de datos de Erlangs. El Erlang es una unidad de medida adimensional donde 1 Erlang puede equivaler a una llamada de una hora o dos llamadas de media hora y así sucesivamente. Con esta

información se puede estimar dónde se produce más actividad telefónica y dónde puede haber más población, pero estimar el número de personas con cierta precisión no es posible. Extracciones de datos de telefonía móvil con información asociada al dispositivo móvil permitieron en posteriores estudios avanzar en la mejora de las estimaciones de presencia de población. Terada et al. (2013) proponen un método compuesto de tres pasos principales para estimar la presencia de la población: (1) estimar el número de dispositivos móviles servidos por cada una de las celdas de telefonía presentes en la red, (2) elevar la muestra de usuarios al total de la población y (3) transformar la información de celdas a zonas de estudio. Sterly et al. (2013) estiman la densidad de población en la región de Costa de Marfil y comparan los resultados con datos de población de fuentes oficiales. Concluyen que la hipótesis adoptada de suponer que los clientes del operador se reparten homogéneamente entre la población no es válida en este caso. Oyabu et al. (2013) estiman los lugares de residencia de la población mediante telefonía móvil y comparan los resultados con datos censales para distintos niveles de agregación. Observan que con niveles de agregación elevados, los resultados son satisfactorios. Para niveles de desagregación elevados, identifican que otras variables como la densidad de población influyen en la calidad de los resultados. Deville et al. (2014) comparan la capacidad de los datos de telefonía móvil para proporcionar información detallada sobre distribución de población con otros métodos de teledetección o uso de datos geoespaciales. Determinan la población en las distintas unidades censales identificando los usuarios de telefonía móvil conectados a cada torre y repartiendo la muestra en función del área de intersección entre el área de Voronoi asociada a la torre y la sección censal. Douglass et al. (2015) infieren información sobre el número de personas en una zona a partir de la actividad telefónica (número de llamadas, SMS, etc.) registrados en dicha zona mediante la expresión '*log wi = b + α log pi'* siendo 'wi' la actividad telefónica en la zona 'i' y 'pi' la población en la zona 'i'. Los parámetros '*b' y 'α'* son parámetros a calibrar. Concluyen que, para áreas densamente pobladas, la estimación de población mediante el modelo propuesto proporciona buenos resultados.

2.4.2 Exposición a la contaminación mediante telefonía móvil

En apartados anteriores se han mostrado distintos ejemplos de estudios que han utilizado los datos de telefonía móvil para el estudio de los patrones de movilidad (González et al. 2008; Song et al. 2010a; Bagrow & Lin 2012), para obtener información de demanda de transporte (White and Wells 2002; Cáceres et al. 2007; Sohn and Kim 2008; Calabrese et al. 2011a) o para analizar la influencia de la red social en la movilidad (Cho et al. 2011; Phithakkitnukoon et al., 2012; Ythier et al. 2013). Del mismo modo, otras áreas de investigación como el turismo (Ahas et al., 2007; Ahas et al.,2008; Eurostat, 2014), la prevención de desastres naturales (Bengtsson et al., 2011) o las áreas de estudios socio-económicos (Eagle et al. 2010, Soto et al.,2011) están empezando a utilizar datos de telefonía para sus estudios. Sin embargo, la aplicación de los datos de telefonía móvil para el análisis de la exposición de la población a la contaminación es un área prácticamente inexplorada. Sólo existen unos pocos estudios publicados recientemente relacionados con esta aplicación.

Dewulf et al. (2016) calculan la exposición diaria a NO_2 utilizando datos de telefonía móvil de aproximadamente 5 millones de personas en Bélgica. Los datos de telefonía móvil se recogen de sondas instaladas en la red móvil, almacenando datos de todos los usuarios activos. Los datos se recopilan de diferentes eventos de red, tales como llamadas, mensajes de texto y sesiones de datos. Además, la información se recoge cada 3 horas cuando no se detecta actividad del teléfono móvil. Se calcula un indicador de presencia de población en intervalos de 15 minutos. Para aquellos usuarios en los que la información de ubicación no está disponible dentro del intervalo de 15 minutos, se considera que el usuario permanece en la ubicación anterior. El estudio estima la exposición individual a la contaminación del aire considerando dos enfoques diferentes: un enfoque estático y un enfoque dinámico. El enfoque estático considera que el usuario permanece en casa a lo largo del día y el enfoque dinámico toma en consideración los patrones de movilidad de las personas. La ubicación del lugar de residencia se estima como la ubicación del usuario a las 4:00 de la mañana en un día específico. Los resultados muestran, en promedio, un

aumento en la exposición a NO_2 si se tienen en cuenta los patrones de movilidad de las personas. Se presentan como limitaciones del estudio la escasa cantidad de datos disponibles (sólo dos días), la falta de variables sociodemográficas asociadas a los datos de telefonía móvil, la falta de información semántica (por ejemplo, el propósito del viaje) y las limitaciones metodológicas a la hora de estimar la ubicación del hogar.

Nyhan et al. (2016) evalúan la exposición de la población a la contaminación en la ciudad de Nueva York utilizando información de presencia de población estimada a partir de datos de telefonía móvil y datos espacio-temporales de niveles de concentración de PM2.5. Se utilizaron datos de tráfico móvil (específicamente datos de tráfico 3G) que incluían llamadas telefónicas, SMS y solicitudes pasivas de datos (por ejemplo, aplicaciones que se ejecutan en segundo plano) de varios operadores para estimar el porcentaje de población respecto del total presente en cada distrito por hora del día. Se utiliza como principal hipótesis que la distribución espacial de los usuarios de telefonía móvil es un buen proxy de la distribución espacial del total de la población. Los resultados obtenidos con este enfoque dinámico se comparan con los obtenidos suponiendo una distribución estática de la población basada en datos del censo. Los resultados muestran una diferencia estadísticamente significativa ($p < 0,05$) entre el enfoque estático y el dinámico en la mayoría de los distritos. Una de las limitaciones destacadas por el estudio es el sesgo potencial en los patrones de movilidad de determinados grupos de población, particularmente aquellos grupos de población que tienen menos probabilidades de utilizar dispositivos móviles (niños y ancianos).

Gariazzo et al. (2016) llevaron a cabo una evaluación dinámica de la exposición a la contaminación en la ciudad de Roma utilizando datos de telefonía móvil y concentraciones de contaminantes obtenidas mediante modelado (NO2, O3 y PM2.5). Se utilizaron datos de telefonía móvil proporcionados en el marco del TIM BIGDATA Challenge 2015. El conjunto de datos comprende el post-procesamiento de datos de tráfico de telecomunicaciones (por ejemplo, llamadas, SMS e Internet) agregados en una cuadrícula. Para cada celda de la cuadrícula se proporciona una estimación del número de

usuarios de telefonía móvil cada 15 minutos. Se considera (al igual que en Dewulf et al. 2016) que si no hay información del usuario disponible en 15 minutos, el usuario permanece en la ubicación anterior detectada. Al igual que Nyhan et al. (2016) se supone que la distribución espacial de los usuarios de telefonía móvil es un buen indicador del porcentaje de población que está presente en cada área. Del mismo modo, se disponía de información de presencia de población que incluía datos de edad y género, pero obtenidos de un conjunto de datos diferente al anteriormente mencionado (sólo llamadas salientes), no permitiendo el análisis conjunto de ambos datasets. Los resultados se compararon con el enfoque estacionario basado en datos del Censo Nacional, detectando subestimaciones significativas de la cantidad de población expuesta a la contaminación.

CAPÍTULO III: METODOLOGÍA

En este capítulo se presenta la metodología desarrollada para la extracción de patrones de actividad y movilidad de la población a partir de datos de telefonía móvil. Esta metodología presenta un carácter generalista y adaptable a distintas características (granularidad temporal y espacial) y tipologías de datos de telefonía móvil. La metodología recoge los principales avances del estado del arte e implementa nuevas mejoras[5]. Del mismo modo, en este capítulo también se presentan las metodologías complementarias desarrolladas para poder llevar a cabo los casos de aplicación prácticos relacionados con la red social y la movilidad (sección 3.2) y la exposición de la población a la contaminación (sección 3.3). Para el primero de estos casos, se muestra un procedimiento para generar redes egocentristas de los usuarios de telefonía móvil a partir de los registros de llamadas telefónicas. Para el segundo caso, se detalla una nueva aproximación para la estimación de presencia de la población a partir de datos de telefonía móvil.

3.1 Determinación de Patrones de Actividad y Movilidad

En esta sección se presenta la metodología propuesta para la extracción de información sobre patrones de actividad y movilidad de la población a partir del análisis de datos de telefonía móvil. Se propone un proceso basado en 5 etapas principales:

1. Pre-procesado, formateo y limpieza de los datos
2. Extracción de patrones de actividad y movilidad
3. Identificación de localizaciones frecuentes
4. Elevación muestral
5. Estadísticas de movilidad

[5] Señalar que este apartado ha sido revisado y actualizado a lo largo de toda la investigación con el objetivo de recoger los últimos avances metodológicos en este campo

En la *Figura 7* se muestra un esquema general de la metodología propuesta[6].

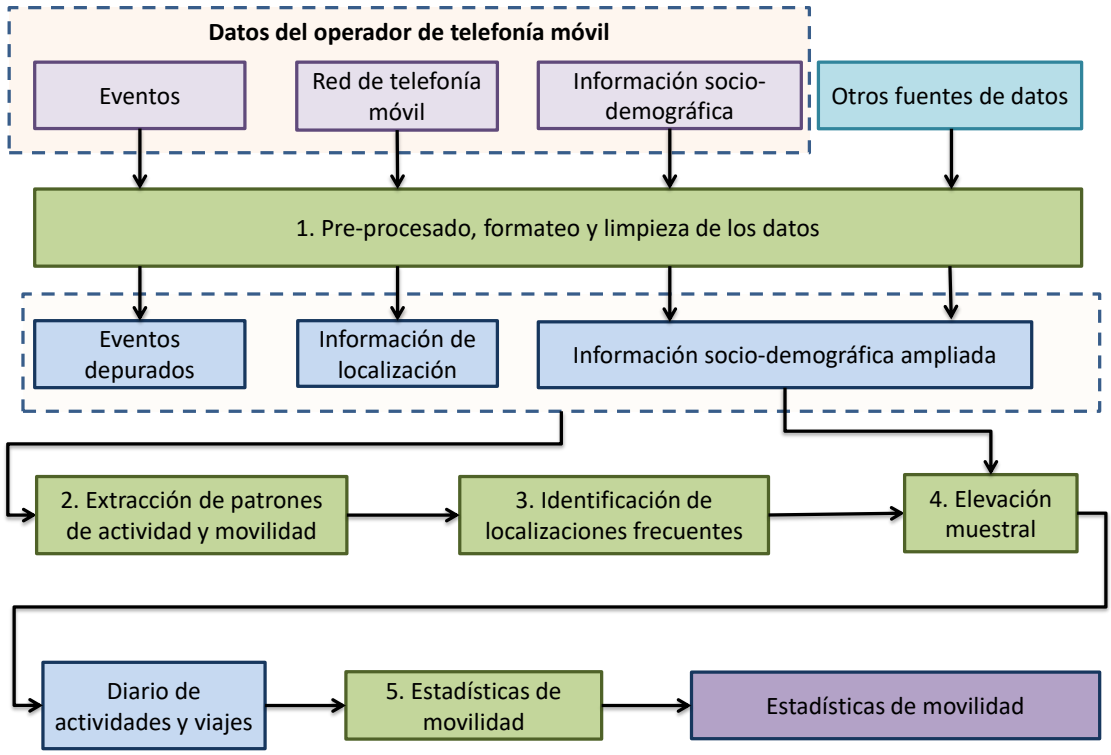


Figura 7. Esquema general del proceso de extracción de patrones de actividad y movilidad

3.1.1 Pre-procesado, formateo y limpieza de los datos

El objetivo principal de esta etapa es la depuración y pre-procesado de los datos de telefonía móvil con el objetivo de minimizar los errores presentes en los datos de partida y otorgarles el formato más adecuado para su posterior análisis. Del mismo modo, en esta etapa también se depuran y pre-procesan otras fuentes de datos necesarias para los análisis (por ejemplo, datos censales).

[6] Dependiendo del estudio, el orden de los pasos 2 y 3 puede intercambiarse, realizando primero la identificación de localizaciones frecuentes y posteriormente calculando los patrones de actividad y movilidad

3.1.1.1 Depuración y pre-procesado de los datos de eventos de red

En primer lugar, de todos los datos presentes en los registros de telefonía móvil, se deben seleccionar aquellos relevantes para el estudio que se quiera llevar a cabo. Para el caso de estudios de movilidad se debe recoger, al menos, la siguiente información:

- **ID anonimizado del usuario principal**: ID único del usuario que realiza el evento (llamada, conexión de datos, actualización en la localización, etc.).
- **Fecha y hora del registro**: fecha y hora asociada al registro. Para el caso de eventos de llamadas, se suele disponer de, al menos, un registro al inicio de la llamada y de un registro al final de la llamada.
- **Información de localización**: puede venir dada por el identificador de celda a la cual se conecta el usuario principal o por las coordenadas estimadas del dispositivo móvil.

Los registros que aportan información en un formato no válido, es decir, que no cumplen con las especificaciones definas por el operador de red, son descartados. Del mismo modo, se comprueba la consistencia en el campo de fecha y hora, comprobando que la fecha sea posible (por ejemplo, si se proporciona información del mes 16 el dato es descartado).

En la *Tabla 3* se muestra un ejemplo de los datos procedentes de los eventos de telefonía móvil a almacenar de cara a posteriores análisis.

Tipo de evento	ID usuario principal	Fecha y hora	Información de localización (ID de celda vs localización precisa)	
Llamada	AJ3zvCRet5QW	2014-10-25 11:45:35	659028426845216	40.416, -3.703
	AJ3zvCRet5QW	2014-10-25 11:52:12	659028426845216	40.419, -3.701
Sesión de datos	AJ3zvCRet5QW	2014-10-25 14:25:48	659025249358745	40.465, -3.689

Tabla 3. Ejemplo de datos procedentes de eventos de telefonía móvil después de su selección, depuración y formateado

3.1.1.2 Depuración y pre-procesado de los datos de la red de telefonía móvil

Este paso es necesario para los casos en los que no se disponga de información triangulada sobre la localización del dispositivo móvil. Los datos de la red de telefonía móvil aportan información sobre las características y localización de las torres de telefonía y sus antenas. En primer lugar, los registros que aportan información en un formato no válido, es decir, que no cumplen con las especificaciones definas por el operador de red, son descartados. Cruzando la información del ID de celda presente en los eventos con los datos de la red de telefonía, se obtiene la localización de la torre asociada a dicha celda. Posteriormente, existen dos aproximaciones posibles al problema de estimación de la localización del dispositivo móvil. La primera aproximación es considerar que la localización de la torre es una buena estimación de la localización real del dispositivo móvil, por lo que sería similar al caso en el que se dispone de información de localización mediante triangulación. La segunda aproximación consiste en estimar el área de cobertura donde seguramente se encuentre el dispositivo a partir de los datos de los emplazamientos de las distintas torres. Para ello, se generan las áreas de Voronoi como una aproximación de la cobertura real de cada torre. En la *Figura 8* se muestra un ejemplo

de áreas de Voronoi asociadas a un conjunto de torres. Una vez definidas las áreas de Voronoi, se puede estimar la localización del dispositivo móvil como un punto dentro del área de Voronoi, como por ejemplo su centroide (este planteamiento vuelve a ser similar al de considerar la localización de la torre o la localización triangulada) o también se puede operar con la información espacial del área de Voronoi.

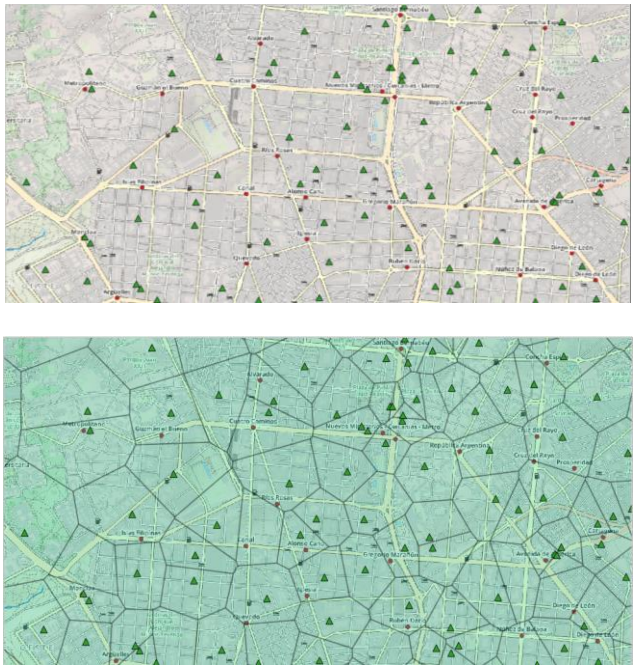


Figura 8. Arriba – ejemplo de emplazamientos de torres representados por un triángulo verde. Abajo – áreas de Voronoi correspondiente a los emplazamientos de las torres

3.1.1.3 Depuración y pre-procesado de los datos socio-demográficos

En primer lugar, se debe definir qué información socio-demográfica presente en los registros del operador es relevante para el objeto del estudio. Generalmente estos registros contienen mucha información que no es útil para el objetivo final del estudio, por lo que no es necesario almacenarla ni depurarla, evitando consumir esfuerzos en esta tarea. Registros que suelen ser útiles a la hora de realizar los análisis son, por ejemplo, aquellos relacionados con la edad y el género de los usuarios. Esta información puede utilizarse para segmentar la información obtenida por edad o por género y también ayudar en los procesos de elevación de la muestra. Es importante señalar que la información de edad y género puede estar asociada al titular de la línea, que no tiene por

qué corresponderse con el usuario del dispositivo móvil[7]. Esta consideración es importante tenerla en cuenta a la hora de analizar y evaluar los resultados. Es previsible que la mayoría de los posibles errores se presenten en tramos de edad de gente joven, cuyos contratos están a nombre de sus padres o tutores legales. A medida que la edad aumenta, es previsible que la información sea de mayor calidad. En casos de incertidumbre en cuanto a la calidad de la información, se recomienda realizar una segmentación no demasiado elevada, acorde con la fiabilidad de los datos.

Como para el resto de fuentes de datos, los datos socio-demográficos deben también ser depurados y formateados de manera eficiente. Para el caso de la edad y el género es recomendable comprobar que los valores que toman las variables son razonables (por ejemplo, para registros con edades no comprendidas entre 12-120 años debería realizarse una depuración posterior o eliminar los datos) y que los valores son consistentes a lo largo del tiempo (es decir, que en registros posteriores generados por el operador el género no cambia y que la edad es compatible con las fechas de las actualizaciones).

En algunas ocasiones, también se dispone de información sobre el código postal de residencia, pero su uso no es aconsejable, ya que el dato puede estar desactualizado y el nivel de agregación no es el más adecuado. Para determinar el lugar de residencia se considera más adecuado basarse en los patrones de comportamiento del usuario extraídos del análisis longitudinal de los datos (ver sección 3.1.3).

3.1.1.4 Depuración y pre-procesado de otras fuentes de datos

Aparte de los datos procedentes de la telefonía móvil, en muchas ocasiones es necesario utilizar otras fuentes de datos complementarias. En función de la aplicación final que se realice, las fuentes de datos necesarias pueden ser muy variadas (información de puestos de trabajo, información sobre usos del suelo, información sobre oferta de transporte, etc.). Como mínimo, es necesario recoger información sobre los marcos muestrales que se utilizarán en el proceso de elevación de la muestra. Esta información también debe ser

[7] La disponibilidad de información sobre el titular y usuario del teléfono móvil depende en gran medida del proceso de recogida de este tipo de información por parte del operador de telefonía móvil.

depurada y pre-procesada de manera eficiente para su integración con el resto de datos procedentes de la telefonía móvil.

3.1.2 Extracción de patrones de actividad y movilidad

El objetivo principal de esta etapa es extraer información sobre las actividades y viajes realizados por los usuarios durante un día concreto. Se define como 'viaje' el desplazamiento entre dos actividades consecutivas, y como 'actividad' una interacción o conjunto de interacciones con el entorno que tienen lugar en una misma localización y que motivan que el individuo se desplace hasta allí.

En primer lugar, es necesario estimar lo que denominaremos estancias. Las estancias son lugares donde el usuario permanece un cierto tiempo. Si la estancia presenta una duración mayor a un cierto umbral, entonces esa estancia se clasificará como actividad. Para detectar los lugares de estancia es necesario depurar la información de localización procedente de los eventos de telefonía móvil. Los eventos de telefonía móvil pueden proporcionar dos localizaciones distintas o dos identificadores de celda distintos sin que el dispositivo se haya desplazado de su ubicación inicial. Por lo tanto, los datos proporcionan distintas estancias cuando en realidad se trata de la misma. Para evitar estos problemas, en el caso de trabajar con información de coordenadas asociadas al evento, se propone utilizar técnicas de clusterización para determinar las estancias. En el caso de trabajar con áreas de cobertura (por ejemplo, áreas de Voronoi), se propone un planteamiento basado en la utilización de velocidades de cambio de señal para filtrar desplazamientos ficticios. En la sección 3.1.2.1 se describe el proceso para la obtención de estancias. Una vez obtenidas las estancias, se llevan a cabo tres análisis adicionales para la obtención de los patrones de actividad y movilidad de los usuarios:

- Eliminación de usuarios no válidos
- Detección de actividades
- Identificación de viajes

En los apartados posteriores se describen en detalle cada uno de estos pasos.

3.1.2.1 Determinación de estancias y filtrado de señales

La mayoría de los estudios presentes en la literatura aplica técnicas de clusterización para estimar las estancias. Existen un gran número de técnicas de clusterización que podrían aplicarse en este caso, proporcionando resultados similares. Entre las distintas técnicas posibles, se propone utilizar la propuesta por Jiang et al. (2013), que se basa en una adaptación de Hariharan & Toyama (2004) para el estudio de señales de GPS, que ha arrojado resultados satisfactorios en varios estudios con telefonía móvil (Çolak et al. 2015, Alexander et al. 2015). A continuación se detalla el algoritmo propuesto:

- sea una secuencia $Di = (di(1), di(2), \dots, di(ni))$ de eventos de telefonía móvil asociados a un usuario 'i', donde $di(k) = (t(k), x(k), y(k))'$ $k = 1, \dots, ni$, y $t(k), x(k)$ e $y(k)$ son el tiempo, la longitud y la latitud del evento k-th del usuario 'i' respectivamente. Primero, se extraen los puntos $di(k)$ que están a una distancia inferior a un cierto umbral respecto de sus eventos sucesivos ($di(k+1), di(k+2), \dots . di(k+m)$). Posteriormente, para reducir los errores de localización por saltos de señales, se asume que $di(k), \dots, di(k+m)$ son observados cuando el usuario está en una localización específica, que puede venir definida, por ejemplo, por el centroide de todas las localizaciones asociados.

Por otro lado, para el caso de utilizar la aproximación basada en áreas de cobertura en vez de coordenadas estimadas de localización, se propone utilizar como área de estancia la propia área de cobertura asociada a la celda de telefonía móvil y filtrar los posibles saltos de señales mediante criterios de velocidad de cambio entre antenas. Se considera que una localización es debida a un salto de señales (y por lo tanto no debe tenerse en consideración) si la velocidad de cambio observada es mayor a un cierto umbral (VC). La velocidad de cambio se define como el cociente entre la distancia y el tiempo entre eventos consecutivos:

$$v_c = \frac{D_{eventos}}{T_{eventos}} \qquad Si\ v_c \geq VC \rightarrow\ salto \qquad [1]$$

La distancia entre eventos puede medirse como la distancia entre las torres de telefonía asociadas a las áreas de cobertura o como la distancia entre los centroides de las áreas de cobertura. El valor de 'VC' deberá escogerse en función de las características de los datos de telefonía móvil. Por ejemplo, Iovan et al. (2013) utilizan una aproximación similar al problema y proponen un valor de umbral de velocidad 'VC' de 200km/h.

3.1.2.2 Eliminación de usuarios no válidos

En este proceso se eliminan de la muestra aquellos usuarios que contengan datos con una granularidad temporal insuficiente como para poder determinar con fiabilidad las actividades y viajes que realizan a lo largo del día. Se define como un usuario no válido a aquel que contenga registros consecutivos que difieran un cierto tiempo 'tr' superior a un cierto umbral temporal 'TR'.

$$Clasificación\ de\ usuarios \begin{cases} tr > TR \rightarrow Usuario\ no\ válido \\ tr \leq TR \rightarrow Usuario\ válido \end{cases} \qquad [2]$$

Este criterio puede aplicarse para el total del día o para distintos periodos del día, tomando valores distintos de 'TR' por periodo. El planteamiento de tomar valores distintos de 'TR' por periodo viene motivado por el hecho de que el número de eventos por unidad de tiempo a lo largo del día para un mismo usuario no es uniforme, presentándose menos eventos normalmente en horario nocturno. Este planteamiento supone una mejora con respecto a estudios previos que solamente consideraban el número de registros totales o un número mínimo de periodos de actividad al día sin considerar la variabilidad de los datos (Song et al. 2010a; Onnela et al. 2011, Schneider et al. 2013). No obstante, es importante señalar que el considerar distintos valores de 'TR' en función de los periodos del día también puede llevar a eliminar de la muestra ciertos usuarios, como por ejemplo personas con trabajos nocturnos si el criterio de selección de

la muestra es más restrictivo durante el periodo diurno. Los valores a adoptar de 'TR' dependerán de las características de los datos de telefonía móvil y del objetivo específico del estudio. Se deberá buscar una solución de compromiso entre la muestra útil que se obtiene una vez descartados los usuarios y el nivel de error que se introduce al considerar ciertos segmentos de usuarios y usuarios con pocos eventos a lo largo del día.

3.1.2.3 Identificación de actividades

El criterio más utilizado en la literatura científica para detectar actividades es el tiempo de estancia en una determinada localización (Calabrese et al. 2013; Schneider et al. 2013; Holleczek et al. 2014, etc.). Se determina que un usuario está realizando una actividad 'A' en una estancia 'E' si el tiempo de estancia observado 'te' es igual o superior a un cierto umbral temporal 'TA'.

$$detección\ de\ actividades \begin{cases} te \geq TA \rightarrow actividad\ detectada \\ te < TA \rightarrow no\ existe\ actividad \end{cases} \quad [3]$$

El tiempo de estancia observado en una determinada localización se mide como la diferencia temporal entre el primer registro y el último registro en dicha localización. Es importante señalar que el tiempo observado entre registros es inferior a la duración real de la actividad, aspecto relevante a considerar a la hora de definir 'TA'. El valor más apropiado de 'TA' dependerá en gran medida del objetivo del estudio. Estudios previos proponen utilizar un tiempo de 10 (Jiang. et al. 2013, Çolak et. al 2015) o 20 minutos (Holleczek et al. 2014). Con este criterio potencialmente se descartan como actividades, por ejemplo, las estancias en intercambiadores de transporte. No obstante, se pueden perder actividades de corta duración como por ejemplo llevar a los niños al colegio. Se debe buscar una solución de compromiso entre considerar actividades de corta duración que en realidad son paradas entre dos etapas de un mismo viaje frente a descartar estancias de corta duración que en realidad sí son actividades.

3.1.2.4 Estimación de viajes

Una vez identificadas las actividades a lo largo del día, el viaje se define como el desplazamiento entre actividades consecutivas. El viaje viene definido por la localización de origen, la localización de destino, la hora de inicio del viaje y la duración del viaje. El origen y el destino del viaje están definidos directamente por la localización de una actividad y su actividad posterior, respectivamente. Tanto la hora de inicio del viaje como la duración del mismo son valores indeterminados que deben estimarse. El problema de determinación de la hora de inicio y la duración del viaje puede definirse mediante el planteamiento de dos restricciones:

$$hora_{inicio\_viaje} = \left[hora_{final_{origen}}, hora_{inicial_{destino}} - duración_{viaje}\right] \quad [4]$$

$$duración_{viaje} = \left[0, \quad hora_{inicial_{destino}} - hora_{final_{origen}}\right] \quad [5]$$

La hora de inicio del viaje deberá tomar un valor comprendido entre el último registro identificado en la actividad de origen (*hora_final_origen*) y la diferencia entre el primer registro identificado en la actividad de destino (*hora_inicial_destino*) y la duración estimada del viaje. La duración máxima del viaje estará determinada como la diferencia entre el primer registro en la actividad de destino y el último en la actividad de origen.

Para estimar la duración del viaje es necesario conocer la distancia del viaje y la velocidad media del mismo. La duración del viaje viene definida por la siguiente expresión:

$$duración_{viaje} = min \begin{cases} d = \dfrac{distancia}{velocidad_{media}} \\ d = hora_{inicial_{destino}} - hora_{final_{origen}} \end{cases} \quad [6]$$

Para obtener los valores de distancia de viaje y velocidad media con precisión es necesario identificar la ruta y el modo de transporte utilizado. La identificación de modo y ruta a partir de los datos de telefonía móvil no siempre es posible, siendo necesario estimar estas variables mediante alguna otra aproximación. Para algunos estudios, una aproximación razonable puede ser estimar la distancia de viaje como la distancia euclídea entre las localizaciones de origen y destino del viaje, asumiendo una ruta en línea recta sobre el plano. Del mismo modo, puede asumirse a modo de simplificación una velocidad media constante para todos los desplazamientos. Una vez definida la duración del viaje, el rango de incertidumbre para la estimación de la hora de inicio del viaje queda reducido. De hecho, si la duración del viaje tomara su valor máximo admisible, la hora del inicio del viaje estaría totalmente determinada, y coincidiría con la hora del último registro en la actividad de origen. Para determinar la hora de inicio del viaje, se propone utilizar una función de probabilidad para asignar la hora dentro de los valores posibles. Con carácter general, se puede utilizar una función uniforme, de modo que todas las horas dentro del intervalo tengan la misma probabilidad de ser escogidas. En caso de disponer de información histórica de encuestas, de datos de aforos de tráfico o billetaje de la zona de estudio, se propone utilizar una función probabilística adaptada a dichos valores.

El resultado final del proceso descrito en 3.1.2 es un conjunto de actividades y viajes que denominamos **diario de actividades y viajes**. La información que proporciona el diario es una secuencia de actividades y viajes con las horas de inicio y finalización de cada actividad. En la *Figura 9* se presenta un ejemplo sobre el proceso de transformación de registros de telefonía móvil a una secuencia de actividades y viajes.

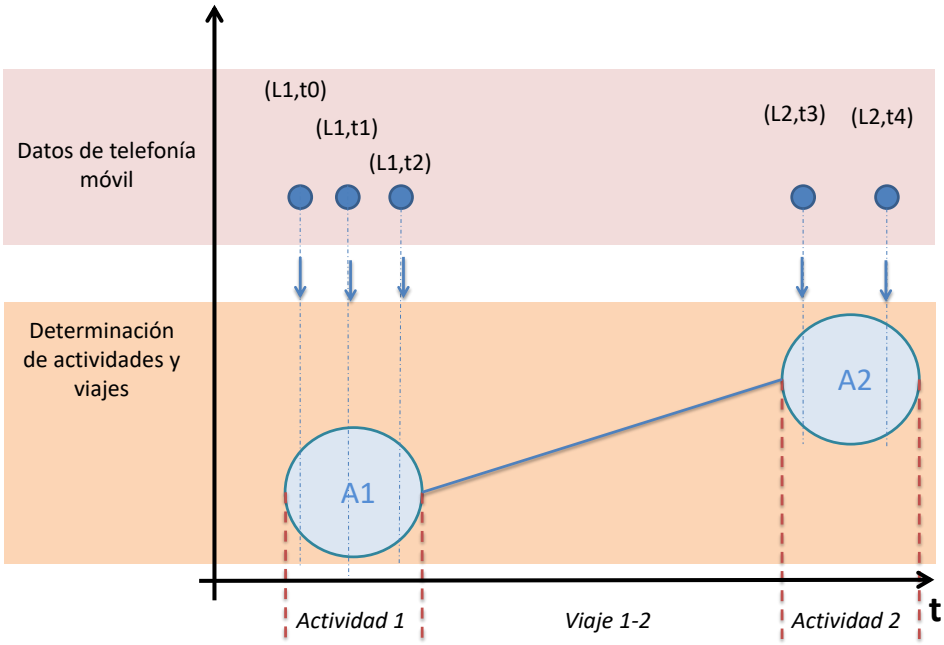


Figura 9. Proceso de transformación de los eventos de telefonía móvil al diario de actividades y viajes

3.1.3 Identificación de actividades frecuentes

El objetivo principal de esta etapa es determinar la tipología de las actividades frecuentes llevadas a cabo por los usuarios a lo largo de un periodo de tiempo determinado. Estudios previos que han analizado las actividades frecuentes de los usuarios se han centrado en el análisis de los lugares de residencia y trabajo (p.ej., Isaacman et al. 2011, Phithakkitnukoon et al. 2012, Chen and Mei 2014). En la metodología aquí presentada se propone considerar adicionalmente actividades frecuentes distintas al lugar de residencia y trabajo, como por ejemplo actividades relacionadas con eventos deportivos (p. ej., ir al gimnasio 3 días a la semana) o actividades de ocio (p.ej., ir al cine). A estas actividades frecuentes las denominaremos “otras actividades frecuentes”.

Se define como actividad frecuente aquella actividad en la que el usuario aparece un mínimo número de días con respecto al total de días analizados de un mismo tipo (por ejemplo, martes), con respecto al total de días laborables (considerando días laborables

los días de lunes a jueves) o con respecto al total de fines de semana (considerando fines de semana la unión de sábados y domingos). El viernes no ha sido integrado ni en día laborable ni en día festivo de manera intencionada, debido a sus características potencialmente mixtas. Otras clasificaciones de días laborables y festivos son posibles. El número mínimo de días (frecuencia mínima) para considerar una actividad como frecuente viene determinado por la siguiente expresión:

$$Frecuencia\_mínima = \alpha \cdot muestra\_de\_días\_analizados \qquad [7]$$

donde 'α' es un coeficiente de reducción y *'muestra_de_días_analizados'* es el total de días de un mismo tipo presentes en la muestra (días específicos de la semana, días laborables y días festivos). A la hora de definir el valor más adecuado para 'α' se deberá considerar la granularidad temporal de los datos de telefonía móvil. El parámetro 'α' puede tomar distintos valores en función del tipo de actividad que se quiera analizar.

3.1.3.1 Identificación del lugar de residencia

Se define como lugar de residencia aquella localización en la que el usuario pernocta durante el periodo de estudio. Una actividad frecuente es considerada como residencia si es la localización más frecuente entre las Xh p.m. y las Yh a.m. para los días de estudio considerados. Para la definición de los valores más apropiados de Xh e Yh se deberán tomar en consideración los patrones de comportamiento de la población del área de estudio. Por ejemplo, Calabrese et. al (2013) toman como valores de Xh e Yh, 6 y 8 respectivamente. Picornell et al. 2015 (publicación basada en el caso de estudio presentado en la sección 4.2 de este documento) obtuvieron resultados satisfactorios para un coeficiente 'α' de residencia de 0,2. Del mismo modo, otros estudios con aproximaciones similares (Alexander et al. 2015) han considerado una frecuencia mínima de una vez por semana para la determinación del lugar de residencia ('α' de 1/7=0,15). Por último, es importante señalar que esta aproximación puede clasificar de manera errónea el lugar de residencia para algunos segmentos de la población con un

comportamiento distinto al de la mayoría (por ejemplo, personas con trabajo nocturno). No obstante, la ventaja de esta aproximación es que es sencilla de implementar y ha demostrado resultados satisfactorios en distintos estudios de movilidad urbana (Picornell et al. 2015, Alexander et al. 2015). En la sección 5.2.1 se presenta una discusión sobre posibles mejoras futuras de esta metodología.

3.1.3.2 Identificación del lugar de trabajo

Se define como lugar de trabajo aquél donde se realiza una actividad recurrente en un periodo temporal distinto al del lugar de residencia. Una actividad frecuente es considerada como trabajo si es la localización más frecuente entre las Xw a.m. y las Yw p.m. para los días de estudio considerados. Al igual que sucede con la definición del lugar de residencia, los parámetros Xw e Yw deberán adoptarse en función de los patrones de comportamiento de los trabajadores en el área de estudio. Por ejemplo, Lenormand et al. (2014) toman valores de Xw=9 e Yw=5. Otros estudios (por ejemplo, Çolak et al. 2015) añaden la restricción de que la localización del hogar sea distinta a la localización de trabajo. En la metodología aquí propuesta, no se añade esta restricción, considerando que pueden existir casos en los que las personas trabajen cerca de su lugar de residencia. Picornell et al. 2015 obtuvieron resultados satisfactorios para un coeficiente 'α' de 0,3 para el análisis del lugar de trabajo. Alexander et al. 2015 también obtuvieron resultados satisfactorios utilizando frecuencias de un día laborable a la semana (equivalente a valores de 'α' de 0,2). Lo comentado en el apartado anterior sobre las limitaciones de la identificación del lugar de residencia para algunos segmentos de la población aplica de la misma manera en este caso. En la sección 5.2.1 se presenta una discusión sobre posibles mejoras futuras de esta metodología.

3.1.3.3 Identificación de otras localizaciones frecuentes

Todas aquellas actividades identificadas como actividades frecuentes distintas al lugar de residencia y trabajo son clasificadas como "otras" actividades frecuentes. Este tipo de clasificación es novedoso con respecto a otros estudios (Çolak et al. 2015, Alexander et al,

2015), que solamente consideran como actividades frecuentes casa y trabajo, englobando al resto de actividades como 'otras', independientemente de que sean o no frecuentes.

3.1.4 Elevación muestral

El objetivo de esta etapa es elevar la información muestral para obtener estadísticas a nivel poblacional. Esta etapa no suele aparecer en la literatura científica, siendo pocos los estudios que hacen referencia a ella (p.ej., Terada et al. (2013) y Alexander et al. (2015)).

Se propone utilizar una elevación basada en la residencia de los usuarios. Se propone aplicar un factor de elevación por usuario que venga determinado por la siguiente expresión:

$$f_{usuario} = \frac{población\ total\ en\ la\ zona\ (tipo\ de\ usuario)}{muestra\ en\ la\ zona\ de\ residencia\ (tipo\ de\ usuario)} \qquad [8]$$

En primer lugar, es necesario dividir el área de estudio en zonas donde la población total en dicha zona sea conocida. Generalmente, se utiliza una zonificación basada en secciones censales o niveles más agregados de población (distrito, municipio, etc.) dado que es el formato empleado por las entidades estadísticas para proporcionar dicha información. Posteriormente, a cada usuario de la muestra se le asigna una zona en función de su lugar de residencia identificado mediante el procedimiento descrito en 3.1.3.1. La suma de todos los usuarios asignados a una misma zona constituye la muestra en dicha zona. La elevación muestral puede realizarse para el total de la población de la zona o segmentado por 'tipo de usuario', entendido como un conjunto de usuarios con unas características socio-demográficas específicas (edad, género, etc.). En función de la información socio-demográfica disponible, la fiabilidad de la misma y el objeto del estudio, se podrá decidir sobre la idoneidad de utilizar un factor de elevación segmentado por características socio-demográficas.

3.1.5 Estadísticas de movilidad

El objetivo de esta etapa es calcular las estadísticas de movilidad relevantes para el estudio a partir de la información del diario de actividades y viajes obtenido anteriormente. El número y tipología de estadísticas de movilidad que se pueden obtener es muy amplio. A continuación se presenta una lista de los indicadores que se consideran más relevantes para los estudios de movilidad. El cálculo de los indicadores a partir de la información de los diarios de actividades y viajes es trivial:

- **Número de viajes por persona:** número de viajes medio por persona en el día.
- **Distribución horaria de los viajes:** número de viajes totales para una franja de tiempo definida**.** La distribución horaria de los viajes puede representarse agregada para distintos periodos del día (hora punta de mañana, valle, punta tarde) o a niveles más desagregados.
- **Distribución de la distancia de los viajes:** número de viajes totales para un rango de distancias definido.
- **Matrices origen-destino:** dada una zonificación, las matrices origen-destino se obtienen como el sumatorio de todos los viajes entre cada par de zonas (considerando cada sentido), incluyendo los viajes con origen y destino la misma zona (viajes intrazona).
- **Propósito del viaje:** tipo de actividad en destino. A partir de los análisis presentados previamente se pueden identificar 4 tipos de propósito de viaje: casa, trabajo, otro frecuente y otro no frecuente (actividad llevada a cabo en el día analizado pero que no ha sido clasificada como actividad frecuente).
- **Perfilado socio-demográfico:** los indicadores antes mencionados pueden segmentarse en función de las características socio-demográficas de las personas.

3.2 Determinación de la Red Social

En esta sección se presenta una metodología para estimar la red social de las personas a partir del análisis de los datos de telefonía móvil. En la *Figura 10* se muestra un esquema general de la metodología propuesta.

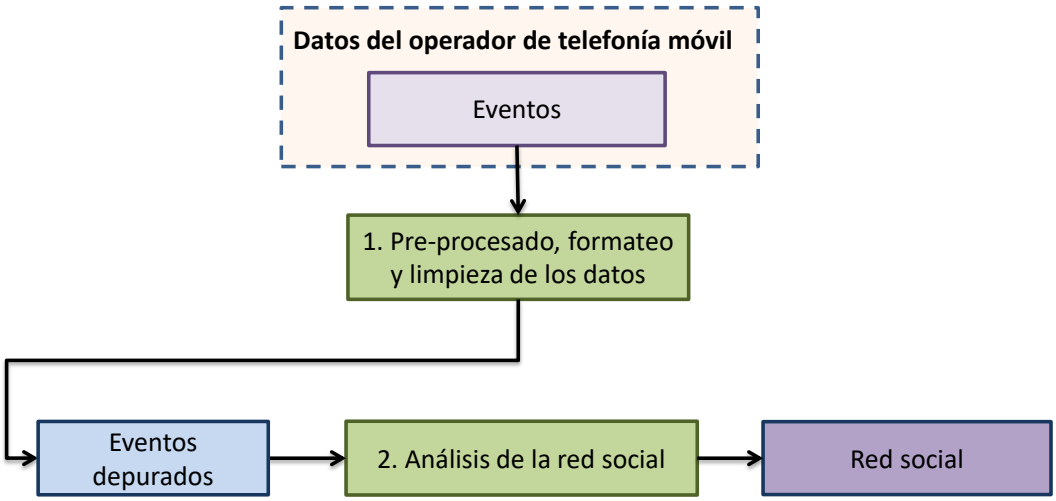


Figura 10. Esquema general del proceso de análisis de la red social

3.2.1 Depuración y pre-procesado de los datos de eventos de red

En primer lugar, de todos los datos presentes en los eventos de telefonía móvil, se deben seleccionar aquellos relevantes para el estudio que se quiera llevar a cabo. Para el caso de estudios de la red social de los usuarios de telefonía, es necesario almacenar los registros que contengan información referente al emisor y receptor de la comunicación (por ejemplo, los registros de llamadas). Para cada evento se debe almacenar, al menos, la siguiente información:

- **ID anonimizado del usuario principal**: ID único del usuario emisor de la comunicación.
- **ID anonimizado del usuario secundario**: ID único del usuario receptor de la comunicación.

Los registros que aportan información en un formato no válido, es decir, que no cumplen con las especificaciones definas por el operador de red son descartados. En la *Tabla 4* se muestra un ejemplo de los datos procedentes de los eventos de telefonía móvil a almacenar de cara a posteriores análisis.

Tipo de evento	ID usuario principal	ID usuario secundario
Llamada	AJ3zvCRet5QW	++/Gjhgyge45
SMS	AJ3zvCRet5QW	/fkRno4O9erT

Tabla 4 .Ejemplo de datos a extraer de los CDRs para estudios de red social

3.2.2 Análisis de la red social

Para la determinación de la red social de cada usuario se propone utilizar un planteamiento de red egocentrista, en la que el usuario principal (*ego*) tiene algún tipo de relación con un conjunto de usuarios (*alters*). Se considera que existe relación entre dos usuarios si existe reciprocidad en las comunicaciones que realizan. Esta condición es habitualmente empleada en la literatura científica (Onnela et al. 2007; Phithakkitnukoon et al. 2012; Chen and Mei 2014). Con esta condición se pretende eliminar comunicaciones unidireccionales realizadas por personas que no pertenecen a la red social del usuario. Por lo tanto, la red social del *ego* está definida por un conjunto de nodos (donde cada nodo es un *alter*) y conexiones bidireccionales entre dichos nodos representando reciprocidad en las comunicaciones. Para medir el nivel o grado de relación entre el *ego* y los *alters* se proporciona a cada conexión un peso proporcional al número de comunicaciones entre el *ego* y el *alter* en relación al total de comunicaciones realizadas por el *ego* dentro de su red social. El grado de relación entre el *ego* y el *alter* se puede expresar de la siguiente manera:

$$Gi = \frac{ci}{\sum_{j=1}^{n} cj} \qquad [9]$$

Siendo 'Gi' el grado de relación entre el *ego* y el *alter* 'i', 'ci' el número de comunicaciones entre el *ego* y el *alter 'i'*, 'n' el número de alters dentro de la red social y 'cj' el número de comunicaciones entre el *ego* y el *alter 'j'.*

En la *Figura 11* se muestra un esquema de red egocentrista con conexiones bidireccionales ponderadas.

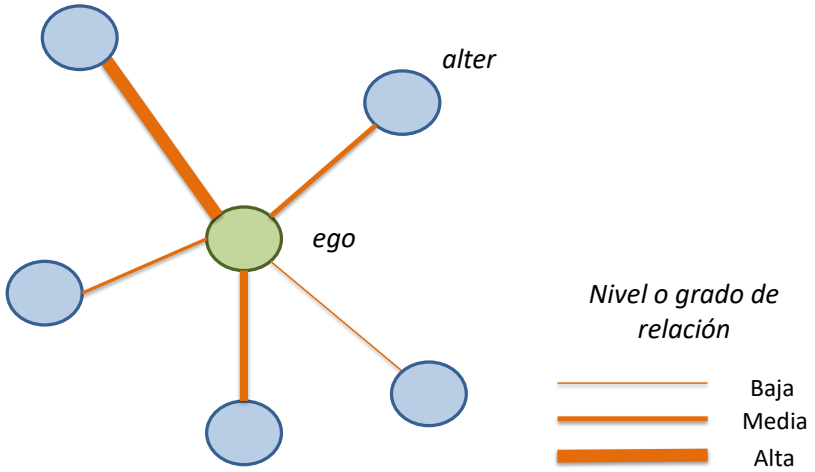


Figura 11. Ejemplo de red egocentrista con conexiones bidireccionales ponderadas

3.3 Determinación de Presencia de Población

En esta sección se propone una metodología para estimar el número de personas presentes en una determinada zona para distintos momentos del día; a este indicador lo denominaremos presencia de población. En primer lugar, es necesario definir qué entendemos por 'presencia'. Existen distintos indicadores de presencia. Por ejemplo, la presencia se puede definir como el número de apariciones únicas de personas en una determinada zona durante un periodo concreto del día; o también, como el número total de personas que permanecen en una determinada zona durante un periodo de tiempo concreto. En el primer caso, todas las personas que crucen la zona de estudio serán contabilizadas una única vez, mientras que en el segundo caso, sólo serán contabilizadas

aquellas personas que durante un tiempo determinado permanezcan en la zona. En este estudio se entiende como **presencia**, el total del número de personas que aparecen en una zona determinada 'Z' durante un periodo de tiempo 'T' afectadas por un factor de estancia 'fe', definiéndose dicho factor de estancia como la relación entre el tiempo de estancia de cada persona en la zona y la duración del periodo considerado. El indicador de presencia puede expresarse matemáticamente de la siguiente manera:

$$presencia_{zona\_T} = \sum_{i=1}^{n} \frac{Tz_i}{Td} \qquad [\ 10\ ]$$

Siendo 'n' el número de personas que aparecen en la zona 'Z' en el periodo 'T', 'Tzi' el tiempo que la persona 'i' permanece en la zona 'Z' durante el periodo 'T' y 'Td' la duración del periodo.

En la *Figura 12* se muestra un esquema general de la metodología propuesta para la estimación de presencia. Como se puede apreciar, el esquema es muy similar al presentado en la sección 3.1, añadiendo el análisis de presencia al final del proceso. El planteamiento que se propone es utilizar la información del diario de actividades y viajes como input para estimar el indicador de presencia. Este es un planteamiento novedoso con respecto a otros estudios de presencia basados en datos de telefonía móvil, en los cuales se analiza la presencia en una zona desde el punto de vista de los registros contabilizados en las antenas o torres de telefonía (p. ej., Gariazzo et al. 2016; Nyhan et al. 2016) y no desde el punto de vista de la movilidad de las personas.

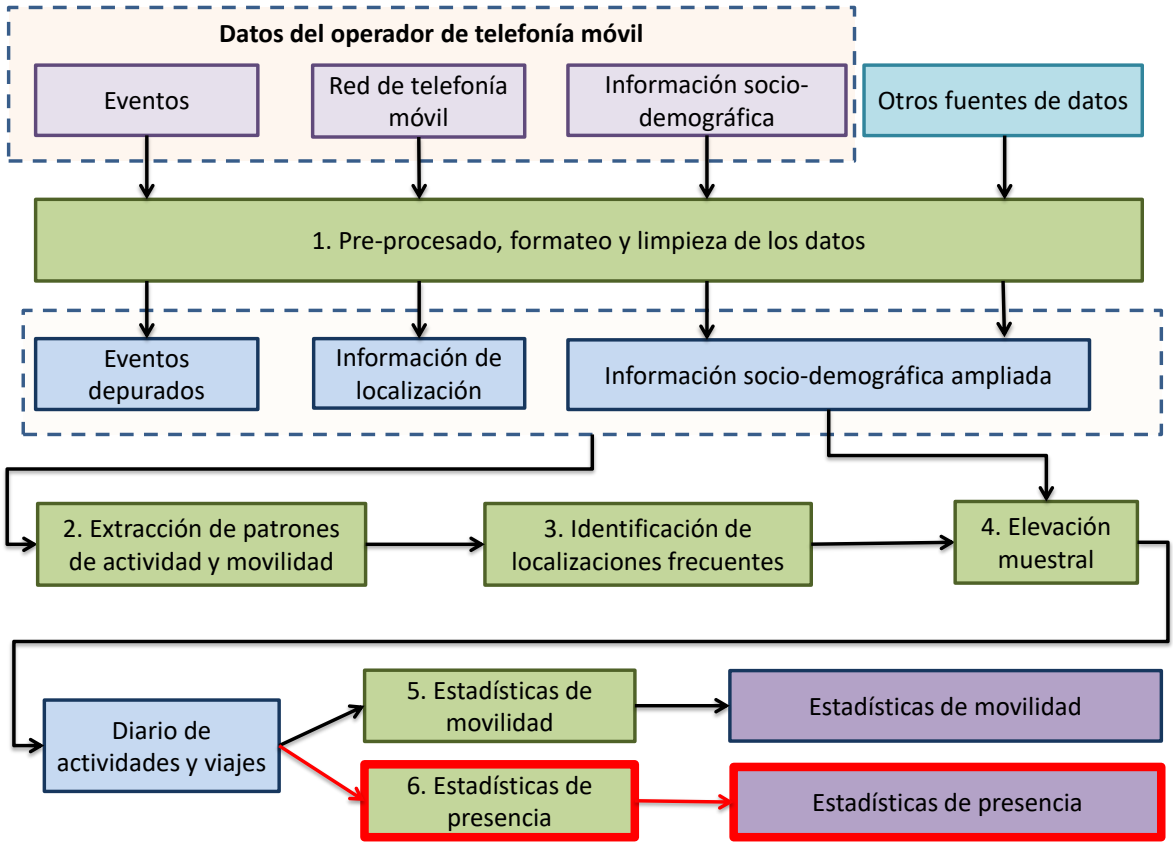


Figura 12. Esquema general del proceso de extracción de información de presencia

3.3.1 Estadísticas de presencia

En este apartado se presenta la metodología para la determinación del indicador de presencia. Como se ha comentado anteriormente, el indicador de presencia puede calcularse mediante la expresión definida en [10]. La zonificación y los periodos de estudio a considerar son datos de entrada del problema. La variable pendiente de determinar para poder estimar el indicador de presencia es el tiempo de estancia de cada persona en cada zona por periodo ('Tzi'). Los pasos a seguir para calcular el tiempo de estancia 'Tzi' son:

- Determinar la localización a lo largo del día
- Determinar el instante de cambio entre zonas
- Calcular el tiempo de estancia por zona

3.3.1.1 Determinación de la localización a lo largo del día

A partir del diario de actividades y viajes es posible estimar la localización de los usuarios a lo largo del día. Para el caso de las actividades, la localización es igual a las coordenadas asociadas a dicha actividad. Para el caso de los viajes, la localización en cada instante depende de la trayectoria del viaje y las velocidades en cada tramo. Para el caso particular de suponer la simplificación de trayectoria recta en el plano y una velocidad constante a lo largo del recorrido, la localización 'L' de un usuario para un cierto instante 't' que se desplaza entre una actividad A (con coordenadas 'C_A' y hora de finalización 't_A') y una actividad B separadas temporalmente un tiempo T_{AB} vendría determinada por:

$$L(t) = C_A + \frac{t - t_A}{T_{AB}} \overrightarrow{AB} \qquad [\ 11\ ]$$

3.3.1.2 Determinación del instante de cambio entre zonas

Para calcular el instante de cambio entre zonas se debe realizar la intersección espacial entre la trayectoria de cada viaje y la zonificación de estudio. El instante de paso de una zona a otra se define como el instante en el cual la localización del usuario coincide con el límite entre dos o más zonas. Conocido el punto de intersección de la trayectoria con los límites de las zonas, el instante de paso se determina como la hora de inicio del viaje más el tiempo transcurrido desde el origen del viaje hasta el punto de intersección. Para el caso de trayectoria recta y velocidad constante, el instante de paso se obtiene despejando 't' de la expresión [11], considerando en L(t) las coordenadas del punto de intersección de la trayectoria con los límites entre zonas. Este proceso se repite, para cada viaje, tantas veces como límites entre zonas intersecte la trayectoria del viaje.

3.3.1.3 Cálculo del tiempo de estancia por zona

El tiempo de estancia de un usuario en una zona 'Tz', para un periodo de tiempo determinado, se define como el sumatorio de los tiempos de estancia asociados a las 'n'

actividades realizadas en dicha zona más el sumatorio de los tiempos de recorrido de los 'm' viajes realizados en dicha zona.

$$T_Z = \sum_{i=1}^{n} T_{ACT} + \sum_{j=1}^{m} T_{VIAJE}$$

La determinación del tiempo de estancia correspondiente a actividades es trivial a partir del diario de actividades y viajes. Para el caso de los viajes, es necesario tener en cuenta la trayectoria del viaje dentro de cada zona. Cada viaje dentro de una zona se puede definir a partir de dos puntos de su trayectoria. A estos puntos los llamaremos puntos característicos. Los viajes se pueden clasificar en función de sus puntos característicos como:

- **Origen + destino**: viajes intrazona. Viajes que no cruzan ningún límite entre zonas
- **Origen o destino + punto de intersección**: viaje con origen o destino la zona que cruza en su trayecto los límites de la zona en un punto.
- **Punto de intersección + punto de intersección**: viaje con origen y destino fuera de la zona que cruza los límites de la zona en dos puntos.

Una vez clasificado el tipo de viaje para cada zona con respecto a los puntos característicos que lo definen, el tiempo de estancia del viaje en la zona viene determinado por la diferencia entre los instantes de sus dos puntos característicos (estos instantes vienen determinados por el diario de actividades y viajes o se pueden determinar siguiendo los pasos descritos en 3.3.1.2). Por ejemplo, un viaje que se inicia dentro de una zona en el instante 't_inicial' y que cruza el límite de dicha zona en el instante 't_cruce', su tiempo de estancia en la zona será la diferencia entre 't_cruce' y 't_inicial'. El proceso completo se repite para cada usuario para obtener el indicador de presencia global por zona en un periodo determinado.

En la *Figura 13* se muestra gráficamente un ejemplo de determinación del tiempo de estancia de un usuario ('Tz') para varias zonas ('Zj') durante un periodo de tiempo determinado 'T'.

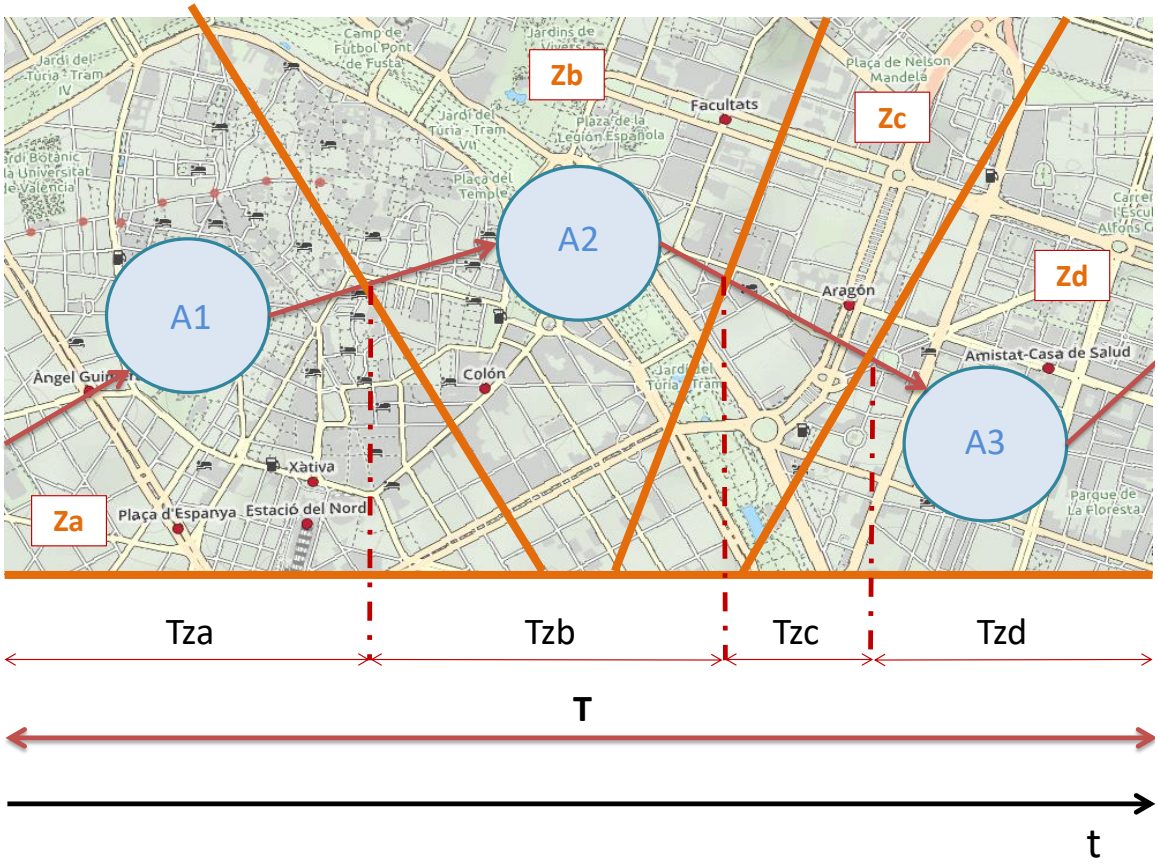


Figura 13. Ejemplo de determinación de los tiempos de estancia para cada zona 'Tz' para un periodo de tiempo 'T'.

CAPÍTULO IV: APLICACIONES PRÁCTICAS

En este capítulo se presentan tres aplicaciones prácticas donde se utilizan una o varias de las metodologías desarrolladas en esta investigación.

Cada una de las aplicaciones prácticas contiene, al menos, los siguientes apartados:

- **Objetivos y alcance del estudio**: descripción de los objetivos principales del estudio, así como la definición del alcance del mismo.

- **Descripción de los datos utilizados**: descripción de las características de los datos empleados en el estudio.

- **Metodología**: descripción de la metodología empleada. En los casos en que se aplique directamente la metodología desarrollada en el Capítulo III, se hará referencia a las secciones específicas de dicho capítulo y se especificarán los parámetros empleados en cada caso. Es importante señalar aquí que la metodología presentada en el Capítulo III recoge todas las mejoras que se han ido identificando a lo largo de esta investigación, algunas de ellas introducidas después de la realización de algunas aplicaciones prácticas. En esta sección de metodología se recogen las posibles diferencias con la metodología detallada en el Capítulo III.

- **Resultados y discusión**: presentación y discusión de los principales resultados del estudio.

- **Limitaciones de los datos de telefonía móvil**: se dedica una sección a comentar las limitaciones identificadas asociadas a los datos de telefonía móvil empleados en el estudio.

- **Conclusiones**: principales conclusiones extraídas del estudio

4.1 Análisis de la Movilidad Urbana en la Región Metropolitana de Barcelona

4.1.1 Objetivos y alcance del estudio

El objetivo principal de este estudio es aplicar la metodología desarrollada para la obtención de estadísticas de movilidad a partir de datos de telefonía móvil en un entorno real. Del mismo modo, otro de los objetivos principales del estudio es validar dicha metodología comparando los resultados obtenidos con estadísticas procedentes de encuestas.

El estudio se centra en la obtención de estadísticas de movilidad para la Región Metropolitana de Barcelona (incluye las comarcas del Alt Penedès, Baix Llobregat, Barcelonès, Garraf, Maresme, Vallès Occidental y Vallès Oriental). En la *Figura 14* se muestra el área de estudio considerada.

Figura 14. Zona de estudio – Región Metropolitana de Barcelona

Dado que uno de los objetivos principales del estudio es comparar los resultados obtenidos con encuestas, el alcance del estudio se define teniendo en consideración la información de movilidad disponible para la zona de estudio. La información sobre movilidad más actualizada del área de estudio corresponde a las encuestas de movilidad en día laborable que se realizan en la Comunidad Autónoma de Cataluña anualmente (EMEF – Enquesta de Mobilitat en día Feiner). En base a la información que proporcionan estas encuestas, se define el alcance del estudio de la siguiente manera:

- **Población de estudio**: residentes en la Región Metropolitana de Barcelona
- **Días a analizar**: día laborable promedio
- **Periodo**: septiembre-octubre 2009. Se selecciona la encuesta del 2009 ya que los datos disponibles de telefonía móvil para este estudio corresponden a dicho año (ver sección 4.1.2.1).
- **Resolución temporal**: demanda de movilidad para el total del día
- **Estadísticas de movilidad**: estadísticas básicas y matrices origen-destino:
    - **número de viajes por persona**
    - **porcentaje de usuarios que no se desplazan en el día**
    - **distribución horaria de los viajes**
    - **matrices origen-destino a nivel comarcal**. Para las matrices origen destino se analizan en concreto los viajes intra-comarcales y los viajes generados por la comarca del Barcelonès al resto de comarcas.

4.1.2 Descripción de los datos

4.1.2.1 Datos de la red de telefonía móvil

Los datos de telefonía móvil utilizados en este estudio consisten en un conjunto de CDRs y datos sobre la infraestructura de la red (en concreto datos sobre la ubicación de las antenas). No se dispone de información socio-demográfica asociada a los CDRs.

Los CDRs contienen información sobre registros de llamadas. Los CDRs utilizados para este estudio fueron recopilados para España por uno de los tres principales operadores de red de telefonía del país. Los CDRs corresponden al periodo temporal entre septiembre y noviembre de 2009. Remarcar que el periodo para el cual se dispone de información coincide con el periodo de la EMEF 2009, lo que es muy conveniente a efectos de comparación de resultados. En total se dispone de 53 días, incluyendo días laborables y fines de semana, proporcionando más de 10.000 millones de registros espacio-temporales para todo el periodo. Para este estudio se han seleccionado únicamente los días laborables. En concreto se dispone 29 días laborables, considerando como días laborables los lunes, martes, miércoles y jueves.

La información que ha sido extraída de los CDRs es la siguiente:

- **ID anonimizado del usuario principal**: usuario que realiza la llamada
- **Hora de inicio:** hora en la que se produce la llamada
- **Duración:** duración de la llamada
- **IDs únicos de las antenas:** Identificadores de las antenas al inicio y final de la llamada a las cuales se conecta el usuario principal

Este información se formatea, se depura y se pre-procesa para su posterior análisis. En la *Tabla 5* se muestra un ejemplo del formato de almacenamiento utilizado para los datos.

Tipo de evento	ID usuario principal	Fecha y hora	ID antena
Llamada	AJ3zvCRet5QW	2009-10-25 11:45:35	659028426845216
	AJ3zvCRet5QW	2009-10-25 11:52:12	659028426845180

Tabla 5. Ejemplo de datos utilizados para el estudio

Por otro lado, a partir de los datos de las ubicaciones de las antenas, se han calculado las áreas de Voronoi correspondientes a cada emplazamiento. La disposición de los emplazamientos de las antenas proporciona una resolución espacial de decenas o cientos

de metros en áreas urbanas densamente pobladas y varios kilómetros en zonas de menor densidad de población. Se ha tomado como localización aproximada dentro del área de Voronoi las coordenadas de su centroide.

Por último, remarcar que ninguno de los participantes en este estudio ha participado en los procesos de encriptación o extracción de los CDRs.

4.1.2.2 Estadísticas de movilidad – EMEF 2009

Para llevar a cabo la validación de la metodología empleada se han utilizado estadísticas procedentes de la EMEF de 2009. La EMEF del 2009 consistió en una encuesta telefónica realizada a 12.682 personas de Cataluña, de las cuales 5.797 pertenecían a la Región Metropolitana de Barcelona. La encuesta tuvo lugar entre los meses de septiembre y octubre. Un resumen de la metodología empleada y de los principales resultados de la EMEF 2009 está disponible a través de la página web de la Autoritat del Transport Metropolità (ATM) de Barcelona (enlace). De todas las estadísticas disponibles, se ha considerado relevante utilizar a efectos de comparación las estadísticas referentes a la Región Metropolitana de Barcelona, en concreto aquellas relacionadas con el número de viajes por persona, el porcentaje de personas que no realizan viajes en el día, la distribución horaria de los viajes y la distribución de los viajes (matrices OD).

4.1.3 Metodología

En esta sección se describe la metodología empleada en este estudio. Los pasos que se han seguido para calcular las estadísticas de movilidad en la zona de estudio son:

1. Identificación del lugar de residencia de los usuarios
2. Determinación del diario de actividades y viajes de los usuarios
3. Elevación muestral
4. Cálculo de estadísticas de movilidad

4.1.3.1 Identificación del lugar de residencia de los usuarios

El objetivo principal de esta etapa del proceso es determinar el lugar de residencia de los usuarios, para seleccionar aquellos usuarios objeto de estudio (residentes en la Región Metropolitana de Barcelona).

Se ha aplicado la metodología descrita en la sección 3.1.3 para obtener el lugar de residencia de los usuarios. Se han utilizado como días de muestra a analizar los 29 días laborables disponibles. Se ha considerado conveniente utilizar como parámetro 'α' un valor de 0,2 y como periodo característico de residencia el periodo comprendido entre las 8 p.m. y las 7 a.m.

4.1.3.2 Determinación del diario de actividades y viajes de los usuarios

En esta etapa el objetivo es determinar los diarios de actividades y viajes para todos los usuarios identificados como residentes en la Región Metropolitana de Barcelona. Para ello se aplica la metodología detallada en la sección 3.1.2 a los 29 días de muestra disponibles y se obtiene como resultado final la media de los resultados de todos los días. Señalar que para este estudio no se ha aplicado ningún filtro para evitar saltos entre señales[8].

En primer lugar, debe definirse el umbral temporal para la selección de los usuarios válidos. Dado que la granularidad temporal de los CDRs de llamadas es muy variable según las horas del día (especialmente muy baja durante el periodo nocturno) se considera adecuado fijar dos umbrales temporales, uno para el periodo nocturno y otro para el resto del día. Se considera como periodo nocturno el periodo comprendido entre las 21:00 y las 06:00. Para dicho periodo se aplica un umbral temporal 'TR_nocturno' de 8 horas. Para el resto del día se considera conveniente utilizar un umbral temporal 'TR_diurno' de 4 horas.

En segundo lugar, es necesario definir el tiempo de estancia en una localización para clasificar dicha estancia como actividad. Dado que el estudio se realiza en entornos

[8] La mejora de filtrado de señal no estaba incorporada en el momento en el que se llevó a cabo el estudio. No obstante, señalar que para el caso de CDRs de llamadas, al disponer de datos de baja granularidad temporal, es de esperar que la influencia del filtro de saltos entre señales sea poco significativa.

urbanos, se considera conveniente utilizar como umbral de actividad 'TA' el valor de 30 minutos.

Por último, es necesario definir los parámetros asociados a la determinación de los viajes. Se considera conveniente realizar la simplificación de que los viajes entre dos actividades se realizan en un trayecto recto en el plano y a velocidad media constante. Como velocidad media se considera conveniente utilizar una velocidad de 15 km/h[9]. Por otro lado, se utiliza una función de probabilidad basada en los datos disponibles en la EMEF 2009 para asignar la hora exacta del viaje. En concreto, se define la función de probabilidad a partir de la información del número de viajes por hora proporcionados por la EMEF 2009, siendo el valor de la probabilidad horaria el número de viajes en dicha hora dividido por el total de viajes en el día.

4.1.3.3 Elevación muestral

La muestra de usuarios de telefonía móvil se elevada a nivel de sección censal utilizando datos del censo 2011 como marco muestral. Se aplica el factor de elevación definido en la sección 3.1.4 sin considerar ningún tipo de segmentación.

4.1.3.4 Cálculo de estadísticas de movilidad

Las estadísticas de movilidad se obtienen del procesamiento de la información proporcionada por los diarios de actividad y viajes de los usuarios. Para este estudio se calculan las siguientes estadísticas:

- **Número de viajes por persona**: número de viajes medio por persona en el día.
- **Porcentaje de personas que no realizan ningún viaje**: se calcula como el número de personas sin viajes con respecto al total de personas.
- **Distribución de los viajes**: se calcula la distribución de los viajes por hora del día. Se considera como criterio de hora del viaje la hora del inicio del viaje. El número

[9] Velocidad media considerada para todos los modos de transporte (pie, transporte público, coche, etc.) bajo la hipótesis de trayectoria recta en el plano.

de viajes en una hora concreta se obtiene como la suma de todos los viajes realizados por los usuarios a esa hora. Los resultados se presentan en porcentaje con respecto al total del día.

- **Viajes intrazona a nivel comarcal**: el número de viajes intrazonales en cada comarca se calcula como el sumatorio de todos los viajes con origen y destino la comarca. La información se presenta en porcentaje con respecto al total de viajes en dicha comarca.

- **Distribución de los viajes de la comarca del Barcelonès:** para calcular la distribución de los viajes se realiza el sumatorio de todos los viajes con origen la comarca del Barcelonès y destino el resto de comarcas de la Región Metropolitana de Barcelona. El resultado se proporciona en porcentaje con respecto al total de viajes con destino fuera de la comarca del Barcelonès.

4.1.4 Resultados y discusión

En esta sección se presentan los resultados obtenidos comparándolos al mismo tiempo con la información proporcionada por la EMEF 2009.

En la *Tabla 6* se muestran los resultados correspondientes a la cantidad de muestra útil disponible, el porcentaje de personas sin viajes en el día y el número medio de viajes por persona. La población total de la Región Metropolitana de Barcelona es de 4,2 millones de personas (año 2009). La muestra de usuarios utilizada por la EMEF 2009 para la Región Metropolitana es de 5.797 personas mientras que la muestra de usuarios útil obtenida de la telefonía móvil después de los procesos de depuración es de 68.247 personas. La muestra útil obtenida con telefonía supone alrededor del 3% de la potencial, debido principalmente a la baja granularidad temporal de los datos; aun así, la muestra obtenida con telefonía móvil es casi 12 veces superior que la utilizada por la encuesta. Respecto al porcentaje de personas que no realizan ningún viaje, la EMEF 2009 y el método propuesto proporcionan valores similares, 9,9% y 13,7% respectivamente. Por otro lado, la EMEF 2009 proporciona unos valores de viajes por persona de 3,72 mientras que el método

basado en telefonía proporciona valores de 3,58, lo que supone unas discrepancias inferiores al 5% en cuanto a generación de viajes.

Estadísticas Básicas	EMEF 2009	Telefonía Móvil
Muestra	5.797	68.247
Porcentaje de personas que no realizan ningún viaje	9.9%	13.7 %
Average number of trips per user	3.72	3.58

Tabla 6. Estadísticas básicas de movilidad y comparativa con EMEF 2009

Adicionalmente, se han calculado y comparado los viajes intra-comarcales y los viajes con origen la comarca del Barcelonès y destino el resto de comarcas, con el objetivo de comparar la información de distribución de viajes que proporcionan ambas metodologías. En la *Figura 15* se muestra una comparativa entre la EMEF 2009 y el método basado en telefonía móvil, comparando el porcentaje de viajes intra-comarcales con respecto al total de viajes generados para cada una de las 7 comarcas de la Región Metropolitana de Barcelona: Alt Penedès, Vallès Occidental, Maresme, Baix Llobregat, Vallès Oriental, Barcelonès y Garraf. Del mismo modo, en la *Figura 16* se muestra la distribución del número de viajes entre el Barcelonès y el resto de las comarcas, proporcionándose la información en porcentaje con respecto al total de viajes externos generados por el Barcelonès. Puede apreciarse como, en ambos casos, la información que proporciona la EMEF 2009 y el método basado en telefonía móvil es muy similar.

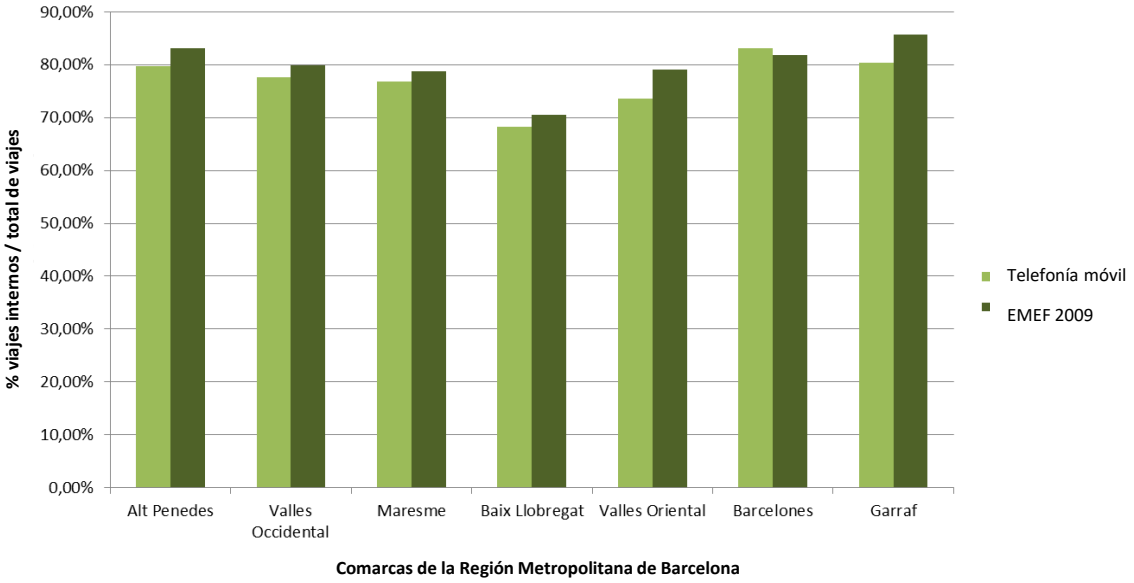


Figura 15. Porcentaje de viajes internos con respecto al total de viajes

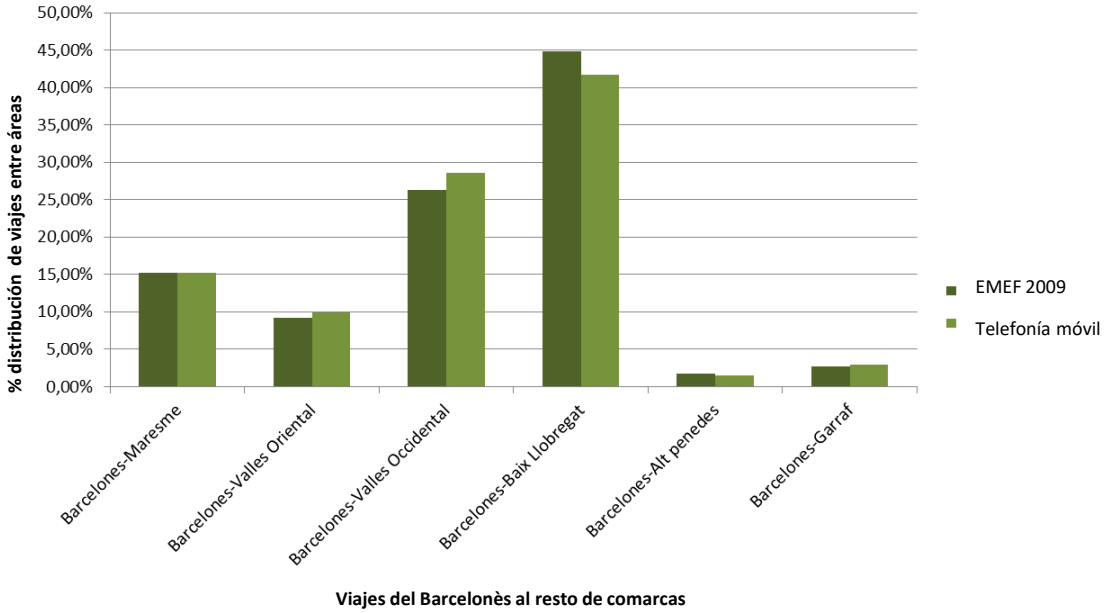


Figura 16. Distribución de viajes desde el Barcelonès al resto de comarcas de la Región Metropolitana de Barcelona

También se ha calculado y comparado la distribución horaria de los viajes según la EMEF 2009 y según el método propuesto. En la *Figura 17* se muestra, para ambos métodos, el porcentaje de viajes en cada hora del día con respecto al total de viajes en el día. Puede observarse como el perfil de viajes en ambos casos es muy similar, con pequeñas discrepancias principalmente en las horas punta. El método de telefonía móvil presenta

un porcentaje mayor de viajes en la hora punta de la mañana (7,81% frente a 7,07%), mientras que la EMEF 2009 presenta una punta por la tarde más temprana y de mayor magnitud (8,47% de los viajes entre las 5 p.m. y las 6 p.m. frente al 7,37% de los viajes entre las 7 p.m. y las 8p.m.). Según la EMEF 2009, el 90,3% de los viajes se producen entre las 7 y las 21 horas; según el método basado en telefonía este porcentaje es del 87,35%, muy similar al proporcionado por la EMEF 2009.

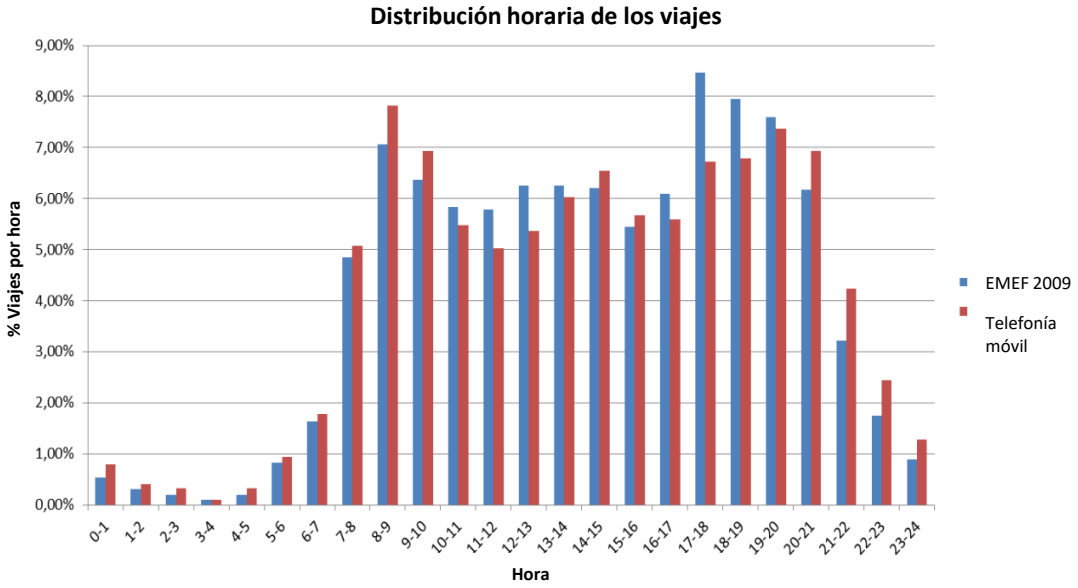


Figura 17. Zona de estudio – Región metropolitana de Barcelona

En este punto es importante señalar un aspecto relevante con respecto a la comparativa entre los perfiles horarios de los viajes. El método de telefonía móvil, según se ha comentado en la sección 4.1.3.2, utiliza la información del perfil horario de los viajes de la EMEF 2009 para calcular la función de probabilidad que determina la hora exacta del viaje. Esto significa que, en periodos del día con baja granularidad temporal de los datos de telefonía móvil, el perfil obtenido mediante telefonía tenderá a reproducir en gran medida el perfil de la EMEF 2009 (especialmente en horario nocturno). No obstante, en periodos donde se dispone de más datos, la influencia de la función de probabilidad será mucho

más reducida. De hecho, este aspecto puede apreciarse claramente en las discrepancias encontradas en las horas punta del día, especialmente en el caso de la punta de la tarde.

Por último, aparte de la comparativa de los indicadores de movilidad, también es relevante evaluar características como el coste económico de ambas metodologías o los plazos de ejecución. Respecto a la EMEF 2009, el presupuesto para llevar a cabo solo los trabajos de campo estaría en torno a los 80.000€ y el plazo de ejecución rondaría el mes[10]. La obtención de estadísticas básicas de movilidad y matrices OD con datos de telefonía móvil con una resolución espacial similar a la de la encuesta podría tener un presupuesto de unos 25.000€ y un plazo de ejecución de unos pocos días o semanas[11]. El método de telefonía móvil supone una reducción importante de costes y de plazos. No obstante, también es importante señalar que la encuesta de la EMEF también recoge otro tipo de información que no se obtiene mediante la metodología propuesta basada en datos de telefonía móvil (motivo detallado de los viajes, modo de transporte, información socio-demográfica, opinión de los usuarios, etc.). Por lo tanto, la metodología basada en datos de telefonía móvil puede complementar y mejorar parte de la información de las encuestas, pero no sustituirla completamente. La principal ventaja del método basado en telefonía móvil es la calidad de sus matrices OD, ya que, al calcularse sobre una muestra de población muy superior, es capaz de identificar potencialmente pares OD que las encuestas convencionales podrían no detectar (la problemática de los 'ceros' en las matrices de movilidad). A medida que aumenta el tamaño de la muestra de población necesaria para el estudio, más atractiva es la solución basada en telefonía móvil. Por ejemplo, para el caso de encuestas domiciliarias con unas muestras de población en torno a 1-3%, con unos costes de ejecución de 1-4 millones de euros y unos plazos de ejecución de años, la telefonía móvil presenta una alternativa muy atractiva (gran tamaño de muestra, costes de pocos cientos de miles de euros, plazo de ejecución de semanas) para el cálculo de matrices OD de calidad (Picornell & Willumsen 2016).

[10] La estimación se ha basado en los datos de la licitación de la EMEF de 2015 (enlace) aplicando un factor corrector para tener en cuenta las diferencias respecto al número de entrevistas de cada trabajo.
[11] Estimación basada en los precios y plazos publicados en Picornell & Willumsen (2016)

4.1.5 Limitaciones de los datos de telefonía móvil

Los datos de telefonía móvil utilizados en el estudio corresponden únicamente a datos de llamadas telefónicas, por lo que solo se dispone de información de localización del dispositivo cuando se recibe o realiza una llamada. Esto conlleva que la granularidad temporal de los datos para un número significativo de usuarios sea muy baja. Esto queda reflejado en el reducido porcentaje de muestra últil (3% respecto del total de usuarios de telefonía móvil) que se obtiene una vez realizado el proceso de depuración de la muestra. Del mismo modo, la baja granularidad temporal hace que la determinación de la hora del viaje sea poco precisa (especialmente en horario nocturno), necesitándose información complementaria de encuestas (y/o aforos, billetaje, etc.) para poder obtener resultados satisfactorios. Por otro lado, al no disponer de información socio-demográfica para corregir la muestra, puede que los resultados obtenidos presenten algún sesgo como consecuencia de que algunos segmentos de la población no estén correctamente representados.

Las principales limitaciones identificadas están asociadas con las características de los datos de telefonía móvil disponibles para este estudio. La mejora de la granularidad temporal de los datos (por ejemplo, mediante el almacenamiento de datos de SMS, sesiones de datos o la recogida de información sobre la localización del dispositivo de manera periódica) eliminaría gran parte de las limitaciones encontradas. Del mismo modo, la disponibilidad de información socio-demográfica fiable de los usuarios permitiría corregir los posibles sesgos presentes en la muestra.

4.1.6 Conclusiones

La metodología propuesta en el Capítulo III para obtener estadísticas de movilidad a partir de datos de telefonía móvil ha sido testada en un entorno real, en concreto en la Región Metropolitana de Barcelona. El objetivo principal de este estudio era calcular diferentes estadísticas de movilidad y, al mismo tiempo, validar la metodología propuesta comparando los resultados obtenidos con información procedente de encuestas. Los

resultados de la comparativa muestran que la metodología basada en datos de telefonía móvil proporciona resultados muy similares a los obtenidos mediante encuestas, demostrando el potencial de los datos de telefonía móvil para capturar patrones de movilidad. Remarcar que la comparativa ha podido llevarse a cabo con datos del mismo periodo temporal, lo que ha sido muy conveniente a efectos de validación de la metodología. La comparativa se ha realizado tanto para estadísticas básicas de movilidad (por ejemplo, número de viajes por persona) como para información sobre distribución de viajes (matrices OD). La muestra útil de usuarios obtenida mediante la telefonía móvil es muy superior (12 veces superior) que la muestra empleada por la encuesta, a pesar de la baja granularidad temporal de los datos de telefonía móvil. Es importante señalar que la comparativa solamente ha podido realizarse a un nivel bastante agregado, debido a las limitaciones asociadas a la información disponible de la EMEF 2009. Niveles superiores de desagregación de la información podrían dar lugar a discrepancias más significativas entre los resultados de ambas metodologías (especialmente en lo referente a distribución de viajes – matrices OD). Por otro lado, también se han comparado las distribuciones horarias de los viajes obtenidas por ambas metodologías, obteniendo resultados muy similares. Las principales discrepancias se observan en la hora punta de la tarde, donde la EMEF 2009 presenta una punta por la tarde más temprana y de mayor magnitud (8,47% de los viajes entre las 5-6 p.m. frente al 7,37% de los viajes entre las 7- 8p.m. según telefonía móvil). Por último, también se ha realizado una comparativa respecto a los costes y plazos de ejecución de ambas soluciones. La solución basada en telefonía móvil supone una reducción de costes y plazos significativa. No obstante, es importante señalar que la información que se obtiene mediante telefonía móvil no sustituye completamente la información recogida mediante encuestas. La principal ventaja de los datos de telefonía móvil es la calidad de sus matrices OD, al estar obtenidas de una muestra muy elevada de usuarios. Esta metodología implica reducciones de costes y plazos muy significativas para encuestas de gran tamaño (encuestas domiciliarias). Por último, señalar que, una interesante futura línea de trabajo podría estar enfocada a la comparación de la metodología basada en datos de telefonía con información de encuestas de mayor calidad

(por ejemplo, encuestas domiciliarias) que permitieran una mejor y más detallada comparación de resultados entre ambas metodologías.

4.2 Análisis del potencial de los datos de telefonía para caracterizar las relaciones entre la red social y los patrones de movilidad

4.2.1 Antecedentes, objetivos y alcance del estudio

El objetivo principal de este estudio[12] es examinar las relaciones entre la red social y los patrones de movilidad de las personas a través del uso de los datos de telefonía móvil. El estudio se centra, por un lado, en el análisis de las localizaciones frecuentes compartidas por los usuarios de una misma red social y, por otro lado, en el análisis de los lugares que visitan de manera conjunta (co-ubicación). El objetivo último es identificar la naturaleza de los lugares compartidos por los usuarios de una misma red social. La mayoría de estudios previos se han centrado en el análisis de la distribución de distancias entre los lugares de residencia de los miembros de la red social y algunos pocos también han analizado la distribución con respecto al lugar de trabajo (por ejemplo, Phithakkitnukoon et al. 2012). Sin embargo, la caracterización de los lugares comúnmente visitados por los miembros de la red social así como la naturaleza de los mismos es un aspecto no explorado aún que se aborda en este estudio. Del mismo modo, estudios previos han analizado los eventos de co-ubicación sin considerar en detalle los patrones de movilidad de los usuarios. Por ejemplo, Calabrese et al. 2011b considera co-ubicación cuando se realiza una llamada entre los miembros de una misma red social y esta se produce en un sitio cercano al del supuesto lugar de encuentro, representando la llamada un acto de coordinación entre los miembros de la red social. Esta hipótesis puede no detectar situaciones en las que no se produce coordinación previa o cuando la coordinación se produce en una zona alejada del lugar de encuentro. Por otro lado, Chen & Mei 2014 consideran que se produce co-ubicación cuando dos personas se encuentran en la misma zona durante un mismo periodo, dividiendo el día en dos periodos, mañana (8 a.m. – 8 p.m) y tarde (8:01 p.m – 7:59 a.m.). La extensión de los periodos considerados hace posible que dos personas que aparecen en dichos periodos en una misma zona no coincidan en el mismo instante. En el presente estudio, de cara a mejorar la estimación de co-ubicación, se identifica la

[12] La versión publicada de este estudio puede encontrarse en el siguiente enlace

ubicación de cada miembro de la red social a lo largo de todo el día de manera independiente y se comparan las ubicaciones de los miembros de una misma red social entre sí. Este planteamiento es independiente de si se producen o no llamadas de coordinación entre los miembros de la red social y permite identificar eventos de co-ubicación con una incertidumbre temporal más reducida que en estudios previos.

El estudio presenta dos alcances diferenciados. Por un lado, para analizar la red social y las localizaciones frecuentes se analizan todos los usuarios móviles de España de uno de los principales operadores del país. Por otro lado, para el estudio de las situaciones de co-ubicación, el estudio se centra en una sub muestra de usuarios, en concreto en los usuarios residentes en el Área Metropolitana de Barcelona.

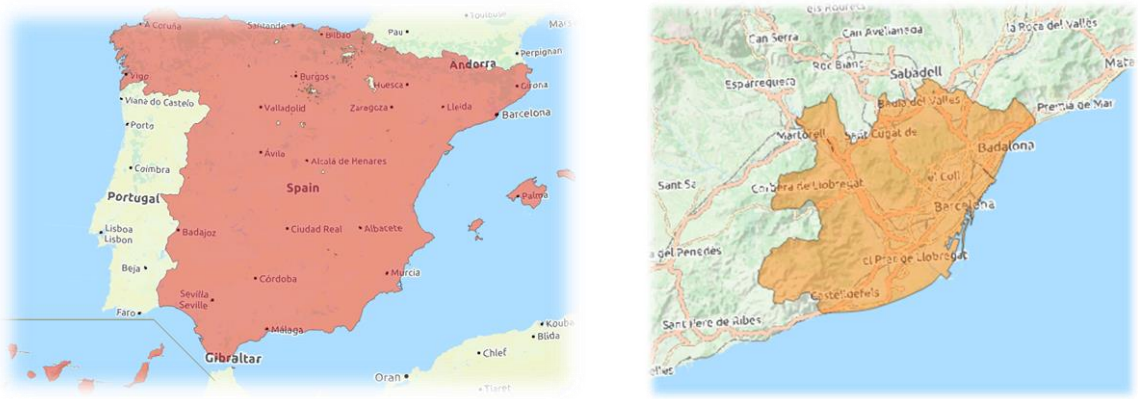


Figura 18. Ámbitos de estudio para el análisis de la red social y la movilidad. Izquierda – Ámbito nacional para el estudio de red social y localizaciones frecuentes. Derecha – Ámbito para el estudio de situaciones de co-ubicación(Área Metropolitana de Barcelona)

4.2.2 Descripción de los datos

Los datos de telefonía móvil utilizados[13] en este estudio consisten en un conjunto de CDRs y datos sobre la infraestructura de la red (en concreto datos sobre la ubicación de las antenas). No se dispone de información socio-demográfica asociada a los CDRs.

[13] Señalar que el corpus de los datos utilizados en este apartado es el mismo que en la sección 4.1.2

Los CDRs contienen información sobre registros de llamadas. Los CDRs utilizados para este estudio fueron recopilados para España por uno de los tres principales operadores de red de telefonía del país. Los CDRs corresponden al periodo temporal entre septiembre y noviembre de 2009. En total se dispone de 53 días, incluyendo días laborables y fines de semana, proporcionando más de 10.000 millones de registros espacio-temporales para todo el periodo. La información que ha sido extraída de los CDRs es la siguiente:

- **ID anonimizado del usuario principal**: usuario que realiza la llamada
- **ID anonimizado del usuario secundario**: usuario que recibe la llamada
- **Hora de inicio:** hora en la que se produce la llamada
- **Duración:** duración de la llamada
- **IDs únicos de las antenas:** Identificadores de las antenas al inicio y final de la llamada

Este información se formatea, se depura y se pre-procesa para su posterior análisis. En la *Tabla 7* se muestra un ejemplo del formato de almacenamiento utilizado para los datos.

Tipo de evento	ID usuario principal	ID usuario secundario	Fecha y hora	ID antena
Llamada	AJ3zvCRet5QW	++/Gjhgyge45	2009-10-25 11:45:35	659028426845216
	AJ3zvCRet5QW	++/Gjhgyge45	2009-10-25 11:52:12	659028426845180

Tabla 7. Ejemplo de datos utilizados para el análisis de la red social y movilidad

Por otro lado, a partir de los datos de las ubicaciones de las antenas, se han calculado las áreas de Voronoi correspondientes a cada emplazamiento. La disposición de los emplazamientos de las antenas proporciona una resolución espacial de decenas o pocos cientos de metros en áreas urbanas densamente pobladas y varios kilómetros en zonas rurales.

Por último, remarcar que ninguno de los participantes en este estudio ha participado en los procesos de encriptación o extracción de los CDRs.

4.2.3 Metodología

En esta sección se describe la metodología empleada en este estudio. Los pasos que se han seguido para analizar las relaciones entre la red social y los patrones de movilidad de la población son los siguientes:

1. Determinación de la red social de los usuarios
2. Identificación de las localizaciones frecuentes de los usuarios
3. Determinación de la presencia de los usuarios a lo largo del día.
4. Análisis de las relaciones entre la red social y las localizaciones frecuentes
5. Análisis de co-ubicación

4.2.3.1 Determinación de la red social de los usuarios

El primer paso del proceso es determinar la red social de los usuarios a partir del análisis de los datos disponibles de telefonía móvil. La metodología seguida para calcular la red social es la explicada en la sección 3.2.2. El resultado que se obtiene al aplicar esta metodología es una red egocentrista ponderada. Se ha utilizado como indicador del grado de relación entre el *ego* y un *alter* el número total de llamadas entre ellos.

4.2.3.2 Identificación de las localizaciones frecuentes de los usuarios

En esta etapa, el objetivo es determinar las localizaciones frecuentes[14] de cada uno de los usuarios. Para ello se han utilizado los 53 días disponibles de la muestra y se ha aplicado la metodología explicada en la sección 3.1.3. A la vista de los datos disponibles, se ha considerado conveniente utilizar como primera aproximación un valor de 'α' de 0,35 para estimar las localizaciones frecuentes. Para los casos específicos de casa y trabajo se ha considerado conveniente utilizar valores de 'α' de 0,2 y 0,3 respetivamente. Del mismo

[14] En este estudio se emplea el término "localizaciones frecuentes" como sinónimo de "actividades frecuentes".

modo, se ha considerado conveniente utilizar como periodo de casa el periodo comprendido entre las 8 p.m. y las 7 a.m., y como periodo de trabajo el periodo comprendido entre las 8 a.m. y las 5 p.m. Para la determinación de los lugares de residencia y trabajo se han utilizado únicamente los días laborables de lunes a jueves.

4.2.3.3 Determinación de la presencia de los usuarios a lo largo del día

Para cada uno de los usuarios presentes en la muestra se ha calculado su diario de actividades y viajes. La metodología seguida en el estudio es similar a la explicada en las secciones 3.1 y 3.3 con algunas pequeñas modificaciones. Las modificaciones vienen motivadas por el hecho de que cuando se realizó este estudio, la metodología de presencia de población explicada en la sección 3.3 no estaba aún desarrollada. A continuación se detalla paso a paso la metodología empleada en este estudio para estimar la presencia del usuario a lo largo del día:

1. Se recoge la información geolocalizada del primer registro del usuario (L0, t0)
2. Se recoge la información geolocalizada del siguiente registro (L1, t1)
3. Si (t1-t0) > TR → la información de localización entre t0 y t1 no está disponible
4. En caso contrario ((t1-t0) <= TR):
    a. Si L0=L1=L → la localización del usuario entre [t0,t1] es L
    b. Si L0≠L1 → t'=f (t0, t1). La localización del usuario es L0 entre [t0, t'] y L1 entre [t',t1].

Siendo 'TR' el tiempo máximo entre dos registros consecutivos para considerar que no existe ninguna localización intermedia relevante; y f(ti,tj) una función de probabilidad que determina el instante en el cual se produce el paso de una localización a otra (se considera que el paso de una localización a otra se produce de manera instantánea). En la *Figura 19* se muestra un ejemplo de la transformación de registros de telefonía a información de presencia a lo largo del día. En el ejemplo se aprecia como el usuario realiza estancias en las localizaciones L0, L1 y L2 pero no se dispone de datos suficientes en el periodo [t1, t2] para aportar información.

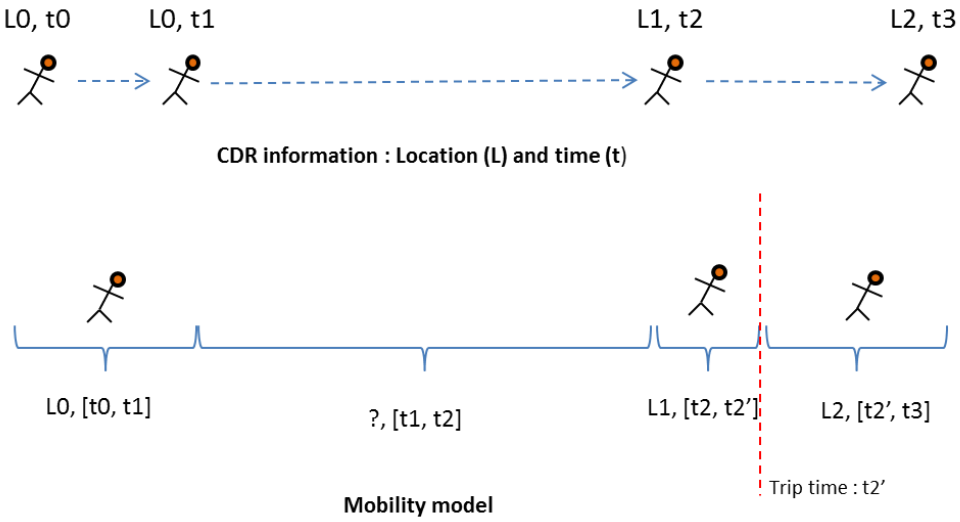


Figura 19. Ejemplo de transformación de registros de telefonía a información de presencia

4.2.3.4 Análisis de las relaciones entre la red social y las localizaciones frecuentes

El objetivo principal de este proceso es analizar la relación entre las localizaciones frecuentes visitadas por el *ego* y aquellas visitadas por su red social. Solamente se consideran en esta etapa los usuarios para los cuales se ha podido identificar su lugar de residencia y su lugar de trabajo. Para cada red egocentrista, se comparan las localizaciones frecuentes del *ego* con las localizaciones frecuentes de los *alters*; identificando las localizaciones comunes y clasificando la relación ego-alter según las tipologías de localizaciones frecuentes compartidas (casa, trabajo y otras). Por lo tanto, se definen un total de 9 posibles relaciones, derivadas de las combinaciones posibles entre las tipologías de localizaciones frecuentes. Señalar que puede darse la situación de que la localización de casa sea igual a la localización del trabajo para un mismo usuario. En estos casos, el tipo de relación se asigna de manera proporcional (por ejemplo, si el usuario 'A' y el usuario 'B' comparten una localización común 'L', y 'L' es la localización de casa y trabajo para el usuario 'A' y es 'otra' localización frecuente para el usuario 'B'; entonces la relación se clasifica como 50% casa-otra y 50% trabajo-otra).

4.2.3.5 Análisis de co-ubicación

El objetivo de este análisis es comparar la información de presencia de todos los usuarios de una misma red social para identificar situaciones de co-ubicación. Se define co-ubicación como la situación en la que dos personas se encuentran ubicadas en un mismo lugar en el mismo instante de tiempo. Para cada red egocentrista se compara la información de presencia del *ego* con los diarios de los *alters* para todos los días de la muestra. Cuando se identifica co-ubicación, esta se clasifica en función de las tipologías de localización del *ego* y del *alter*. En este caso se consideran 4 tipologías de localización, las tres asociadas a localizaciones frecuentes y otra clasificada como 'no frecuente', para los casos en los que la localización no forma parte de las localizaciones frecuentes. Por lo tanto, existen 16 posibles tipologías de co-ubicación, asociadas a las posibles combinaciones de las 4 tipologías de localización. El análisis de co-ubicación se ha realizado sobre un subconjunto de la muestra, en concreto, se han analizado únicamente los usuarios cuya residencia se ha identificado en el Área Metropolitana de Barcelona (estos usuarios definirán el conjunto de egos de la muestra) y sus contactos sociales (independientemente de su lugar de residencia). Por último, al igual que en el análisis anterior, los usuarios para los cuales no se ha identificado casa y trabajo han sido descartados de la muestra.

4.2.4 Resultados y discusión

En esta sección se presentan y discuten los resultados del estudio. En concreto se presentan los resultados correspondientes a la generación de la red social, al análisis de las localizaciones frecuentes, a la estimación de la información de presencia, y a la relación entre la red social y los patrones de actividad y movilidad de la población. En los resultados se emplea el símbolo 'σ' para representar la desviación típica de las variables analizadas.

4.2.4.1 Estadísticas de la red social

Para la totalidad de la muestra de usuarios (alrededor de 24 millones de usuarios), se han generado redes egocentristas en función de las llamadas entre los usuarios. El número medio de alters por ego identificado es de 9,31, con una desviación típica de 17,19 (de ahora en adelante, para representar la desviación típica de una variable, la media se acompañará del símbolo 'σ'; en este caso sería σ=17,19). La media de llamadas entre dos usuarios (variable que mide el grado de relación entre los usuarios) es de 21. Por otro lado, señalar que el 90% de los egos tiene menos de una llamada por día con cada alter.

4.2.4.2 Estadísticas de localizaciones frecuentes

Del análisis de las localizaciones frecuentes se obtiene que la media de localizaciones frecuentes por usuario es de 3,47 (σ=2,83). Considerando que la muestra de usuarios que se ha utilizado para realizar estas estadísticas cumple la condición de que al menos las localizaciones de casa y trabajo han sido identificadas, en media, cada usuario tiene 1,5 localizaciones frecuentes adicionales, las cuales podrían estar asociadas a actividades sociales. Este resultado muestra que existen otras localizaciones frecuentes aparte de las normalmente consideradas (casa, trabajo) cuya importancia (considerada como el número de localizaciones de un cierto tipo con respecto del total de localizaciones del usuario) es similar a las relacionadas con las actividades casa y trabajo y, en algunos casos, incluso más importantes (basándonos en los resultados de desviación típica de la variable). Este resultado refuerza la idea de que considerar otras localizaciones frecuentes aparte de casa y trabajo es esencial para capturar correctamente los patrones de movilidad de los usuarios. En la *Figura 20* se muestra la distribución de 'otras' localizaciones frecuentes con respecto a la tipología de día en que se realizan estas actividades. La mayoría de estas localizaciones frecuentes se han identificado los martes, los jueves y los viernes. Remarcar que las actividades frecuentes en días laborables y días en fin de semana se identifican en base al total de días analizados de ese mismo tipo. Por lo tanto, es posible clasificar por ejemplo una actividad como frecuente en base a los lunes pero no frecuente en base a

días laborables. Por otro lado, también hay localizaciones frecuentes que lo son respecto de un día concreto de la semana y respecto a los días laborables y los fines de semana.

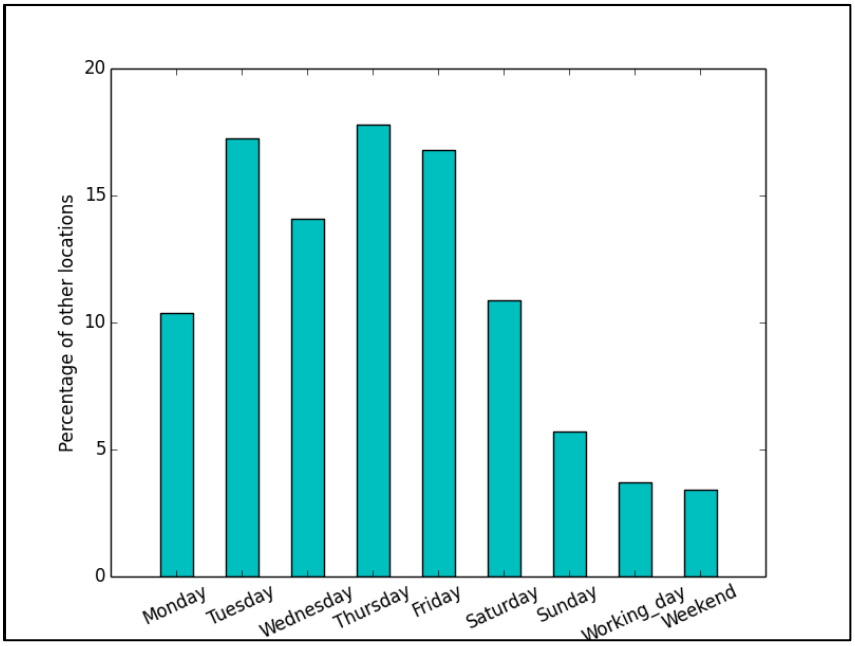


Figura 20. Distribución de 'otras' localizaciones frecuentes según la tipología de día

De cara a validar la metodología empleada para la determinación del lugar de residencia, se ha realizado un análisis de correlación comparando los resultados obtenidos con la distribución de la población española para todo el territorio nacional. Los resultados se han comparado a nivel de provincias (52 provincias), mostrando una alta correlación (R^2 = 0,93). Del mismo modo, para validar los resultados obtenidos en cuanto a las relaciones casa-trabajo, se ha realizado un análisis de correlación comparando los resultados obtenidos con los datos de residencia y puestos de trabajo del censo de 2011 para el Área Metropolitana de Barcelona. La información se analiza a nivel municipal (36 municipios) obteniéndose unos resultados satisfactorios (R^2 = 0,99). En la *Figura 21* y en la *Figura 22* se muestran los análisis de correlación realizados para el caso del estudio del lugar de residencia y para el caso del estudio de las relaciones casa-trabajo respectivamente. A la vista de los resultados, puede decirse que los coeficientes 'α' empleados para la determinación de las localizaciones frecuentes de casa y trabajo parecen apropiados. El coeficiente 'α' para el caso de 'otras' localizaciones frecuentes es más difícil de validar,

debido a que no existen estadísticas relevantes con respecto a estas variables. Sin embargo, a la vista de que los coeficientes de casa y trabajo han aportado resultados satisfactorios, un coeficiente 'α' de 0,35 para el caso de 'otras' localizaciones frecuentes parece razonable. Es importante señalar que los coeficientes 'α' propuestos en este estudio son adecuados para la resolución temporal y espacial de los datos de telefonía utilizados en este caso, y que otros coeficientes 'α' similares podrían aportar resultados igualmente satisfactorios. Para determinar el rango de valores 'α' apropiados para cada tipo de localización frecuente sería necesario realizar un estudio de sensibilidad del parámetro 'α', lo cual queda fuera del alcance de este estudio.

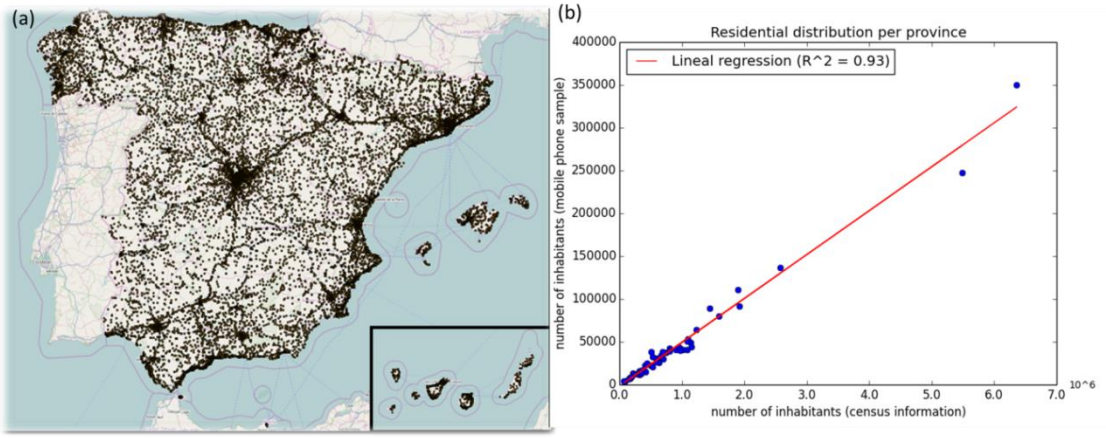


Figura 21. (a) Distribución del lugar de residencia basada en los datos de telefonía móvil (b) Análisis de correlación entre el censo de población y los resultados con telefonía móvil

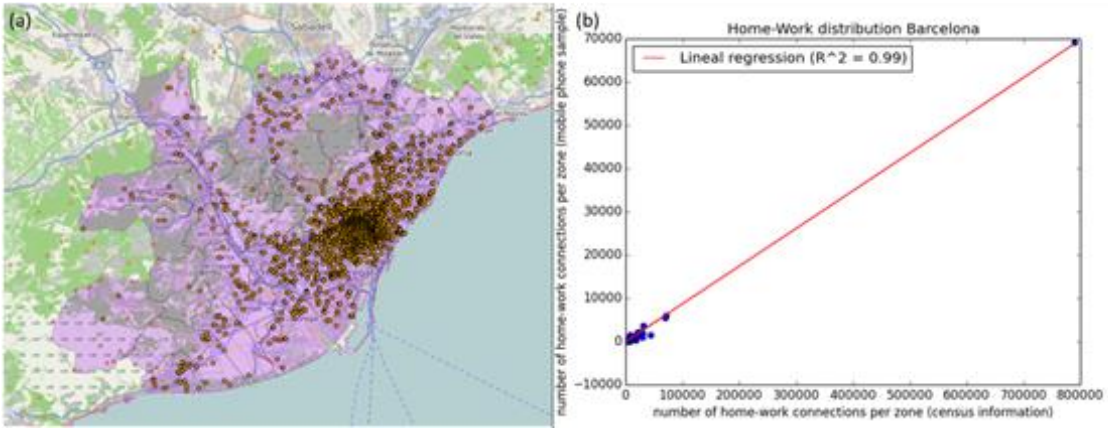


Figura 22. (a) Distribución espacial de los lugares de residencia y puestos de trabajos para el Área

Metropolitana de Barcelona obtenidos mediante telefonía móvil (b) Análisis de correlación entre la información de residencia y puestos de trabajo del censo 2011 y los obtenidos mediante telefonía móvil

4.2.4.3 Información de presencia de población

Se ha calculado la información de presencia para todos los residentes en el Área Metropolitana de Barcelona y para la totalidad de la red social asociada a la misma (independientemente de su lugar de residencia). En total, se han analizado alrededor de 250.000 usuarios. La función de probabilidad f(ti,tj) para determinar la hora del viaje (hora de cambio de una localización a otra) se ha especificado a partir de la información de la encuesta de movilidad en día laborable del año 2009 ('Enquesta de Mobilitat en Día Feiner', EMEF 2009). La probabilidad de viajar a una hora específica es igual al cociente entre el número de viajes en esa hora frente al total de viajes en el día. Se ha considerado apropiado asignar al parámetro 'TR' (tiempo máximo entre dos registros consecutivos para considerar que no existe ninguna localización intermedia relevante) el valor de 4 horas.

Los resultados del modelo propuesto proporcionan, en media, 4,6 horas de información sobre la localización de los usuarios para días laborables y 2 horas de información para fines de semana (2,5 horas los sábados y 1,5 horas los domingos). En la *Figura 23* se muestra la distribución del tiempo de información medio sobre localización para cada hora del día. Puede apreciarse como en horario nocturno se dispone de mucha menos información que durante el horario diurno, debido a que la actividad telefónica de llamadas se realiza mayoritariamente en horario diurno. Del mismo modo, para días laborables, se aprecian dos puntas en el día en las cuales se dispone de mayor información, que están asociadas a las 12 y a las 18 horas. En el caso de fines de semana también se aprecia esta tendencia pero las puntas de información se producen con un cierto retranqueo con respecto a los días laborables, siendo las horas de mayor disponibilidad de información alrededor de las 13 y las 20 horas.

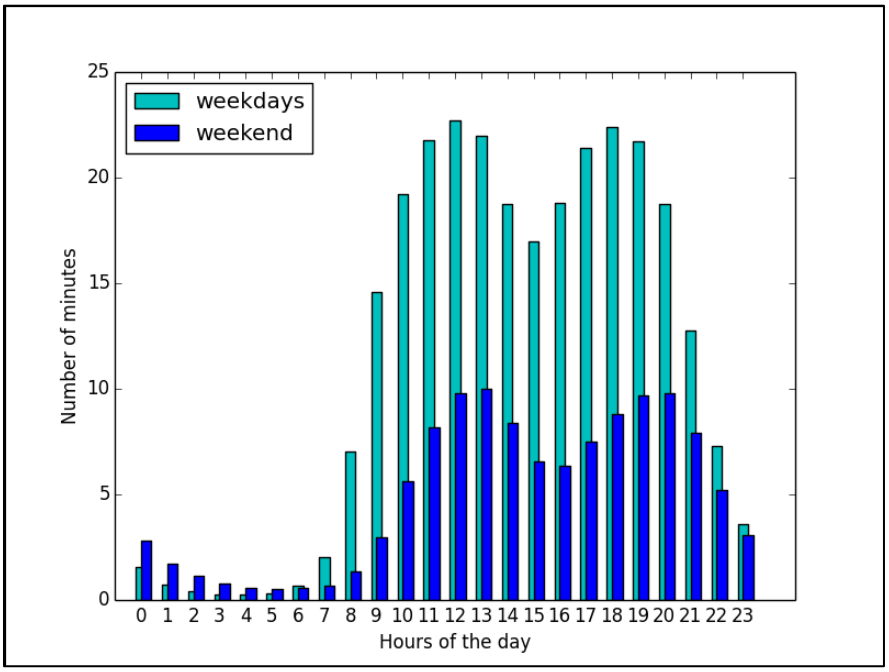


Figura 23. Número de minutos de información de presencia media por usuario para cada hora del día y para distintos tipos de días (laborables (weekdays) y fines de semana (weekend))

4.2.4.4 Análisis de las localizaciones frecuentes compartidas por la red social

Para el total de 24 millones de redes egocentristas calculadas, sólo aquellas en las que ha sido posible identificar la casa y el trabajo del ego han sido consideradas (alrededor de 2,3 millones de redes). Del mismo modo, sólo los alters con casa y trabajo han sido considerados. Para cada red egocentrista, alrededor del 17% (σ=13%) de los alters proporcionan dicha información.

Los resultados muestran que cada ego comparte, en media, al menos una localización frecuente con el 61,23% (σ=36,88%) de los alters. Este resultado muestra una alta relación entre la red social y las localizaciones frecuentes visitadas por sus miembros. Los egos comparten 1,36 (σ=0,96) localizaciones frecuentes con cada alter. Considerando que cada usuario tiene de media 3,47 localizaciones frecuentes (ver sección 4.2.4.2), cada ego comparte el 40% de dichas localizaciones con cada uno de sus alters. Este resultado muestra que no solamente el número de alters que comparten localizaciones frecuentes

con el ego es relevante, sino que también es relevante el número de localizaciones frecuentes en común entre el ego y los alters. Además, se observa una alta correlación lineal positiva (R^2 = 0,97) entre la media de llamadas y el número de localizaciones frecuentes compartidas, lo que sugiere que, a mayor grado de relación entre el ego y el alter (medido como el número de llamadas entre el ego y el alter) mayor número de localizaciones frecuentes en común. En la *Figura 24* se muestra la correlación entre el número medio de llamadas entre dos usuarios y el número de localizaciones frecuente en común.

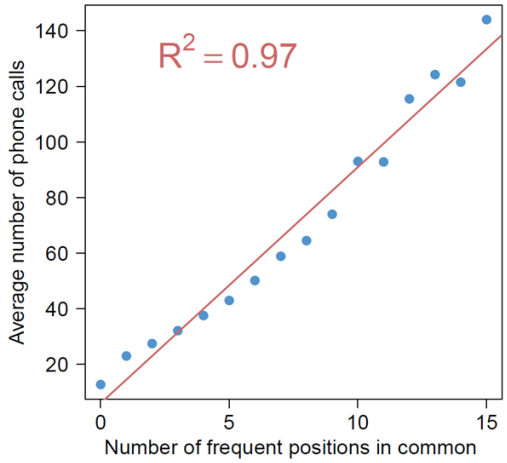


Figura 24. Correlación entre el número medio de llamadas entre dos usuarios y el número de localizaciones frecuente en común

Para cada una de las localizaciones frecuentes compartidas entre el ego y los alters se ha identificado el tipo de relación en función del tipo de localización compartida (casa, trabajo, otra). Por ejemplo, si el ego reside en la localización 'L' y el alter visita frecuentemente al ego en su domicilio (por lo que la localización 'L' es una localización del tipo 'otra' frecuente para el alter), el tipo de relación entre ellos se clasificara para este caso como casa-otra. Se han identificado y analizado un total de 9,6 millones de interacciones ego-alter. En la *Tabla 8* se muestra la distribución del tipo de relaciones entre el ego y los alters identificadas. Desde el punto de vista del ego, la mayoría de las interacciones con los alters se producen en 'otras' localizaciones frecuentes (54%) y en menor media en casa y en el trabajo (21% y 24% respectivamente). El tipo de relación más

común entre el ego y los alters es aquella en la que ambos comparten 'otra' localización frecuente distinta de sus casas y trabajos. Se podría considerar que las relaciones sociales son aquellas en las que al menos una de las localizaciones frecuentes se clasifica como 'otra'. Los resultados muestran que al menos el 27% (σ = 32%) de las 'otras' localizaciones frecuentes del ego son compartidas con los alters, siendo el 14%, 16% y 70% las probabilidades de que esa localización sea la casa, el trabajo u 'otra' localización frecuente del alter respectivamente. Se dice que 'al menos' ese porcentaje es compartido porque sólo un porcentaje de los alters proporciona información sobre su presencia y, por tanto, es de esperar que otras localizaciones frecuentes compartidas no se hayan identificado y que el resultado obtenido subestime el porcentaje de localizaciones frecuentes compartidas.

Ego / Alters	Casa	Trabajo	Otra	Total
Casa	7.41 %	6.42 %	7.62 %	21.45%
Trabajo	6.42 %	9.07 %	8.58 %	24.07%
Otra	7.62 %	8.58 %	38.28 %	54.48%

Tabla 8 Distribución del tipo de relación ego-alter para localizaciones frecuentes

4.2.4.5 Resultados del análisis de co-ubicación

El análisis de co-ubicación se ha realizado para el Área Metropolitana de Barcelona durante los 53 días de datos disponibles. El número medio de localizaciones distintas por usuario a lo largo de los 53 días es de 58,3 (σ=51,29). Remarcar que estas localizaciones no tienen por qué ser lugares donde el usuario está realizando una actividad, también pueden ser localizaciones a lo largo de un viaje. Desde el punto de vista del ego, el número de localizaciones en común por alter es de 8,75 (σ=3,94), lo que corresponde al 15% de las localizaciones del ego. De esas localizaciones en común, 1,22 (σ=0,74) corresponden a localizaciones en las que se ha identificado co-ubicación. Para esas localizaciones en las que se ha identificado co-ubicación se producen, en media, un total

de 4,36 (σ=4,17) eventos de co-ubicación a lo largo del periodo analizado. Es importante señalar que, al igual que sucedía en el análisis de las localizaciones frecuentes compartidas por la red social, el número de eventos de co-ubicación está probablemente subestimado, ya que no se dispone de información de presencia para las 24 horas del día.

Para cada localización en la que se ha producido co-ubicación, el tipo de relación entre el ego y el alter se ha clasificado en función de los tipos de localización compartida (casa, trabajo, otra frecuente y otra no frecuente). Alrededor de 1,4 millones de relaciones entre ego-alter han sido identificadas y analizadas. En la *Tabla 9* se muestra la distribución de los tipos de relación para eventos de co-ubicación entre el ego y los alters. Desde el punto de vista del ego, un número significativo de localizaciones en las que se ha producido co-ubicación corresponden a localizaciones no frecuentes (43%) y 'otras' localizaciones frecuentes (30%); y en menor medida a casa (11%) y trabajo (16%). La mayoría de los lugares donde se produce co-ubicación corresponden a localizaciones frecuentes (57%) siendo 19,5%, 28,5% y 52% la probabilidad de que se produzca co-ubicación en casa, trabajo y 'otras' localizaciones frecuentes respectivamente cuando la co-ubicación se produce en localizaciones frecuentes. Estos porcentajes son muy similares a los obtenidos en el análisis de red social y localizaciones frecuentes en común (21,5%, 24%, 54,5% para casa, trabajo y 'otras' localizaciones frecuentes respectivamente). A primera vista, este resultado puede parecer obvio ya que cuanto mayor sea el número de localizaciones frecuentes en común de un mismo tipo, mayor será la probabilidad de que se produzcan situciones de co-ubicación en ese tipo de localizaciones. No obstante, dado que el análisis de red social y localizaciones frecuentes no tiene en cuenta la restricción de co-ubicación, puede darse el caso que un ego y un alter compartan varias localizaciones frecuentes pero que nunca coincidan en el mismo instante. Este resultado refuerza la hipótesis de que 'otras' localizaciones frecuentes pueden asociarse a lugares donde las personas de una misma red social interactúan (donde se producen eventos de co-ubicación) y, por tanto, sugiere que la distribución del tipo de relación ego-alter del análisis de localizaciones frecuentes en común puede ser utilizado como proxy de la probabilidad de co-ubicación del ego cuando la co-ubicación se produce en localizaciones frecuentes. Esto es un

aspecto muy relevante ya que el cálculo de localizaciones frecuentes en común es mucho más sencillo y menos costoso computacionalmente que el análisis de co-ubicación.

Por otro lado, aunque la co-ubicación se produce en su mayoría en localizaciones frecuentes, la co-ubicación en localizaciones no frecuentes es prácticamente igual de relevante (43%). Desde el punto de vista del ego, cuando se produce una situación de co-ubicación en una localización no frecuente, hay una probabilidad del 8,1%, 11,3%, 22,8% y 57,8% que esa localización corresponda a la casa, trabajo, 'otra' localización frecuente y otra localización no frecuente del alter respectivamente. Las localizaciones no frecuentes pueden considerarse como destinos raramente visitados por los usuarios. Estas localizaciones es difícil que sean recogidas por los actuales modelos de transporte, los cuales, generalmente, sólo toman en consideración los costes generalizados del viaje (tiempo, coste económico, etc.) y omiten la influencia de la red social. En casi la mitad de los casos (42%), situaciones de co-ubicación asociadas a localizaciones no frecuentes del ego están relacionadas con localizaciones frecuentes del alter. El tipo de relación ego-alter más frecuente es aquella en la que ambos comparten una localización no frecuente (24,81%). Este tipo de relación entre localizaciones no frecuentes puede verse como una decisión conjunta tomada entre el ego y el alter para reunirse en un lugar distinto de sus respectivas localizaciones frecuentes.

Ego / Alter	Casa	Trabajo	Otra	No frecuente	Total
Casa	2.51%	2.14%	2.92%	3.48%	11.05%
Trabajo	2.14%	5.15%	4.11%	4.86%	16.26%
Otra	2.92%	4.11%	12.94%	9.78%	29.75%
No frecuente	3.48%	4.86%	9.78%	24.81%	42.94%

Tabla 9. Distribución del tipo de relación ego-alter en eventos de co-ubicación

4.2.5 Limitaciones de los datos de telefonía móvil

Aunque los resultados muestran un alto potencial de los datos de telefonía móvil para proporcionar información relevante sobre las relaciones entre la red social y los patrones de movilidad de la población, estos datos no están exentos de limitaciones.

En primer lugar, es necesario remarcar las limitaciones asociadas a la propia naturaleza de los datos de telefonía móvil. El hecho de utilizar datos de llamadas entre los usuarios para determinar su red social inevitablemente omite otros tipos de relaciones que se producen por otras canales de comunicación, como por ejemplo las relaciones cara a cara o mediante mensajería. Además, puede que se estén clasificando de manera errónea algunas relaciones como relaciones sociales, como por ejemplo relaciones de trabajo de carácter esporádico. Estas limitaciones intrínsecas pueden llevar a errores en la estimación de algunas variables de la red social (por ejemplo, número de contactos) y pueden introducir sesgos en algunos análisis (por ejemplo, cuando se consideran de manera errónea interacciones que en realidad no son de carácter social).

Aparte de estas limitaciones intrínsecas, la resolución espacial y temporal de los datos disponibles tiene una influencia significativa en los resultados:

- La resolución temporal de los datos puede definirse como la cantidad de datos disponibles por unidad de tiempo. Para este estudio, solamente se disponía de información cuando el teléfono móvil recibía o realizaba una llamada. Por lo tanto, puede darse el caso de que alguna localización frecuente no sea identificada o sea mal clasificada como no frecuente si el número de llamadas realizadas o recibidas en dicha localización no es un buen proxy del tiempo de permanencia del usuario en dicha localización. Del mismo modo, algunas interacciones sociales pueden no identificarse si no se dispone de información cuando están teniendo lugar (por ejemplo, situaciones de co-ubicación).

- La resolución espacial de los datos determina el grado de precisión en la estimación de la localización del usuario. En este estudio, la resolución espacial se

corresponde con el tamaño del área de Voronoi asociada a cada emplazamiento de antenas. Esto proporciona una incertidumbre espacial de decenas o pocos cientos de metros en zonas urbanas y varios kilómetros en zonas rurales. Este estudio considera que se produce un evento de co-ubicación cuando los usuarios se encuentran en la misma área de Voronoi, lo que puede llevar a sobreestimar este tipo de eventos de co-ubicación, especialmente en zonas de baja densidad de población. Además, desde un punto de vista de las relaciones sociales, es importante señalar que los eventos de co-ubicación no aseguran que exista interacción social entre los usuarios. El hecho de que dos personas se encuentren en la misma zona en el mismo instante de tiempo no aporta la certeza de que se esté produciendo una interacción social entre ellas.

La mejora en la resolución temporal y espacial de los datos proporcionaría resultados de mayor calidad. La resolución temporal puede mejorarse utilizando otro tipo de registros aparte de las llamadas, como por ejemplo mensajes de texto y principalmente registros de sesiones de datos. Por otro lado, la resolución espacial podría mejorarse utilizando técnicas de triangulación de señal o datos de redes WiFi o información de GPS del dispositivo cuando fuera posible.

Por último, la representatividad de la muestra de los usuarios de telefonía móvil es un aspecto clave. La muestra tiene que ser de un tamaño suficiente y estar homogéneamente distribuida entre la población con el objetivo de minimizar los posibles sesgos. Los datos de telefonía móvil utilizados en este estudio recogen una muestra de aproximadamente el 50% de la población española de 2009 y se encuentra homogéneamente distribuida por el territorio como se ha demostrado en la comparativa con datos censales. No obstante, como los datos no disponen de información socio-demográfica, algunos segmentos de la población pueden no estar correctamente representados. Este problema podría resolverse en estudios posteriores si se dispusiera de información socio-demográfica (por ejemplo, edad, género, etc.) asociada a los usuarios de telefonía móvil.

En resumen, los datos de telefonía móvil abren una oportunidad para entender mejor las relaciones entre la red social y los patrones de movilidad de la población. No obstante, las características y limitaciones de los datos comentadas anteriormente deben ser tenidas en cuenta a la hora de analizar e interpretar los resultados.

4.2.6 Conclusiones

Es ampliamente reconocido que la red social influye de manera significativa en los patrones de viaje de las personas. La mayoría de las decisiones sobre dónde realizar una actividad están influenciadas por la red social. El objetivo principal de este estudio es contribuir a un mejor entendimiento sobre de qué manera la red social influye en los patrones de movilidad de las personas, analizando la naturaleza de los lugares (casa, trabajo, otras localizaciones frecuentes, otras localizaciones no frecuentes) compartidos por los contactos de una misma red social. Actualmente esta información se recoge de encuestas que son costosas de realizar. El uso de los datos de telefonía móvil proporciona la oportunidad de recoger información de movilidad y red social de manera conjunta, superando algunas de las limitaciones presentes en las encuestas. A diferencia de otras fuentes de datos como Facebook o Twitter, los datos de telefonía móvil tienen la ventaja de proporcionar información más relevante sobre las interacciones cara-a-cara, proporcionando al mismo tiempo información sobre una parte muy significativa de la red social del usuario. Del análisis conjunto de la red social con (1) las localizaciones frecuentes y (2) los patrones de presencia de la población, se han obtenido estadísticas relevantes sobre la naturaleza de las localizaciones compartidas por los usuarios de una misma red social. Los resultados refuerzan la hipótesis de que ‘otras’ localizaciones frecuentes pueden ser consideradas como lugares donde potencialmente se producen interacciones entre usuarios de una misma red social. Además, se ha observado que la mayoría de los eventos de co-ubicación tienen lugar en localizaciones frecuentes clasificadas como ‘otras’ y en localizaciones no frecuentes de los usuarios. De hecho, la relación de co-ubicación ego-alter más común es aquella en la que ambos se encuentran en una localización no frecuente.

Los resultados obtenidos en este estudio pueden ayudar a mejorar la definición de los actuales modelos de transporte basados en actividades, aportando nuevas variables a considerar a la hora de simular dónde y cuándo se realizan las distintas actividades. Por ejemplo, puede que una persona realice una actividad en una localización no óptima con respecto al coste de transporte incurrido o al atractivo de la zona; no obstante, dicha actividad puede venir explicada por la influencia de la red social. En la *Figura 26* se muestra un ejemplo de posibles actividades realizadas por dos personas de una misma red social considerando y sin considerar la información de la red social. La mejora en los modelos de transporte puede contribuir a una mejor evaluación de las políticas públicas asociadas con servicios de movilidad compartida, como por ejemplo el transporte bajo demanda o el carpooling.

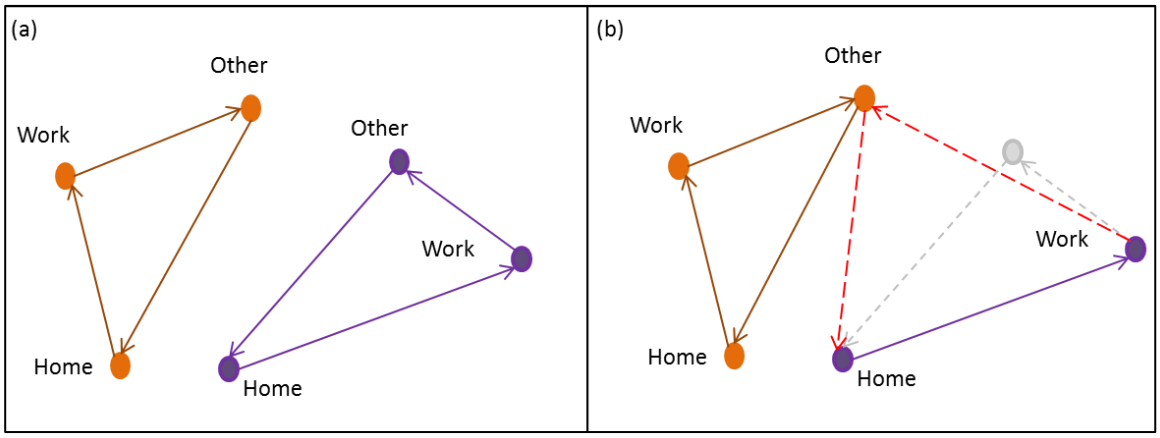


Figura 25. Diario de actividades y viajes de 2 agentes de la misma red social durante un día estándar: (a) resultados del modelo sin considerar la influencia de la red social y (b) resultados del modelo considerando la influencia de la red social.

Por último, remarcar que, a pesar del potencial que han mostrado los datos de telefonía móvil para proporcionar información sobre las relaciones entre la red social y los patrones de movilidad de las personas, se deben tener en cuenta las limitaciones que pueden presentar estos datos, como su alta heterogeneidad espacio-temporal o la falta de información socio-demográficos. La fusión con otras fuentes de datos (por ejemplo datos socio-demográficos procedentes de estadísticas públicas) es un planteamiento interesante de cara a rellenar los huecos de información no cubiertos por la telefonía móvil, siendo

esta una de las futuras líneas de investigación más interesantes en este campo. Otra línea de investigación interesante puede estar relacionada con el análisis de la distribución de las distancias de los viajes de carácter social. Especialmente interesante es el caso en el que el ego y el alter comparten una localización no frecuente, con el objetivo de explorar si existe algún tipo de decisión conjunta en busca de maximizar el beneficio mutuo.

4.3 Análisis de la exposición de la población a la contaminación

4.3.1 Antecedentes, objetivos y alcance del estudio

Estudios previos sobre exposición a la contaminación mediante telefonía móvil (Dewulf et al. 2016; Nyhan et al. 2016; Gariazzo et al. 2016) suelen tomar en consideración una o más de las siguientes hipótesis: (1) si no se dispone de información de localización del usuario en un periodo concreto del día, se supone que este no se ha desplazado y se encuentra ubicado en la última localización detectada y (2) la distribución espacio-temporal de los usuarios de telefonía móvil activos es representativa de la distribución real de la población. La validez de estas hipótesis depende en gran medida del alcance del estudio (el nivel agregación espacial y temporal de los resultados a estimar) y de las características de los datos de telefonía (especialmente su resolución espacial y temporal). La primera hipótesis solamente se podrá considerar valida si la granularidad temporal de los datos es suficientemente elevada o si la zonificación de estudio presenta un alto grado de agregación (cuando la mayoría de las actividades de las personas tienen lugar dentro de cada una las zonas definidas). Respecto a la segunda hipótesis, su validez dependerá de la penetración del operador (u operadores) en el mercado de telefonía móvil y como sus clientes se distribuyen entre los diferentes segmentos de la población. Por otro lado, la mayoría de los estudios de presencia de población suelen plantear un enfoque basado en la torre o celda de telefonía, aforando todos los dispositivos móviles que se conectan a ella. Este planteamiento, al no considerar los patrones de movilidad de las personas, no permite detectar saltos en la red de telefonía (cambio de conexión del dispositivo de una celda a otra sin que el dispositivo se haya desplazado), ni tampoco estimar con precisión el tiempo de estancia del dispositivo en la celda (si el dispositivo está realizando un desplazamiento, permanecerá menos tiempo en la celda). El presente estudio trata de superar las limitaciones identificadas en estudios previos planteando un enfoque basado en la movilidad de las personas, identificando la presencia de la población como resultado de sus actividades y viajes a lo largo del día. Del mismo modo, se plantean técnicas de filtrado de usuarios y procedimientos de expansión de la muestra para adaptarse a

distintas tipologías de datos y para corregir posibles sesgos asociados a la tipología de clientes del operador.

El objetivo principal de este estudio es aplicar la metodología desarrollada de estimación de presencia de población (ver sección 3.3) para el análisis de la exposición de la población a la contaminación, en concreto a la exposición al NO_2. Para ello se han utilizado datos de telefonía móvil para la estimación de presencia de población y datos de concentración de contaminantes obtenidos mediante modelado[15]. Otro de los objetivos principales del estudio es comparar los resultados obtenidos mediante esta nueva metodología con los resultados proporcionados por los métodos estáticos tradicionales.

El estudio se realiza en el entorno del municipio de Madrid, abarcando también algunos municipios de los alrededores. Se define un área de estudio de 40kmx44km dividiendo el territorio en celdas de $1km^2$ (1760 celdas), consistente con el modelo de contaminación del aire disponible para la ciudad de Madrid. En la *Figura 26* se muestra el ámbito del área de estudio y la cuadrícula definida.

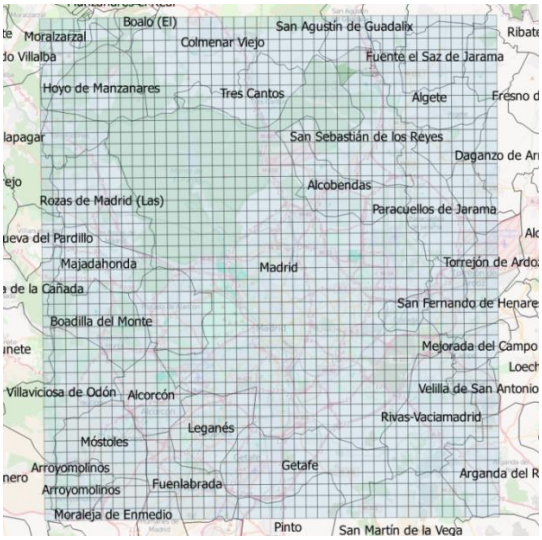


Figura 26. Área de estudio – División del territorio en cuadrículas de $1Km^2$

[15] La parte del estudio relacionada con el cálculo de las concentraciones de contaminantes ha sido llevada a cabo por el Departamento de Ingeniería Química Industrial y Medioambiente de la Universidad Politécnica de Madrid (UPM)

Los cálculos se han realizado para el periodo de noviembre de 2014. Dentro de dicho periodo se ha seleccionado el día 17 de noviembre por presentar unos patrones representativos de niveles de contaminación del aire y por ser un día en el que los resultados del modelo de dispersión de contaminantes empleado mostraban un ajuste satisfactorio. Dado que el área de estudio presenta un alto carácter atractor de viajes, se ha considerado conveniente utilizar como población de estudio no solamente la población residente en el área de estudio si no también la población residente en las provincias limítrofes a la provincia de Madrid: Ávila, Segovia, Guadalajara, Cuenca y Toledo (ver *Figura 27*)

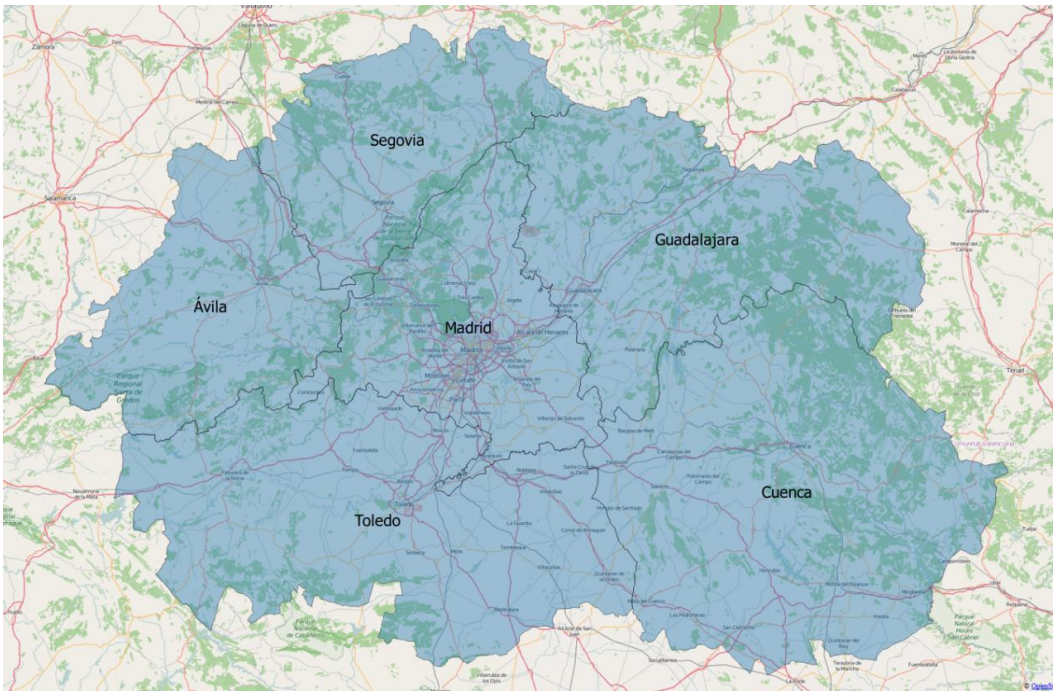


Figura 27. Ámbito espacial de la población de estudio considerada

4.3.2 Descripción de los datos

Como fuente de datos principal se han utilizado datos anonimizados de telefonía móvil proporcionados por uno de los principales operadores de red de España, con una cuota de mercado alrededor de un tercio de la población. Estos datos contienen las posiciones geolocalizadas de los dispositivos móviles tanto para eventos activos (llamadas, SMS, etc.)

como para eventos pasivos (handover, actualización de la red, etc.), que se corresponden con la actividad registrada durante los días laborables del mes de noviembre de 2014. Los datos de telefonía disponibles proporcionan una granularidad temporal muy elevada, que permite determinar con alto nivel de detalle la localización del dispositivo a lo largo del día. En cuanto a la granularidad espacial, se dispone de información de localización del dispositivo móvil a nivel de celda de telefonía, lo que supone una precisión espacial de decenas/cientos de metros en ciudad y varios kilómetros en zonas rurales. Los datos proporcionados por el operador también incluyen información socio-demográfica, como la edad o el género. El único dato socio-demográfico que se ha utilizado para este estudio es el dato de edad, como entrada en el proceso de elevación muestral. Los registros fueron debidamente anonimizados por el operador, para proteger la privacidad de los usuarios de acuerdo con la legislación vigente.

4.3.3 Metodología

En la *Figura 28* se muestra un esquema de la metodología seguida para la estimación de los indicadores de exposición de la población a la contaminación. El proceso se compone de los siguientes pasos:

- Cálculo de presencia de población
- Simulación de la concentración de NO_2
- Cálculo del indicador de exposición a la contaminación
- Comparativa con metodologías tradicionales

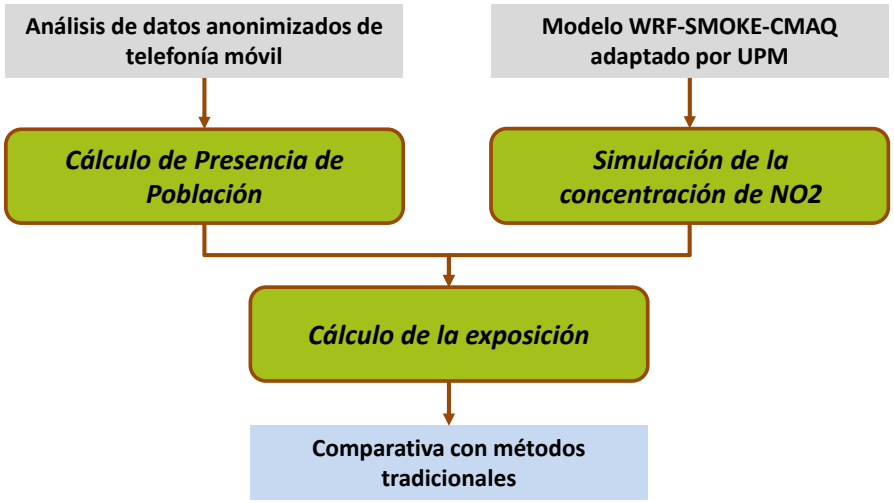


Figura 28. Esquema metodológico para el cálculo de la exposición a la contaminación

4.3.3.1 Cálculo de presencia de población

Para calcular el indicador de presencia de población se han utilizado las metodologías propuestas en las secciones 3.1 y 3.3. A continuación se describen los pasos que se han seguido y los parámetros utilizados en las distintas fases del proceso:

- **Identificación de actividades frecuentes**: el objetivo principal de esta etapa es identificar el lugar de residencia de las personas para seleccionar la muestra objeto de estudio. En este caso se han seleccionado los usuarios con residencia en Madrid y en sus provincias limítrofes. Se ha utilizado como muestra de días a analizar todos los días laborables (L-J) del mes de noviembre de 2014. Se ha considerado conveniente utilizar como parámetro 'α' una valor de 0,2 y como periodo característico de residencia el periodo comprendido entre las 8 p.m. y las 7 a.m.

- **Extracción de patrones de actividad y movilidad**: En primer lugar, para evitar los saltos de señal, se utiliza una velocidad de cambio 'VC' de 200km/h. En segundo lugar, se descartan aquellos usuarios que presenten una granularidad temporal baja en sus registros. Se toma como umbral temporal 'TR' el valor de 4 horas para todo el día salvo para el periodo nocturno (definido de 8 p.m. a 7a.m.) en el que se

considera un 'TR' de 8 horas. Posteriormente, se define una estancia mínima 'TA' de 30 minutos para considerar actividad. Finalmente, para la determinación del viaje, se asumen las hipótesis de trayectoria recta y velocidad constante (asignada como 15 km/h[16]) y se utiliza una función de probabilidad uniforme para determinar la hora exacta del viaje.

- **Elevación muestral**: la muestra se elevada a nivel de sección censal utilizando datos del censo 2011 como marco muestral. Se aplica el factor de elevación definido en 3.1.4 para cada sección censal segmentando por edad (se considera solamente el segmento de población mayor a 16 años).

- **Presencia de población**: se aplica la metodología descrita en la sección 3.3.1 para obtener el indicador de presencia para cada una de las celdas de $1Km^2$ y para cada hora del día.

4.3.3.2 Cálculo de concentraciones de NO_2

El cálculo de las concentraciones de NO_2 ha sido llevado a cabo por el Departamento de Ingeniería Química Industrial y Medioambiente de la Universidad Politécnica de Madrid (UPM). Para realizar los análisis se ha empleado el modelo WRF-SMOKE-CMAQ adaptado por la UPM. El modelo de calidad del aire proporciona la concentración de dióxido de nitrógeno ($^{\mu g}/_{m^3}$ NO_2) con resolución temporal horaria y espacial de acuerdo con la malla de celdas de 1 km^2 presentada anteriormente en la *Figura 26*.

4.3.3.3 Cálculo del indicador de exposición a la contaminación

Una vez calculados los indicadores de presencia de población y los niveles de concentración de NO_2 (ambos con la resolución espacial de $1km^2$ y por franjas horarias) se procede a calcular el indicador de exposición. El indicador de exposición se define como el producto de la concentración de NO_2 (C_NO_2) de cada zona por el indicador de presencia

[16] Velocidad media considerada para todos los modos de transporte (pie, transporte público, coche, etc.) bajo la hipótesis de trayectoria recta en el plano.

en cada zona (P_T), dando lugar a un indicador "exp_NO_2" que se ha utilizado para evaluar el grado de exposición de la población al NO_2:

$$\exp\_NO_2 = C_{NO_2} * P_T \qquad [12]$$

4.3.3.4 *Comparativa con metodologías tradicionales*

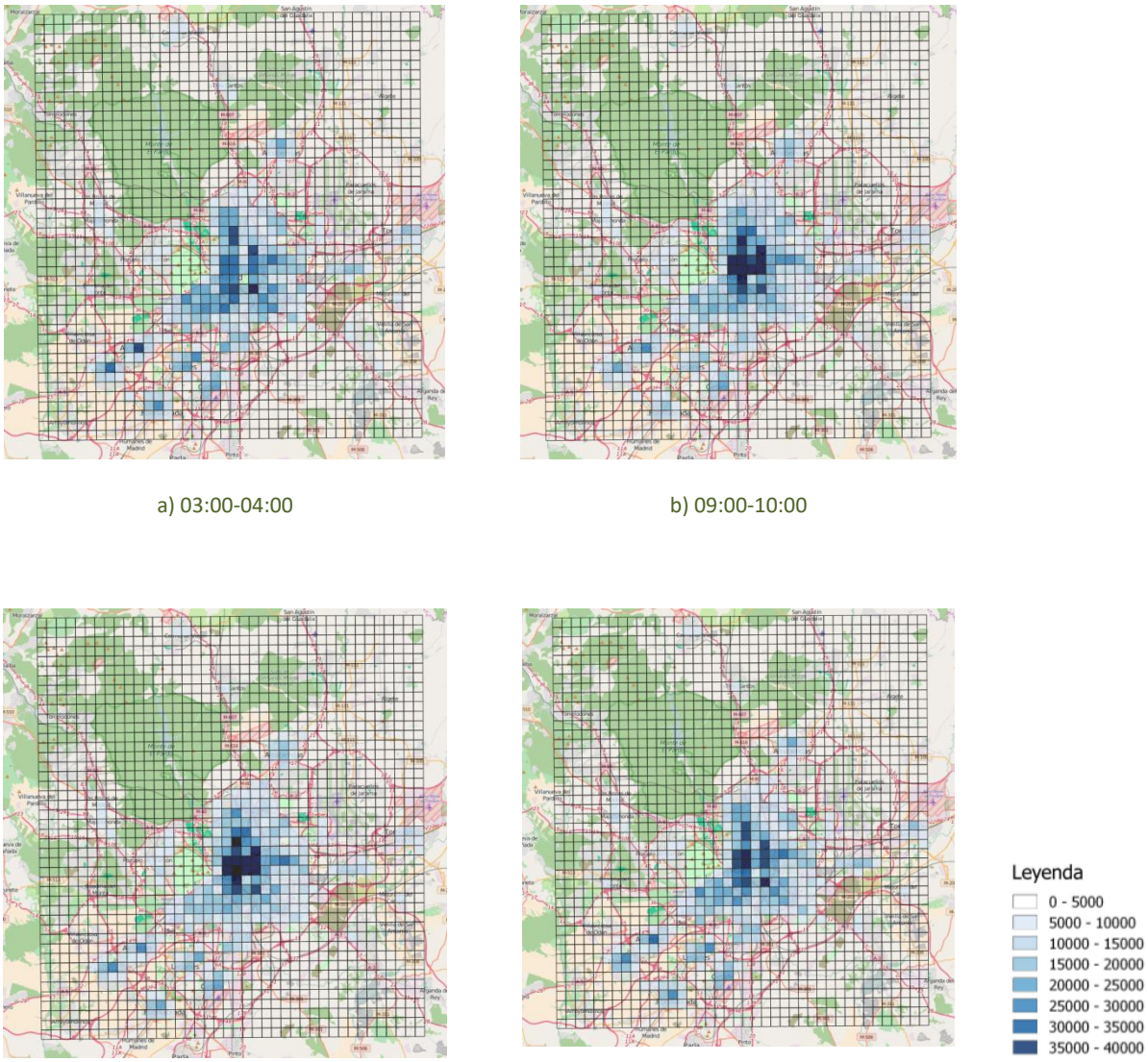
Los estudios tradicionales suelen estimar el indicador de exposición utilizando información sobre la residencia de la población, ya que no se suele disponer de información fiable y actualizada sobre la presencia de población en la ciudad a lo largo del día. En este estudio se ha comparado la aproximación tradicional (aproximación estática) con la aproximación basada en información de datos de presencia a partir de datos de telefonía móvil (aproximación dinámica). Señalar que en la aproximación estática utilizamos como indicador de presencia el número de residentes en cada zona basándonos en datos del censo 2011.

4.3.4 Resultados y discusión

4.3.4.1 *Indicador de presencia*

El indicador de presencia se calcula para cada zona y para cada hora del día. En la *Figura 29* se muestra, para distintas horas del día, los valores del indicador de presencia obtenidos en las distintas zonas. Se puede observar como la distribución de la población cambia a lo largo del día de manera significativa, con la zona centro de Madrid ejerciendo como zona de atracción para los municipios de los alrededores. En la *Figura 30* se presenta, para cada hora del día: (1) el indicador de presencia agregado para la totalidad del área de estudio, (2) la población censada en el área de estudio y (3) el indicador de presencia asociado a los no residentes en el área de estudio (personas residentes en las provincias limítrofes de Madrid). Se puede observar como a primeras y últimas horas del día, los datos censales y el indicador de presencia son muy similares. Esto sugiere que el modelo de presencia de población está capturando correctamente los patrones de comportamiento de la población. A medida que avanza el día, se observa un aumento en el indicador de

presencia, explicado por la llegada de no residentes al área de estudio. En las horas centrales del día, el porcentaje de no residentes respecto al total de la población supone alrededor del 8%, reflejando la importancia de considerar a la población no residente en los estudios de exposición a la contaminación en grandes áreas metropolitanas atractoras.



a) 03:00-04:00

b) 09:00-10:00

c) 15:00-16:00

d) 21:00-22:00

Figura 29. Indicador de presencia para distintas horas del día

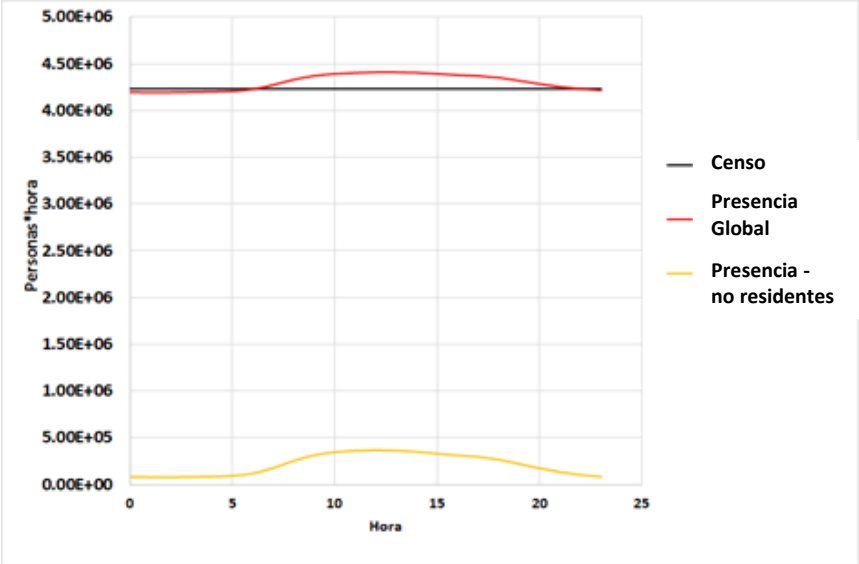


Figura 30. Evolución de la presencia de la población en el área de estudio a lo largo del día

4.3.4.2 Simulación de la concentración de NO2

El análisis se ha realizado para el día 17 de noviembre de 2014. Se trata de un día con un nivel de concentración representativa del mes, típico de condiciones otoñales, en el que la media de concentración del conjunto de estaciones que conforman la red de vigilancia de la calidad del aire del Ayuntamiento de Madrid se situó en 42,0 µg/m^3, muy similar a la media del mes (41,5 µg/m^3). Se trata de un lunes, que presenta la típica distribución de concentración de NO_2 con dos máximos relativos, uno matutino coincidiendo con la hora punta de la mañana y otro vespertino al que característicamente se asocia el valor máximo diario. Los valores de concentración de NO_2 para cada celda de 1 km^2 y por hora del día se han estimado mediante el sistema WRF-SMOKE-CMAQ. El sesgo total relativo del modelo es del 12,4% con un error relativo global del 21,3%. Pese a ser desviaciones considerables, son razonables para un sistema de modelización de este tipo y se estima que la precisión es suficiente para el objeto del estudio. En la *Figura 31* se presenta la concentración media de NO_2 estimada para la zona de estudio. Señalar que los indicadores aquí estimados se utilizarán tanto para la aproximación dinámica (datos de telefonía) como para la estimación estática (datos censales).

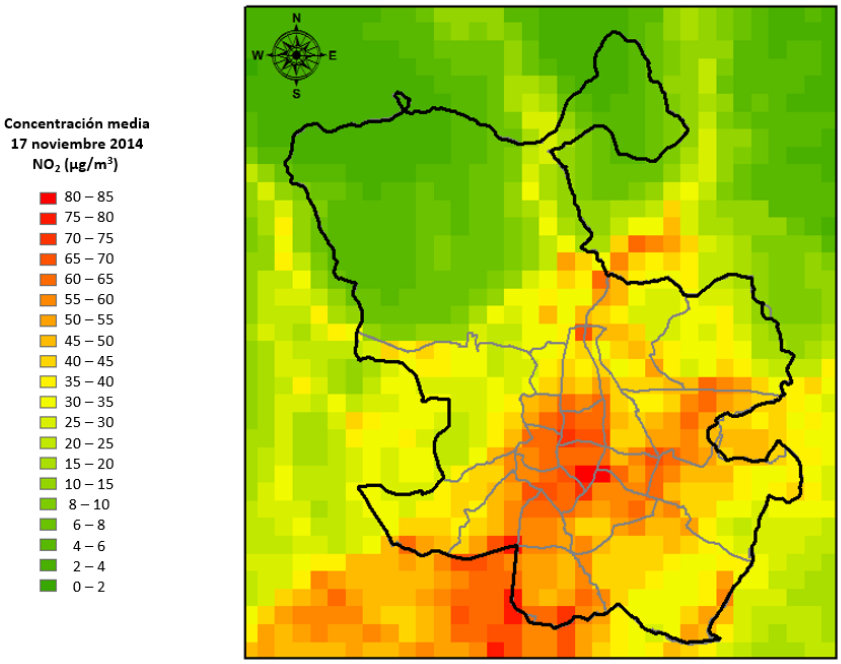


Figura 31. Concentración media de NO_2

4.3.4.3 *Cálculo del indicador de exposición a la contaminación*

El indicador de exposición se calcula como el producto del indicador de presencia y la concentración de NO_2 para cada celda y para cada hora del día. En la *Figura 32* se muestra el indicador de exposición a la contaminación para el caso dinámico y para el caso estático, considerando la zona de estudio de manera agregada (sumando los valores de todas las celdas). Puede observarse como los valores obtenidos para el caso dinámico y estático no difieren significativamente, con apenas un 3-4% de variación. En primera instancia, cabría esperar resultados bastante distintos entre ambas metodologías, al proporcionar los datos de telefonía móvil información de mayor calidad sobre la distribución de la población a lo largo del día. El motivo principal de que los resultados sean tan similares para este nivel de agregación es debido a que los flujos principales de movilidad dentro del área de estudio suceden entre zonas con niveles de concentración de NO_2 similares. Si se analizan los resultados con mayor nivel de desagregación (por celda de 1 km^2) también se observa a simple vista una distribución general para el total de la

zona de estudio similar (ver *Figura 33*); no obstante, analizando celda a celda en detalle se pueden observar discrepancias significativas (ver *Figura 34*).

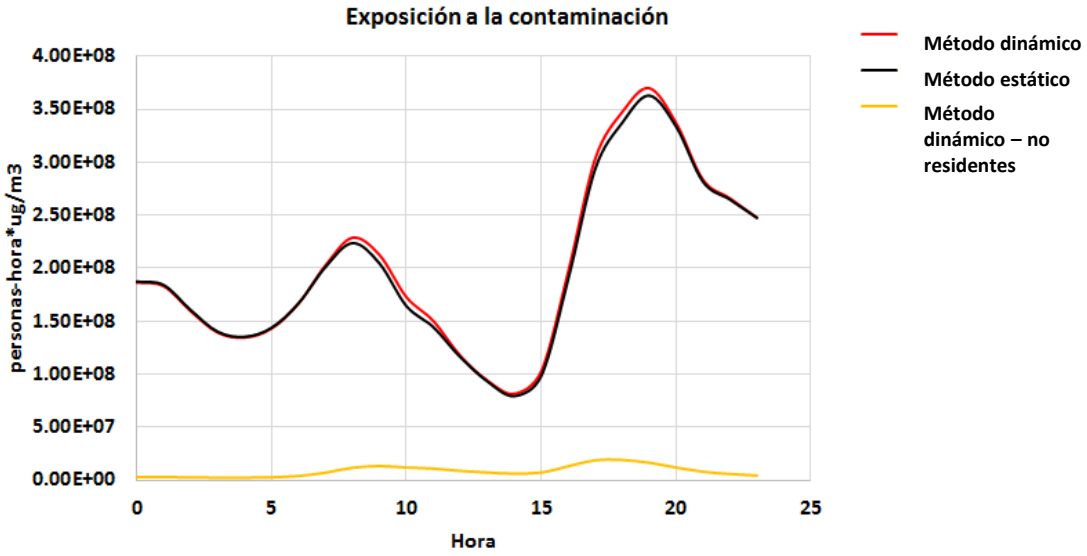


Figura 32. Indicador de exposición agregado para el total del área de estudio

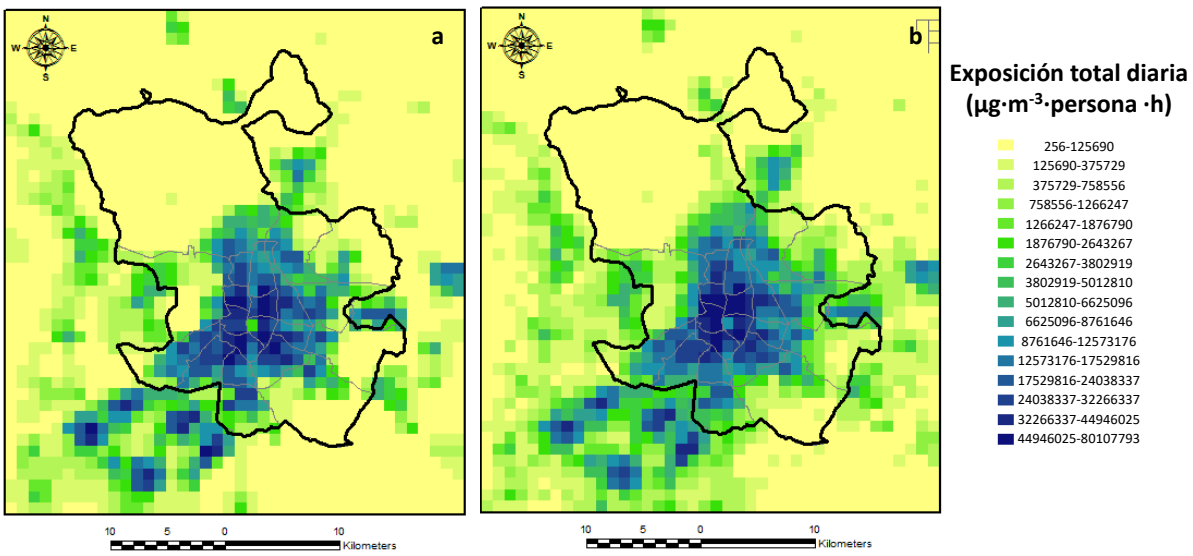


Figura 33. Distribución espacial de la exposición a la contaminación total diaria para el área de estudio considerando aproximación estática (a) y aproximación dinámica (b)

En la *Figura 34* se presenta, para las zonas de mayor exposición a la contaminación total diaria, el ratio entre los valores calculados por la metodología dinámica y por la

metodología estática. Se observan zonas donde el ratio es bastante elevado, llegando a mostrar diferencias entre un método y otro de hasta un 500%.

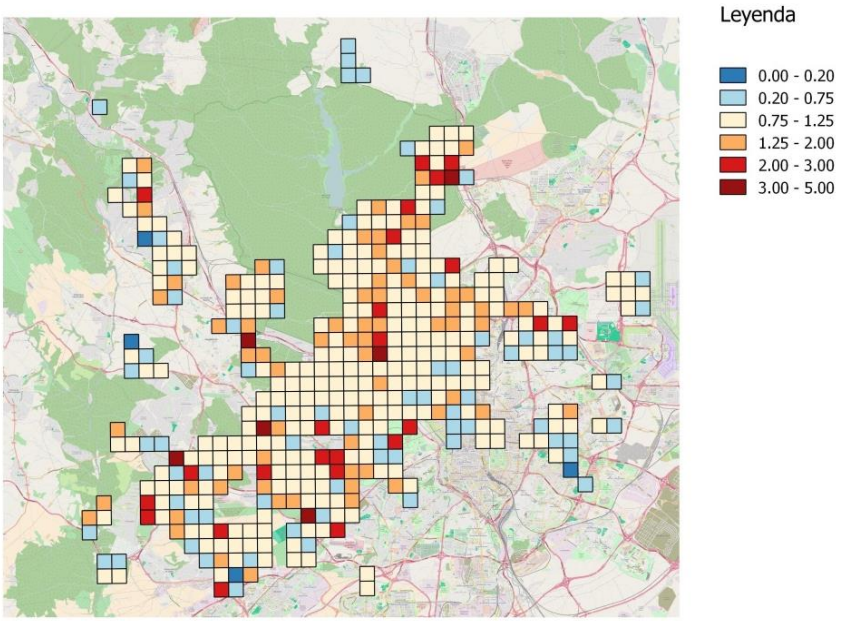
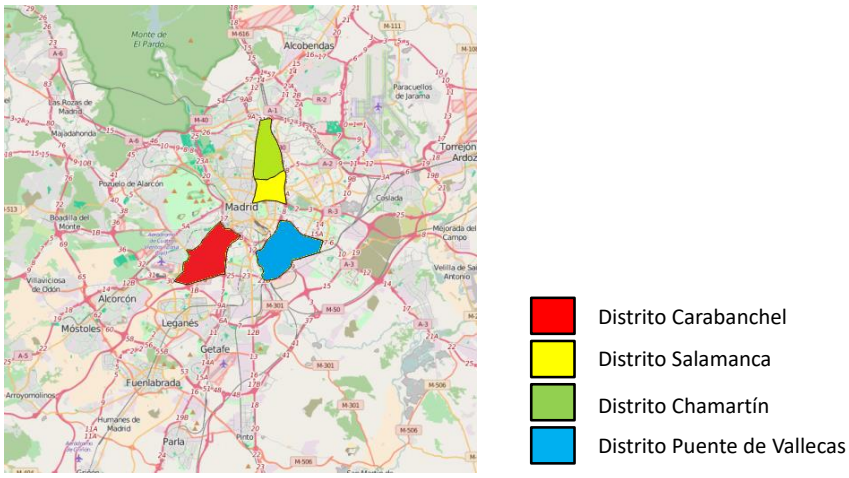
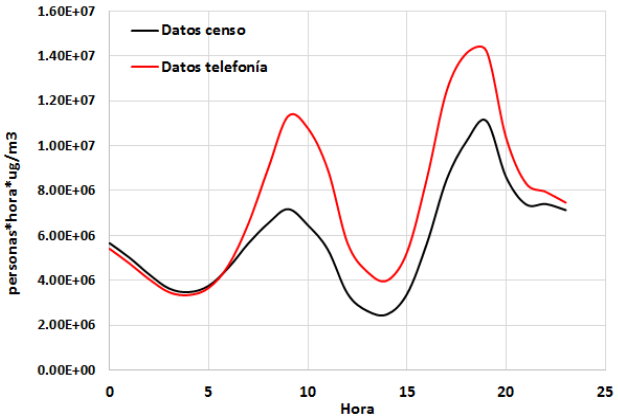


Figura 34. Ratio entre indicadores de exposición(aproximación dinámica vs aproximación estática)

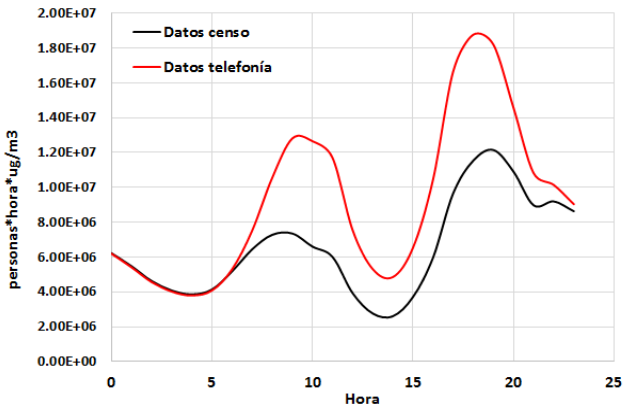
Con el objetivo de estudiar con mayor profundidad las diferencias encontradas, se ha realizado un análisis en detalle para varios distritos del municipio de Madrid con patrones de actividad a priori diferentes. En concreto, se han estudiado los distritos de Chamartín, Salamanca, Carabanchel y Puente de Vallecas. En la *Figura 35* se muestra el indicador de presencia a lo largo del día calculado para cada distrito aplicando la metodología estática y dinámica. Puede apreciarse como para los distritos de Chamartín y Salamanca los valores del indicador de exposición son mucho más altos en la aproximación dinámica, debido a que son zonas de gran afluencia de gente (diferencia de casi un 70% más en el caso de Chamartín y casi un 100% más en el caso del distrito de Salamanca). Por el contrario, los distritos de Carabanchel y Puente de Vallecas presentan valores del indicador de exposición menores en el caso de la aproximación dinámica. Por lo tanto, puede apreciarse como la aproximación estática, al no capturar los patrones reales de movilidad de la población, tiende a sobrestimar la exposición de la población a la contaminación en unos casos y a subestimarla en otros. En estudios previos (Nyhan et al. 2016; Gariazzo et

al. 2016) también se identificaba un aumento del indicador de exposición en algunas zonas mediante la metodología dinámica; sin embargo, no se hacía referencia a los casos en los que la aproximación dinámica proporciona valores inferiores a la aproximación estática. Es importante también señalar y tener en cuenta estos casos a la hora de realizar una evaluación global de la exposición de la población a la contaminación.

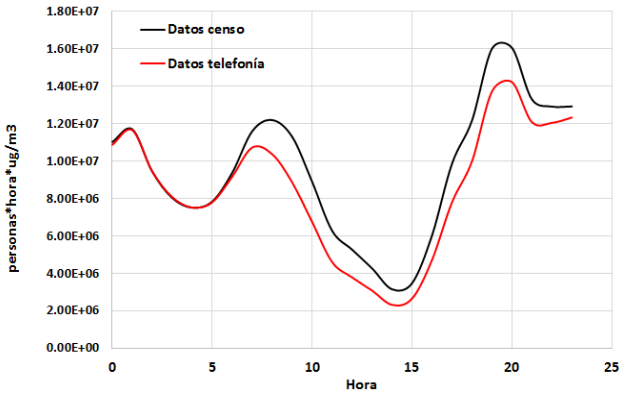




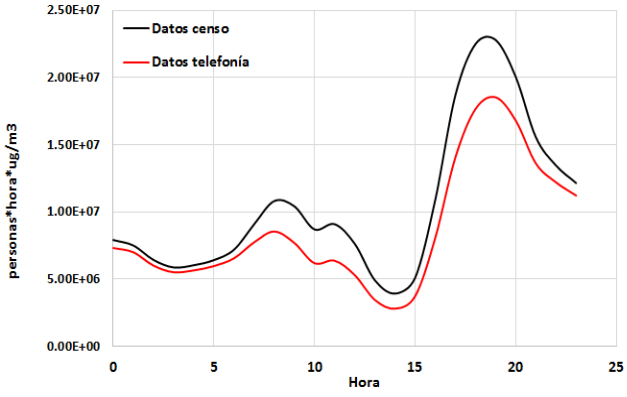
a) Exposición a la contaminación en el distrito Chamartín



b) Exposición a la contaminación en el distrito Salamanca



c) Exposición a la contaminación en Carabanchel



d) Exposición a la contaminación en Puente de Vallecas

Figura 35. Comparativa entre los métodos dinámicos (telefonía móvil) y estáticos (censo) para distintos distritos de Madrid

4.3.5 Limitaciones de los datos de telefonía móvil

Los datos de telefonía móvil disponibles para este caso de estudio proporcionaban una granularidad temporal bastante elevada y un grado de precisión espacial adecuada y compatible con el requerido por el modelo de calidad del aire. Las principales limitaciones identificadas en este caso hacen referencia a la calidad de los datos socio-demográficos disponibles. En un número elevado de casos no se disponía de información socio-demográfica de los usuarios (por errores en el registro o porque el usuario presenta un contrato de prepago) y en los casos en los que se disponía de información, esta suele estar referenciada al titular del contrato, que no tiene por qué coincidir con el usuario del dispositivo móvil. Este hecho puede tener influencia en el proceso de elevación muestral y, del mismo modo, limitar posibles estudios en los que se requiera realizar segmentación de la población por género/edad. Por otro lado, la muestra de usuarios solamente contiene información de personas con dispositivo móvil, por lo que aquellos segmentos de población que no tengan dispositivo móvil (principalmente niños pequeños) no están representados en la muestra, y se requiere de estudios complementarios para analizar la exposición a la contaminación de este segmento de población. Estas limitaciones mencionadas anteriormente justifican que en el presente estudio se haya realizado el proceso de elevación muestral considerando solamente el segmento de población mayor de 16 años de manera agregada.

4.3.6 Conclusiones

De cara a evaluar los efectos que la contaminación del aire tiene en la salud de las personas es necesario llevar a cabo estudios sobre el grado de exposición de la población a la contaminación. Una de las principales limitaciones que presentan estos estudios es la falta de información sobre la presencia de la población en la ciudad. En este estudio, se ha propuesto una metodología para medir la presencia de la población basada en el análisis de datos de telefonía móvil. Se ha presentado una metodología novedosa que estima la presencia de la población como resultado de sus actividades y viajes realizados a lo largo del día, a diferencia del planteamiento tradicional basado en el aforamiento de los

usuarios de telefonía móvil a nivel de torre o celda de telefonía. Esta metodología presenta una serie de ventajas, como la eliminación de saltos de señal o una mejor estimación del tiempo de estancia en cada zona. Del mismo modo, se han propuesto técnicas de filtrado de usuarios y métodos de expansión para superar algunas de las limitaciones identificadas en estudios previos. Los resultados obtenidos muestran el potencial de los datos de telefonía para capturar los patrones de comportamiento dinámico de las personas, pudiendo capturar información tanto de los residentes como de los visitantes a una ciudad. El método basado en telefonía (aproximación dinámica) se ha comparado con el método tradicional basado en datos censales (aproximación estática). Se ha observado que, a nivel agregado para toda el área de estudio, los resultados de exposición a la contaminación obtenidos mediante la aproximación estática y dinámica son muy similares. Sin embargo, analizando los resultados con mayor nivel de desagregación, se observan diferencias significativas entre ambos métodos. El método tradicional, al no considerar los patrones de actividad y movilidad de la población, tiende a subestimar de manera significativa el indicador de exposición a la contaminación en zonas de gran actividad económica (no asociadas al hogar). Esta conclusión también ha sido observada en estudios previos (Nyhan et al. 2016; Gariazzo et al. 2016). Por otro lado, el método estático tiende a sobreestimar la exposición a la contaminación en zonas de carácter mayoritariamente residencial, debido a que un número considerable de personas se desplazan fuera de su zona de residencia para trabajar. Los patrones observados reproducen lo que en un principio podría esperarse, no obstante, la principal aportación de esta metodología es que permite estimar de manera cuantitativa estas variaciones.

Aunque los datos de telefonía han mostrado un gran potencial para capturar los patrones de comportamiento de la población, también presentan una serie de limitaciones que es importante considerar. Para este caso en concreto, la principal limitación observada viene dada por la calidad de los datos socio-demográficos disponibles. En muchos casos existe falta de información y en otros existe cierta incertidumbre sobre su validez, al estar asociada la información al titular del contrato y no al usuario del dispositivo móvil. Por otro lado, existen segmentos de la población que no disponen de dispositivos móviles

(especialmente niños) para los cuales no es posible utilizar esta metodología. Otras limitaciones intrínsecas como la resolución espacial de los datos tiene menor importancia en este caso, ya que la propia resolución espacial para la cual se disponible de información sobre concentración de contaminantes también es limitada.

Como conclusión, puede decirse que la aproximación estática para el cálculo de la exposición utilizada tradicionalmente en los estudios de impacto en salud (e.g. Boldo et al., 2011) puede proporcionar información razonablemente precisa en estudios con un nivel de agregación elevado (varias decenas de kilómetros), sin embargo, para estudios a una escala más fina es fundamental considerar los patrones de actividad y movilidad de la población. Una futura línea de trabajo interesante podría ser la mejora de los datos socio-demográficos mediante el análisis de los patrones de comportamiento de los usuarios (identificando el segmento de la población al que pertenece en función de su uso del teléfono, sus patrones de movilidad, etc.) y mediante la fusión con otras fuentes de datos (por ejemplo, datos socio-demográficos asociados al lugar de residencia). Del mismo modo, otra línea de investigación interesante puede ser el análisis de la exposición individual en vez de la exposición de la población, para poder identificar grupos de personas en riesgo por lugar de residencia y poder dimensionar los servicios sanitarios necesarios asociados a dicha zona. La metodología propuesta en este estudio es de aplicación directa para estudios de exposición individual.

CAPÍTULO V: CONCLUSIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

5.1 Conclusiones

En esta Tesis Doctoral se ha desarrollado y validado una metodología para la extracción de patrones de actividad y movilidad de la población a partir de datos de telefonía móvil. Esta metodología ha sido aplicada en tres casos prácticos para: (1) determinar estadísticas básicas de movilidad y matrices origen-destino, (2) analizar las interacciones entre la red social y la movilidad y (3) evaluar la exposición de la población a la contaminación. Las principales contribuciones metodológicas que ha aportado esta investigación son las siguientes:

- Definición de una metodología detallada que abarca todo el proceso de análisis de los datos, desde su pre-procesamiento hasta el cálculo de indicadores.
- Definición de una solución adaptable a distintas características espacio-temporales de los datos de telefonía móvil.
- Procedimiento para la identificación de localizaciones/actividades frecuentes de diversa naturaleza.
- Procedimiento para la estimación precisa de la hora del viaje.
- Procedimiento para la selección de muestra útil de usuarios.
- Procedimiento para la expansión de la muestra al total de la población.
- Procedimiento para la estimación de presencia de población a partir de los patrones de actividad y movilidad de las personas.

La hipótesis principal en la que se ha basado esta investigación es que la información que proporcionan los registros de actividad telefónica es una alternativa real a las actuales encuestas, permitiendo la obtención de patrones de actividad y movilidad de una manera más eficaz y eficiente. Podría decirse que esta hipótesis ha sido validada parcialmente.

Los resultados del estudio no dejan lugar a duda sobre el potencial de los datos de telefonía móvil para obtener información relevante sobre estadísticas básicas de

movilidad (número de viajes por persona, distribución de la distancia de los viajes, etc.) y matrices origen-destino. Esta novedosa metodología permite realizar estudios sobre una muestra de usuarios muy superior a la que permiten las encuestas. Del mismo modo, los costes y plazos de ejecución asociados a la recogida y análisis de la información son muy inferiores a los de las encuestas. Por ejemplo, una encuesta domiciliaria, para una ciudad de 3-5 millones de habitantes y un tamaño de muestra del 1-3% de la población puede tener unos costes de entre 1 y 4 millones de euros y unos plazos de ejecución de años. Un análisis similar mediante datos de telefonía móvil presenta un tamaño de muestra muy superior, unos costes muy inferiores y un plazo de semanas (Picornell & Willumsen 2016). Otra ventaja de los datos de telefonía móvil es que permiten analizar información histórica, pudiendo analizar un día específico de distintos meses o años pasados. Mediante las encuestas no es posible por lo general analizar un día específico, sólo un día "promedio", debido a los plazos de tiempo necesarios para la recogida de la muestra. Además, las encuestas requieren haber planificado los trabajos con antelación para poder recoger información de un día o periodo específico, aspecto que mediante la telefonía móvil no es necesario. Seguramente, una de las principales ventajas que aportan los datos de telefonía móvil sea la calidad de las matrices origen-destino que se pueden obtener de los mismos. Debido a la gran muestra de usuarios de la cual se dispone, es posible capturar viajes que no son recogidos mediante encuestas, minimizando el problema de los 'ceros' en las matrices de viajes. No obstante, aunque las ventajas que aporta la telefonía móvil son evidentes, también es necesario señalar las limitaciones que presenta. Algunas limitaciones vienen derivadas de la precisión espacio/temporal de los datos. Por un lado, la precisión espacial de los datos implica que en algunos casos no sea posible identificar la movilidad de corta distancia en ciudad o la movilidad intra-municipal en municipios con poca población, debido a la precisión espacial que proporcionan las antenas y a su distribución no homogénea en el territorio (decenas/cientos de metros en ciudad y varios kilómetros en zonas de baja densidad de población). Por otro lado, la resolución temporal conlleva que, en algunos casos, pueda existir incertidumbre con respecto a la hora de inicio o fin de las actividades, y como consecuencia, se puedan producir errores en la

estimación de los instantes de inicio y fin de los viajes. Otro aspecto relevante es la representatividad de la muestra de usuarios de telefonía móvil. Es importante evaluar si la muestra de usuarios del operador de telefonía móvil es representativa de los distintos segmentos de la población. Esto tiene implicaciones relevantes en el proceso de elevación muestral, pudiendo dar lugar a sesgos en los resultados si no se dispone de suficiente muestra de los diferentes segmentos de población. En este sentido, también es importante evaluar la calidad de los datos socio-demográficos (por ejemplo, edad y género) proporcionados por el operador que se utilizan en el proceso de segmentación de la muestra. Errores en estos datos también podrían generar sesgos en los resultados. Por otro lado, es importante señalar que a través únicamente de los datos de telefonía móvil no es posible recoger toda la información que actualmente se recopila mediante encuestas. Información detallada sobre el motivo preciso de los viajes, el modo y la ruta en ámbitos urbanos, información socio-demográfica detallada de las personas o la opinión de los usuarios sobre la calidad del transporte público, son algunos ejemplos de variables que aportan las encuestas y que, actualmente, no pueden extraerse directamente de los datos procedentes de telefonía móvil. Del análisis de las ventajas y limitaciones de los datos de telefonía móvil se deduce la necesidad de utilizar diferentes metodologías y distintas fuentes de datos para poder obtener una visión completa de la movilidad en las ciudades. La solución pasa por identificar qué fusión o combinación de fuentes de datos es la más adecuada para obtener información sobre movilidad de manera eficaz y eficiente.

Con respecto al análisis de la red social y la movilidad, en esta investigación se ha demostrado el valor que los datos de telefonía móvil pueden aportar. La información extraída de los datos de telefonía móvil ha permitido caracterizar las localizaciones frecuentes que visitan los miembros de una misma red social. Del mismo modo, ha permitido identificar dónde y cuándo los miembros de una misma red social se reúnen (co-ubicación). En comparación con las encuestas convencionales que se realizan para analizar la red social y la movilidad de manera conjunta, los datos de telefonía móvil aportan un tamaño de muestra muy superior (millones de personas frente a cientos o pocos miles de personas). Los resultados obtenidos refuerzan la hipótesis de que ‘otras’

localizaciones frecuentes distintas de casa y trabajo pueden ser consideradas como lugares donde potencialmente se producen interacciones entre las personas de una misma red social. Además, se ha observado que la mayoría de los eventos de co-ubicación tienen lugar en 'otras' localizaciones frecuentes y en localizaciones no frecuentemente visitadas por las personas. Los resultados obtenidos en este estudio pueden ayudar a mejorar la definición de los actuales modelos de transporte basados en actividades, aportando nuevas variables a considerar a la hora de simular dónde y cuándo se realizan las distintas actividades. La mejora en estos modelos de transporte puede contribuir a una mejor evaluación de las políticas públicas asociadas a servicios de movilidad compartida, como por ejemplo el transporte bajo demanda o el carpooling. A pesar del potencial que han demostrado los datos de telefonía móvil para proporcionar información sobre las relaciones entre la red social y los patrones de movilidad de las personas, es importante tener en cuenta las limitaciones que esta metodología presenta. Por un lado, pueden existir relaciones sociales por otros medios de comunicación que no sean llamadas telefónicas, por lo que parte de la red social del usuario puede no estar representada en este tipo de análisis. Este aspecto es particularmente relevante con la proliferación de aplicaciones de mensajería instantánea como WhatsApp. Por otro lado, las características espacio-temporales de los datos influyen de manera significativa en los resultados. Para medir la frecuencia de aparición en una localización o situaciones de co-ubicación, la resolución temporal de los registros de telefonía afecta de una manera determinante. Si la granularidad temporal es baja, pueden no clasificarse como localizaciones frecuentes algunas que sí lo sean o no detectar eventos de co-ubicación que realmente están teniendo lugar. Por otro lado, la resolución espacial que actualmente aportan los datos de telefonía móvil (decenas/cientos de metros en ciudad) hace imposible determinar a ciencia cierta si dos personas están interactuando socialmente en un mismo lugar o simplemente se encuentran en la misma zona. Por último, también es importante señalar que la información socio-demográfica disponible es generalmente de menor calidad que la obtenida mediante encuestas. La mejora esperada en la calidad de los datos de telefonía móvil (resolución espacio-temporal) así como la fusión con otras fuentes de datos (por

ejemplo datos socio-demográficos procedentes de estadísticas públicas) puede ayudar a superar las limitaciones que actualmente presenta esta metodología.

En esta investigación también se ha mostrado el potencial de los datos de telefonía móvil para mejorar las estimaciones de exposición de la población a la contaminación. Se ha demostrado cómo la información de patrones de actividad y movilidad de la población permite estimar mejor la ubicación de las personas a lo largo del día, y por consiguiente, mejorar las estimaciones de exposición a la contaminación. La gran diferencia de esta metodología con estudios tradicionales es que, en este caso, se utiliza información dinámica sobre la ubicación de las personas a lo largo del día, mientras que tradicionalmente se suele utilizar información censal sobre el lugar de residencia (información estática) para llevar a cabo este tipo de análisis. De los resultados del estudio se desprende que la aproximación estática para el cálculo de la exposición utilizada tradicionalmente en los estudios de impacto en salud puede proporcionar información razonablemente satisfactoria en estudios con un cierto nivel de agregación (varias decenas de kilómetros), sin embargo, para estudios a una escala más detallada es fundamental considerar los patrones de actividad y movilidad de la población. A pesar de las evidentes ventajas que aportan los datos de telefonía para este tipo de estudios, éstos no están exentos de limitaciones. Para este caso en concreto, la principal limitación observada viene dada por la calidad de los datos socio-demográficos disponibles. En muchos casos existe falta de información y en otros existe cierta incertidumbre sobre su validez. Por otro lado, existen segmentos de la población que no están representados al no tener dispositivos móviles (especialmente niños). Esta limitación es extensible a cualquier aplicación con datos de telefonía móvil. La fusión con otras fuentes de datos socio-demográficos es fundamental para superar estas limitaciones.

Los desarrollos realizados en el marco de esta investigación, así como los resultados satisfactorios obtenidos, motivaron la constitución de la empresa Kineo Mobility Analytics S.L. (www.kineo-analytics.com) en el año 2015 para la explotación comercial de dichas soluciones. Kineo Mobility Analytics se constituye como una empresa conjunta de

Nommon Solutions and Technologies S.L. y Luis Willumsen (www.luiswillumsen.com). Desde entonces hasta el día de hoy, la metodología sigue desarrollándose para dar respuesta a las limitaciones identificadas y se está aplicando con éxito en distintos estudios donde se requiere información sobre los patrones de actividad y movilidad de la población.

En resumen, en esta investigación se ha demostrado el potencial de los datos de telefonía móvil para proporcionar información relevante sobre los patrones de actividad y movilidad de la población. La metodología propuesta permite superar muchas de las limitaciones de los métodos convencionales de obtención de información sobre movilidad, como las muestras de tamaño reducido, los costes económicos o los plazos de ejecución. Para algunos aspectos en concreto (como la información de matrices origen-destino o los indicadores de presencia de población) los datos de telefonía pueden aportar información de mayor calidad que los métodos convencionales. No obstante, también existe cierta información (opinión de los usuarios, perfilado socio-demográfico detallado, etc.) que no puede extraerse de los datos de telefonía móvil. Por lo tanto, la información que se obtiene de la telefonía móvil es, a día de hoy, complementaria y no totalmente sustitutiva de las encuestas. La solución para obtener una visión completa y robusta sobre los patrones de actividad y movilidad de la población pasa necesariamente por la combinación de distintas metodologías y la fusión con distintas fuentes de datos.

La metodología propuesta en este estudio para la extracción de patrones de actividad y movilidad de la población a partir de datos de telefonía móvil proporciona la posibilidad de entender mejor los patrones de movilidad de la población, abre las puertas a la mejora de los modelos de movilidad y de transporte actuales, y permite llevar a cabo aplicaciones prácticas que antes no eran posibles por la falta de información. El uso de los datos de telefonía móvil para estudios de movilidad ya es una realidad a día de hoy , y está siendo explotado comercialmente por distintas empresas (Airsage en EEUU, Kineo Mobility Analytics en España, Positium en Estonia, etc.). La mejora previsible en la calidad de los datos de telefonía móvil así como las futuras mejoras metodológicas para su análisis,

permiten vaticinar que los datos de telefonía móvil se posicionarán en el corto plazo como una de las fuentes de datos más relevantes para estudios de movilidad.

5.2 Futuras líneas de investigación

Los resultados obtenidos y las conclusiones extraídas del presente estudio abren la puerta a un gran número de futuras líneas de investigación. Estas líneas de trabajo futuro se detallan a continuación, diferenciando entre aquellas relacionadas con la mejora de los datos y con aspectos metodológicos de aquellas relacionadas con aplicaciones futuras de la metodología desarrollada.

5.2.1 Datos y metodología

- **Mejora de la precisión espacial y temporal de los datos de telefonía móvil**: desde el punto de vista del operador de red, mejoras relacionadas con el aumento de la granularidad temporal y precisión espacial de los datos (esta última especialmente) podrían repercutir de manera significativa en la calidad de los resultados obtenidos. Estas mejoras pasan por investigar qué sensores disponibles son capaces de permitir una gestión eficiente de la red y, al mismo tiempo, una recogida de datos de mayor calidad. La metodología propuesta en esta investigación está pensada para poder adaptarse a estos cambios en la calidad de los datos sin necesidad de llevar a cabo modificaciones significativas.

- **Análisis de sensibilidad de los parámetros del modelo**: la metodología propuesta presenta una serie de parámetros que deben ser establecidos a la hora de determinar los diarios de actividades y viajes de los usuarios de telefonía móvil (por ejemplo, el tiempo mínimo para considerar que una estancia representa una actividad). La idoneidad de los valores de los parámetros empleados en las aplicaciones prácticas ha sido comprobada en algunos casos mediante la comparación de los resultados con estadísticas oficiales. No obstante, un futuro trabajo de investigación relevante sería llevar a cabo un análisis de sensibilidad de

los valores de los parámetros del modelo para analizar con mayor detalle la influencia que éstos tienen en los resultados finales.

- **Mejora en la identificación del lugar de residencia y de trabajo de la población**: tanto el lugar de residencia como el lugar de trabajo se determinan como la actividad más frecuente dentro de unos intervalos del día predefinidos. Esta sencilla aproximación ha demostrado resultados muy satisfactorios en la mayoría de los estudios realizados (Lenormand et al. 2014, Picornell et al. 2015, Alexander et al. 2015). No obstante, esta metodología podría clasificar de manera equivocada el lugar de residencia como el lugar de trabajo y viceversa para aquellas personas que tengan trabajos nocturnos, trabajos por turnos, etc.. Información sobre usos del suelo, horarios de apertura de centros de trabajo y número de puestos de trabajo por zona podría ayudar a superar estos problemas.

- **Mejora en el criterio de selección de la muestra**: el criterio de selección de la muestra de usuarios válidos se basa en identificar usuarios con una frecuencia de registros superior a un cierto umbral para distintos periodos del día. Este umbral se define como un valor único para toda la población. Esto puede descartar usuarios con unos patrones de registros de actividad diferentes a la mayoría, como por ejemplo trabajadores nocturnos que pudieran tener menos actividad telefónica durante el día. De cara a mejorar el criterio de selección de la muestra, una interesante futura línea de investigación podría estar dirigida a la definición de distintos umbrales adaptados a los distintos grupos de población en función de sus patrones de registros.

- **Mejora en la estimación de la localización de las actividades**: como se ha comentado anteriormente, la resolución espacial es una de las limitaciones que presentan los datos de telefonía móvil. Para mejorar la ubicación exacta de las actividades dentro del área de incertidumbre que proporcionan los datos de telefonía móvil una línea de investigación interesantes es la utilización de

información sobre usos del suelo y puntos de interés para mejorar las estimaciones de localización.

- **Clasificación de tipos de actividad/propósitos de viaje**: la metodología actual contempla 4 posibles tipos de actividad o propósitos de viaje: casa, trabajo, otros frecuentes y otros no frecuentes. La distinción entre otras actividades frecuentes y otras actividades no frecuentes ya representa una novedad con respecto a otros estudios. No obstante, su clasificación más detallada según recurrencia o tipo de actividad (deporte, ocio, etc.) es aún un aspecto pendiente de investigar. El análisis de la recurrencia y horarios de dichas actividades así como la fusión con datos de uso del suelo y oferta de ocio disponible (por ejemplo, concierto en el estadio del equipo de fútbol de la ciudad) puede ayudar a caracterizar mejor los tipos de actividades realizadas.

- **Mejora de la información socio-demográfica**: la información socio-demográfica procedente de datos de telefonía móvil suele ser escasa y, en muchos casos, puede contener errores derivados de que la información disponible hace referencia al titular del contrato y no al usuario del teléfono móvil, que no tendrían por qué ser la misma persona. Estudios dirigidos a validar e inferir las características socio-demográficas de la muestra mediante la fusión con datos socio-demográficos disponibles (por ejemplo, datos de nivel de ingresos por residencia) y mediante el análisis de los patrones de comportamiento (por ejemplo, identificar a grupos de estudiantes a través de sus patrones de actividad vacacional de larga duración durante periodos no lectivos) son necesarios para mejorar una de las limitaciones más relevantes que presentan a día de hoy los datos de telefonía móvil.

- **Estimación de modo y ruta en ámbitos urbanos**: debido a las limitaciones intrínsecas actuales de los datos de telefonía, la identificación de ruta y modo en ámbitos urbanos no siempre es posible, resultando especialmente difícil para viajes de muy corta distancia/duración donde la oferta de transporte sea muy similar y variada. En este caso, parece que la solución a este problema podría

orientarse hacia la fusión con otras fuentes de datos, en lugar de intentar sobreexplotar las capacidades de los datos de telefonía móvil. Una fuente de datos prometedora en este sentido son los datos procedentes de la tarjeta inteligente de transporte público, que aportan información muy relevante sobre la movilidad en transporte público. Distintos estudios han demostrado el potencial que tiene esta fuente de datos para obtener información sobre demanda en transporte público (Wang at al. 2011, Munizaga & Palma 2012, Munizaga et al. 2014). Del mismo modo, la fusión con aforos de tráfico, registros de aparcamientos públicos, datos de sistemas de bicicleta público, etc. puede ayudar a determinar el reparto modal y ruta de los viajes.

- **Validación mediante comparación con grandes encuestas:** la metodología de esta investigación ha sido comparada con datos de encuestas de un periodo temporal similar al periodo temporal de los datos de telefonía móvil disponibles. Las únicas encuestas disponibles presentaban un tamaño de muestra reducido. LA realización de comparativas con encuestas de mayor calidad (por ejemplo, encuestas domiciliarias de 1-3% de la población) sería muy relevante para evaluar más en detalle las ventajas y limitaciones de los datos de telefonía móvil. Especialmente relevante es la comparativa de las matrices OD, variable fundamental para los estudios de movilidad.

- **Validación a nivel individual de la metodología:** la mayoría de estudios que realizan validaciones sobre la idoneidad de los algoritmos desarrollados para extraer patrones de actividad y movilidad de la población a partir de datos de telefonía móvil se basan en la comparación con encuestas. Estas validaciones se realizan, en la mayoría de los casos, con un nivel de agregación considerable, debido a las limitaciones prácticas de las encuestas. Como señala Chen et al. (2016), puede que a nivel individual se estén produciendo errores pero que a nivel agregado los resultados sean satisfactorios. No obstante, es importante señalar en este punto que si se obtienen resultados agregados similares a los de las encuestas

en lo relativo a matrices OD y otras métricas como por ejemplo el número de viajes por persona o la distribución de distancia de los viajes, es de esperar que sean los resultados a nivel individual también sean satisfactorios. Cuantas más variables relevantes se comparen, mayor seguridad en la validez de los resultados agregados. Por ejemplo, Picornell & Willumsen (2016) validan las matrices OD obtenidas mediante telefonía móvil comparando el flujo de tráfico derivado de las mismas con datos de aforos en distintos puntos de la carretera. Para validar las soluciones a nivel individual, se podría recurrir a datos de GPS o a la realización de encuestas de una sub-muestra de usuarios de telefonía móvil, tomando como cierta esta información y comparándola con la información extraída de los datos de telefonía móvil. Procedimientos similares se han empleado con anterioridad con datos de dispositivos GPS y encuestas para validar los resultados obtenidos mediante GPS.

5.2.2 Aplicaciones prácticas

- **Definición de una metodología integrada para recoger información de movilidad en ciudades**: una de las conclusiones más señaladas en este estudio es que ninguna fuente de datos por sí sola puede aportar una imagen completa sobre la movilidad en la ciudad. Es necesario la fusión de distintas fuentes de datos para poder completar las distintas piezas de información necesarias. Actualmente, los procedimientos de recogida de información de movilidad mediante encuestas presentan un alto grado de estandarización; sin embargo, no existen metodologías orientadas a cómo se deben utilizar e integrar las nuevas fuentes de datos disponibles (telefonía móvil, tarjetas de transporte público, etc.) en estudios de movilidad. La definición de una metodología que integre el uso de encuestas, aforos de tráfico, datos de telefonía móvil, datos de tarjeta inteligente de transporte, etc. de una manera eficaz y eficiente para la gestión continua de la movilidad en las ciudades representa sin duda, a día de hoy, una de las líneas de trabajo futuras más interesantes.

- **Mejora de los modelos de elección de actividades**: la información extraída del análisis de la red social y la movilidad a partir de datos de telefonía móvil puede ayudar a mejorar los actuales modelos de transporte basados en actividades, aportando nuevas variables a considerar a la hora de simular dónde y cuándo se realizan las distintas actividades. Una línea de trabajo interesante está relacionada con la definición y validación de nuevos modelos de elección de actividad que aprovechen la información masiva proporcionada por los datos de telefonía móvil.

- **Aplicación a estudios de exposición individual:** una línea de investigación interesante derivada del estudio de exposición de la población a la contaminación es el estudio de exposición a nivel individual. Estos estudios ya se realizan en la actualidad utilizando normalmente datos de GPS para un grupo reducido de personas (p.ej. Dons et al. 2011). También empiezan a plantearse los primeros estudios que utilizan datos de telefonía móvil (Dewulf et al. 2016). El análisis de la exposición individual puede ayudar a identificar grupos de personas expuestas a altos niveles de contaminación y clasificarlas por lugar de residencia, lo que permitiría dimensionar mejor los servicios sanitarios necesarios en las distintas áreas residenciales. La metodología propuesta en este estudio es de aplicación directa para estudios de exposición individual.

- **Aplicación a estudios en tiempo real**: actualmente, la mayoría de las aplicaciones prácticas de los datos de telefonía móvil se basan en extraer información para estudios de planificación, analizando datos del pasado para evaluar actuaciones presentes o futuras. Un nuevo campo interesante son las aplicaciones en tiempo real dirigidas al ámbito de la operación. Esto implica, por un lado, la necesidad de mejoras importantes de rendimiento tanto en la extracción y pre-procesado de los datos como en el proceso de extracción de patrones de actividad y movilidad a partir de los mismos. Por otro lado, también implica la necesidad de aplicar modelos predictivos que determinen el comportamiento futuro (por ejemplo, el destino del viaje) y que con cierta periodicidad (por ejemplo, cada 15 minutos)

vayan actualizándose en función de los datos reales que se van recogiendo y analizando. Aplicaciones interesantes pueden ser aquellas relacionadas con gestión de tráfico o aplicaciones en el ámbito de la seguridad ciudadana, analizando en tiempo real las aglomeraciones de personas en la ciudad.

ANEXO I - Artículo científico: “*Exploring the potential of phone call data to characterize the relationship between social network and travel behavior*”

Artículo publicado en la revista Transportation: “Picornell, M., Ruiz, T., Lenormand, M., Ramasco, J.J. , Dubernet, T. and Frías-Martínez, E.: *Exploring the potential of phone call data to characterize the relationship between social network and travel behaviour*. Transportation 42, 647-668. (2015)”

La version final publicada de este artículo está disponible en:

https://link.springer.com/article/10.1007/s11116-015-9594-1

Exploring the potential of phone call data to characterize the relationship between social network and travel behavior

Miguel Picornell ● Tomás Ruiz ● Maxime Lenormand ● José J. Ramasco ● Thibaut Dubernet ● Enrique Frías-Martínez

Abstract

Social network contacts have significant influence on individual travel behavior. However, transport models rarely consider social interaction. One of the reasons is the difficulty to properly model social influence based on the limited data available. Non-conventional, passively collected data sources, such as Twitter, Facebook or mobile phones, provide large amounts of data containing both social interaction and spatiotemporal information. The analysis of such data opens an opportunity to better understand the influence of social networks on travel behavior. The main objective of this paper is to examine the relationship between travel behavior and social networks using mobile phone data. A huge dataset containing billions of registers has been used for this study. The paper analyzes the nature of co-location events and frequent locations shared by social network contacts, aiming not only to provide understanding on why users share certain locations, but also to quantify the degree in which the different types of locations are shared. Locations have been classified as frequent (*home*, *work* and *other*) and non-frequent. A novel approach to identify co-location events based on the intersection of users' mobility models has been proposed. Results show that *other* locations different from *home* and *work* are frequently associated to social interaction. Additionally, the importance of non-frequent locations in co-location events is shown. Finally, the potential application of the data analysis results to improve activity-based transport models and assess transport policies is discussed

Keywords travel behavior, social network, mobile phone, Call Detail Record, activity-based modelling

---

Miguel Picornell (corresponding author)
Nommon Solutions and Technologies, Madrid, Spain.
email: miguel.picornell@nommon.es
phone: +34 918 388 597

Tomás Ruiz.
Universitat Politècnica de València, València, Spain

Maxime Lenormand
Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), 07122 Palma de Mallorca, Spain

José J. Ramasco
Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), 07122 Palma de Mallorca, Spain

Thibaut Dubernet
Institute for Transport Planning and Systems (IVT), ETH Zurich, 8093 Zurich, Switzerland

Enrique Frías-Martínez
Telefonica Research, Madrid, Spain

Introduction

Travel behavior is nearly always modelled as a set of independent decisions across travelers. This approach provides satisfactory results for regularly-scheduled or very inelastic activities, like work trips, but ignores the fact that intra- and extra-household interactions play a key role in many other trips and activities (e.g., leisure trips) that are planned jointly and/or depend on the trips and activities of the social contacts. The concept of a 'full individual daily pattern', which constitutes the core of the original activity-based approach, needs to be expanded to account for the influence of the social network. A key issue is incorporating realistic geographic social networks into agent-based models, which makes it necessary to characterize the form and statistical properties of the underlying social structures and the strengths of their influences. The analysis of new data sources, such as online social networks or mobile phone data, can help improve the understanding of the interdependencies and co-evolution of the social networks and the activity-travel patterns. In recent years, there has been an increasing interest in studies related to human mobility patterns (e.g., Brockmann et al. 2006; Gonzalez et al. 2008; Song et al. 2010a; Gould 2013) and social networks (e.g. Onnela et al. 2007; Lazer et al. 2009; Carrasco et al. 2008a; Clifton 2013), some of them using different spatiotemporal information from non-conventional, passively collected data sources (e.g., GPS, mobile phones, Twitter, etc.). However, only a few studies have analyzed both aspects at the same time using mobile phone data records (Calabrese et al. 2011a; Cho et al 2011; Phithakkitnukoon et al. 2012; Chen and Mei 2014). The main objective of this paper is to examine the relationship between travel behavior and social networks using mobile phone data. The paper focuses on the analysis of the characteristics of the locations shared by social contacts, aiming to understand and quantify why and in which degree those locations are shared.

The structure of the paper is as follows: first, a review of previous work related to the interaction between social network, travel behavior and the use of mobile phones is presented. Secondly, the scope and contributions of the paper are shown. In the third place, the characteristics of the dataset used in this study are described. Fourthly, the methodology followed to obtain users' social network and travel behavior from mobile phone data and to analyze the relationship between them is explained. Then, the results and main findings are presented, and their application to inform activity-based models and assess mobility policies such as carpooling is explained. Finally, the main conclusions of the paper and further research avenues are discussed.

Literature review

Social networks and travel behavior

It has been recognized that the characteristics of people's social network influence social activities and related travels (Axhausen 2005; Arentze and Timmermans 2006; Carrasco and Miller 2006). There is an increasing number of transport studies that are including social networks as an important factor to improve travel demand models. Earlier applications of social networks in transport planning and travel behavior studies date from the beginning of the present millennium. Dugundji and Walker (2005) derived a mode choice model using various static associative social networks that group

individuals by several statistics. Paez and Scott (2005) presented a similar approach to estimate the share of telecommuting at a firm in consideration of peer pressure to appear at one's desk. Carrasco and Miller (2006) explicitly included social networks in a conceptual model of social activity-travel behavior. Marchal and Nagel (2006) allowed cooperative agents in a microsimulation to share information with each other about activity locations and about other agents, in order to optimize trip chains. Arentze and Timmermans (2006) presented a framework for a multi-agent microsimulation that produces a dynamic social network which evolves together with activity-travel patterns. Hackney et al. (2006) also studied interdependencies between social networks and travel behavior. Silvis et al. (2006) found relations between number of trips and locations visited, and the social network size and number of repeated contacts. Molin et al. (2007) analyzed the influence of the size and composition of the social network on travel demand. Arentze and Timmermans (2008) focused on direct effects of social networks on activity-travel interactions. Carrasco et al. (2008a) studied the spatial distribution of social activities, focusing on the home distance of individuals. Carrasco et al. (2008b) explored the relationships between travel behavior, ICT use and social networks. Carrasco and Miller (2009) studied the effects of characteristics of individuals' personal networks and interactions on activity frequency. Hackney and Marchal (2009) developed a microsimulation model which incorporated a social network on top of a daily activity scheduler. More recently, Hackney and Marchal (2011) and Ronald et al. (2012a, 2012b) have taken into consideration the role of social networks in travel behavior using an agent-based approach. Habib and Carrasco (2011) analyze the effects of social networks on the timing and duration of activities. Van den Berg et al. (2013) studied the effects of social networks and telecommunications on activity-travel patterns. Moore et al. (2013) studied links between personal networks, time use and geographical location of people. Sharmeen et al. (2013, 2014) analyzed face-to-face social interaction and geographic accessibility.

Although significant theoretical advances have been made in understanding how the social network influences travel behavior, data availability is still a significant limitation for this kind of studies. As remarked by Van den Berg et al. (2013), only a few data collection efforts have been made so far in order to incorporate social networks in models of travel demand. Furthermore, data is usually collected through personal surveys which are limited in sample size (hundreds of users) and period of time (few days). For instance, Carrasco et al. (2008c) carried out a survey of 350 people and in-depth interviews of a subsample of 84, and Van den Berg et al. (2013) performed a survey combining a questionnaire and a 2-days social interaction diary, obtaining 747 responses (response rate 20%). On the other hand, non-conventional passively collected data from Twitter, Facebook or mobile phones, which provide relevant information on social relations and user's location data, can open an opportunity to deal with data limitation problems. In contrast to surveys, these new data sources provide location information as well as social interaction of millions of users during long periods of time. In terms of mobility, mobile phone data is one of the best sources to obtain spatiotemporal information for a long period of time covering a big percentage of the entire population (Lane et al. 2010). Additionally, when analyzing social networks, mobile phone data have the advantage of providing more relevant face-to-face personal relationship information compared to other data sources such as Twitter or Facebook (Phithakkitnukoon et al. 2012). Therefore, mobile phones seem to be one of

the most appropriate data sources to simultaneously analyze social network and travel behavior. At this point, it is worth mentioning that, as well as data from surveys, mobile phone data have their own limitations and drawbacks (e.g. limited socio-demographic information available due to privacy issues), which will be discussed at the end of this paper.

Travel behavior and mobile phone data

Recent studies from the human and social research area have demonstrated the usefulness of mobile phone data to study travel behavior. González et al. (2008), Song et al. (2010a, 2010b), and Bagrow and Lin (2012) have demonstrated that human mobility is highly structured and governed by certain patterns. Slim and Ahas (2010) used mobile phone positioning data to identify individuals' residential locations in Estonia. Ahas et al. (2010) also monitored the movements of suburban commuters in the city of Tallinn, Estonia. Mobile phone positioning data has been used to study how people move during social events (Calabrese et al. 2010). Song et al. (2010a) studied the predictability of human mobility from location data of GSM tower IDs. Becker et al. (2011) identified residential location of daily workers and the late-night revelers in the city of Morristown, New Jersey, USA, in order to understand daily flows of people in and out of city. Isaacman et al. (2011) used Call Detail Records (hereafter CDRs) to identify locations where people spend most of their time. They validated the algorithms used, derived via logistic regressions, by comparing their results to ground truth data provided by a group of volunteers. The algorithms identified home and work sites with median errors under one mile. Do and Gatica-Pérez (2012) developed algorithms to predict user mobility using various types of data collected from mobile phones of 153 volunteers during 17 months (GPS, WiFi APs, calling logs, etc.). In the transport field, research interest in relation to mobile phone-based data has been concentrated on using mobile phones as probes for estimation of aggregate level traffic parameters, such as travel time and travel speed (Bar-Gera 2007), mode share (Wang et al. 2010; Doyle et al., 2011), origin-destination matrices (White and Wells 2002; Cáceres et al. 2007; Sohn and Kim 2008; Calabrese et al. 2011b) and traffic flows (Cáceres et al. 2012). Reviews of current practices using mobile phone as traffic probes can be found in Yim (2003), Rose (2006), Cáceres et al. (2008) and Steenbruggen et al. (2011).

Social networks, travel behavior and mobile phone call data

Communication information from mobile phone data can be used to infer social network structures. For example, Eagle et al. (2009) used call logs, Bluetooth devices in proximity, cell tower IDs, application usage and phone status collected from mobile phones of 94 volunteers to study friendship behaviors. Mobile phones were equipped with software applications that recorded and sent the data to a central server. The analysis of the mobile phone data was compared with self-reported data. They found that friendship is related to in-role communication and proximity (those interactions likely to be associated with work, e.g. proximity at work), as well as with extra-role communication and proximity (those interactions that are unlikely to be associated with work, such as Saturday night proximity). Using just the extra-role communication factor from that analysis, it was possible to accurately predict 96% of symmetric non-friends (subjects who work together but neither considers the other a friend) and 95% of symmetric friends; in-role communication produced a similar accuracy. Thus they could

accurately predict self-reported friendships based only on objective measurements of behavior. Landline and mobile phone data were used by Sobolevsky et al. (2013) as a proxy for interactions to identify community regions. They detected coherent areas, and most of their boundaries closely follow existing political or socio-economic borders.

A considerable number of studies have used mobile phone data to either analyze social network or travel behavior. However, studies using mobile phone data to jointly analyze social networks and travel behavior are scarce. Phithakkitnukoon et al. (2011) identified residential locations of individuals with mobile phone positioning data and quantified the strength of social ties based on call duration. They found that residential migration can affect the strength of social ties over time: strong ties persist through a migration, while weak ties tend to disappear. In a subsequent study (Phithakkitnukoon et al. 2012), the authors found that 80% of individuals' mobile phone traces were within the 20 km proximity of their nearest social ties' residential locations. Calabrese et al. (2011) used a subset of mobile phone data from 1 million users in Portugal to study the relationship between their telecommunications patterns and physical locations. They found that there was a strong positive correlation between the call frequency between two individuals and the frequency of co-location occurrences. Cho et al. (2011) studied social travel using cell phone location data (estimated from the nearest cell phone tower of both the persons making and receiving the call), and data from two online location-based social networks. Ythier et al. (2013) used data from phone calls, sms logs and GPS of 111 people to investigate the influence of communication and social contacts on travel behavior. They found that people tend to travel in a similar manner as those they are socially connected to (consistently with the social network and travel literature) and that communication use is a complement to physical travel (consistently with the telecommunication and travel literature). Chen and Mei (2014) identified social ties and characterized basic mobility patterns using a mobile phone dataset of around 425.000 users with both location information and calling information for a large urbanized city in China.

Scope and paper contribution

The use of mobile phone data to analyze social network and travel behavior interaction is gaining interest due to its potential to identify social and travel behavior patterns based on a large sample of individuals. This source of data has the advantage of being collected passively, with no human errors, no non-response and no fatigue/attrition. The main purpose of this paper is to contribute to efforts in this area by focusing on two main aspects: (1) the relationship between social network and frequent locations visited by social network individuals, and (2) the analysis of co-location, defined as the events in which two individuals of the same social network are in the same place at the same time. Note that when analyzing frequent locations of the social network, co-location is not strictly required.

With respect to frequent locations, research on social networks and travel behavior has mainly focused on home locations, taking the spatial proximity of residential locations as a proxy of social interaction, although some studies have also addressed the distribution of the work location of the social network (e.g., Phithakkitnukoon et al. 2012). Also, mobile phone data have been analyzed to infer home and work locations, defined as the most frequent locations in a certain period of time. In this paper, in

addition to home and work locations, a methodology is proposed to identify other frequent locations. Additionally, instead of analyzing the spatial distribution of frequent locations of the social network, the focus is on the analysis of the nature (i.e. home, work, other location) of the locations shared by the individuals of the social network, aiming to understand if the reason why the user is at that location is influenced by its social contacts.

Regarding co-location, few studies have faced this problem through the use of mobile phone data. Calabrese et al. (2011) identified a co-location event when two individuals who are in the same area (different from users' home and work) call each other, which is seen as coordination between them to meet in a nearby area. Chen and Mei (2014) concluded that other attributes related to mobility patterns such as co-location are equally or even more strongly related to social interaction than the spatial distribution of residential locations of the social network; co-location is assumed to occur when two individuals are in the same place during a time frame, dividing every day into two time frames, daytime (8 am - 8 pm) and night-time (8:01 pm - 7:59 pm). In the present paper, a novel methodology to analyze co-location is proposed. For each individual, a mobility model identifying locations visited along the different days of the sample is defined. By crossing the mobility models of the individuals, a co-location event is identified if two users of the same social network are in the same place at the same time. The methodology proposed allows the identification of co-location events even if there is not phone communication between the individuals and with a high temporal resolution. Similarly to frequent locations analysis, one of the main objectives is to analyze the nature of co-location events.

To the best of our knowledge, the dataset used for this study (described in detail below) is the largest one considered so far to analyze the interaction between the social network and travel behavior

Dataset

The mobile phone data used for this study consists of a set of Call Detail Records (CDRs). CDRs are generated when a mobile phone connected to the network makes or receives a phone call or uses a service (e.g., SMS, MMS, etc.). For invoicing purposes, the information regarding the time and the Base Transceiver Station (BTS) tower to which the user was connected when the call was initiated and ended is logged, providing an indication of the geographical position of the user at certain moments. No information about the exact position of a user in the area of coverage of a BTS is known. Also, no information about the location of the cell phone is known or stored if no interaction is taking place. The CDRs used in this study were collected for Spain, comprising anonymous call information for around 24 million users, accounting for more than 50% of the 2009 Spanish population. The CDRs cover a period of time from September to November 2009 consisting of 53 days (including weekdays and weekends) which provide more than 10 billion spatiotemporal registers. From the information contained in each CDR, the following call information was extracted: caller's anonymous ID, callee's anonymous ID, day of the call, time when the call starts, duration of the call, caller's connected tower when the call starts and caller's connected tower when the call ends. Users' positions are collected from BTS towers around Spain, leading to a location accuracy of few hundreds of meters in urban areas

and several kilometers in rural areas due to the different density of towers. In order to preserve privacy, original records were encrypted. Additionally, all the information presented in this paper is aggregated. No contract or demographic data were available for this study. None of the authors of this study participated in the encryption or extraction of the CDRs.

Methodology

In this section we explain the methodology followed to: (1) determine the social network of the users, (2) identify the frequent locations visited by each user, (3) develop user mobility models, (4) analyze the interactions between social network and frequent locations and (5) analyze co-location events.

Social network

The determination of the social network of each user is based on an egocentric network approach, leading to a network of users (alters) with whom the main user (ego) has some relation. It has been considered, as in other similar studies (Onnela et al. 2007; Phithakkitnukoon et al. 2012; Chen and Mei 2014), that a relationship between two different users only exists if the phone communication between them is reciprocal. Therefore, the social network of the ego is defined as a set of nodes (one node per user) and undirected connections or links between them representing reciprocal calls. In order to measure the strength of these relations, links have been weighted by the total numbers of calls between users.

Frequent locations

User frequent locations are defined as those places repetitively visited by the user along a certain period of time. Previous studies identifying frequent locations based on mobile phone data have mainly focused on home and work locations (e.g., Isaacman et al. 2011, Phithakkitnukoon et al. 2012, Chen and Mei 2014). In this paper, other relevant locations are additionally considered. Some frequent locations are hard to identify due to their particular spatiotemporal characteristics: for instance, it seems that a place where a person goes swimming all Mondays should be considered as a frequent location; however, if the frequency is measured on a weekly or monthly basis (i.e., 7 or 30 days) this location would probably be wrongly discarded. To give response to this problem and maximize the identification of relevant frequent locations, different criteria have been defined. A location is considered frequent if the user appears at that location a minimum number of days on a single day basis; on a working day basis, considering working days from Monday to Thursday; or on a weekend basis, considering weekend from Saturday to Sunday. Fridays have intentionally not been classified neither as working days nor weekend due to their particular mixed characteristics. The minimum number of days (minimum frequency) to consider a location as a frequent location is determined by the following expression:

$$Minimum\_frequency = \alpha \cdot total\_sample\_days$$

where ‘α’ is a reduction coefficient and *‘total_sample_days’* is the total number of days of a certain type present in the sample (e.g., total number of Mondays, total number of

working days, etc.). The alpha coefficient determines the ratio between the minimum number of appearances on days of a certain type (single day, working day, etc.) and the total number of days of that type present in the sample. It is important to note that the frequency of appearance of a user at a certain location is in most cases underestimated, since the user will only appear in that position if he/she makes or receives a call. Therefore, these considerations about the nature of the data have to be taken into account when selecting the value of the alpha coefficient. As a first approach, an alpha coefficient of 0.35 has been considered adequate to estimate frequent locations.

Additionally, frequent locations have been classified into three different groups: *home*, *work* and *other*. *Home* and *work* locations are estimated considering only working days. A frequent location is classified as *home* if it is the most frequent location between 8 p.m. and 7 a.m. Similarly, *work* location is considered the most frequent location between 8:00 a.m. and 5 p.m. Finally, all other frequent locations different from *home* and *work* are classified as *other* locations. Note that a single location can be classified simultaneously as *home* and *work* location. In contrast to *other* frequent locations, as *home* and *work* locations present time restrictions, it seems reasonable that lower alpha coefficients should be considered in these cases. Moreover, as the effective hours (hours when there exists a high probability of making or receiving a call) considered for *home* locations are lower than those considered for *work* locations, the alpha coefficient considered for *home* locations should be lower. Under these considerations, alpha coefficients of 0.2 and 0.3 have been considered appropriate to estimate *home* and *work* locations respectively.

Mobility model for co-location analysis

CDRs provide, on average, spatiotemporal information of each user every several hours. This level of detail could be useful for analyzing individual daily mobility partners such as *home* and *work* trips; however, when analyzing co-location events between an individual and its social network, more detailed information is needed. To respond to this limitation, a mobility model which expands the spatiotemporal information present in the CDRs providing an estimation of the position of the user along the day has been developed (see Fig. 1). The mobility model for each user is defined as follows:

1. *CDR of the user = User's location and time information (L0, t0)*
2. *Next CDR of the user = User's location and time information (L1, t1)*
3. *if (t1 – t0) > T_threshold --> Location information missed from t0 to t1*
4. *else:*
    - *if L0 = L1 = L --> User location between [t0, t1] is L*
    - *else --> t' = f(t0, t1) / User location is L0 between [t0, t'] and L1 between [t', t1].*

T_threshold represents the maximum time distance between 2 instances (t0, t1) to consider that no relevant intermediate locations exist between those instances; and *f (ti, tj)* is a probability function that determines the time when a trip is performed.

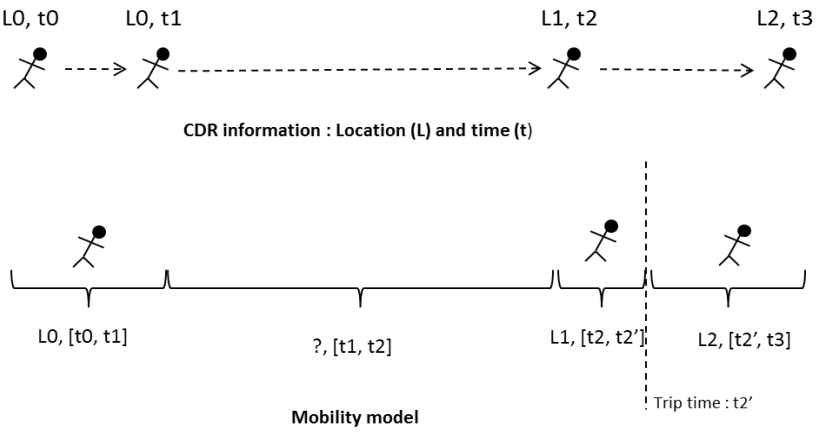


Fig. 1 Comparison between the information provided by CDRs and the information provided by the mobility model

Social network and frequent locations analysis

The main objective of this analysis is to explore the relation between the frequent locations visited by the ego and those visited by its social network. Only users whose *home* and *work* locations have been identified are considered for the analysis (around 2,300,000 users). For each egocentric network, the frequent locations of the ego are compared with the frequent locations of the alters. The common frequent locations are identified and the type of relation between those locations is classified according to the characteristics of the locations shared. There are 9 possible types of relations derived from the combination of the three possible types of frequent locations [*home, work, other*]. As mentioned before, *home* and *work* could correspond to the same location; in these cases, the type of relation is proportionally assigned (e.g., user 'A' and user 'B' share a common location 'L', for 'A' location 'L' is simultaneously *home* and *work*, and for 'B' it's *other* location; the relation type will be classified as 50% *home-other* and 50% *work-other*).

Co-location analysis

The mobility models of the different users belonging to the same social network are compared to identify co-location events. In contrast to the previous analysis, both frequent and non-frequent locations are considered. The different locations are classified as *home*, *work*, *other* and *non-frequent*. For each egocentric network, the different locations of the ego and the alters are compared along the different days of the sample. Co-location is identified and classified according to the characteristics of the locations shared. There are 16 possible types of co-locations derived from the 4 types of locations. The co-location analysis has been performed using a subset of the whole dataset covering the users living in the metropolitan area of Barcelona and their social network (independently of the place of residence), leading to a sample of around 250,000 users. Note that, as in the previous analysis, the mentioned sample only considers users whose *home* and *work* locations have been identified.

Results and discussion

Social network statistics

From the whole sample of CDRs, around 24 million of egocentric networks have been identified. The average number of alters per ego is 9.31 with a standard deviation of 17.19. The average number of phone calls between two users (considered as a proxy of the strength of the social relation) is 21. 90% of the egos have less than a phone call per day with each alter.

Frequent locations statistics

For each user of the sample, the frequency of appearance of his/her locations has been calculated. The average number of frequent locations per user is 3.47 with a standard deviation of 2.83. Considering that the minimum number of frequent locations is two (*home* and *work*), on average every user has 1.5 *other* locations which could be associated to social activities. This result shows that there are other frequent locations, apart from the commonly considered *home* and *work*, whose importance (measured as the number of locations) is similar to the non-social ones and in some cases even more important (based on the standard deviation results). This result supports the idea that considering other activities different from *home* and *work* is essential to properly capture users' travel behavior. Figure 2 shows the distribution of *other* locations according to the day of the week. Most of the *other* frequent locations have been identified on Tuesdays, Thursdays and Fridays. Moreover, there are some locations that are at the same time frequent on a working day basis and on a weekend basis.

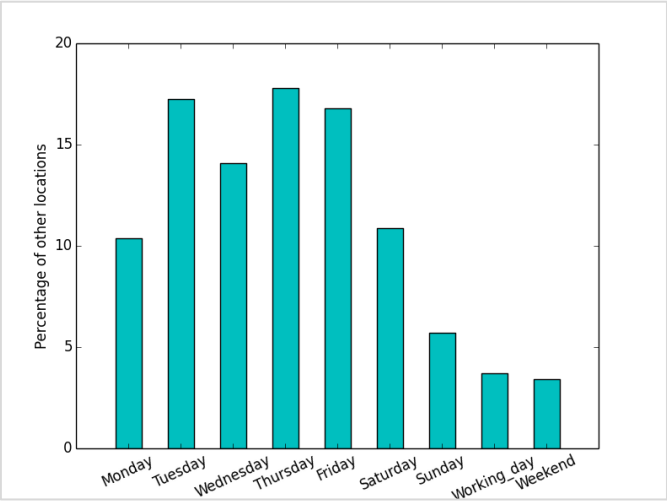


Fig. 2 Distribution of other frequent locations according to the day of the week

To validate the methodology used to estimate *home* locations, a correlation analysis comparing the results and the 2009 population distribution of Spain has been performed. The results are compared at province level (52 provinces), showing a high correlation, with $R^2 = 0.93$ (see Fig. 3). Similarly, to validate the relation between *home* and *work* locations (commuting trips) a correlation analysis comparing the results obtained and the 2011 census for the metropolitan area of Barcelona has been carried

out. The results are compared at municipal level (36 municipalities) providing an R^2 = 0.99 (see Fig 4). According to correlation results, the alpha coefficients used to estimate *home* and *work* locations seem to be adequate. The alpha coefficient for *other* locations is more difficult to validate because of the fact that there are no relevant statistics available. However, since *home* and *work* locations coefficients seem to be adequate, a value of 0.35 for *other* locations seems reasonable. It is important to highlight that the alpha coefficients proposed are appropriate for the temporal and spatial resolution of this dataset; and that other similar alpha coefficients applied to this dataset may also lead to good results. To determine which range of alpha values is appropriate for each type of location, a sensitivity or robustness analysis would be needed, being this analysis out of the scope of the present paper.

Mobility model

Mobility models have been defined for the residents of the metropolitan area of Barcelona and their social networks (250,000 users). The probability function *f(ti,tj)* used to determine the time when the trip is performed has been extracted from the trip statistics presented in the 2009 mobility survey of the metropolitan area of Barcelona ('Enquesta de Mobilitat en Día Feiner', EMEF 2009). A *T_threshold* of 4 hours has been considered appropriate for the analysis. User mobility models provide on average 4.6 hours of location information on weekdays and 2 hours on weekends (2.5 hours on Saturdays and 1.5 hours on Sundays), distributed as shown in figure 5.

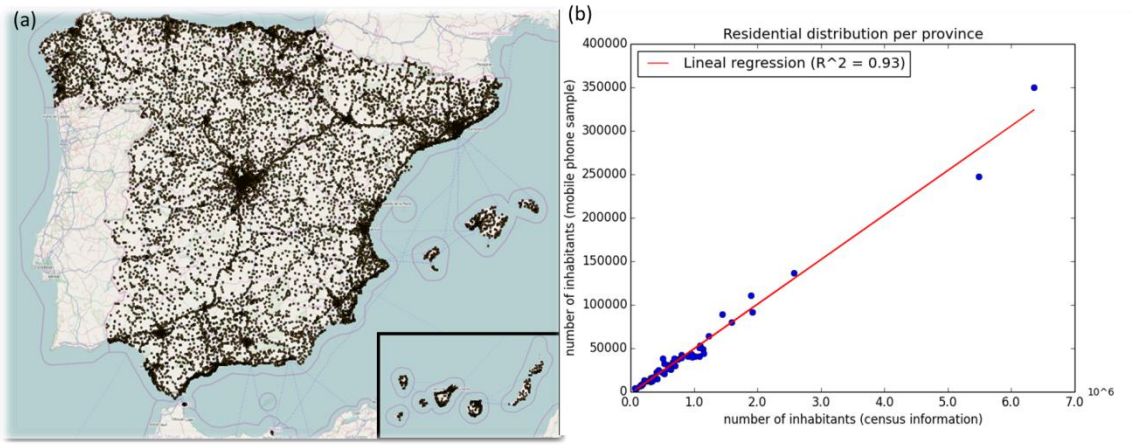


Fig. 3 (a) Home distribution based on mobile phone data analysis (b) Correlation analysis between 2009 census population information and mobile phone results

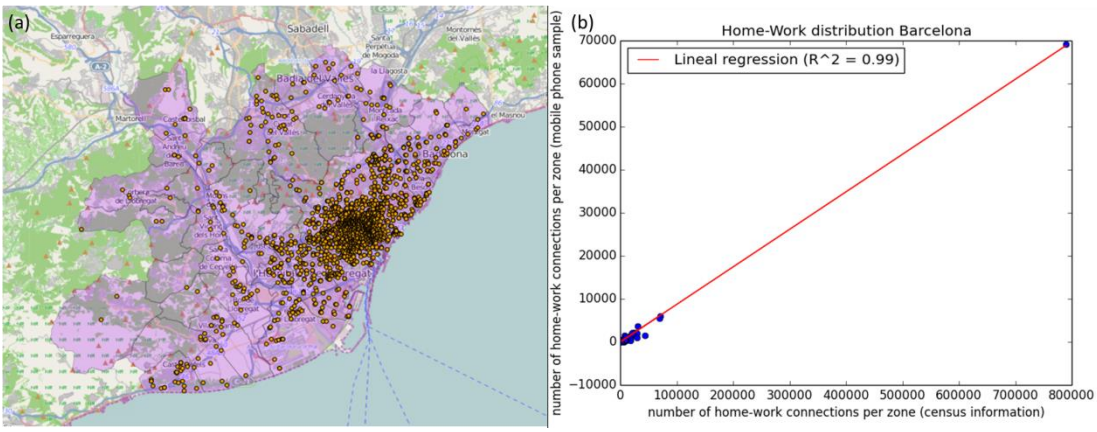


Fig. 4 (a) Home and work locations of the metropolitan area of Barcelona obtained from mobile phone data analysis (b) Correlation analysis between 2011 census Barcelona information and mobile phone results

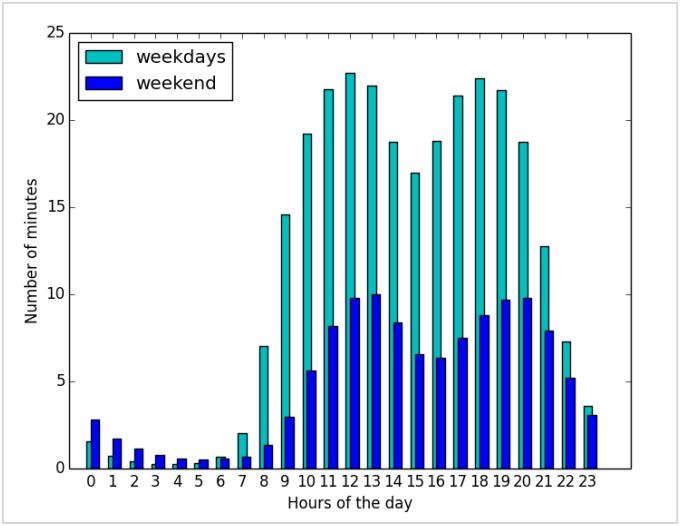


Fig. 5 Time coverage of mobility models on weekdays and weekend

Social network and frequent locations interaction results

From the 24 million egocentric networks in the sample, only those in which the ego has *home* and *work* location information are considered for the analysis (around 2.3 million). Similarly, only alters with information about their frequent locations are considered (for each egocentric network, around 17% of the alters provide that information, with a standard deviation of 13%). Results show that each ego shares, on average, at least one frequent location with 61.23 % of the alters (standard deviation of 36.88%), suggesting a significant relationship between the social network and the frequent locations visited by the users. Egos share 1.36 frequent locations with each of the alters, with a standard deviation of 0.94. Considering that users have 3.47 frequent locations, each ego shares 40% of those locations with each of the alters of the network. This result shows that not only the number of alters who share common locations with the ego is significant but also the number of common locations between

the ego and each alter. Moreover, it is observed a strong positive correlation (R^2=0.97) between the average number of phone calls and the number of frequent locations in common; suggesting that as the strength of the relation increases (number of phone calls between the ego and the alters), so does the number of common locations (see Fig. 6).

For each of the common locations between the ego and the alters, the type of interaction between them is identified and classified according to the types of the shared locations. For example, if an ego lives in location A, and the alter visits the ego frequently (so position A is an *other* frequent location for the alter), the type of interaction between them will be classified as *home-other*. A total of 9.6 million interactions ego-alter were identified and analyzed. Table 1 shows the distribution of the types of interaction between the ego and the alters. From the ego's point of view, most of the interactions with the alters occur in *other* frequent locations (54%) and to a lesser extent in *home* and *work* locations (21% and 24%, respectively). The most common type of interaction (38.28 %) between the ego and the alters is one in which they share an *other* location different from their home and work.

Social interactions can be associated to those where are least one of the locations is classified as *other.* Results show that at least 27% (standard deviation 32%) of the ego's *other* locations are shared with alters, being 14%, 16% and 70% the probabilities that the location corresponds to the *home*, *work* and *other* location of the alter respectively. It is said 'at least' because only a percentage of the alters provide location information and therefore some shared frequent locations may be missing, and consequently the percentage of *other* locations shared is probably underestimated.

Ego / Alters	Home	Work	Other	Total
Home	7.41 %	6.42 %	7.62 %	21.45%
Work	6.42 %	9.07 %	8.58 %	24.07%
Other	7.62 %	8.58 %	38.28 %	54.48%

Table 1. Distribution of the ego-alter frequent locations interaction types.

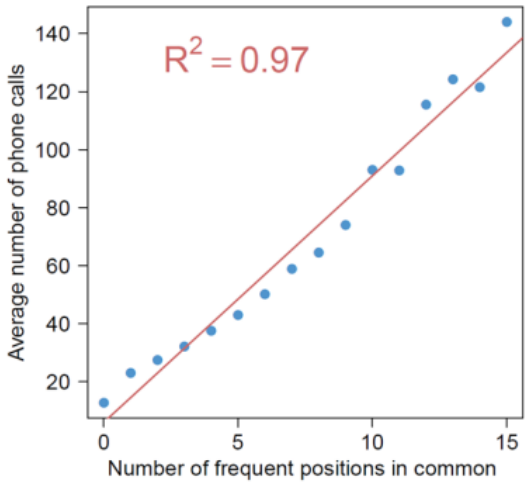


Fig. 6 Correlation between the average number of phone calls between two users and the number of frequent positions in common

Co-location analysis results

Co-location events have been analyzed for the metropolitan area of Barcelona during a period of 53 days from September to October 2009. The average number of appearances per user along the 53 days is 58.3 with a standard deviation of 51.29. Those locations are not necessarily places where the user performs an activity, but they could also be locations along a certain trip. From an ego's perspective, the number of locations shared per alter is 8.75 (standard deviation of 3.94), corresponding to 15% of the ego's locations. From those 8.75 common locations, 1.22 locations (standard deviation 0,74) are co-located locations. Each of those co-located locations has produced on average 4.36 co-location events along the sample, with a standard deviation of 4.17. It is important to note that, as it happened when analyzing common frequent locations, co-location events are probably underestimated since mobility models do not cover the 24 hours of the day.

For each co-located location, the type of the interaction between the ego and the alter has been classified according to the types of the location shared (*home, work, other* and *non-frequent* location). Around 1.4 million interactions ego-alter were identified and analyzed. Table 2 shows the distribution of the types of interaction between the ego and the alters. From the ego's point of view, a significant number of co-located locations are *non-frequent* locations (43%) and *other* frequent locations (30%), and to a lesser extent *home* (11%) and *work* (16%) locations.

Most of the places where egos co-locate are ego's frequent locations (57%), being 19.5%, 28.5% and 52% the probability of co-location at ego's *home, work* and *other* locations respectively when co-location occurs at a frequent location. These percentages are quite similar to those obtained when analyzing the types of interaction of ego's frequent locations (21.5%, 24% and 54.5% for *home, work* and *other* locations respectively). At first glance, this result might seem obvious since the more the frequent location interactions of a specific type, the higher the probability of co-locating in this type of interaction. However since the determination of *other* frequent locations does not consider the variable time, it could be the case that an ego and an alter sharing several *other* positions do not co-locate in any of them. This result also supports the hypothesis that *other* frequent locations could be associated to places where individuals of the same social network interact (co-locate), which suggests that the distribution of the type of interaction ego-alter when analyzing frequent locations could be used as a proxy of the probability of an ego to co-locate in its different types of frequent locations when co-location occurs in a frequent location. This is of considerable importance since the calculation of the type of interaction ego-alter considering frequent locations is simpler and less time consuming than the co-location analysis.

On the other hand, although co-location in frequent locations is majority, *non-frequent* locations are almost equally important (43%). When an ego co-locates at a *non-frequent* location, there is a probability of 8.1%, 11.3%, 22.8% and 57.8% that this location corresponds to the alter's *home, work, other and non-frequent* location respectively. *Non-frequent* locations can be seen as destinations rarely visited by the users. These locations will be hardly explained by transport models which only consider generalized travel costs (time, economic cost, etc.) and omit the influence of the social network. In almost half of the cases (42%), when the ego co-locates at a *non-frequent* location, it is because that location corresponds to a frequent location of

the alter. In the rest of the cases, both (the ego and the alter) shared a *non-frequent* location. The most common type of ego-alter interaction (24.81%) is one in which they share a *non-frequent* location. This type of interaction (*non-frequent-non-frequent*) can be seen as a joint decision between the ego and the alter to decide a place to meet different from their frequent locations.

Ego / Alter	Home	Work	Other	Non-frequent	Total
Home	2.51%	2.14%	2.92%	3.48%	11.05%
Work	2.14%	5.15%	4.11%	4.86%	16.26%
Other	2.92%	4.11%	12.94%	9.78%	29.75%
Non-frequent	3.48%	4.86%	9.78%	24.81%	42.94%

Table 2. Distribution of the ego-alter co-location interaction types.

Mobile phone data discussion: characteristics and limitations

Although it has been shown that mobile phone call data have the potential to provide rich information about the interaction between social networks and travel behavior, they are not free of drawbacks and limitations.

First, it is important to remark the limitations associated to the fundamental nature of mobile phone data. The use of mobile phone call data to identify social relationships inevitably misses other kind of interactions conducted through other communication channels, such as face to face or e-mail interactions. Moreover, it may misidentify certain interactions as social, such as sporadic work relationships. These intrinsic limitations may lead to errors in the estimation of some social network variables (e.g., number of social contacts) and may introduce bias in some analysis (e.g., when non-social interactions are considered in the analysis).

Apart from these intrinsic limitations, the temporal and the spatial resolution of the data highly influence the results:

- The temporal resolution of the data can be defined as the quantity of data available per unit of time. For this study, information is only available when the mobile phone user makes/receives a call. Therefore, it could be the case that a frequent location is missed or misclassified as non-frequent location if the number of calls made or received at that location is not a good proxy of the time spent by the user at that location. Likewise, some social interactions may be missed if information is not available when that social interaction is occurring.

- The spatial resolution of the data determines the accuracy in the estimation of the user position. In this study, the spatial resolution corresponds to the size of the Voronoi area associated to each BTS, which varies from few hundreds of meters in urban areas to kilometers in less populated areas. This research assumes that two users co-locate if they are in the same Voronoi area, which may lead to overestimate co-location, especially in less populated areas. Additionally, from a social perspective, it is important to remark that co-location does not ensure social interaction. Just because two users are in the same area

at the same time, it is not possible to know with certainty that a social interaction between them is taking place.

The improvement of the temporal and spatial resolution of the data will lead to more accurate results. Temporal resolution can be improved by recording other type of registers apart from calls, such as text messages or Internet connections, while spatial resolution can be enhanced by means of signal triangulation, WiFi or GPS information.

Finally, apart from data characteristics (intrinsic and extrinsic), the representativity of the mobile phone data sample is a key question. The sample has to be of enough size and homogeneously distributed among the population in order to minimize bias. The mobile phone data analyzed for this study accounts for more than 50% of the 2009 Spanish population and it is homogeneously distributed across the territory, as the comparison with census information confirms. However, as socio-demographic information is missing, some population profiles may not properly be represented in the sample. For future studies, this problem could be mitigated if the dataset provided by the mobile phone operator included some basic sociodemographic parameters available to the mobile phone company, such as age, gender, etc.

In summary, mobile phone data open an opportunity to better understand the relationship between social network and travel behavior. However, the characteristics and limitations mentioned above have to be considered when analyzing and interpreting the results in order to devise how to effectively use mobile phone data to complement the insights gained from traditional surveys.

Applications to the transport sector

Activity-based models: improvement of travel behavior modelling

The results obtained from the joint analysis of users' social network and travel behavior provide relevant information to enrich activity-based models. Indeed, one important challenge for operational daily mobility models is the prediction of location choice for *discretionary* (as opposed to mandatory) activities: while home and work locations can typically be obtained from reliable sources, such as census, the high flexibility of discretionary types makes them much more difficult to handle, the tendency being to *underestimate* traveled distances for those purposes.

In the past, various approaches have been proposed to tackle this problem. The first one, building on the classical *random utility framework*, proposes to account for *unobserved heterogeneity* in location characteristics and individual tastes using *random error terms* for each agent-location pair (see *e.g.* Horni 2013). This approach yields pretty good results, but has two main drawbacks:

- it substitutes an explanation of *why* individuals travel further than expected by random noise,
- the choice of the location being independent across agents, it is unable to represent *joint traveling to a joint activity*. Not only does this represent a substantial part of travel, but it is of prime interest for forecasting the impact of policies aiming at effecting *car occupancy.*

For those reasons, a second approach to discretionary activity location choice has been proposed, which takes into account the *willingness to pass time with social contacts* in the utility an agent derives from its daily plan (Axhausen 2005). The basic idea is the following: the choices of an agent result from a tradeoff between the benefits it derives from performing activities and the generalized cost (money, time, etc.) of the associated trips. For instance, the MATSim software platform (www.matsim.org) considers utility-maximizing agents, trying to get the most of their day given travel times (influenced by others via congestion). The basic utility function simply separately scores activity performance and travel, and sums the resulting values:

$$V = \sum_i V_i perf + \sum_j V_j leg$$

where $V_i perf$ is the reward (normally positive) to perform an activity, and $V_j leg$ the penalty (normally negative) of traveling. As long as the marginal utility of travel time is lower than the marginal utility of performing an activity, agents have an incentive to perform shorter trips. If the utility derived from an activity is allowed to vary depending on *who* participates, however, agents may get an incentive to travel further to meet social contacts — possibly reproducing the tendency elicited by the analysis in the previous section. This *joint* location choice is a first, necessary step, to include *joint travel to joint activities* in a simulation framework.

This comes however at high cost: the universal destinations choice set is enormous, and the multi-objective aspect of the problem requires the usage of non-traditional *solution concepts* to represent joint decisions (see Dubernet and Axhausen 2014, for a comparison of two solution concepts to simulate household mobility, or Ronald et al. 2012 and Ma et al. (2011, 2012) for rule-based simulated bargaining approaches). The interaction patterns between the social network and travel behavior obtained from the mobile phone data analysis may however help to make this kind of simulations tractable, for instance using the following steps:

- Consider as possible destinations the frequent locations of an agent's social contacts, since results show that co-location frequently occurs in those locations.
- Look into the intersection between isotims of the social network (line of equal transport cost), in order to find possible destinations (probably non-frequent destinations) that maximize users utility. This proposal is based on the fact that co-location events also take place at non-frequent locations.
- Use the kind of patterns obtained from the mobile phone data analysis to calibrate/validate the model.

Figure 7 shows an example of how the introduction of social interaction in activity-based models could influence the results. As results show, there is a significant probability that individuals of the same social network share *other* frequent locations. If no social interaction is considered, agents will take their own decisions and select the *other* location as a function of the benefits and costs they get. However, if social interaction is considered, there is a probability that an agent chooses the same *other* location as a contact of its social network even when that decision implies more generalized travel costs.

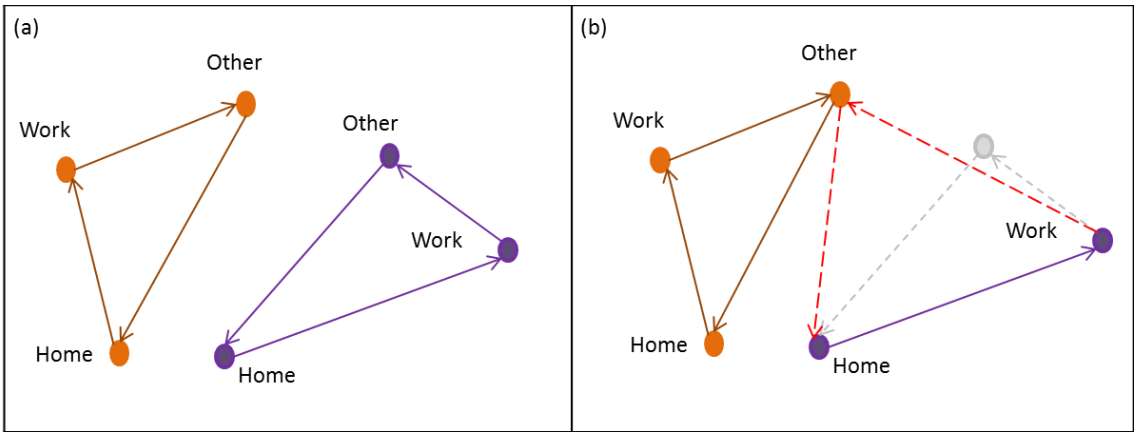


Fig. 7 Diary of activities and trips of 2 agents of the same social network during a standard day: (a) Model results without considering social network influence; (b) Model results considering social network influence

Transport policy applications

There are several policy applications where social interaction together with travel behavior information can be useful for policy planning and assessment purposes. A better understanding of the influence of the social network on travel behavior and the availability of transport modelling tools taking into account these considerations are important for evaluating transport policies where social interaction is relevant. Such transport policies are usually related with services in which transport resources (vehicles) are shared. Some examples of policies where the approach proposed in this paper can be useful are shown below:

- Transport on demand: transport on demand aims to minimize the underutilization of the transport services by dynamically adjusting supply to demand. For example, it is possible to identify frequent locations where people go out on Saturday night and determine their places of residence. Depending of the distribution of homes and the time variance of return trips, the impact (congestion, safety, etc.) of a new collective transport which maximizes the usage (number of passengers) and minimizes the cost (minimum route) could be assessed.
- Carpooling: it is usually recognized that people belonging to the same social network are more conducive to sharing transport resources. In the case of carpooling, social interaction is especially important for several reasons: the driver will probably not be a professional driver; people may feel uncomfortable sharing a car with people they don't know, etc. From the information of the home locations of the social network and the possible destinations (frequent or not), it is possible to evaluate if sharing a car could be beneficial for them.

Conclusions

It is widely recognized that social contacts have a significant influence on individual's travel behavior. Most decisions about where to perform an activity are related to the social network. This paper contributes to better understand the way social network

influences travel behavior by analyzing the nature (*home, work, other, non-frequent*) of the locations shared by social contacts using mobile phone data, showing the potential of this non-conventional data source to provide relevant information on both social interaction and travel behavior.

From the crossing analysis of social networks with frequent locations and mobility models, relevant statistics about mobility patterns and the nature of locations shared by social contacts have been obtained. The results support the hypothesis that *other* frequent locations of individuals can be considered as potential places where users of the same social network interact. Moreover, it has been shown that most of the co-location interactions are those related to ego's *non-frequent* and *other* frequent locations. Indeed, the most common type of ego-alter interaction is one in which they share *non-frequent* locations. Additionally, the potential value of these results to inform activity-based models and assess transport policies in which transport resources are shared has been discussed.

Despite the potential of mobile phone data to provide rich information about the interaction between social networks and travel behavior, a number of drawbacks and limitations shall be taken into account, such as the high spatio-temporal heterogeneity of the data or the lack of socio-demographic information. These shortcomings and limitations have been analyzed in depth. Data fusion with other data sources is a promising approach to fill the gaps of information (such as socio-demographic gaps) as well as to validate the results.

This research has thrown up many questions in need of further investigation. An interesting future line of research is the analysis of the length distribution of the trips derived from the social activities distinguishing between the different types of interactions ego-alter (e.g. *non-frequent* - *other*). Especially interesting is the case in which both users share a *non-frequent* location, aiming to explore if there is any kind of joint decision among them looking for a mutual benefit.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement n° 318367 (EUNOIA project) and nº 611307 (INSIGHT project). The work of ML has been funded under the PD/004/2013 project, from the Conselleria de Educación, Cultura y Universidades of the Government of the Balearic Islands and from the European Social Fund through the Balearic Islands ESF operational program for 2013-2017.

References

Ahas, R., Aasa, A., Silm, S. and Tiru, M. Daily rhythms of suburban commuters' movements in the tallinn metropolitan area: Case study with mobile positioning data. Transportation Research Part C, 18, 2010, 45–54.

Arentze, T. and H. J. Timmermans (2006). Social Networks, Social Interactions and Activity-Travel Behavior: A Framework for Micro-Simulation. *Paper presented at the 85th Annual Meeting of the Transportation Research Board*, January 2006, Washington, D.C.

Arentze, T. and Timmermans, H. (2008) Social Networks, Social Interactions, and Activity-Travel Behavior: A Framework for Microsimulation. Environment and Planning B: Planning and Design, 35, 1012-1027.

Axhausen, K.W. (2005) Social Networks and Travel: Some Hypotheses. In: Donaghy, K.P., Poppelreuter, S. and Rudinger, G., Eds., Social Aspects of Sustainable Transport: Transatlantic Perspectives, Ashgate, Aldershot, 90-108.

Bagrow, J. P. and Lin, Y.-R. Mesoscopic structure and social aspects of human mobility. PloS one, 7, 5, 2012, 1-11.

Bar-Gera, H. (2007). Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times:A case study from israel. Transportation Research Part C, 15, 2007, 380–391.

Becker, R. A., Cáceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A. and Volinsky, C. A tale of one city: Using cellular network data for urban planning. Pervasive Computing, IEEE 10, 4, 2011, 18-26.

Brockmann D., Hufnagel L., and Geisel T., Nature 439, 462 (2006).

Caceres, N., Wideberg, J. P. and Benitez, F. G. (2007). Deriving origin–destination data from a mobile phone network. IET Intelligent Transport Systems, 1, 1, 2007, 15-26.

Caceres, N., Wideberg, J. P. and Benitez, F. G. (2008). Review of traffic data estimations extracted from cellular networks. IET Intelligent Transport Systems, 2, 3, 2008, 179–192.

Caceres, N., Romero, L. M., Benitez, F. G. and Castillo, J. M. D. (2012). Traffic flow estimation models using cellular phone data. IEEE Transactions on Intelligent Transportation Systems, 13, 3, 2012, 1430-1441.

Calabrese, F., Pereira, F. C., Lorenzo, G. D., Liu, L. and Ratti, C. (2010) The geography of taste: Analyzing cell-phone mobility and social events. Proceedings of IEEE International Conference on Pervasive Computing, 2010.

Calabrese F, Smoreda Z, Blondel VD, Ratti C (2011a) Interplay between Telecommunications and Face-to-Face Interactions: A Study Using Mobile Phone Data. PLoS ONE 6(7): e20814. doi:10.1371/journal.pone.0020814.

Calabrese, F., Lorenzo, G. D., Liu, L. and Ratti, C.(2011b) Estimating origin-destination flows using mobile phone location data. Pervasive Computing, IEEE, 10, 4, 2011, 36–44.

Carrasco, J.A. and Miller E. J. (2006), "Exploring the propensity to perform social activities: Social networks approach," *Transportation*, 33: 463-480.

Carrasco, J.A., Hogan B., Wellman B., and Miller E. J.(2008a), "Collecting social network data to study social activity-travel behaviour: An egocentric approach," Environment and Planning B, 35(6), 961-980

Carrasco, J.A., Hogan B., Wellman B., and Miller E. J.(2008b), "Agency in social activity and ICT interactions: The role of social networks in time and space," Tijdschrift voor Economische en Sociale Geografie (Journal of Economic & Social Geography), 99(5), 562-583.

Carrasco, J.A., Miller E. J., Wellman B.(2008c). How far and with whom do people socialize? Empirical evidence about the distance between social network members. Transportation Research Record: Journal of the Transportation Research Board

Carrasco, J.A. and E. J. Miller (2009), "The social dimension in action: A multilevel, personal networks model of social activity frequency," Transportation Research Part A, 43(1), 90-104.

Chen, C. and Mei, Y (2014) Does distance still matter in facilitating social ties? The roles of mobility patterns and the built environment. Presented at 93rd TRB Annual Meeting.

Cho E., Myers S.A., Leskovek J. (2011). Friendship and mobility: user movement in location-based social networks. KDD ‘11 Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 1082-1090.

Clifton, K.J. (2013)The social context of travel behavior. In Transport Survey Methods: Best Practice for Decision Making, Ed. Zmud et al, pp. 441-448.

Do T. and Gatica-Perez D. (2012). Contextual conditional models for smartphone-based human mobility prediction. In Proc. ACM Int. Conf. on Ubiquitous Computing, Pittsburgh, Sep. 2012.

Doyle, J., Hung, P., Kelly, D., Mcloone, S. and Farrell, R. (2011). Utilising mobile phone billing records for travel mode discovery. ISSC 2011, Trinity College Dublin, 2011, June 2011.

Dubernet, T. and K. W. Axhausen (2014) Solution Concepts for the Simulation of Household-Level Joint Descision Making in Multi-Agent Travel Simulation Tools, paper presented at the 14th Swiss Transport Research Conference (STRC), Ascona, 2014

Dugundji, E. and J. Walker (2005). Discrete Choice with Social and Spatial Network Interdependencies: An Empirical Example Using Mixed GEV Models with Field and “Panel” Effects. Transportation Research Record 1921, pp. 70-78.

Eagle N., Pentland A., and Lazer D. (2009), "Inferring Social Network Structure using Mobile Phone Data", Proceedings of the National Academy of Sciences (PNAS) 106(36), pp. 15274-15278.

González, M. C., Hidalgo, C. A. and Barabási, A.-L. (2008). Understanding individual human mobility patterns. Nature, 453, 2008, 779-782.

Gould, J. (2013) Cell phone enabled travel surveys: the medium moves the message. In Transport Survey Methods: Best Practice for Decision Making, Ed. Zmud et al, pp. 51-70.

Habib, K.N., and J.A. Carrasco (2011), "Investigating the role of social networks in start time and duration of activities: A trivariate simultaneous econometric model," Transportation Research Record: Journal of the Transportation Research Board, 2230, 1-8.

Hackney, Jeremy K. and Kay W. Axhausen (2006) An agent model of social network and travel behavior interdependence, paper presented at the 11th International Conference on Travel Behaviour Research, Kyoto, August 2006.

Hackney, J. and Marchal, F. (2009). A model for coupling multi-agent social interactions and traffic simulation, in: TRB 2009 Annual Meeting, 2009.

Hackney J., Marchal F. (2011). A coupled multi-agent microsimulation of social interactions and transportation behavior. Transportation Research Part A 45, 296–309

Horni, A. (2013) Destination choice modeling of discretionary activities in transport microsimulations, Ph.D. Thesis, ETH Zurich, Zurich

Isaacman S.,Becker R., Caceres R., Kobourov S., Martonosi M., Rowland J., and Varshavsky A.(2011). Identifying important places in people's lives from cellular network data. In Proc. Int. Conf. on Pervasive Computing, San Francisco, Jun. 2011.

Lane N. D., Miluzzo E., Lu H., Peebles D., Choudhury T., and Campbell A. T. (2010). A survey of mobile phone sensing. Communications Magazine, IEEE, 48(9):140–150.

Lazer D., Pentland A., Adamic L., Aral S., Barabasi, A.-L., Brewer D., Christakis N., Contractor N., Fowler J., Gutmann M., Jebara T., King G., Macy M., Roy D., Van Alstyne M., Science 323, 721 (2009).

Ma, H., N. Ronald, T. A. Arentze and H. J. P. Timmermans (2011) New credit mechanism for semicooperative agent-mediated joint activity-travel scheduling, Transportation Research Record, 2230, 104–110.

Ma, H., T. A. Arentze and H. J. P. Timmermans (2012) Incorporating selfishness and altruism into dynamic joint activity-travel scheduling, paper presented at the 13th International Conference on Travel Behaviour Research (IATBR), Toronto, July 2012

Marchal and Nagel (2006) allowed cooperative agents in a microsimulation to share information with each other about activity locations and about other agents, in order to optimize trip chains.

Molin, E.J.E., Arentze, T.A. & Timmermans, H.J.P. (2007). Social activities and travel demands : a model-based analysis of social-network data. Transportation Research Record, 2082, 168-175

Moore, J., J.A. Carrasco, and A. Tudela (2013), “Exploring the links between personal networks, time use, and the spatial distribution of social contacts,” Transportation, 40(4), 773-788.

Onnela J-P, Saramaki J, Hyvonen J, Szabo G, Lazer D, et al. (2007) Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci USA 104: 7332–7336.

Páez A, Scott D M, 2007, "Social influence on travel behavior: a simulation example of the decision to telecommute" Environment and Planning A 39(3) 647 – 665.

Phithakkitnukoon S, Calabrese F, Smoreda Z, Ratti C (2011). Out of Sight Out of Mind: How Our Mobile Social Network Changes During Migration. In: Proceedings of the IEEE International Conference on Social Computing. Cambridge, MA, USA, pp. 515–520.

Phithakkitnukoon S, Smoreda Z, Olivier P (2012) Socio-Geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data. PLoS ONE 7(6):e39253. doi:10.1371/journal.pone.0039253.

Ronald, N.A., Arentze, T.A. & Timmermans, H.J.P. (2012a). Modeling social interactions between individuals for joint activity scheduling. Transportation Research Part B, 46, 276-290.

Ronald, N.A., Dignum, V., Jonker, C., Arentze, T.A. & Timmermans, H.J.P. (2012b). On the engineering of agent-based simulations of social activities with social networks. Information and Software Technology, 54(6), 625-638.

Rose, G.(2006) Mobile phones as traffic probes: Practices, prospects and issues. Transport Reviews, 26, 3, 2006, 275-291.

Sharmeen, F., Arentze, T. and Timmermans, H. (2013) A Multilevel Path Analysis of Social Network Dynamics and The Mutual Interdependencies Between Face-to-Face and ICT Modes of Social Interaction in The Context of Life-Cycle Events. In: Roorda, M.J. and Miller, E.J., Eds., Travel Behaviour Research: Current Foundations, Future Prospects, Lulu Press, Toronto, 411-432.

Sharmeen, F., Arentze, T.A. and Timmermans, H.J.P. (2014) Dynamics of Face-To-Face Social Interaction Frequency:Role of Accessibility, Urbanization, Changes in Geographical Distance and Path dependence. Journal of Transport Geography, 34, 211-220

Silm, S. and Ahas, R. The seasonal variability of population in estonian municipalities. Environment and Planning A, 42, 2010, 2527-2546.

Silvis, J., Niemeier, D. and D’Souza, R. (2006). Social Networks and Travel Behavior: Report from an integrated travel diary, paper presented at the 11th International Conference on Travel Behaviour Research, Kyoto, August 2006.

Sobolevsky S, Szell M, Campari R, Couronné T, Smoreda Z, et al. (2013) Delineating Geographical Regions with Networks of Human Interactions in an Extensive Set of Countries. PLoS ONE 8(12): e81707.

Sohn, K. and Kim, D. (2008). Dynamic origin–destination flow estimation using cellular communication system. IEEE Transactions on Vehicular Technology, 57, 5, 2008, 2703-2713.

Song, C., Koren, T., Wang, P. and Barabási, A.-L.(2010a) Modelling the scaling properties of human mobility. Nature Physics, 6, 2010, 818-823.

Song, C., Qu, Z., Blumm, N. and Barabási, L.-L. .(2010b) Limits of predictability in human mobility. Science, 327, 5968, 2010, 1018-1021.

Steenbruggen, J., Borzacchiello, M. T., Nijkamp, P. and Scholten, H. (2011). Mobile phone data from gsm networks for traffic parameter and urban spatial pattern assessment: A review of applications and opportunities. GeoJournal, 2011, DOI 10.1007/s10708-011-9413-y.

Van den Berg, P., Arentze, T. and Timmermans, HJP (2013) A path analysis of social networks, telecommunication and social activity–travel patterns. Transportation Research Part C 26 (2013) 256–268.

Wang, H., Calabrese, F., Lorenzo, G. D. and Ratti, C. (2010). Transportation mode inference from anonymized and aggregated mobile phone call detail records. 13th International IEEE Annual Conference on Intelligent Transportation Systems, 2010, 318-323.

White, J. and Wells, I. (2002). Extracting origin destination information from mobile phone data. Road TranSport Information and Control, 2002, 19-21 March 2002.

Yim, Y. (2003). The state of cellular probes. California PATH Working Paper, 2003, UCB-ITS-PRR-2003-25.

Ythier, J., Walker, J.L. and Bierlaire, M (2013) The influence of social contacts and communication use on travel behavior: a smartphone-based study. In: Transportation Research Board Annual Meeting.

ANEXO II - Artículo científico: "*Population dynamics based on mobile phone data to improve air pollution exposure assessments*"

Enviado para revisión a la revista: *Journal of Exposure Science and Environmental Epidemiology*

Population dynamics based on mobile phone data to improve air pollution exposure assessments

Abstract:

Air pollution is one of the greatest challenges facing cities today, being road transport one of the main contributors to pollutants such as NOx o PM. In order to efficiently evaluate which are the most appropriate policies to reduce the impact of road transport, it is essential to conduct rigorous population exposure assessments. One of the main limitations associated to those studies is the lack of information about population distribution in the city along the day (population dynamics). The pervasive use of mobile devices in our daily lives opens new opportunities to gather large amounts of anonymised, passively-collected geolocation data allowing the analysis of population activity and mobility patterns. This study presents a novel methodology to estimate population dynamics from mobile phone data based on a user-centric mobility model approach. This methodology was tested in the city of Madrid (Spain) to evaluate population exposure to NO2. A comparison with traditional census-based methods shows relevant discrepancies at disaggregated levels and highlights the need to incorporate mobility patterns into population exposure assessments.

Keywords: population exposure, population dynamics, mobile phone data, air pollution

Authors: M. Picornell (1), T.Ruiz (2), R. Borge (3), P.García (1), D. de la Paz (3) and J.Lumbreras (3)

(1) Kineo Mobility Analytics S.L., Madrid, 28043, Spain;

(2) Universitat Politècnica de València, València, 46022, Spain;

(3) Department of Chemical and Environmental Engineering, Technical University of Madrid (UPM), Madrid, 28006, Spain;

Corresponding author email: kineo@kineo-analytics.com

1. INTRODUCTION

Air pollution is one of the greatest challenges facing cities today. More than 80% of people living in urban areas that monitor air pollution are exposed to air quality levels that exceed the World Health Organization limits (WHO 2016). Pollutants are generated from a wide range of sources, including industry, transport, agriculture, waste management and households. Road transport plays an important role in cities, being one of the main contributors to nitrogen oxides (NOx) and particulate matter (PM) (EEA 2014). A large number of epidemiological studies have demonstrated that exposure to air pollution increases the risk of suffering severe diseases such as lung cancer or chronic and acute respiratory diseases (Curtis et al., 2006; Latza et al., 2009). Seeking to reduce the negative effects of road transport, cities are encouraging a shift towards more sustainable

modes of transport by fostering public transport, car-sharing, cycling or walking. Likewise, several cities have already implemented policies related to vehicles access restriction, parking management or traffic calming. In order to evaluate the impacts of those policies, it is essential to conduct population exposure assessments to air pollutants (Henneman et al., 2017). Population exposure estimations rely on both pollutants concentration and population presence. The estimation of pollutants concentration within urban areas is usually based on data collected from air quality monitoring stations and modelling techniques to improve the spatio-temporal resolution of the information (Jerret et al., 2005; Yuval et al., 2013). Recent studies focused on new methodologies to assess urban population exposure, based on alternative technologies and mobile monitoring approaches (Castell et al., 2017; Mead et al., 2013; Van den Bossche et al., 2015). However, there is still a need to define methods to consistently assess population exposure to air pollution at city scale.

Information about the spatial distribution of population along the day (from now on population dynamics) is usually collected from surveys, census data or administrative registers. While modelling techniques to estimate pollutants concentrations have experienced significant improvements in the last decades (Solazzo et al., 2012; Ching, J.K.S., 2013; Baklanov et al., 2014; de la Paz et al., 2016; Santiago et al., 2017), information about population dynamics is still one of the big limitations associated to population exposure assessments (Jerret et al., 2005). The most common approach is to consider home location as a proxy of population location (Huynh et al. 2006; Hoek et al., 2008; Anenberg et al., 2010; Boldo et al. 2011; Cesaroni el al. 2013; Brunekreef et al. 2015) and constitutes the basis for most air quality health impact assessment studies (Kunzli et al., 2000; Pope et al., 2006) However, several studies have reported important discrepancies between personal exposure and exposure at residence (Avery et al. 2010; Setton et al., 2011; Shekarrizfard et al., 2017). It has been demonstrated that travel behavior significantly influences on exposure to air pollution (Beckx et al. 2009, Dons et al. 2011; Lefebvre et al., 2013). Some studies have taken into account travel behavior by using activity-based models (Burke et al. 2001; Hatzopoulou et al.2010; Panis et al. 2010). One of the main limitations of those models is the availability of quality data to calibrate and validate them. In several cases, available data is unreliable, scarce, and/or outdated due to the practical limitations of surveys (expensive-to-collect, time consuming, small samples, inaccurate responses, etc.). Borge et al. (2016) estimate dynamic population exposure through detailed pedestrian fluxes modelling based on the social force approach (Fellendorf and Vortisch, 2010). This approach however, is computationally expensive and requires intensive input data so it is only applicable to small modelling domains within an urban area.

The pervasive use of mobile devices in our daily lives and the availability of low-cost sensing devices (e.g. bluetooth, WiFi detectors) opens new opportunities to gather large amounts of

anonymised, passively-collected geolocation data; allowing the analysis of population behavioral patterns. There are several studies analyzing population activity and mobility patterns based on GPS and mobile phone Apps (Zheng et al. 2008; Zignani et al. 2010; Hasan et al. 2013). The main advantage of those data sources is the potentially high spatial and temporal resolution that they can provide. However, the main limitations are related to the usually small samples available (based on small groups of volunteers or specific population segments) and/or the low temporal resolution of collected data (conditioned by battery consumption or data collection procedures). Similarly, sensing devices such as Bluetooth or WiFi have been used to detect population presence and infer population patterns (Van Londersele et al. 2009; Versichele et al. 2012; Naini et al.2012). This approach provides rich information with a high spatial resolution. Nonetheless, it is usually spatially limited due to infrastructure installation and maintenance needs and requires a specific mobile device configuration (e.g. WiFi or Bluetooth switched-on). Recent studies (Kontokosta & Johson 2017) pointed out the problem of non-unique MAC identifiers which limits its application to longitudinal analysis. Data collected from mobile phone network operators (from now on, mobile phone data) provides medium-high temporal and spatial resolution, wide spatial coverage and huge samples; overcoming some of the limitations identified in other data sources. Although spatial resolution is much lower than other solutions such as GPS or WiFi sensors, it is high enough to estimate population dynamics at city scale. Several studies analyzed population dynamics based on mobile phone data (Ratti et al. 2006; Reades et al. 2007; Terada et al. 2013; Deville et al. 2014), but they have only recently been applied to evaluate individual and population exposure to air pollution. Dewulf et al. 2016 analyzed individual exposure to NO2 using mobile phone data from 5 million mobile phone users living in Belgium during two days. Results show, on average, an increase in exposure to NO2 if user mobility patterns are taking into account. Nyhan et al. 2016 evaluated population-weighted exposure to air pollution in New YorK City using population dynamics estimated from mobile phone data and spatiotemporal PM2.5 concentration levels. Similarly, Gariazzo et al. 2016 analyzed population exposure to different air pollutants (NO2, O3, PM2.5) in the city of Rome using population statistics provided in the frame of the TIM BIGDATA Challenge 2015.

Previous studies analyzing population exposure through mobile phone data usually presented a tower-based approach and estimated population dynamics considering at least one of the following hypothesis: (1) if no location information is available, user remains in the previous location detected, and (2) the relative distribution of active mobile phone subscribers is a proxy of the actual distribution of the entire population in the study area. Depending on the mobile phone data characteristics and mobile phone network operator penetration ratio, the aforementioned hypotheses may be not always valid. Additionally, other limitations pointed out by previous studies are as follows: i) lack of semantic (trip purpose, mode of transport, etc.) and socio-demographic information (Dewulf et al. 2016), ii) errors in counting and positioning population due to gaps in

mobile phone data (Gariazzo et al. 2016), and iii) possible bias in the mobility patterns of individuals who are less likely to carry mobile devices (young and elderly) and to travel on a daily basis (Nyhan et al. 2016). This paper aims to contribute to a better evaluation of population exposure to air pollution by overcoming some of the limitations highlighted by previous studies. It presents a novel methodology to estimate population dynamics based on a user-centric mobility model approach instead of the traditional tower-based approach. Additionally, valid user selection and sample expansion methods are proposed to improve population dynamic estimations. The new methodology was tested in the city of Madrid (Spain), evaluating the population exposure to NO2 during a standard working day and comparing results with those obtained from static-based (census based) methodologies. The structure of the paper is as follows: first, it is presented a definition of the type of mobile phone data considered in this study and a description of the new methodology proposed to estimate population dynamics; secondly, the methodology proposed is tested in the city of Madrid (Spain) and the results are compared to traditional census-based methodologies and, finally, the conclusions and future research lines associated with this research are shown.

2. METHODOLOGY

2.1 Mobile phone data

In this study, the term "mobile phone data" makes reference to the data collected by mobile phone network operators (MNOs) when mobile devices interact with the mobile network. This data have been traditionally stored for billing purposes, collecting information every time the mobile phone device makes or receives a phone call, sends a SMS or connects to the Internet. Other network events from which data can also be collected are for example turning-on/off of mobile devices, handover events or automatic location updates. Information about the time and the antenna to which the device is connected is stored, providing an indication of the geographical location of the device at certain moments. The temporal resolution of the data depends on the type of events stored by the MNOs. Datasets containing data session events and/or periodic location updates provide high temporal resolution information. On the other hand, the spatial resolution is usually associated to the coverage area of the antennas, providing a location accuracy of few hundred of meters in urban areas and several kilometers in rural areas. In some cases, triangulation information is used to improve location accuracy. Apart from the spatio-temporal data, basic socio-demographic information such as gender or age associated to the mobile device is sometimes available from the MNO client database.

2.2 Population Dynamics

Most of the studies estimating population dynamics from mobile phone data presented a tower-based approach, identifying the number of unique active mobile devices connected to a specific tower and assuming that the relative distribution of the sample at different periods of the day is representative of the actual population distribution. In this paper, a user-centric mobility model is

presented to estimate population dynamics as the result of individual activities and trips performed by the users along the day, taking into account sample selection and expansion considerations. The methodology proposed is divided in four main steps:

- Identification of home location
- Mobility model - determination of activities and trips
- Extrapolation to the whole population
- Calculation of population dynamics

Identification of home location

Several studies have already dealt with the problem of estimating frequent locations from mobile phone data (e.g. Isaacman et al. 2011, Phithakkitnukoon et al. 2012; Chen and Mei 2014). The following formula was used in this study to identify frequent locations as presented by Picornell et al. (2015):

$$minimum\_frequency = \alpha \cdot sample\_days \; [1]$$

A location is considered as a frequent location if the number of appearances in that location during a specific time period (sample days) is greater than a minimum threshold (minimum frequency). Home location is defined as the most frequent location between X p.m. and Y a.m. during working days. X and Y values are selected depending on the population patterns of the area under study. Likewise, the 'α' parameter is defined depending on the characteristics of mobile phone data. Picornell at al. (2015) obtained satisfactory results using a value of 0.2 for identifying home location from mobile phone call data.

Mobility model - determination of activities and trips

A mobility model based on a sequence of activities and trips was proposed. Activities were identified as those locations where people spend time and trips were defined as the displacement between two consecutive activities. The methodology to determine activities and trips is explained below.

1. **Remove non-valid users**. Non-valid users were defined as those users with consecutive registers that differ a time 'tr' greater than 'TR'. This criterion intends to remove users with long periods of time without information. In contrast to other approaches, if location information is not available during a period 'TR', the user is removed instead of considering that he/she remains in the same location.

2. **Identify activities**. All the activities performed by valid users along the day were identified. A user is performing an activity 'A' in a location 'L', if the user spends a time 'ta' greater than 'TA' in that location.
3. **Determine trips**. Trips were determined as displacements between activities. The origin and destination of the trip was defined by the activities' locations. Trip time (Eq. 2) was defined as any time within the following interval:

$$trip_{time} = \left[origin_{last-register}, destination_{first-register} - trip_{duration}\right] \quad [2]$$

Where '*origin_last-register*' is the last mobile phone register observed in the origin activity, '*destination_first-register*' is the first register observed in the destination activity, and '*trip_duration'* is the estimation of the time spent to travel from the origin to the destination. As it is not always possible to determine mode and route from mobile phone data in urban areas, as a simplification, a linear trajectory and constant speed was used to estimate trip duration. Once trip duration is estimated, trip time was randomly assigned within the possible options.

Figure 1 shows and example of how mobile phone data is transformed into activity and trip information.

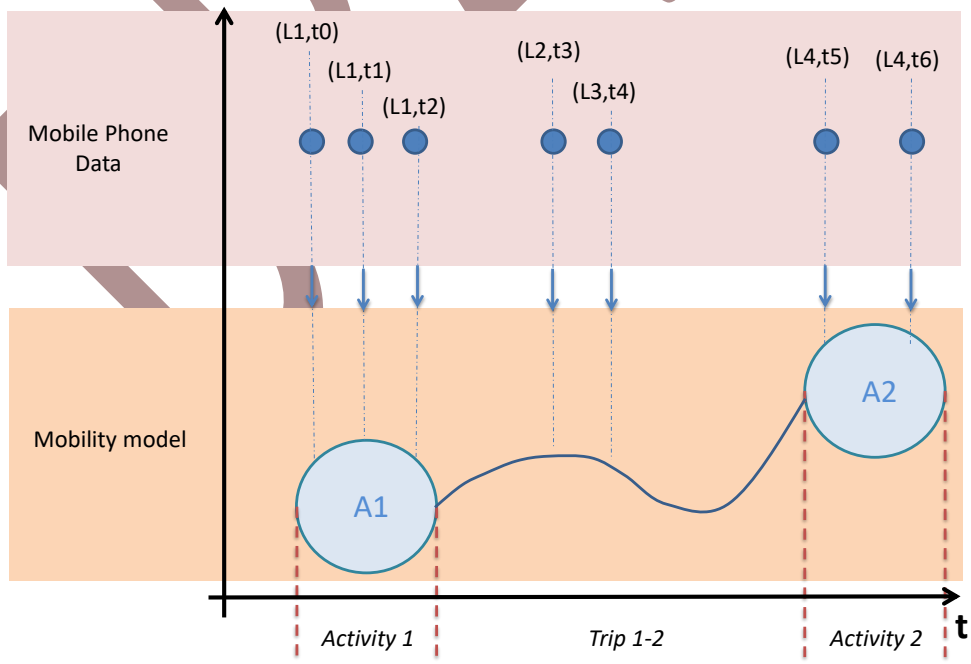


Figure 1. From mobile phone data to activity and trip information

Extrapolation to the whole population

The mobile phone sample was extrapolated to provide information about the whole population, and an expansion factor based on residence and socio-demographic information was applied. Users were classified by zone of residence, range of age and gender. For each group, a different expansion factor was applied. This approach intends to correct the possible non-homogenous sample distribution among the population. The expansion factor was calculated as follows:

$$expansion_{factor} = \frac{population(residence\_zone,\ age,\ gender)}{sample(residence\_zone, age, gender)} \quad [3]$$

Calculation of population dynamics

The number of people in a specific zone during a specific period of time (PZT) can be calculated as the number of people in that zone affected by the percentage of time spent in that zone in relation to the period duration:

$$P_{ZT} = \sum_{i=1}^{N} \frac{Tzi}{Td} \quad [4]$$

where 'N' is the size of the population under study, 'Tzi' is the time spent by the user 'i' in the zone 'Z' during the period 'T', and 'Td' is the duration of the period 'T'. Figure 2 shows an example of the time spent in different zones (TZ) by a user 'i' during a period 'T'.

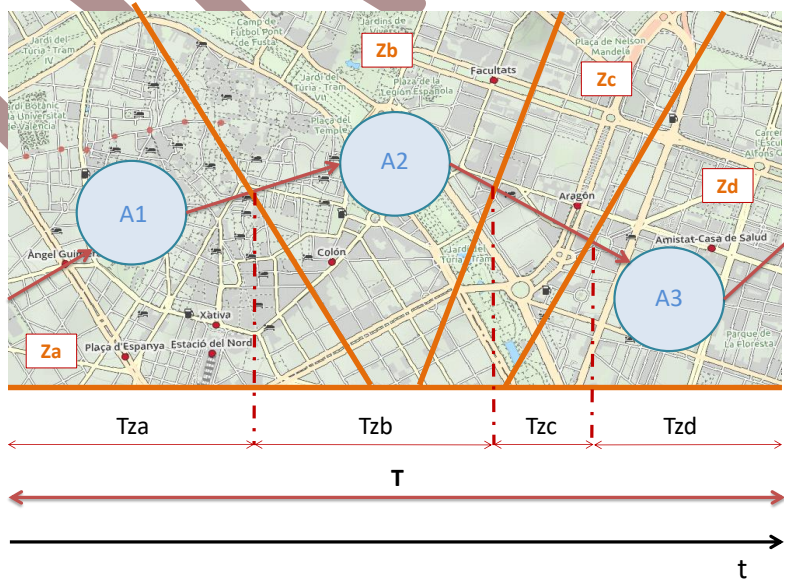


Figure 2. Example of the time spent in different zones (Tz) during a period T extracted from the mobility model information

3. CASE STUDY

The methodology proposed was tested in the city of Madrid (Spain) to measure the exposure of population to NO_2. The study area (see Figure 3) was divided in cells of 1x1km, covering a total area of 40x44km accordingly to the Air Quality Model (AQM) domain and discretization available for the city of Madrid. The analysis was performed for the 17th of November of 2014, which was a standard day in terms of population mobility and a representative day in terms of NO_2 levels. A model assessment for this particular day shows a satisfactory performance so temporal variation and spatial concentration gradients predicted by the model can be deem reasonable and fit for purpose. Since a significant number of people commute to Madrid from surrounding areas, the population of those areas (specifically Segovia, Ávila, Toledo, Guadalajara and Cuenca) has also been considered in the analysis, totalizing broadly 8 million people.

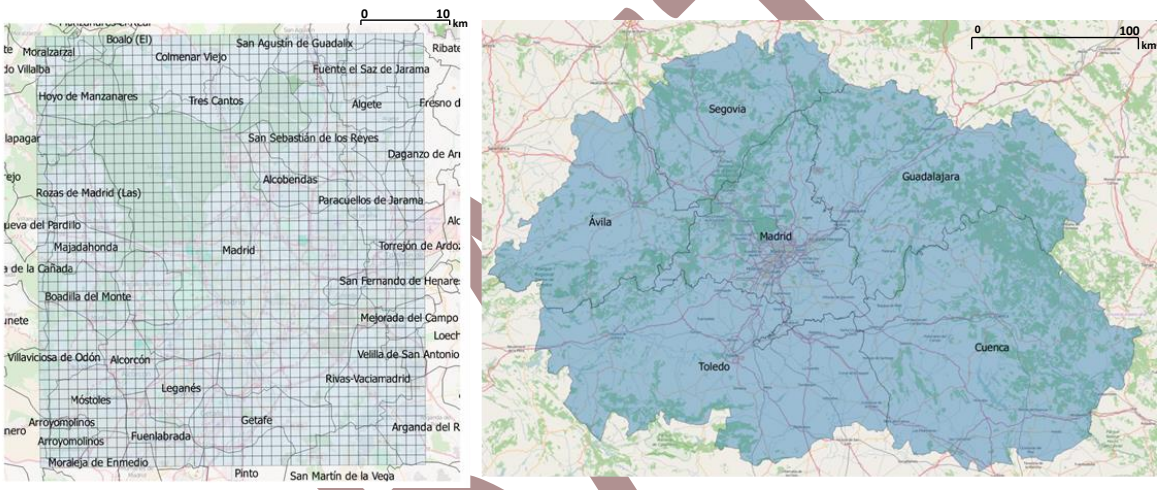


Figure 3. Study area: Madrid and surrounding provinces

Mobile Phone Dataset

The mobile phone data used for this study consisted of a set of Call Detail Records (CDRs). CDRs comprise information about calls, SMS and data sessions. Information regarding time and Base Transceiver Station (BTS) tower where the mobile phone device was located was logged, providing an indication of the geographical position of the user at certain moments. The density of towers provides a location accuracy of few hundreds of meters in urban areas and several kilometers in rural areas. CDRs were collected for the region of Madrid and its surrounding provinces, covering a period of time from October to November 2014. In order to preserve privacy, original records were encrypted and all the information extracted was aggregated. None of the authors of this study participated in the encryption or extraction of the CDRs.

Air Quality Information

Eulerian mesoscale 3D AQMs consistently describe the different physical and chemical processes that determine air quality at urban scales. Although model estimates involve uncertainties and

errors, unlike observations from air quality monitoring stations, a validated and properly fed model can provide consistent and meaningful information of ambient air concentration for a whole city throughout a given period. The simulation of ambient concentration in this study relies on the WRF (Skamarock and Klemp, 2008) – SMOKE (Institute for the Environment, 2009) – CMAQ (Byun and Schere, 2006) modelling system and a detailed bottom-up emission inventory specifically develop for this area (Borge et al., 2014) to produce reliable ambient air quality estimations. NO_2 outputs with 1h temporal resolution and 1 km^2 spatial resolution were used for this analysis. This pollutant was selected because it is currently the main concern from the air quality point of view in Madrid and the primarily focus of plans and measures. In addition, the modelling system used has been extensively tested and assessed for this pollutant (Borge et al., 2014 and references within).

Mobility model specification – population dynamics

Home location was estimated as the most frequent location between 8 p.m. and 7 a.m. during working days in the period of October and November of 2014. Alpha parameter ('α') was set to 0.2 accordingly to previous validated studies (Picornell et al. 2015). In order to remove non-valid users, a time threshold between registers (TR) of 8 hours during the night period (defined from 8 pm to 7 am) and 4 hours during the rest of the day (from 7 am to 8 pm) were set. Similarly, activity time threshold (TA) was set to 30 minutes. Trip duration was estimated considering a linear trajectory between the origin and the destination and a constant speed of 15 km/h[1] for all trips. The exact trip time was estimated using a uniform probability function. The expansion factor (see [3]) was applied to all valid users grouping them by census track and age range (it was considered only one age segment, population over 16 years). Population dynamics were computed for each 1 km^2 cell and per each hour of the day. As other reasonable parameter values (e.g. α = 0.3 or TA = 25 minutes) may lead to slightly different results, a sensibility analysis would be of interest to evaluate parameters' influence in final results. It will be part of future studies.

Population exposure indicator

The population exposure to NO_2 in a specific area has been computed as the product of NO_2 concentration and the number of persons per hour in that zone ($\frac{\mu g}{m^3} * person * hour$). From now on, we will refer to that indicator as "exposure indicator". It should be noted that this index intends to assess general exposure levels since it does not take into account individual exposure patterns or accumulated dose.

Results and discussion

[1] Road average speed in the city of Madrid in 2014 was 24 km/h according to city council open database (link). Since distance is measured as a linear trajectory and other modes (bus, metro, walking, etc.) may have lower average speeds, it was considered appropriate to define an average speed of 15 km/h for all kind of trips.

Figure 4 shows the 24-h average NO_2 ground level concentration predicted by the modelling system. The exposure indicator has been calculated for the 1760 cells and for each hour of the day. Figure 5 shows an example of its evolution in each cell at different periods of the day (4 a.m., 9 a.m. and 12:00 p.m.). At the beginning of the day (4 a.m.), NO_2 levels are lower, and Madrid population is mainly located at residential areas. During the morning peak (9 a.m.), most of the population from the metropolitan area and surrounding provinces commute to the city centre. Both population and NO_2 concentration significantly increase at the city centre, leading to higher values of exposure. At midday, NO_2 levels decrease while population remains concentrated in the city centre.

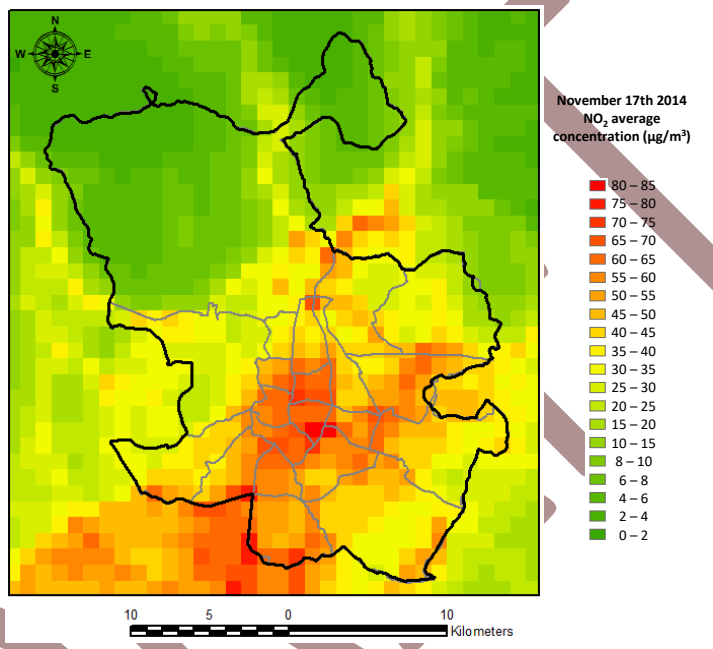


Figure 4. NO2 average 24-h concentration

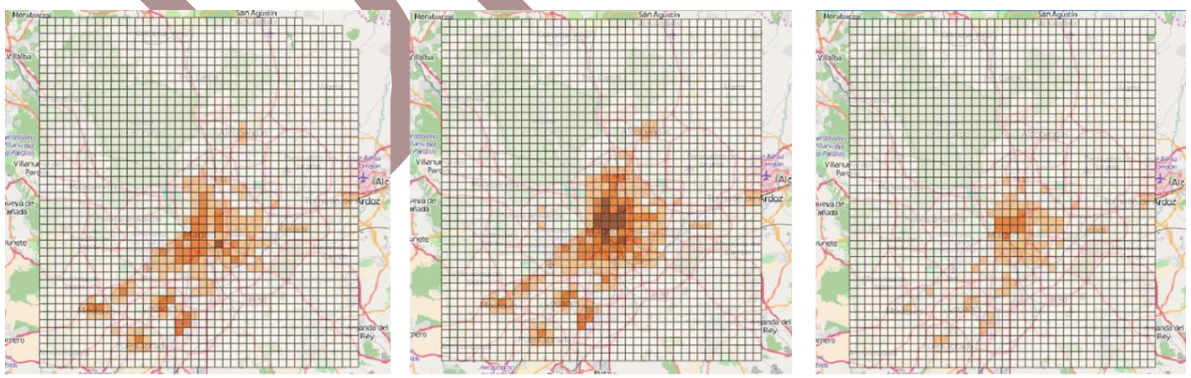


Figure 5. Exposure indicator at different times of the day

Results obtained with this novel methodology were compared to those obtained with the traditional static-based approach, and illustrated in Figures 6 and 7. Results show that both methodologies provide similar average results for the whole area, as total exposure difference throughout the day is below 4%. The main reason for such similarity is negligible population fluxes on the area boundaries since most of the movements correspond to round trips that start and end inside the modelling domain. Moreover, most of the trips are performed between areas with similar NO_2 levels. However, important discrepancies can be observed at more disaggregate levels. Figure 8 shows the evolution of the exposure indicator along the day at four different districts of the city of Madrid: Salamanca, Puente de Vallecas, Chamartín, and Carabanchel. The exposure indicator was calculated using the static and the dynamic approaches. It can be seen that, for the case of the Chamartín and Salamanca districts, the static approach significantly underestimates the exposure to NO_2. On the other hand, for the Carabanchel and Puente de Vallecas districts, the static approach overestimates the exposure indicator. These results are explained by the different district activity patterns, being Chamartín and Salamanca business attractor districts and Carabanchel and Puente de Vallecas mainly residential areas. Note that both methodologies provide similar results at first hours of the day, since most of the people are located at home location during these hours.

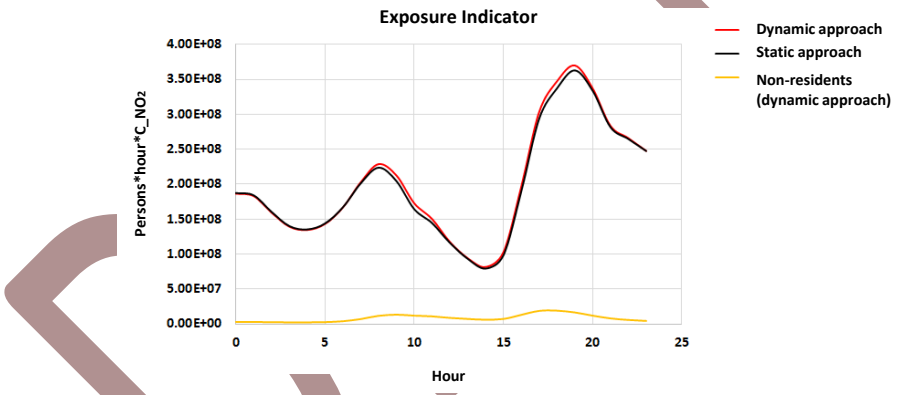


Figure 6. Exposure Indicator evolution for the whole study area. Static versus dynamic approach

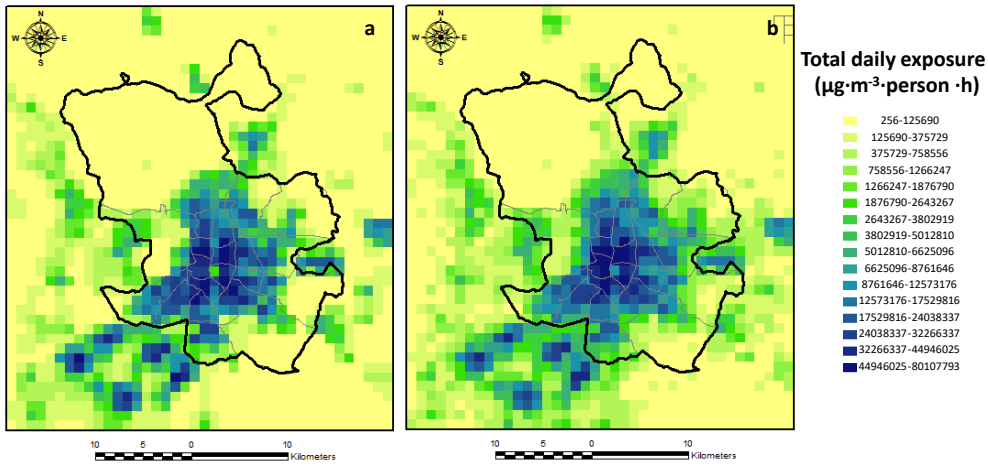


Figure 7. Total exposure throughout the day estimates: a) static approach and b) dynamic approach

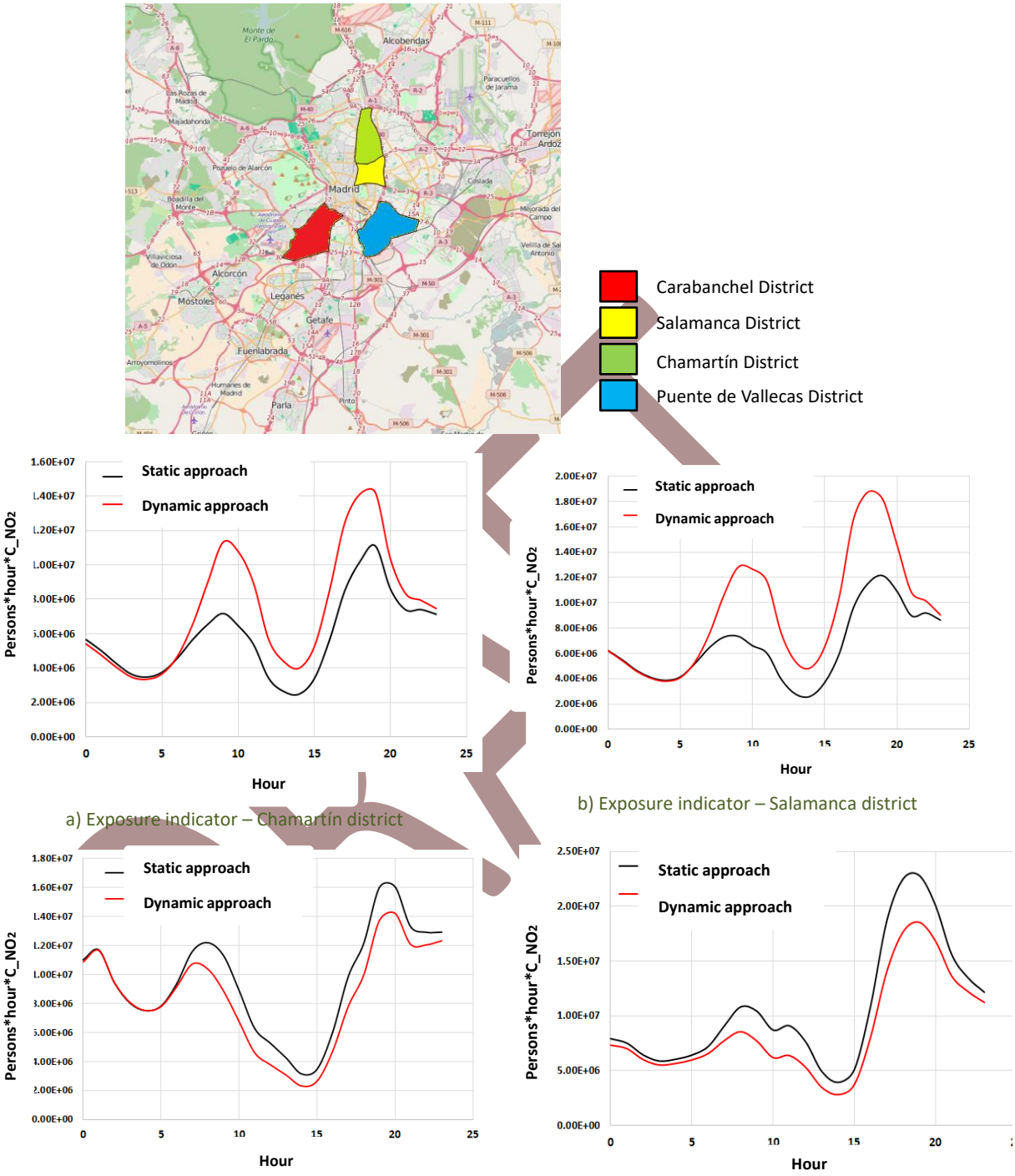


Figure 8. Exposure indicator evolution for different Madrid districts. Comparison between static and dynamic approaches.

4. CONCLUSIONS

Air pollution is one of the greatest challenges facing cities today. In order to evaluate effectiveness of alternative policies and measures to reduce air pollution impact, population exposure assessments are needed. However, population exposure assessments are limited by the lack of reliable information about population distribution along the day in the city (population dynamics). In this study, a novel methodology to estimate population dynamics based on mobile phone data was presented to improve population exposure assessments. A user-centric mobility model approach was proposed in contrast to traditional tower-based approaches. Sample filtering and expansion methods were proposed to overcome some of the limitations pointed out by previous studies. The methodology proposed was tested in the city of Madrid (Spain) to evaluate population exposure to NO2. Results show that, for spatially aggregated analysis, the conventional static methodology (census based) and the dynamic methodology (mobile phone based) provide similar results. However, relevant discrepancies were found when analyzing results at more disaggregated levels; being the consideration of mobility patterns essential to determine the actual population exposed to air pollution.

Although the study demonstrated the potential of mobile phone data to capture population dynamics and improve population exposure assessments, this approach is not free of drawbacks and limitations. One of the limitations is associated with the socio-demographic information available. Socio-demographic information from mobile phone data is usually scarce and normally refers to the contract holder, who may not be the same person as the mobile phone user. This may influence in the extrapolation process and may lead to errors when analyzing results by socio-demographic profiles. Additionally, there are segments of the population who do not have mobile phone (mainly children), so complementary studies are needed for evaluating air pollution exposure of those population segments.

This research opens new opportunities for further investigation. The user-centric mobility model approach brings the possibility of classifying the presence of population by type of activity (home, work, leisure, etc.) and trips. In the case of trips, mobile phone data and other data sources (e.g. smart card data) may help estimating the mode of transport. A better understanding of activities (e.g. staying at home) and trips (e.g. commuting by metro) performed by population will potentially lead to a better evaluation of air pollution exposure. Additionally, another interesting research line is the one related to the performance of individual exposure assessments, which would be enabled by the proposed approach.

5. ACKNOWLEDGEMENTS

This study was supported by the Madrid City Council and the TECNAIRE-CM (innovative technologies for the assessment and improvement of urban air quality) scientific programme funded by the Directorate General for Universities and Research of the Greater Madrid Region (S2013/MAE-2972)

REFERENCES

Anenberg, S.C., Horowitz, L.W., Tong, D.Q., West, J.J., 2010. An estimate of the global burden of anthropogenic ozone and fine particulate on premature human mortality using atmospheric modeling. Environmental Health Perspectives 118, 1189–1195.

Avery, C. L.; Mills, K. T.; Williams, R.; McGraw, K. A.; Poole, C.; Smith, R. L.; Whitsel, E. A. Estimating error in using ambient PM2.5 concentrations as proxies for personal exposures: a review. Epidemiology 2010, 21, 215−223 (2010).

Baklanov, A., Schlünzen, K., Suppan, P., Baldasano, J., Brunner, D., Aksoyoglu, S., Carmichael, G., Douros, J., Flemming, J., Forkel, R., Galmarini, S., Gauss, M., Grell, G., Hirtl, M., Joffre, S., Jorba, O., Kaas, E., Kaasik, M., Kallos, G., Kong, X., Korsholm, U., Kurganskiy, A., Kushta, J., Lohmann, U., Mahura, A., MandersGroot, A., Maurizi, A., Moussiopoulos, N., Rao, S., Savage, N., Seigneur, C., Sokhi, R., Solazzo, E., Solomos, S., Sørensen, B., Tsegas, G., Vignati, E., Vogel, B., Zhang, Y., 2014. Online coupled regional meteorology chemistry models in Europe: current status and prospects. Atmos. Chem. Phys. 14 (1), 317e398.

Beckx C, Int Panis L, Arentze T, Janssens D, Torfs R, Broekx S, Wets G.: A dynamic activity-based population modelling approach to evaluate exposure to air pollution: methods and application to a Dutch urban area. Environ Impact Assess Rev. 2009;29:179–85 (2009).

Boldo, E., Linares, C., Lumbreras, J., Borge, R., Narros, A., García-Pérez, J., Fernández-Navarro, P., Pérez-Gómez, B., Aragonés, N., Ramis, R., Pollán, M., Moreno, T., Karanasiou, A., López-Abente, G.: Health impact assessment of a reduction in ambient PM2.5 levels in Spain. Environment International 37, 342-348 (2011).

Borge, R., Lumbreras, J., Pérez, J., de la Paz, D., Vedrenne, M., de Andrés, J.M., Rodríguez, M.E., 2014. Emission inventories and modeling requirements for the development of air quality plans. Application to Madrid (Spain). Science of the Total Environment 466-467, 809-819.

Borge, R., Narros, A., Artíñano, B., Yagüe, C., Gómez-Moreno, F.J., de la Paz, D., Román-Cascón, C., Díaz, E., Maqueda, G., Sastre, M., Quaassdorff, C., Dimitroulopoulou, C., Vardoulakis, S., 2016. Assessment of microscale spatio-temporal variation of air pollution at an urban hotspot in Madrid (Spain) through an extensive field campaign. Atmospheric Environment 140, 432–445.

Brunekreef B, Hoek G, Schouten L, Bausch-goldbohm S, Fischer P, Armstrong B, Hughes E, Jerrett M, Brandt P Van Den. Effects of long-term exposure on respiratory and cardiovascular mortality in the Netherlands: the NLCS-AIR study. 2009. http://www.n65.nl/NCLS-AIR-Study-2009.pdf. Accessed 5 June 2015. (2015)

Burke, J. M.; Zufall, M.; Özkaynak, H. A population exposure model for particulate matter: case study results for PM2.5 in Philadelphia, PA. J. Exposure Anal. Environ. Epidemiol. 2001, 11, 470−489 (2001).

Byun, D.W., Schere, K.L., 2006. Review of the governing equations, computational algorithms and other components of the models-3 community multiscale air quality (CMAQ) modeling systems. Applied Mechanics Review 59 (2), 51–77.

Castell, N., Dauge, F.R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., Bartonova, A., 2017. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environment International 99, 293-302.

Cesaroni, G., Baldoni, C., Gariazzo, C., Stafoggia, M., Sozzi, R., Davoli, M., Forastiere, F.: Long term exposure to urban air pollution and mortality in a cohort of more than a million adults in Rome. Environ. Health Perspect. 121 (3), 324e331.(2013)

Chen, C. and Mei, Y (2014) Does distance still matter in facilitating social ties? The roles of mobility patterns and the built environment. Presented at 93rd TRB Annual Meeting.

Ching, J.K.S., 2013. A perspective on urban canopy layer modeling for weather, climate and air quality applications. Urban climate 3, 13-39.

Curtis L, Rea W, Smith-Willis P, Fenyves E, Pan Y. Adverse health effects of outdoor air pollutants. Environ Int 2006;32(6):815–30 [Aug].

De la Paz, D., Borge, R., Martilli, A., 2016. Assessment of a high resolution annual WRF-BEP/CMAQ simulation for the urban area of Madrid (Spain). Atmospheric Environment 144, 282-296.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., ... & Tatem, A. J.: Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences, 111(45), 15888-15893 (2014).

Dewulf, B.; Neutens, T.; Lefebvre, W.; Seynaeve, G.; Vanpoucke, C.; Beckx, C.; Weghe, N.V. Dynamic assessment of exposure to air pollution using mobile phone data. Int. J. Health Geogr. 2016, 15, 1–14.

Dons, E.; Int Panis, L.; Van Poppel, M.; Theunis, J.; Willems, H.; Torfs, R.; Wets, G.: Impact of time-activity patterns on personal exposure to black carbon. Atmos. Environ. 2011, 45, 3594−3602 (2011).

European Environmental Agency (EEA): https://www.eea.europa.eu/signals/signals-2013/infographics/sources-of-air-pollution-in-europe/view (2014)

Fellendorf, M., Vortisch, P., 2010. Microscopic Traffic Flow Simulator VISSIM. Fundamentals of Traffic Simulation. Springer, 63-93.

Gariazzo, C., Pelliccioni, A., Bolignano, A., 2016. A dynamic urban air pollution population exposure assessment study using model and population density data derived by mobile phone traffic. Atmos. Environ. 131, 289–300

Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013, August). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In Proceedings of the 2nd ACM SIGKDD international workshop on urban computing (p. 6). ACM.

Hatzopoulou, M.; Miller, E. J. Linking an activity-based travel demand model with traffic emission and dispersion models: Transport's contribution to air pollution in Toronto. Transport. Res.D: Tr. E 2010, 15, 315−325 (2010).

Henneman, L.R.F., Liu, C., Mulholland, J.A., Russell, A.G., 2017. Evaluating the effectiveness of air quality regulations: A review of accountability studies and frameworks. Journal Of The Air & Waste Management Association 67, 144-172

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., Briggs, D., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric Environment 42, 7561-7578.

Huynh M, Woodruff TJ, Parker JD, Schoendorf KC.: Relationships between air pollution and preterm birth in California. Pediatr Perinat Epidemiol. 2006;20:454–61 (2006).

Institute for the Environment. SMOKE v2.7 user's manual. Chapel Hill, NC: University of North Carolina; [Available online at: http://www.smoke-model.org/version2. 7/html/ch01.html], 2009.

Isaacman S.,Becker R., Caceres R., Kobourov S., Martonosi M., Rowland J., and Varshavsky A.(2011). Identifying important places in people's lives from cellular network data. In Proc. Int. Conf. on Pervasive Computing, San Francisco, Jun. 2011.

Jerret, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglou, T., Morisson, J., Giovis, C., 2005. A review and evaluation of intraurban air pollution exposure models. Journal of Exposure Analysis and Environmental Epidemiology 15, 185-204

Kontokosta, C. E., & Johnson, N.: Urban phenology: Toward a real-time census of the city using Wi-Fi data. Computers, Environment and Urban Systems, 64, 144-153 (2017).

Kunzli, N., Kaiser, R., Medina, S., Studnicka, M., Chanel, O., Filliger, P., Herry, M., Horak Jr., F., Puybonnieux-Texier, V., Quenel, P., Schneider, J., Seethaler, R., Vergnaud, J.C., Sommer, H., 2000. Public-health impact of outdoor and traffic-related air pollution: a European assessment. Lancet 356, 795–801

Latza, U., Gerdes, S. & Baur, X. (2009). Effects of nitrogen dioxide on human health: systematic review of experimental and epidemiological studies conducted between 2002 and 2006. International Journal of Hygiene and Environmental Health, 212, 271–287.

Lefebvre, W., Degrawe, B., Beckx, C., Vanhulsel, M., Kochan, B., Bellemans, T., Janssens, D., Wets, G., Janssen, S., De Vlieger, I., 2013. Presentation and evaluation of an integrated model chain to respond to traffic-and health-related policy questions. Environmental Modelling & Software 40, 160-170.

Mead, M.I., Popoola, O.A.M., Stewart, G.B., Landshoff, P., Calleja, M., Hayes, M., et al.,2013. The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. Atmospheric Environment 70, 186-203.

Naini, F. M., Dousse, O., Thiran, P., & Vetterli, M.: Population size estimation using a few individuals as agents. In Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on (pp. 2499-2503). IEEE.(2011)

Nyhan, M., Grauwin, S., Britter, R., Laden, F., McNabola, A., Misstear, B., Ratti, C., 2016. 'Exposure Track' - The Impact of Mobile Device Based Mobility Patterns on Quantifying Population Exposure to Air Pollution. Environmental Science & Technology

Panis, L.I.: New directions: air pollution epidemiology can benefit from activity-based models. Atmos. Environ. 44, 1003e1004 (2010).

Phithakkitnukoon S, Smoreda Z, Olivier P (2012) Socio-Geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data. PLoS ONE 7(6):e39253. doi:10.1371/journal.pone.0039253.

Picornell M, Ruiz T, Lenormand M, Ramasco J, Dubernet T, Frias-Martinez E (2015), Exploring the potential of phone call data to characterize the relationship between social network and travel behavior, Transportation, 42, pp. 647-668.

Pope 3rd, C.A., Dockery, D.W., 2006. Health effects of fine particulate air pollution: lines that connect. J. Air Waste Manage. Assoc. 56, 709–742. Roorda-Knape, M.C., Janssen, N.A.H., de Hartog, J., van Vliet, P.H.N., Harssema, H., Brunekreef, B., 1998. Air pollution from traffic in city districts near major motorways. Atmos. Environ. 32, 1921–1930.

Ratti C, Pulselli RM, Williams S, Frenchman D: Mobile landscapes: using location data from cell phones for urban analysis. Environment and Planning B 33: 727.C. (2006)

Reades, J., Calabrese, F., Sevtsuk, A., & Ratti, C.: Cellular census: Explorations in urban data collection. IEEE Pervasive Computing, 6(3). (2007)

Santiago, J.L., Borge, R., Martin, F., de la Paz, D., Martilli, A., Lumbreras, J., Sanchez, B., 2017. Evaluation of a CFD-based approach to estimate pollutant distribution within a real urban canopy by means of passive samplers. Sci. Total Environ. 576, 46–58.

Setton, E., Marshall, J.D., Brauer, M., Lundquist, K.R., Hystad, P., Keller, P., CloutierFisher, D., 2011. The impact of daily mobility on exposure to traffic-related air pollution and health effect estimates. Journal of Exposure Science and Environmental Epidemiology 21 (1), 42-48.

Shekarrizfard, M., Faghih-Imani, A., Tetreault, L.F.,vYasmin, S., Reynaud, F., Morency, P., Plante, C., Drouin, L., Smargiassi, A., Eluru, N., Hatzopoulou, M., 2017. Modelling the Spatio-Temporal Distribution of Ambient Nitrogen Dioxide and Investigating the Effects of Public Transit Policies on Population Exposure Environmental Modelling & Software 91, 186-198.

Skamarock, W. C., and Klemp, J. B.: A time-split nonhydrostatic atmospheric model for weather research and forecasting applications, Journal of Computational Physics, 227, 3465-3485, 2008.

Solazzo, E., Bianconi, R., Vautard, R., Appel, K.W., Moran, M.D., Hogrefe, C., Galmarini, S., 2012. Model evaluation and ensemble modelling of surface-level ozone in Europe and North America in the context of AQMEII. Atmos. Environ. 53, 60-74.

Terada, M., Nagata, T., & Kobayashi, M.: Population estimation technology for mobile spatial statistics. NTT DOCOMO Techn. J, 14, 10-15 (2013).

Van den Bossche, J., Peters, J., Verwaeren, J., Botteldooren, D., Theunis, J., De Baets, B., 2015. Mobile monitoring for mapping spatial variation in urban air quality: development and validation of a methodology based on an extensive dataset. Atmospheric Environment 105, 148-161.

Van Londersele, B., Delafontaine, M., & Van de Weghe, N.: Bluetooth Tracking. GIM International. 23-25. (2009).

Versichele, M., Neutens, T., Delafontaine, M. & Van de Weghe, N.:The use of Bluetooth for analyzing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. Applied Geography, 32(2), 208-220 (2012).

WHO (2016) Global Urban Ambient Air Pollution Database

Yuval, B.S., Broday, D.M., 2013. Data–driven nonlinear optimization of a simple air pollution dispersion model generating high resolution spatiotemporal exposure. Atmos. Environ. 79, 261–270

Zheng, Y., Li, Q., Chen, Y., Xie, X., & Ma, W. Y.: Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing* (pp. 312-321). ACM (2008).

Zignani, M., & Gaito, S. (2010, October). Extracting human mobility patterns from gps-based traces. In *Wireless Days (WD), 2010 IFIP* (pp. 1-5). IEEE.

Bibliografía

Ahas, R., Aasa, A., Mark, Ü., Pae, T., Kull, T.: Seasonal tourism spaces in Estonia: case study with mobile positioning data. Tourism Management 28(3): 898–910 (2007)

Ahas, R., Aasa, A., Roose, A., Mark, Ü., Silm, S.: Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. Tourism Management 29(3): 469–486 (2008)

Ahas, R., Aasa, A., Silm, S., Tiru, M.: Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: case study with mobile positioning data. Transp. Res. Part C 18, 45–54 (2010)

Alexander L., Jiang S., Murga M. and González M.: Origin-Destination trips by purpose and time of day inferred from mobile phone data, Transportation research part C. Emerging Technologies, vol. 58 , pp. 240-250, (2015).

Arentze, T.,Timmermans, H. J.: social networks, social interactions and activity-travel behavior: a framework for micro-simulation. Paper presented at the 85th annual meeting of the Transportation Research Board, Washington, D. C., Jan 2006 (2006)

Arentze, T., Timmermans, H.: Social networks, social interactions, and activity-travel behavior: a framework for microsimulation. Environ. Plan. 35, 1012–1027 (2008)

Aurenhammer F.: Voronoi Diagrams – A Survey of a Fundamental Geometric Data Structure. ACM Computing Surveys, 23(3):345–405 (1991).

Autoritat del Transport Metropolità (ATM): Enquesta de Mobilitat en Día Feiner (2009). http://doc.atm.cat/ca/_dir_emef/emef2009/files/assets/basic-html/page-1.html

Avery, C. L.; Mills, K. T.; Williams, R.; McGraw, K. A.; Poole, C.; Smith, R. L.; Whitsel, E. A. Estimating error in using ambient PM2.5 concentrations as proxies for personal exposures: a review. Epidemiology 2010, 21, 215−223 (2010).

Axhausen, K.W.: Social networks and travel: some hypotheses. In: Donaghy, K.P., Poppelreuter, S., Rudinger, G. (eds.) Social Aspects of Sustainable Transport: Transatlantic Perspectives, pp. 90–108. Ashgate, Aldershot (2005)

Bagrow, J.P., Lin, Y.-R.: Mesoscopic structure and social aspects of human mobility. PLoS One 7(5), 1–11 (2012)

Bar-Gera, H.: Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: a case study from israel. Transp. Res. Part C 15(2007), 380–391 (2007)

Bassolas A, Lenormand M, Tugores A, Gonçalves B & Ramasco JJ: Touristic site attractiveness seen through Twitter. EPJ Data Science 5, 12. (2016)

Bayir, M.A., Demirbas, M., Eagle, N.: Mobility profiler: a framework for discovering mobility profiles of cell phone users. Pervasive Mobile Comput. 6, 435–454 (2010)

Becker, R.A., Cáceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C.: A tale of one city: using cellular network data for urban planning. Pervasive Comput. IEEE 10(4), 18–26 (2011)

Beckx C, Int Panis L, Arentze T, Janssens D, Torfs R, Broekx S, Wets G.: A dynamic activity-based population modelling approach to evaluate exposure to air pollution: methods and application to a Dutch urban area. Environ Impact Assess Rev. 2009;29:179–85 (2009).

Bengtsson L, Lu X, Thorson A, Garfield R, von Schreeb J.: Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. PLoS Med. 2011, 8 (8): e1001083-10.1371/journal.pmed.100108 (2011).

Boldo, E., Linares, C., Lumbreras, J., Borge, R., Narros, A., García-Pérez, J., Fernández-Navarro, P., Pérez-Gómez, B., Aragonés, N., Ramis, R., Pollán, M., Moreno, T., Karanasiou, A., López-Abente, G.: Health impact assessment of a reduction in ambient PM2.5 levels in Spain. Environment International 37, 342-348 (2011).

Briggs, D., Fecht, D., & De Hoogh, K. (2007). Census data issues for epidemiology and health risk assessment: experiences from the Small Area Health Statistics Unit. Journal of the Royal Statistical Society: Series A (Statistics in Society), 170(2), 355-378.

Brunekreef B, Hoek G, Schouten L, Bausch-goldbohm S, Fischer P, Armstrong B, Hughes E, Jerrett M, Brandt P Van Den. Effects of long-term exposure on respiratory and cardiovascular mortality in the Netherlands: the NLCS-AIR study. 2009. http://www.n65.nl/NCLS-AIR-Study-2009.pdf. Accessed 5 June 2015. (2015)

Burke, J. M.; Zufall, M.; Özkaynak, H. A population exposure model for particulate matter: case study results for PM2.5 in Philadelphia, PA. J. Exposure Anal. Environ. Epidemiol. 2001, 11, 470−489 (2001).

Caceres, N., Wideberg, J.P., Benitez, F.G.: Deriving origin–destination data from a mobile phone network. IET Intell. Transp. Syst. 1(1), 5–26 (2007)

Caceres, N., Wideberg, J.P., Benitez, F.G.: Review of traffic data estimations extracted from cellular networks. IET Intell. Transp. Syst. 2(3), 179–192 (2008)

Caceres, N., Romero, L.M., Benitez, F.G., Castillo, J.M.D.: Traffic flow estimation models using cellular phone data. IEEE Trans. Intell. Transp. Syst. 13(3), 1430–1441 (2012)

Calabrese, F., Pereira, F. C., Lorenzo, G. D., Liu, L., Ratti, C.: The geography of taste: analyzing cell-phone mobility and social events. In: Proceedings of IEEE International Conference on Pervasive Computing (2010)

Calabrese, F., Lorenzo, G.D., Liu, L., Ratti, C.: Estimating origin-destination flows using mobile phone location data. Pervasive Comput. IEEE 10(4), 36–44 (2011a)

Calabrese, F., Smoreda, Z., Blondel, V.D., Ratti, C.: Interplay between telecommunications and face-to-face interactions: a study using mobile phone data. PLoS One 6(7), e20814 (2011b). doi:10.1371/journal.pone.0020814

Calabrese F., Colonna M., Lovisolo P., Parata D., and Ratti C.: "Real-time urban monitoring using cell phones: A case study in Rome," IEEE Trans. Intell. Transp. Syst., vol. 12, no. 1, pp. 141–151(2011c).

Calabrese, F., Diao, M., Lorenzo, G.D., Ferreira Jr., J., Ratti, C.: Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transp. Res. Part C 26, 301–313, (2013).

Carrasco, J.A., Miller, E.J.: Exploring the propensity to perform social activities: social networks approach. Transportation 33, 463–480 (2006)

Carrasco, J.A., Miller, E.J., Wellman, B.: How far and with whom do people socialize? Empirical evidence about the distance between social network members. Transp. Res. Rec. 2076, 114–122 (2008a)

Carrasco, J.A., Hogan, B., Wellman, B., Miller, E.J.: Collecting social network data to study social activitytravel behaviour: an egocentric approach. Environ. Plan. B 35(6), 961–980 (2008b)

Carrasco, J.A., Hogan B., Wellman B., Miller E. J.: Agency in social activity and ICT interactions: The role of social networks in time and space, Tijdschrift voor Economische en Sociale Geografie (J. Eco. Soc. Geogr.), 99(5), 562–583 (2008c)

Carrasco, J.A., Miller, E.J.: The social dimension in action: a multilevel, personal networks model of social activity frequency. Transp. Res. Part A 43(1), 90–104 (2009)

Cesaroni, G., Baldoni, C., Gariazzo, C., Stafoggia, M., Sozzi, R., Davoli, M., Forastiere, F.: Long term exposure to urban air pollution and mortality in a cohort of more than a million adults in Rome. Environ. Health Perspect. 121 (3), 324e331.(2013)

Chen, C., Mei, Y.: Does distance still matter in facilitating social ties? The roles of mobility patterns and the built environment. Presented at 93rd TRB annual meeting (2014)

Chen, C., Ma, J., Susilo, Y., Liu, Y. & Wang, M.: The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation Research Part C: Emerging Technologies, 68, 285-299. (2016)

Cho E., Myers S.A., Leskovek J.: Friendship and mobility: user movement in location-based social networks. In: KDD '11 Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1082–1090 (2011)

Çolak, S.; Alexander, L.P.; Alvim, B.G.; Mehndiratta, S.R.; González, M.C.: Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities ,Transportation Research Record: Journal of the Transportation Research Board,,2526,126-135,2015,Transportation Research Board of the National Academies (2015)

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., ... & Tatem, A. J.: Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences, 111(45), 15888-15893 (2014).

Dewulf, B.; Neutens, T.; Lefebvre, W.; Seynaeve, G.; Vanpoucke, C.; Beckx, C.; Weghe, N.V. Dynamic assessment of exposure to air pollution using mobile phone data. Int. J. Health Geogr. 2016, 15, 1–14 (2016).

DG REGIO: Cities of tomorrow - Challenges, visions, ways forward (2010).

Do T., Gatica-Perez D.: Contextual conditional models for smartphone-based human mobility prediction. In: Proceedings ACM International Conference on Ubiquitous Computing, Pittsburgh, Sept (2012)

Dons, E.; Int Panis, L.; Van Poppel, M.; Theunis, J.; Willems, H.; Torfs, R.; Wets, G.: Impact of time-activity patterns on personal exposure to black carbon. Atmos. Environ. 2011, 45, 3594−3602 (2011).

Douglass, R. W., Meyer, D. A., Ram, M., Rideout, D., & Song, D.: High resolution population estimates from telecommunications data. EPJ Data Science, 4(1), 4 (2015).

Doyle, J., Hung, P., Kelly, D., Mcloone, S., Farrell, R.: Utilising mobile phone billing records for travel mode discovery. ISSC 2011, Trinity College Dublin, June (2011)

Dugundji, E., Walker, J.: Discrete choice with social and spatial network interdependencies: an empirical example using mixed GEV models with field and ‘‘panel’’ effects. Transp. Res. Rec. 1921, 70–78 (2005)

Eagle, N., Pentland, A., Lazer, D.: Inferring social network structure using mobile phone data. Proc. Natl. Acad. Sci. (PNAS) 106(36), 15274–15278 (2009)

Eagle N, Macy M and Claxton R (2010). Network diversity and economic development. Science 328, 1029 (2010)

European Commission (EC): Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the regions “Action Plan on Urban Mobility” (COM(2009) 490 final)(2009).

European Commission (EC): White Paper. Roadmap to a Single European Transport Area – Towards a competitive and resource efficient transport system.(COM(2011) 144 final)(2011).

Eurostat: Feasibility study of the use of mobile positioning data for tourism statistics' Consolidated Report Eurostat Contract No 30501.2012.001–2012.452 (2014)

Foth M.: Handbook of Research on Urban Informatics: The Practice and Promise of the RealTime City. IGI Publishing (2008).

García-Palomares J.C., Gutiérrez J. and Mínguez C.: Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. In Applied Geography, 63: 408-417 (2015).

Gariazzo, C., Pelliccioni, A., Bolignano, A.: A dynamic urban air pollution population exposure assessment study using model and population density data derived by mobile phone traffic. Atmos. Environ. 131, 289–300 (2016)

Gonzalez M., Hidalgo C., and Barabasi A.-L.: Understanding individual human mobility patterns, Nature, vol. 453, no. 7196, pp.779–782 (2008).

Hackney, Jeremy K., Kay W. Axhausen: An agent model of social network and travel behavior interdependence. Paper presented at the 11th international conference on Travel Behaviour Research, Kyoto, Aug (2006)

Hackney, J., Marchal, F.: A model for coupling multi-agent social interactions and traffic simulation, in: TRB 2009 annual meeting (2009)

Habib, K.N., Carrasco, J.A.: Investigating the role of social networks in start time and duration of activities: a trivariate simultaneous econometric model. Transportation Research Record: Journal of the Transportation Research Board 2230, 1–8 (2011)

Hariharan, R., Toyama, K.: Project lachesis: parsing and modeling location histories. Geogr. Inform. Sci., 106–124 (2004).

Hatzopoulou, M.; Miller, E. J. Linking an activity-based travel demand model with traffic emission and dispersion models: Transport's contribution to air pollution in Toronto. Transport. Res.D: Tr. E 2010, 15, 315−325 (2010).

Health Effects Institute: Traffic-related Air Pollution: a Critical Review of the Literature on Emissions, Exposure, and Health Effects. Special Report #17, 2009-05-04 (2009).

Holleczek T., Yu L., Kang Lee J., Senn O., Ratti C. and Jaillet P.: Detecting weak public transport connections fromcellphone and public transport data, *Proceedings of the 2014 International Conference on Big Data Science and Computing* (2014).

Huynh M, Woodruff TJ, Parker JD, Schoendorf KC.: Relationships between air pollution and preterm birth in California. Pediatr Perinat Epidemiol. 2006;20:454–61 (2006).

Iovan, C., Olteanu-Raimond, A.-M., Couronné, T., Smoreda, Z.: Moving and calling: mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies. Geogr. Inform. Sci. Heart Europe Lect. Notes Geoinform. Cartogr., 247–265 (2013).

Isaacman, S.,Becker, R., Caceres, R., Kobourov, S., Martonosi, M., Rowland, J., Varshavsky, A.: Identifying important places in people's lives from cellular network data. In: Procedings International Conference on Pervasive Computing, San Francisco, June (2011)

Jiang, S., Fiore, G. A., Yang, Y., Ferreira Jr, J., Frazzoli, E., & González, M. C. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing (p. 2). ACM.(2013)

Kontokosta, C. E., & Johnson, N.: Urban phenology: Toward a real-time census of the city using Wi-Fi data. Computers, Environment and Urban Systems, 64, 144-153 (2017).

Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T.: A survey of mobile phone sensing. Commun. Mag. IEEE 48(9), 140–150 (2010)

Lenormand M, Picornell M, Garcia Cantú O, Tugores A, Louail T, Herranz R, Barthelemy M, Frías-Martínez E & Ramasco JJ: Cross-checking different sources of mobility information. PLoS ONE 9, e105184 (2014).

Le Tertre, A.; Quenel, P.; Eilstein, D.; Medina, S.; Prouvost, H.; Pascal, L.; Boumghar, A.; Saviuc, P.; Zeghnoun, A.; Filleul, L.; Declercq, C.; Cassadou, S.; Le Goaster, C. Short-term effects of air pollution on mortality in nine French cities: a quantitative summary. Arch. Environ. Health 2002, 57, 311−319 (2002).

Marchal, F., Nagel, K.: Allowed cooperative agents in a microsimulation to share information with each other about activity locations and about other agents, in order to optimize trip chains (2006)

Molin, E.J.E., Arentze, T.A., Timmermans, H.J.P.: Social activities and travel demands : a model-based analysis of social-network data. Transp. Res. Rec. 2082, 168–175 (2007)

Moore, J., Carrasco, J.A., Tudela, A.: Exploring the links between personal networks, time use, and the spatial distribution of social contacts. Transportation 40(4), 773–788 (2013)

Munizaga, M., Palma, C.: Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C* 24 (2012) 9

Munizaga, M., Devillaine, F., Navarrete, C. and Silva, D., "Validating travel behavior estimated from smartcard data," Transp. Res. C, Emerg. Technol., vol. 44, pp. 70–79, Jul. 2014

Naini, F. M., Dousse, O., Thiran, P., & Vetterli, M.: Population size estimation using a few individuals as agents. In Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on (pp. 2499-2503). IEEE.(2011)

Nyhan, M., Grauwin, S., Britter, R., Laden, F., McNabola, A., Misstear, B., Ratti, C.,: 'Exposure Track' - The Impact of Mobile Device Based Mobility Patterns on Quantifying Population Exposure to Air Pollution. Environmental Science & Technology (2016)

Oliver N., Matic A. and Frías-Martínez, E.: Mobile network data for public health: opportunies and challenges, Frontiers in Public Health, (2015).

Onnela JP, Arbesman S, González MC, Barabási AL, Christakis NA: Geographic constraints on social network groups. PLoS One 6(4):e16939 (2011)

Omori, T.; Fujimoto, G.; Yoshimura, I.; Nitta, H.; Ono, M.: Effects of particulate matter on daily mortality in 13 Japanese cities. J. Epidemiol. 2003, 13, 314−322 (2003).

Ortuzar, J.D., Willumsen L.G.: Modelling transport (2011).

Oyabu, Y., Terada, M., Yamaguchi, T., Iwasawa, S., Hagiwara, J., and Koizumi, D.: Evaluating Reliability of Mobile Spatial Statistics. Docomo Technical Journal, 14(3), 16–23. (2013).

Páez, A., Scott, D.M.: Social influence on travel behavior: a simulation example of the decision to telecommute. Environ. Plan. A 39(3), 647–665 (2007)

Panis, L.I.: New directions: air pollution epidemiology can benefit from activity-based models. Atmos. Environ. 44, 1003e1004 (2010).

Phithakkitnukoon, S., Calabrese, F., Smoreda, Z., Ratti, C.: Out of sight out of mind: how our mobile social network changes during migration. Proceedings of the IEEE International Conference on Social Computing, pp. 515–520. Cambridge University Press, Cambridge (2011)

Phithakkitnukoon, S., Smoreda, Z., Olivier, P.: Socio-geography of human mobility: a study using longitudinal mobile phone data. PLoS One 7(6), e39253 (2012). doi:10.1371/journal.pone.0039253

Picornell, M., Ruiz, T., Lenormand, M., Ramasco, J.J. , Dubernet, T. and Frías-Martínez, E.: Exploring the potential of phone call data to characterize the relationship between social network and travel behaviour. Transportation 42, 647-668.(2015)

Picornell, M. and Willumsen, L.: "Transport Models and Big Data Fusion: Lessons from Experience", Proceedings of the European Transport Conference (2016).

Ratti C, Pulselli RM, Williams S, Frenchman D: Mobile landscapes: using location data from cell phones for urban analysis. Environment and Planning B 33: 727.C. (2006)

Reades, J., Calabrese, F., Sevtsuk, A., & Ratti, C.: Cellular census: Explorations in urban data collection. IEEE Pervasive Computing, 6(3). (2007)

Ricciato F., Widhalm P., Craglia M. and Pantisano F.: Estimating population density distribution from network-based mobile phone data (2015).

Ronald, N.A., Arentze, T.A., Timmermans, H.J.P.: Modeling social interactions between individuals for joint activity scheduling. Transp. Res. Part B 46, 276–290 (2012a)

Ronald, N.A., Dignum, V., Jonker, C., Arentze, T.A., Timmermans, H.J.P.: On the engineering of agentbased simulations of social activities with social networks. Inf. Softw. Technol. 54(6), 625–638 (2012b)

Rose, G.: Mobile phones as traffic probes: practices, prospects and issues. Transp. Rev. 26(3), 275–291 (2006)

Schwartz, J. The effects of particulate air pollution on daily deaths: A multi-city case-crossover analysis. Occup. Environ. Med. 2004, 61, 956−961 (2004).

Sharmeen, F., Arentze, T., Timmermans, H.: A multilevel path analysis of social network dynamics and the mutual interdependencies between face-to-face and ICT modes of social interaction in the context of life-cycle events. In: Roorda, M.J., Miller, E.J. (eds.) Travel Behaviour Research: Current Foundations, Future Prospects, pp. 411–432. Lulu Press, Toronto (2013)

Sharmeen, F., Arentze, T.A., Timmermans, H.J.P.: Dynamics of face-to-face social interaction frequency: role of accessibility, urbanization, changes in geographical distance and path dependence. J. Transp. Geogr. 34, 211–220 (2014)

Schneider C. M., Belik V., Couronne T., Smoreda Z., and Gonzalez M. C.: Unravelling Daily Human Mobility Motifs."Journal of The Royal Society Interface 10, no. 84 (April 24, 2013): 20130246–20130246 (2013).

Silm, S., Ahas, R.: The seasonal variability of population in estonian municipalities. Environ. Plan. A 42, 2527–2546 (2010)

Silvis, J., Niemeier, D., D'Souza, R.: Social networks and travel behavior: report from an integrated travel diary. Paper presented at the 11th international conference on Travel Behaviour Research, Kyoto, Aug (2006)

Sohn, K., Kim, D.: Dynamic origin–destination flow estimation using cellular communication system. IEEE Trans. Veh. Technol. 57(5), 2703–2713 (2008)

Song, C., Qu, Z., Blumm, N., Barabasi, A.-L.: Limits of predictability in human mobility. Science 327(5968), 1018–1021 (2010a)

Song, C., Koren, T., Wang, P., Barabasi, A.-L.: Modelling the scaling properties of human mobility. Nat. Phys. 6(2010), 818–823 (2010b)

Soto V, Frias-Martinez V, Virseda J and Frias-Martinez E.: Prediction of Socioeconomic Levels using Cell Phone Records. DOI: 10.1007/978-3-642-22362-4_35 (2011).

Steenbruggen, J., Borzacchiello, M.T., Nijkamp, P., Scholten, H.: Mobile phone data from gsm networks for traffic parameter and urban spatial pattern assessment: A review of applications and opportunities. GeoJournal 78, 223–243 (2011). doi:10.1007/s10708-011-9413-y

Sterly,H.; Hennig, B; Dongo, K.: "Calling Abidjan" – Improving Population Estimations with Mobile Communication Data (IPEMCODA). Retrieved from ResearchGate.net (2013)

Terada, M., Nagata, T., & Kobayashi, M.: Population estimation technology for mobile spatial statistics. NTT DOCOMO Techn. J, 14, 10-15 (2013).

United Nations, Department of Economic and Social Affairs, Population Division: World Urbanisation Prospects : The 2009 Revision (2010).

Van den Berg, P., Arentze, T., Timmermans, H.J.P.: A path analysis of social networks, telecommunication and social activity–travel patterns. Transp. Res. Part C 26(2013), 256–268 (2013)

Van Londersele, B., Delafontaine, M., & Van de Weghe, N.: Bluetooth Tracking. GIM International. 23-25. (2009).

Versichele, M., Neutens, T., Delafontaine, M. & Van de Weghe, N.:The use of Bluetooth for analyzing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. Applied Geography, 32(2), 208-220 (2012).

Wang, H., Calabrese, F., Lorenzo, G. D., Ratti, C.: Transportation mode inference from anonymized and aggregated mobile phone call detail records. In: 13th international IEEE annual conference on intelligent transportation systems, 318–323 (2010)

Wang, W., Attanucci, J., & Wilson, N.H.: "Study of Bus Passenger Origin Destination and Travel Behavior Using Automated Data Collection Systems in London". *90th TRB Annual Meeting*, Washington, D.C. (2011).

White, J. and Wells, I.: Extracting origin destination information from mobile phone data. Road transport information and Control, 19–21 Mar (2002)

Widhalm P., Yang Y., González M., Ulm M. and Athavale S.: Discovering urban activity patterns in cell phone data, Transportation, vol. 42, no. 4, pp. 597-623 (2015).

WHO: Global Urban Ambient Air Pollution Database (2016).

Yim, Y.: The state of cellular probes. California PATH Working Paper, UCB-ITS-PRR-2003-25 (2003)

Yin, J.; Gao, Y.; Du, Z.; Wang, S.: Exploring multi-scale spatiotemporal twitter user mobility patterns with a visual-analytics approach. ISPRS Int. J. Geo-Inf. 2016, 5 (2016).

Ythier, J., Walker, J.L., Bierlaire, M.: The influence of social contacts and communication use on travel behavior: a smartphone-based study. In: Transportation Research Board annual meeting (2013)