D4.1 Analysis of mobility patterns: interdependence of mobility with city structure and social networks

Project acronym : EUNOIA
Project title : Evolutive User-centric Networks for Intraurban Accessibility
Grant Agreement number : 318367
Funding Scheme : Collaborative project
Project start date / Duration: 01 Oct 2012 / 24 months
Call Topic : FP7-ICT-2011-8
Project web-site : http://www.eunoia-project.eu/
Deliverable : D4.1 Analysis of mobility patterns: interdependence of mobility with city structure and social networks
Edition 1
Date 25/11/2014
Status Final
Dissemination Level : Public
Authoring and Approval

Prepared by
Thomas Louail (CEA-IPhT - Researcher)
Oliva García-Cantú (Nommon – Researcher)
Maxime Lenormand (IFISC – Researcher)
Miguel Picornell (Nommon – Researcher)
Oliver O’Brien (CASA – Researcher)

Reviewed by
Marc Barthélemy (CEA-IPhT - Principal Researcher)
José Ramasco (Deputy Scientific Coordinator)
Ricardo Herranz (Deputy Management Coordinator)

Approved by
Maxi San Miguel (Project Coordinator)

Record of Revisions

<table>
<thead>
<tr>
<th>Edition</th>
<th>Date</th>
<th>Description/Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edition 1 Draft 1</td>
<td>30/09/2013</td>
<td>Initial version</td>
</tr>
<tr>
<td>Edition 1 Draft 2</td>
<td>25/10/2013</td>
<td>Second draft with:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Addition of the sections corresponding to the analysis of the functional networks of cities, the OD matrices extraction algorithms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Addition of a preliminary and global analysis of London OD matrix to illustrate what will be done with the OD matrices of Spanish metro areas during the second year.</td>
</tr>
<tr>
<td>Edition 1 Draft 3</td>
<td>30/10/2013</td>
<td>Addition of the visualisations of flows in London (section 4.3)</td>
</tr>
<tr>
<td>Edition 1 Draft 4</td>
<td>11/11/2013</td>
<td>Integrates the comments and corrections proposed by RH, JJR and MP</td>
</tr>
<tr>
<td>Edition 1 Draft 5</td>
<td>24/11/2014</td>
<td>Integration of results from Year 2</td>
</tr>
<tr>
<td>Edition 1</td>
<td>25/11/2014</td>
<td>Minor editorial corrections</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approved for submission to the EC</td>
</tr>
</tbody>
</table>
# Table of Contents

1. INTRODUCTION ................................................................................................................................. 1
   1.1. ORGANISATION OF THE DOCUMENT ......................................................................................... 2

2. FROM MOBILE PHONE DATA TO THE SPATIAL STRUCTURE OF CITIES ........................................... 4
   2.1. GENERAL FEATURES ................................................................................................................ 5
   2.2. GLOBAL ANALYSIS .................................................................................................................. 7
   2.3. HOTSPOTS ANALYSIS .............................................................................................................. 8
   2.4. MODELING THE POLYCENTRIC TRANSITION OF CITIES ....................................................... 11

3. COMPARING AND MODELLING LAND USE ORGANISATION IN CITIES ............................................... 12

4. ANALYSIS OF SOCIAL NETWORKS AND JOINT TRIPS .................................................................... 15
   4.1. USE OF SOCIAL NETWORKS DURING TRAVEL TIME .......................................................... 15
   4.2. SOCIAL NETWORK AND TRAVEL BEHAVIOUR ................................................................. 18

5. VISUALISATION AND MAPPING OF INDIVIDUAL MOBILITY PATTERNS .................................... 22
   5.1. STATION AND STOP LOCATIONS AND IDENTIFIERS .............................................................. 22
   5.2. VEHICLE COUNTS .................................................................................................................. 23
   5.3. INDIVIDUAL BICYCLE SHARING SYSTEM JOURNEYS BY USER ........................................ 24
   5.4. METRO JOURNEY FLOWS ..................................................................................................... 25

6. PERSPECTIVES .................................................................................................................................... 27

ANNEX I. REFERENCES .............................................................................................................................. 28

ANNEX II. RESEARCH PAPERS PUBLISHED IN THE FRAME OF WP4 .................................................... 29

ANNEX III. ABBREVIATIONS AND ACRONYMS .................................................................................... 30

ANNEX IV. ANALYSIS OF URBAN ACTIVITY LEVELS FROM PHONE DATA ......................................... 31
   IV.1 DESCRIPTION OF INPUT DATA ................................................................................................. 31
   IV.2 URBAN AREAS UNDER STUDY ............................................................................................... 31
   IV.3 DESCRIPTION OF DATA PROCESSING .................................................................................... 32

ANNEX V. METHODS: MOBILE PHONE DATA DESCRIPTION AND DETERMINATION OF HOTSPOTS .... 33
   V.1 PHONE DATA – GENERALITIES ................................................................................................. 33
   V.2 IDENTIFICATION OF HOTSPOTS .............................................................................................. 34
   V.3 INFLUENCE OF THE SPATIAL SCALE OF AGGREGATION ...................................................... 37
1. **Introduction**

EUNOIA Work Package 4 ‘Data Analysis’ has been dedicated to the analysis of large sets of individual human mobility data in several European cities. As introduced in our position paper (EUNOIA, 2012), the data analysis pursued several goals:

- Characterising mobility patterns in different types of cities, i.e. with different geographical, structural, political, institutional, socio-economic, and cultural characteristics;
- Uncovering the relations between the spatial organisation of activity centres and land-uses in the city, and the *structure* of mobility in the city;
- Understanding the interplay between the structure of the individuals’ social networks and social characteristics, and their mobility patterns.

Another important aspect was to mine these mobility data in order to discern general and local features of urban mobility patterns, such as the existence and the organisation of activity centres in cities. To extract useful knowledge from the large data sets at our disposal, our objective was to make use of the standard statistical analysis and data mining methods, as well as new spatial analysis methods recently developed in the field of complex networks and spatial networks (Barthélemy, 2011). We also considered, if necessary, to try to improve existing methods for extracting useful, coarse-grained information from large amounts of spatial data. And finally, to propose simple theoretical models of cities, that would be reasonable candidates to explain some of the patterns of the dynamical spatial organisation of cities.

This document and document D4.2 ‘Typology of urban mobility patterns’ present the main results of the work we have performed to fulfil these objectives.

In WP4 we first focused on the analysis of aggregated locational data generated by mobile phones in cities. When records of individual’s mobile phone activity are aggregated for each area of the city, these counts tell us how many individuals are communicating in each area of the city, at successive moments of the day. They thus allow designing new measures of the dynamical, spatial organisation the city, notably the identification of activity *hotspots*. These aggregated, location data also permit to automatically detect land uses, by identifying areas of the city where the ‘rhythms’ of communication activity are correlated in time. Also, from a methodological point of view, it is interesting to precisely analyse the usage of ICT devices in situation of mobility, notably to assess to what extent geolocated data produced by ICT devices provide precise looks on the trajectories of individuals. To that extent we studied the coverage of road networks in 20 European countries by mining a vast database of tweets collected during one year and sent by individuals in situation of mobility. Finally, during the project’s first year, we also started the analysis of the mobility patterns, first by proposing visualisations of mobility trips in London, by making use of data provided by London’s bike sharing system. The aforementioned analyses and their main results are discussed in the present document.

In the project’s second year we more specifically focused on mobility patterns and on the structure of mobility in cities, making use of the multiple sources of individual mobility data collected in WP3. The most important and regular type/motivation of mobility inside urban areas is commuting (journey to work). To study the organisation of mobility at the city scale we then decided to focus on commuting flows in cities. Commuting flows come as OD (Origin-Destination) matrices, a classic tool of urban transport planning. The OD matrix
describes a network, directed and weighted, and in the general case this mobility network is time-dependent. These matrices are usually extremely difficult and costly to measure, and it is only recently, thanks to technological advances such as the GPS or the democratisation of ICT devices, that we can reconstruct OD matrices from the geolocated data produced by the individuals’ devices. The OD matrix encapsulates the complete information on the commuting flows in the city. A first important analysis is then to ensure that different OD matrices constructed from different mobility data sources provide indeed the same information on the structure of mobility in the city, whether they are the result of surveys conducted by Local Authorities and Census Offices, or the result of computational methods filtering huge databases of passive data produced by individuals’ mobile phones, tweets, or transport cards. Secondly, the OD matrix contains the complete information on the flows in a city, and for that reason it fails to immediately provide clear, synthetic and useful information of the structure of mobility in the city. One of our goals was then to propose methods to extract synthetic and stylised facts that would help to better understand the collective mobility patterns observed at the city scale. We proposed a method and we applied it to thirty cities in Spain, and also in London, in order to compare and classify these cities on the basis of their commuting structure. During the project’s second year, we also had access to an additional source of individual data, credit cards data. In addition to the time and space information associated to each transaction of a given customer, these data also inform us on demographic and social characteristics of the owner of the card. Thus, in a statistical approach, they allow to assess the influence of demographic and social characteristics on mobility in the city, if any. The methods, measures, results and proposed typology related to mobility patterns in cities are gathered in document D4.2, ‘Typology of mobility patterns’.

1.1. Organisation of the document

As explained in the previous paragraph, the results of WP4 are disseminated in the two documents D4.1 and D4.2. In this document we present the work performed to make use of urban mobility data to highlight properties of the city itself, including the number, location and organisation of hotspots in the city, and the detection and characterisation of land-use patterns in the city. The two main sources of data used in the work presented in this document were (a) individual, anonymous mobile phone records provided by a major Spanish mobile phone operator; and (b) geolocated tweets at the European scale collected during an entire year. These data have allowed us to address questions that could not be addressed with traditional data sources, resulting from transport surveys.

The data give us snapshots of the spatial distribution of individuals in the city at successive points in time, allowing for information about the spatial distribution of activities at different moments of the day. These digital traces illustrate various properties of the city, and the goal of our analyses was to extract relevant information on the dynamical organisation of cities, resulting from the movements of individuals. In particular, we challenged the following issues:

- The determination of hotspots in cities: can we find general and robust methods to identify the most important locations in a city at a given time? How are these activity centres organised in the city space? Do these centres change during the day? Is their spatial organisation the same from one city to another, or can we observe different classes of cities?
The classification of cities forms: can we use these vast amounts of data to design urban morphological indicators that capture the polycentric degree of a city? Can we propose relevant and useful classifications of city forms based upon these morphological indicators?

The uncovering of the functional networks in the city: can we use the individuals’ communication traces to automatically detect various kinds of land-use in a city? How many different land-uses can be identified? Does this number depend on population size of the city? What is the spatial organisation of these land-uses in the city space? Can we identify common patterns in cities?

In the following sections we propose quantitative answers to these questions. In the next section we explain how we made use of spatially aggregated mobile phone data to design new morphological indicators on cities, and a new general method to detect activity *hotspots* in the city. We studied their number, the statistical distribution of their sizes, their temporal persistency, and their spatial organization in the thirty largest Spanish cities. Then in Section 3 we present our work related to the characterisation and the modelling of land uses in cities, by using methods issued from the field of complex networks. In section 4 'Analysis of social networks and join trips', we analyse the use of Internet social networks while travelling and the social network influence on travel behaviour. We then underline the complementary point of view provided by mobility measures, and we first show in section 5 how visualizations of trips that make use of new mobility data (in this case data from London's bikes sharing system) reveal core-periphery structures, basins of attraction, of favoured pathways in the urban space. These new visualisations of trips in the urban space provide a transition to the second part of the WP4 report, D4.2 'Typology of urban mobility patterns'.
2. From mobile phone data to the spatial structure of cities

The quantitative study of the morphological properties of cities has been for long a classic research subject in quantitative geography and spatial economy. Questions of interest include the characterization and the comparison of cities through their density landscape, their spatial extension, the clustering degree of their activity centres, etc. Until the late 2000’s, these quantitative comparisons of urban forms essentially relied on transport and mobility surveys, and also on remote sensing data. These two data sources give an estimation of the density of individuals and land uses in the city at a rather precise spatial scale, but they offer a ‘static information’ that is subject to many bias. The interviewee is typically asked about the trips she/he made the day before the interview. Additionally data resulting from surveys also have a limited temporal resolution – typically covering one single weekday - and they are costly to get. Early studies in urban geography had developed the idea of measuring the evolving density of individuals at various hours of the day and in different parts of the city, in order to measure morphological and socio-economic evolution of cities during a typical, average weekday. Additionally, traffic surveys in cities worldwide have long provided a general knowledge of the typical timing of urban mobility. However, given the nature of the collected data, their quantity, their resolution, and their cost, these studies could not properly investigate some interesting questions related to the dynamical properties of the spatial structure of cities. Such questions include:

- How much does a city shape change through the course of one day?
- Where are located the city’s hotspots? Do these hotspots appear and disappear depending on the hour of the day, or are they always the same?
- How are these hotspots spatially organised in the city space?
- Is the hierarchy and the spatial organization of hotspots robust through time?
- Is there some kind of typical distance(s) characterising the core of permanent hotspots, or ‘backbone’, of each city?

These questions are important to build quantitative typology of cities and their spatial structure, an important ingredient in the construction of a science of cities. Mobile phone data encapsulate the spatial information about individuals and how it evolves during the day. These datasets thus give us the opportunity to answer the aforementioned questions, and to characterize quantitatively the spatial structure of cities. In (Louail et al., 2014), we have addressed these questions using mobile phone data for a set of the 31 Spanish urban areas mapped on Figure 1.
Figure 1. The location and spatial extension of the 31 Spanish urban areas with more than 200,000 inhabitants in 2011. The set of cities analysed in this article includes very different types of cities such as central cities, port cities and cities on islands. These urban areas are very diverse in terms of geographical location, area, population size and density. (NB: the municipalities included in each urban area are those included in the AUDES database).

2.1. General features

When we trace the time evolution along the day of the number of individuals using their mobile phone, we see that the curve follows the same pattern in every city. Figure 2 shows the average number of mobile phone users per hour according to the day of the week for six of the Spanish cities we studied. We first see that globally the number of phone users is significantly higher during the weekdays than during the weekends, except at night time. From 11pm to 8am, the number of users is relatively low; it reaches a minimum at 5am during weekdays and at 7am during the weekend. For all cities we observe two activity peaks, one at 12am during weekdays (1pm during the weekend) and another one at 6pm during weekdays (and at 8pm during the weekend).
In order to properly compare the temporal patterns obtained in different cities, we can rescale the curves of different cities [cf. Figure 3(a)] by dividing the number of users at each hour by the total number of users in the city during the day. The resulting normalised curves are presented in Figure 3(b). These rescaled curves all ‘collapse’ on a single curve, and suggest the existence of a single ‘urban rhythm’ common to all cities. The data collapse is very good in the morning, while in the afternoon we observe a little more variability from one city to another. It is interesting to note that in four urban areas located in the western part of Andalusia (Sevilla, Granada, Cordoba and Jerez de la Frontera) the activity restarts later in the afternoon, around 5pm, one hour later than in the other cities.

**Figure 2.** Evolution of the number of users in six Spanish urban areas, according to the hour of the day, for each day of the week.

**Figure 3.** Time evolution of the number of mobile phone users per hour during an average weekday. (a) Total number of unique mobile phone users per hour (shown here for the eight biggest Spanish cities). (b) Rescaled numbers of unique users per hour for 31 cities. Each value U_{i(t)} is equal to the number of phone users in city i at time t, N_{i(t)}, divided by the total number of phone users in i during the entire day. The good collapse suggests the existence of an urban rhythm common to all cities.
2.1.1. Two complementary approaches to measure urban forms: global weighted indicators versus hotspots analysis

The aggregated mobile phone data give access to the local density \( \rho_i(t) \) of users at a location \( i \) and at a time \( t \). The difficulty is then to study this complex object that displays variation in time and space. We considered two main directions to tackle this problem. The first one is to define global indicators that are calculated by considering the entire city and by weighting each part of the city by the density of users. The second approach consists in identifying local maxima of the density function \( \rho_i(t) \), or in other words, the city’s hotspots. There are pros and cons in each method. Looking at hotspots is convenient since it provides a clear picture of the important locations in the city, but contains a part of arbitrariness in their determination. On the other hand, working with weighted indices does not require identifying hotspots but at the cost of producing results that are more difficult to interpret. These two approaches can however be seen as complementary since they highlight different properties of the city. Weighted indices inform us about the global properties of a given city, while the hotspots give us a more local look to focus on the 'heart' of the city. We successively applied the two methods.

2.2. Global analysis

We first measured the average weighted distance \( D_V(t) \) between individuals in the city at successive hours of the day. This information provides an interesting indicator about the evolving form of a city during the day. Figure 4(a) shows the evolution of the normalised average, weighted geographical distance between individuals during a typical weekday.

![Figure 4](image.png)

**Figure 4.** Time evolution of the average distance \( D_V(t) \) between phone users in the city, and the values of the dilatation index \( \mu = \frac{\max_t D_V(t)}{\min_t D_V(t)} \) for the 31 Spanish urban areas studied. (a) Illustration of the time evolution of \( D_V \) in three urban areas: Madrid, Sevilla and Zaragoza. This distance \( D_V(t) \) is equal to the average of the distances between each pair of cells weighted by the density of individuals in each of the cells. The resulting distance is then divided by the typical spatial size of the city (given by the square root of the city’s area) in order to compare the curves across cities. (b) Rank-size distribution of the dilatation index \( \mu \) in the 31 metropolitan areas.
For all cities we observe the same typical pattern [see Figure 4(a)]: the curve has two peaks, one around 7 am, when individuals switch on their mobile phones, probably at home or when they are in the transportation system. We then see a decrease of the distance $D_V(t)$ (the city ‘collapses’), displaying spatial concentration of individuals during the middle of the day, which corresponds to the activity period for most individuals (workers/students). During the afternoon we see a second, smaller peak dispersed over 4-5pm, when people start going back home. This afternoon peak is less pronounced, suggesting a higher variety of mobility behaviours at the end of the day. The interesting feature of these curves is the variation of amplitude that informs us about the importance of the ‘spatial collapse’ phenomenon. From this curve it is then natural to calculate for each city a ‘dilatation coefficient’ defined as $\mu = \frac{\max_{t} D_V(t)}{\min_{t} D_V(t)}$.

We show in Figure 4(b) the values of this dilatation coefficient for the 31 Spanish cities. We can roughly distinguish three groups of cities. For the first group with a value of $\mu$ around 1, the average distance stays approximately constant throughout the day. This means that whatever the hour of the day, the spatial spread of the high-density locations does not change significantly. High-density locations correspond to different activities depending on the moment of the day, and a small value of the dilatation coefficient implies that daytime activity places (workplaces, schools, leisure places) are neither more or less spatially concentrated than residences. Homes and activity places are more entangled, supporting the picture of more ‘mixed’ cities, such as Madrid for example. In the opposite case of large values, the spatial organization of the different high-density locations changes significantly along the day. A typical example would be a monocentric city where individuals are localized in the CBD during the day and where residences are spread all around the centre. In our set, Zaragoza for example is representative of this type of cities. For the intermediate group the cities display a less marked behaviour, probably resulting of a combination of monocentric and polycentric features.

### 2.3. Hotspots analysis

We have proposed a general method to determine hotspots among a distribution of densities (see Annex 4 for details). The important point is that once that we have a robust and parameter-free method to define hotspots, we can then determine the number of hotspots in each city. Using two well-defined criteria associated with our method, named ‘Average’ and ‘Loubar’ (see Annex 4 for details), for each city we have calculated the number of hotspots at each hour of the day, then calculated the average over the day. We can then see how these average numbers of hotspots scale with the population size of the cities. Figure 5 displays the number of hotspots versus the population size for the set of the 31 Spanish urban areas under analysis. The linear fit with bi-logarithmic scale on the axes indicates a power law relationship and the value of the slope ($< 1$) both confirms recent results of the literature (Louf et al., 2013), and highlights a clear sub-linear relation between the number of hotspots in a city and its population size. Practically speaking, that means that if one picks up two cities A and B such that A is twice larger than B, then the number of hotspots in A will likely be bigger than the number of hotspots in B, but less than two times bigger.

---

1 This measure is motivated by recent theoretical and empirical work (Louf, 2013) that has highlighted a clear sub-linear relation between the population size of cities and their number of activity centres (defined as employment hotspots). For
D4.1 Analysis of mobility patterns: interdependence of mobility with city structure and social networks

Edition 1

© EUNOIA Consortium

Page 9 of 38

Figure 5. Scatter plot and model fit line of $H$ the number of hotspots vs. $P$ the population size of the city, for the 31 Spanish cities studied. Each point in the scatterplot corresponds to the average of the number of hotspots determined for each one-hour time bin of the 7a.m. to 11p.m. time period, considered for the 5 weekdays. The linear relationship on a log-log plot indicates the power-law relationship between the two quantities, with an exponent value $\beta < 1$, indicating that the number of centres in this set of cities grows sub-linearly with the city’s population size.

Following the determination of hotspots in cities, another feature we inspected is the stability of the hotspots and the evolution of the importance of these hotspots in the city according to the hour of the day. This measure is related to the evolution of the hierarchy of places in the city. For each city we counted the number of one-hour time bins during which each city’s area has been a hotspot. We can then calculate the 'lifetime' of hotspots (measured in number of one-hour bins). The statistical distribution of the hotspots lifetimes in cities reveals the importance of 'permanent' hotspots, i.e. locations which are activity centres during the entire day. Each city has its number of important locations, those that form the 'heart' of the city. In addition to the permanent hotspots we also observe two other main groups: a set of intermediate hotspots (with lifetime of the order half a day) and 'intermittent' hotspots that are present only a few hours per day. The permanent hotspots are the most important locations in the city in terms of density of individuals. An interesting question is whether their rank (according to the density) is constant or changes during the day. Our results show that the heart of the cities is indeed very stables both in space and in time, whatever their population size - see (Louail et al., Sci rep., 2014) calculations and detailed discussion.

Finally another important question about hotspots in cities concerns their spatial organization. We calculated how distant these permanent hotspots are from each other in each of the 30 Spanish cities, when compared to the typical size of the city, given by the square root of the city's area. We show in Figure 6 the rank-plot of our 'compactness coefficient' defined as

$$ C_i = \frac{<D_{per}(t)>}{\sqrt{A_i}}. $$

the U.S., it has been shown that the number of activity centres $N_a$ (determined from employment data) scales as $N_a \propto P^\beta$ with $\beta \approx 0.64$. 
where $< D_{\text{per}(i)} >$ is the average distance between permanent hotspots in city $i$, and $\sqrt{A_i}$ is the square root of the area of city $i$. This indicator tells us how much the core of the city is sprawled all over the city's space. It is thus a measure of the compacity of the city core: for cities with values around 0, the permanent hotspots are very close to one another, when compared to the spatial extension of the urban area. On the contrary, a value close to one indicates that these always-crowded places are spread all over the whole city space. It is interesting to note in Figure 6(b) that the compacity of a city seems to increase with the population size. At least for a large subset of cities, we indeed observe this trend, which is consistent with the idea that the larger the city, the more spread are the hotspots (and the more polycentric it tends to be).

**Figure 6.** Different spatial structures of city cores. Rank plot of our compactness coefficient among the 31 urban areas. (b) Compacity coefficient versus population size. We observe a trend (at least for a large subset of cities, the corresponding fit is shown as a guide to the eye). (c) and (d) The spatial organization of the 1 km$^2$ permanent hotspots determined by the Loubar method (see Annex 4), in the urban areas of Bilbao (~950,000 inhabitants) and Vigo (~370,000 inhabitants). These figures reveal two types of spatial organization: polycentric in the case of Bilbao (c), whose permanent hotspots are not contiguous and more spread over the space of the urban area, and clearly compact and monocentric in the case of Vigo (d).
2.4. Modeling the polycentric transition of cities

The results presented in the previous section provide empirical evidence that many urban areas experience a spatial decentralisation as they grow and expand, and a transition toward a polycentric organisation of activity centres. In a parallel study held at CEA, we have proposed a stochastic, out-of-equilibrium model of city growth, which explains the appearance of subcentres as an effect of traffic congestion. We have shown that congestion triggers the instability of the monocentric regime, and that both the number of subcentres and the total commuting distance travelled by individuals scale sublinearly with the population size of the city. These predictions are in agreement with data gathered for around 9000 US cities between 1994 and 2010. We have then verified this prediction on another completely different dataset, by counting the activity hotspots in Spanish cities, determined from individuals’ mobile phone records, as discussed in the previous section. The description of the model and a complete discussion of the implications are presented in (Louf and Barthélemy, 2013).
3. Comparing and modelling land use organisation in cities

Land use patterns across urban environments play a major role in the experience that residents and visitors have of a city. They have an effect on the liveability of neighbourhoods and even on the health of the local residents. Land use and transportation display a well-established feedback relation. Transport demand depends on the location of residence and business areas, while the presence of new transport lines or facilities such as metro stations can substantially modify the land use mixing in a given area of the city. These ideas lie behind the development of the so-called Land Use Transport Interaction (LUTI) models, which are commonly employed in transport planning around the globe.

An important issue regarding land use refers to the methods employed to estimate it. City Hall registers, surveys or satellite images have been used in the past to this end. The emergence of geo-located ICT technologies introduces extra capabilities to directly measure the use that citizens make of each urban space. The information is exhaustive in terms of spatial and temporal resolution, allowing for the detection of concentrations of people second by second along days, weeks and months. As long as the data is geolocated, different sources are valid to estimate land use. Information from mobile phone call records, geolocated tweets or FourSquare has been considered in the literature. Different data sources have been compared, finding a consistent agreement among the estimations on human concentrations and mobility obtained from different ICT data, as well as with the estimations obtained with more traditional techniques.

Such wealth of information together with the ability to access massive data brought by the Internet era allows the systematic comparison of features across cities. This analysis can lead to the discovery and confirmation of properties that have been hypothesized to be common to all cities, and also to scaling laws providing insights into the way a property scales with city size. Some examples of these properties include number of patents filed, employment, GDP per capita, business diversity, consumption of resources, length of road networks, etc. The finding of these laws raises the hope of the existence of a coherent framework for city science.

In this work, we explore land use patterns in the five most populous urban areas of Spain. Land use information is obtained from mobile phone records using a new framework based on network theory (Figure 7). First, we divide each case study using a grid composed of square (500x500 m²). Then, we calculate for each cell an activity profile in terms of phone calls along time during the days of the week. A Pearson correlation coefficient is calculated between the activities of every pair of cells, obtaining a correlation matrix describing the level of similarity between activity profiles. Then the matrix formed by correlations over a threshold value is used to define an undirected weighted network, which is clustered using community detection techniques. It is interesting to note that in the five cities, between 98 and 100% of the cells are covered with only 4 groups:

1. **Residential**, which is characterized by low activities from 8am to 5-6pm. For the cells composing this group, the activity peaks around 7-8am and during the evening. In the weekend, the activity is almost constant;
2. **Business**, where the activity is significantly higher during the weekdays than during the weekends. Furthermore, it concentrates from 9am to 6-7pm;
3. **Logistics/Industry**, where, as for Business, the activity is higher during the weekdays. We observe a large peak between 5am and 7am followed by a smaller peak around 3pm. This cluster can be related to transport and distribution of goods;
4. **Nightlife**, which is characterized by high activity during the night hours (1am-4am), especially during the weekends. During the weekdays, these areas show higher activity between 9am and 6pm, as for the Business cluster, which may be hinting a certain level of mixing in the land use.

![Image](https://example.com/image.png)

**Figure 7.** Steps of the method to detect land use. We calculate for each cell an activity profile in terms of phone calls along time during the days of the week. A Pearson correlation matrix between cell activities is computed. Then the matrix formed by correlations over a threshold value is used to define an undirected weighted network, which is clustered using community detection techniques.

Once defined the clusters, we can study how the cells in each cluster are organized in the city’s space in order to perform a systematic comparisons of land use distribution across the five cities at different scales. We have define an entropy index to characterize the land use spatial organization. Let us consider a frame containing the full urban area, which is, in turn, sub-divided in a certain number $D^2$ of equal divisions. Each of these subdivisions, $B_i$, intersects the elementary cells so a certain fraction of area falls in each of the land use types: $f^R_i$ in the Residential cluster, $f^B_i$ in Business, $f^L_i$ in Logistics, and $f^{NL}_i$ in Nightlife. An entropy index, $E_i$, can be defined for $B_i$ as

$$E_i = - \sum_{a} f_i^a \ln f_i^a,$$

where $a$ runs over the four clusters. The entropy $E_i$ is then averaged over all the divisions to obtain a global metric for the city at a given scale $E(D)$. $E(D)$ tends to zero if the land use within the divisions becomes uniform, as occurs for instance at large $D$ (small spatial scales). On the other extreme, when $D \to 1$, $E(D)$ converges to a fixed value describing the full city. Figures 8(a) displays $E(D)$ for the 5 case studies, the curves are similar across cities, recalling the shape of scaling functions.

Urban models in the literature are typically built with elaborated mechanisms. However, the shape of $E(D)$ can be explained by a simple model inspired by Schelling’s segregation model. A land use type is initially assigned to each of the cells of a lattice, respecting the distribution of the four types found in each city in such a way that
$E(D=1)$ coincides with observations. A satisfaction index is defined for each cell considering its own type and that of its neighbours. Essentially, the satisfaction increases when a cell is surrounded by cells of the same type. Some land uses attract each other as, for instance, Residential and Business, while others repel as Residential and Logistics. A global satisfaction index is computed by summing all cells satisfaction. The model is updated choosing random pairs of cells and interchanging their land use if the exchange increases the global satisfaction level. This process is performed until the satisfaction reaches a stationary state. Calibrating a single parameter, which measures the tolerance of Residential and Business cells to neighbouring cells of other types except Logistics, we can reproduce the observed $E(D)$ in the real urban areas (see Figure 8(b)). We have included a null model in which the land use types are distributed at random, keeping the real proportions, to show that ignoring the interactions between the different land use types does not allow the reproduction of the curves obtained with the data.

![Figure 8](image_url)

Figure 8. Statistical properties of the land use patterns. (a) Average entropy index as function of the lateral number of divisions (inverse scale) $D$. (b) A model inspired in the Schelling’s segregation model has been introduced. The blue curve corresponds to the entropy as a function of $D$ averaged over all the cities; the red curve corresponds to our model results and the green curve is the outcome of a random null model. The red and green curves display the average over 30 realizations and the error bars with the minimum and maximum observed values.

To conclude, in this study we showed that a network approach can be fruitfully used to characterize urban land uses using mobile phone data. We identified four main land use patterns for five cities that differ in terms of shape, number of inhabitants and area. This analysis revealed some common spatial organisation patterns in the five cities. Finally, we proposed a model inspired in Schelling’s segregation able to explain and reproduce these results with simple interaction rules for land use types.
4. Analysis of social networks and joint trips

4.1. Use of social networks during travel time

An increasing number of geo-located data are generated everyday through mobile devices. This information allows for a better characterization of social interactions and human mobility patterns. Indeed, several data sets coming from different sources have been analysed during the last few years. Some examples include cell phone records, credit card use information, GPS data from devices installed in cars, geolocated tweets, or Foursquare data. This information led to notable insights in human mobility at individual level, but it makes also possible to introduce new methods to extract origin-destination tables at a more aggregated scale, to study the structure of cities and even to determine land use patterns.

In this work, we analyse a Twitter database containing over 5 million geo-located tweets from 39 European countries with the aim of exploring the use of Twitter in transport networks. Two types of transportation systems are considered across the continent: highways and trains. Figure 9(a), Figure 9(b) and 9(c) display maps of geo-located tweets, roads and railways, respectively. A close look at the three maps reveals that while tweets concentrate in cities, there are a number of tweets following the main roads and train lines. In this sense, even roads that go through relatively low population areas can be clearly discerned such as those on Russia connecting the main cities, the area of Monegros in Spain, North of Zaragoza or the main roads in the center of France.

![Maps of geo-located tweets, roads and railways. (a) Geo-located tweets on a map. (b) Highway network. (c) Railway network.](image-url)

Figure 9. Maps of geo-located tweets, roads and railways. (a) Geo-located tweets on a map. (b) Highway network. (c) Railway network.
To identify the tweets on the road/rail between September 2012 and November 2013, we have considered all the tweets geo-located less than 20 meters away from a highway (both directions) or a railway. Then, each tweet on the road/rail is associated with the closest segment of road/rail. Using this information, we can compute the percentage of road and rail segments covered by the tweets (hereafter called highway coverage and railway coverage). A segment of road or rail is covered by tweets if there is at least one tweet associated with this segment.

The percentage of segments (i.e., km) covered by the tweets in Europe is 39% for the highway and 24% for the railway. The highway coverage is better than the railway coverage probably because the number of passenger-kilometres per year, which is the number of passengers transported per year times kilometres travelled, on the rail network is lower. However, the coverage is very different according to the country. Indeed, in Figure 10 we can observe that western European countries have a better coverage than countries of Eastern Europe except Turkey and, to a lesser extent, Russia.

**Figure 10.** Highway and railway coverage. (a)-(b) Locations of the geo-located tweets on the road (a) and rail (b). (c)-(d) Segments of road (c) and rail (d) covered by the tweets. The red segments represent the segments covered by the tweets.
Figure 11 shows the top 20 European countries ranked by highway coverage (Figure 11(a)) and railway coverage (Figure 11(b)). The two countries with the best highway and railway coverages are the United Kingdom and the Netherlands. The tweets cover 97% of the highway system in UK and 89% in Netherlands. On the other hand, the tweets cover up to 80% of the railway network in the UK and 78% in Netherlands. Inversely, the country with the lowest coverage is Moldavia with a highway coverage of 2.5% and a railway coverage of 1%.

![Figure 11. Top 20 of the countries ranked by highway coverage (a) and railway coverage (b).](image)

The first factor to take into account to understand such differences is the penetration rate defined as the ratio between the number of Twitter users and the number of inhabitants of each country. In fact, as it can be observed in Figure 12(a) and Figure 12(b), as a general trend, the coverage of both highway and railway networks is positively correlated with the penetration rate. And, as a consequence, a positive correlation can also be observed between the highway coverage and the railway coverage (Figure 12(c)). However, these relationships are characterised by a high dispersion around the regression curve. Note that the dispersion is higher than what it can look in a first impression because the scales of the plots of Figure 12 are logarithmic. For the two first relationships the mean absolute error is around 7.5% and for the third one the mean absolute error is around 6.5%. This implies that divergences on the geo-located Twitter penetration do not fully explain the coverage differences between the European countries. Part of this dispersion could be due to cultural differences among European countries when using Twitter in transportation.

Disparity in coverage between countries can neither be satisfactorily explained by differences in fares or accessibility to mobile data technology. For example, two countries as France and Spain are similar in terms of highway infrastructure, mobile phone data fares and accessibility, but the geo-located Twitter penetration rates are very different as also are their highway coverage 77% in Spain and 55% in France. Besides penetration rates, divergences in coverage might be the product of cultural differences among European countries when using Twitter in transportation.
4.1 Analysis of mobility patterns: interdependence of mobility with city structure and social networks

Figure 12. Highway and railway coverage as a function of the penetration rate of Twitter in the European countries studied. (a) Highway coverage as a function of the penetration rate. (b) Railway coverage as a function of the penetration rate. (c) Railway coverage as a function of the highway coverage. Red lines are linear regression curves applied to the log-log plots. This means that the slope corresponds to the exponents of power-law relationships. The slope in (a) is around 0.6, 0.8 in (b) and 1 (a linear relation) in (c).

4.2 Social network and travel behaviour

Social network contacts have significant influence on individual’s travel behaviour. However transport models rarely consider this 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 behaviour. One of the goals of the EUNOIA is to examine the relationship between travel behaviour and social networks using mobile phone data, aiming to contribute to the understanding of how the social network influences individual mobility.

4.2.1 Dataset

A huge dataset containing billions of registers has been used for this study. 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 invoice purposes, the information regarding the time and the Base Transceiver Station (BTS) tower where the user was located 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. CDRs 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 spatio-temporal registers.
4.2.2. Identification of social trips

In order to identify social trips, first it is needed to determine the social network from mobile phone data. An egocentric network approach has been used for this purpose. It is assumed that a relationship exists between two different users if the phone communication between them is reciprocal. Therefore, the social network of an ego comprises all alters who have at least one reciprocal call with that ego.

Once the social network of each ego has been identified, it is necessary then to estimate the locations in which the ego is performing an activity, either a social activity or not. If an ego and alter are in the same place at the same time, it is assumed that they are performing a social activity. In order to identify social trips (trips leading to a social activity) and activity performance, mobile phone data has been used. CDRs provide, in average, spatiotemporal information of each user every several hours. This level of detail could be useful for analysing individual daily commuting trips; however, when analysing social events between an individual and its social network, more detailed information is needed. In order to obtain spatio-temporal information on the existing gap between consecutive CDRs, first, only users with enough calls (equivalent to enough registers) have been selected, leading to a sample of users with significant registers along the day. Secondly, aiming to estimate the time spent in each location, trips between consecutive and different locations has been estimated using a trip probabilistic function. Thereby, the time spent and the period of the day the user remains at a specific location is estimated. Jointly analysing the activity and mobility information of all the users of the same social network, social activities can be identified.

This kind of analysis provides rich information about the relationship between social network and travel behaviour: identification of city social activity hot spots, type of location share by social contacts (home, work...), distance travelled to meet a friend in compared to regular trips, etc.

4.2.3. Application to activity based modelling

The results obtained from the joint analysis of user’s social network and travel behaviour 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 travelled 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 do 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 trade-off 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-maximising 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:

\[ V = \sum_i V_{\text{perf}} + \sum_j V_{\text{leg}} \]

where \( V_{\text{perf}} \) is the reward (normally positive) to perform an activity, and \( V_{\text{leg}} \) 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, as well as the literature. 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 social network and travel behaviour 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 agent’s social contacts.
- 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.
- Use the kind of patterns obtained from the mobile phone data analysis to calibrate/validate the model.
- Figure 13 shows an example of how the introduction of the 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.
Figure 13. 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.
5. Visualisation and mapping of individual mobility patterns

A number of visualisations and maps have been created, to help understand the contents and potential benefits of the various datasets uploaded to the EUNOIA data repository as part of WP3. The graphics presented here focus on London, and make use of the variety of datasets made available by the city’s public authority for transport.

5.1. Station and Stop Locations and Identifiers

London’s complex and interleaving public transport structures have resulted in various organisation identifiers for infrastructure, such as stops. The various datafiles supplied by Transport for London (TfL), the Department of Transport (DfT) and other bodies, and used for the EUNOIA project, often use different IDs for the same structure. The map of Figure 14 has been created to catalogue graphically the various IDs in use and also to detect structure such as stations which are close by on different (or even the same) datasets that are actually the same but are duplicated and located close by. For example, Waterloo station can be considered to either be a single transport interchange, or have an underground component managed by Transport for London and an overground component managed by Network Rail. Nearby bus-stops may also take the name of the station. Different parts of other large stations, which at one time were owned by different railway companies, may still be identified differently, as happens to sections of Victoria Station for some datasets.

![Figure 14](image)

*Figure 14:* Bus stops and railway/metro stations plotted on a map of London, with an example stop selected and ID data associated with that stop displayed. The background map is OpenStreetMap which contains data that is copyright OpenStreetMap contributors.
5.2. Vehicle Counts

For London, there are two sources of vehicle counts (road traffic passing fixed counting points) – Transport for London’s Traffic Census, and the Department of Transport’s Traffic Counts.

TfL’s Traffic Census is more recent (April 2013) and covers three time periods – 7-10am, 4-7pm and 6am-8pm. It covers main roads, some smaller roads, and some cycle-only routes. However it only covers approximately 170 points, in central London only. Bicycles, motorbikes, cars, taxis, vans, HGVs and buses are all separately counted.

The DfT’s Traffic Counts are most recently available for 2012, with similar data available annually back to 2000. The data are only available as a single figure for each 24-hour day. It covers around 3300 recording locations, for major roads only, across the whole of London. Further data for roads outside London is also available. The DfT counts don’t distinguish between cars and taxis, but do distinguish between various kinds of HGVs (by axle count).

The TfL data and the 2012 London data from DfT have both been incorporated into the EUNOIA data repository.

This visualisation takes the directional version of the DfT data and plots it, with arrows indicating the general direction concerned, and overlapping coloured arrows showing the vehicle types. Clicking on an arrow reveals the actual measured numbers in that location, for each vehicle type.

Figure 15: Annual averaged daily traffic (AADT) data recording points, by direction, for major roads on London’s road network, as measured by the Department of Transport. The area of the arrow represents the amount of traffic of that type. In this example, blue = cars, red = bicycles, yellow = vans. The traffic data is Crown Copyright. The background map is based on data that is copyright OpenStreetMap contributors.

There is also a version that has been created, that shows the equivalent data from TfL.
5.3. **Individual Bicycle Sharing System Journeys by User**

This map shows all journeys on the Barclays Cycle Hire bicycle sharing system in central London, as made by a single registered user. The user ID is configurable; an example user is shown here. The data consists of approximately 16 million journeys made on the Barclays Cycle Hire system between November 2010 and January 2013. Transport for London makes the data available periodically through its developer portal. The dataset has been processed to remove journeys with missing start/end details, and then added to the EUNOIA data repository.

The routes taken between start and end stations are not known as the bicycles are not equipped with GPS receivers or other route detectors, so journeys are shown with straight lines, with the thickness of each line corresponding to the number of journeys between the two points. The visualisation does not show direction of travel.

It is likely to be possible to create individual travel diaries for some of the most regular users, by studying the times of day they make regular journeys.

It should be noted that the system’s extent only covers central London. Many multi-modal commuter journeys in London end at main-line termini stations which are typically located away from the central business districts of the cities, the bikes then being used to complete the journey to work.

The total number of journeys used by the system on a day typically numbers around 30000 journeys (typical summer 2013 values), whereas 150000 people commute to work by bicycle (2011 census). This latter figure is across the whole city, but bulks of such commuter journeys are concentrated in inner London.
D4.1 Analysis of mobility patterns: interdependence of mobility with city structure and social networks

© EUNOIA Consortium
Page 25 of 38

Figure 16. Journeys by a single user “3554816” of the Barclays Cycle Hire system in central London. Thicker lines correspond to more frequent journeys. The primary journey made by this user is between Waterloo railway station (south-west) and Wenlock Road in Hoxton, a distance of approximately four kilometres. The basemap contains data which is copyright OpenStreetMap contributors. The bicycle sharing system data is copyright Transport for London.

5.4. Metro Journey Flows

This graphic has been created by James Cheshire at CASA using the Oyster Card sample dataset from Transport for London uploaded to the EUNOIA data repository. The dataset is for 5% of journeys in a week in November 2009, the week did not have significant disruptions to the network and so is considered to be typical.

Flows between each pair of stations are shown. The thickness of the lines corresponds to the overall flow between those stations. Colours are used to highlight the larger flows. Numbers correspond to the total number of journeys in the week concerned, within the 5% sample. Routes between start/end stations are not known, although these are normally predictable given the relatively limited topology of most parts of the network.

Only stations on the “London Underground” metro network are shown. The visualisation does not include flows from the Docklands Light Railway or from railway lines. It also does not include any other modes of transport, such as road-based transportation. The network has also changed slightly since 2009, with the East London Line being replaced by a new “Overground” service.
“Oyster card” is the electronic ticketing system used on most Transport for London public transport services. The map does not include journeys that use paper tickets, therefore, or journeys where one of the start or end journeys is missing. In practice, the vast majority of journeys carried out on the London Underground are by Oyster card, with completed start and end records, therefore the map can be considered to be comprehensive.

The graphic reveals the expected concentration of flows into the centre of London from all directions served by the London Underground.

The graphic was produced using the Gephi software application. Two tables, representing the node locations, and the edge values (flows), were created and loaded into the application.

![Figure 17: Journeys on the London Underground, using the Oyster card 5% one-week sample dataset from November 2009 included within the EUNOIA data repository, as visualised by Gephi.](image)
6. Perspectives

In this report we have summarised results related to the spatial organisation of cities. These results include new indicators of city shape that benefit from the spatial and temporal resolution of the digital traces left by citizens when they use their mobile phones or other electronic devices. Our analyses of Spanish cities have provided a proof of the potentiality of these new data sources to better characterise and understand patterns common to the Spanish cities we studied (urban rhythms, number and importance of land-use types, scaling of the number of hotspots). Some measures also revealed different shapes, and permit to cluster cities according to some of their dynamical morphological features (compactness coefficient, spatial organisation of permanent hotspots).

Our results raise new questions that could be addressed in future analyses and future projects. First, according to the continuous digitalisation of urban life, it is expected that scientists get increasing access to rich mobility datasets, probably containing additional information, and covering cities from different regions of the globe. Hence a first natural pursuit of the work would be to replicate the measures performed in our study in cities located in different countries and continent, to cluster cities according to their morphological patterns and enhance the precision of our understanding of daily urban dynamics. It is to note that the morphological properties illustrated on Spanish cities might be specific to European cities, whose urbanisation is old, and notably much older than cities in other continents. It would also be interesting to see if the dynamics are similar in cities of recently urbanised and rapidly developing regions. In this respect, repeating the measures proposed on cities worldwide in which mobile phone datasets are available would bring invaluable and necessary information to the emerging science of cities.

Mobile phone data provides at the same time social interaction and travel patterns information. However, 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. Fusion with other data sources is a promising approach to fill the gaps (such as socio-demographic gaps) as well as to validate the results.

An inevitable direction for further studies will also be to bridge the existing knowledge about centrality patterns in cities with those revealed by new data sources. This should include very soon summaries of the numerous results obtained in the recent years by the very active research community that focuses on the analysis of urban dynamics by mining ICT data. A confrontation of these results with those which have been obtained in the past by quantitative urban geography and spatial economics with mobility surveys and remote sensing data will also be required to not waste research time to reinvent the wheel on more modern data.

New research avenues are also probably at the crossroads of the analysis of multi-layered, infrastructure networks, and human mobility networks. For example centrality measures extracted from the road network structure has been shown recently to be correlated with economic activity. It will be necessary to understand how the structural properties of the multi-layer and multi-modal urban transport networks compare with the patterns extracted from individuals’ mobility networks revealed by ICT data. At this stage, a thorough review of the numerous recent results on urban mobility networks and transport networks would be critically needed to help distinguish robust results and vaguely understood problems, and draw new research avenues.

---

Annex I. References


Lenormand et al. (Under review) Comparing and modelling land use organization in cities.

Lenormand et al. (2014) Tweets on the road. Plos ONE 9(8), e105407.

Louail T. et al. (2014) From mobile phone data to the spatial structure of cities, Scientific reports 4:5276.

### Annex II. Research papers published in the frame of WP4

<table>
<thead>
<tr>
<th>Spatial scale</th>
<th>Temporal scale</th>
<th>Hour</th>
<th>Day</th>
<th>Week</th>
<th>Month to Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>One city</td>
<td></td>
<td>• London tube study (Bosredon et al.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Cross-checking different sources of mobility information (Lenormand et al., Plos One, 2014)</td>
<td></td>
<td></td>
<td>• Influence of social characteristics on the way people move and spend their money (Lenormand et al.)</td>
</tr>
<tr>
<td>Several cities of the same country</td>
<td>• From mobile phone data to the spatial structure of cities (Louail et al., Sci. Rep., 2014)</td>
<td>• Uncovering the spatial structure of mobility networks (Louail et al.)</td>
<td>• Comparing and modeling land use organisation in cities (Lenormand et al.)</td>
<td>• Understanding the influence of the social network on travel behavior (Picornell et al.)</td>
<td></td>
</tr>
<tr>
<td>Several cities worldwide</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Legend**
- Published
- Under review
- In preparation

**Figure 18:** Ten research articles have resulted from the various studies performed in the framework of EUNOIA WP4. We have tried to better understand urban mobility at several spatial and temporal scales, ranging from a study of the structure of the flows in the London tube during a week (with 'tap-in/tap-out' transport card data available for every fifteen minutes period, during an entire week) to a study of mobility patterns at the European scale, using geolocated tweets collected during an entire year. The published papers are in green, papers currently under review are in blue, and running studies are in red.
Annex III. Abbreviations and acronyms

AUDES: Áreas Urbanas De ESpaña. Harmonized delimitation of Spanish metropolitan areas based upon functional criteria. Aggregation of municipalities is decided on the basis of the proportion of residents that work and commute to the metropolitan area;

BTS (tower): Base Transceiver Station. Fixed antenna that facilitates the linkage of the mobile phones to the mobile communication network. In CDR, the geolocation of a user is in fact the geolocation of the BTS tower to which its mobile phone is currently linked.

CBD: Central Business District.

CDR: Call Detail Record, also known as call data record. Data record produced by a telephone exchange or other telecommunications equipment that contains attributes that are specific to a single instance of a phone call or other communication transaction that was handled by that facility or device.

OD: Origin-Destination matrix.

TfL: Transport For London.

UTM: Universal Transverse Mercator, a standard and universal system of cartographic projection.
Annex IV. Analysis of urban activity levels from phone data

IV.1 Description of input data

The data provided by the phone operator correspond to all the CDRs of the calls made in Spain for 55 sample days within the period September- November 2009.

The data used for the analysis include the following information:

- Telephone of origin (encrypted)
- Telephone of destination (encrypted)
- Starting day and time of the call
- Duration of the call in seconds
- Mobile phone operator for the telephone of origin
- Mobile phone operator for the telephone of destination
- Coordinates of the base transceiver station(s) (BTS) to which the phone operator user is connected when the call starts and when the call ends:
  - If the CDR corresponds to a call between a client of the phone operator and a client of another operator, then the BTS corresponds to client of the phone operator.
  - If both users, caller and receiver, are clients of the phone operator, it is not possible to know to which of them corresponds the BTS appearing in the CDR.

IV.2 Urban areas under study

We have chosen to analyse metropolitan areas, rather than municipalities, in order to capture most daily trips between cities and the surrounding area. The study has been initially focused on six metropolitan areas in Spain, selected to display a variety of sizes and population densities:

<table>
<thead>
<tr>
<th>Metropolitan Area</th>
<th>Population</th>
<th>Area</th>
<th>Pop. Density</th>
<th>List of municipalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Madrid</td>
<td>5,512,495</td>
<td>1,935.97</td>
<td>2,847.82</td>
<td>MADRID, ALCOBENDAS, ALCORCÓN, BOADILLA DEL MONTE, BRUNETE, COLMENAR VIEJO, COSLADA, FUENLABRADA, GETAFE, LEGANÉS, MAJAHDONDA, MEJORADA DEL CAMPO, MÓSTOLES, PARACUELLOS DE JARAMA, PARLA, PINTO, POZUELO DE ALCORCÓN, RIVAS-VACIAMADRID, LAS ROZAS, SAN FERNANDO DE HENARES, SAN SEBASTIÁN DE LOS REYES, TORREJÓN DE ARDOZ, TRES CANTOS, VELLÍN DE SAN ANTONIO, VILLANUEVA DE LA CAÑADA, VILLANUEVA DEL PARDILLO, VILLAVICIOSA DE ODÓN</td>
</tr>
<tr>
<td>Barcelona</td>
<td>3,218,071</td>
<td>634</td>
<td>5,075.77</td>
<td>BARCELONA, HOSPITALET DE LLOBREGAT (L'), BADALONA, SANTA COLOMA DE GRAMENET, CORNELLÀ DE LLOBREGAT, SANT BOI DE LLOBREGAT, VILADECANS, CASTELLDEFELS, CERCANYOLA DEL VALLÈS, ESPLUGUES DE LLOBREGAT, GAVÀ, SANT FELIU DE LLOBREGAT, RIPOLLET, SANT ADRIÀ DE BESÓS, MONTCADA I REIXAC, SANT JOAN DESPÍ, BARBERÀ DEL VALLÈS, SANT VICEÇ DELS HORTS, SANT ANDREU DE LA BARCA, MOLINS DE REI, SANT JUST DESVERN, BADIA DEL VALLÈS, CORBERA DE LLOBREGAT, CASTELLBISBAL, PALLEJA, MONTGAT, CERVELLÓ, SANTA COLOMA DE CERVELLÓ, TIANA, BEGUES, TORRELLES DE LLOBREGAT, PAPIOL (EL), SANT CLIMENT DE LLOBREGAT, CERVELLÓ, SANT BOI DE LLOBREGAT, SAN CUGAT DEL VALLÈS, PRAT DE LLOBREGAT (EL), PALMA DE CERVELLÓ (LA)</td>
</tr>
</tbody>
</table>
### IV.3 Description of data processing

The objective of the analysis was to estimate population concentration in different areas, for different days and at different times of the day. Data processing has included the following tasks:

- For each day, we have filtered those CDRs where the BTS to which the user is connected at the beginning of the call is located within the metropolitan area under study, i.e. within any of the municipalities specified in section 3.
- The selected time interval is one hour, i.e. we aim at estimating the average number of users within a particular zone from 00.00 to 01.00, from 01.00 to 02.00, etc.
- The number of users per area have been counted according to the following criteria:
  - each person shall count only once per hour;
  - the number of registered positions of a particular person within an hour and a particular cell, divided by the number of total registered positions for that person (in any cell) within that hour, is taken as a proxy of the portion of that hour during which the user is located within that particular cell.
- Finally, we add all the activity per person, per BTS position and per hour.

---

<table>
<thead>
<tr>
<th>Metropolitan Area</th>
<th>Population</th>
<th>Area (km²)</th>
<th>Pop. Density (Inhab./km²)</th>
<th>List of municipalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valencia</td>
<td>1,549,855</td>
<td>628.81</td>
<td>2,464.74</td>
<td>ALAQUAS, ALBAL, ALBALAT DELS SORELLS, ALBORAYA, ALBUIXECH, ALCÀSSER, ALDAIA, ALFAR, ALFARA DEL PATRIARCA, ALMASSERA, BENETUSSE, BENIPARRELL, BONREPÓS I MIRAMBELL, BURIASSOT, CATARROJA, XIRIVELLA, QUART DE POBLET, PUIG, EMPERADOR, FOIOS, GODELLA, LLOCNOU DE LA CORONA, MANISES, MASSALFASSAR, MASSAMAGRELL, MASSANASSA, MELIANA, MISLATA, MONCADA, MUSEROS, PAIPORTA, PATERNA, PICANYA, PICASSENT, POBLA DE FARNALS (LA), PUCOL, RAFLBUÑOL/RAFLBUNYOL, ROCAPORT, BENAGÉBER, SEDAVID, SILLA, TAVERNES BLANQUES, TORRENT, VALENCIA, VINALESA</td>
</tr>
<tr>
<td>Bilbao</td>
<td>908,916</td>
<td>500.2</td>
<td>1,817.10</td>
<td>ABANTO Y CIÉRVANA, ALONSOTEGI, ARRANKUDIA, ARRIGORRIAGA, BARAKALDO, BARRIA, BASAURI, BERANGO, BILBAO, ZEBERIO, ABANTO Y CIÉRVANA-ABANTO ZIERBENA, DERIO, ETXEBARRI, ERANDIO, GALKAKAO, GORLIZ, GETXO, LARRABETZU, LEOIA, LEMOIZ, LEZAMA, LOIU, UGAO-MIRABALLES, MUSKIZ, ORTUELLA, PLENTZIA, PORTUGALETE, SANTURTZI, SEXTAO, SONDICA, SOPELANA, URDULIZ, VALLE DE TRÁPAGA, ZAMUDIO, ZARATAMO</td>
</tr>
<tr>
<td>Zaragoza</td>
<td>747,377</td>
<td>2,289.64</td>
<td>326.42</td>
<td>ALFAJARÍN, CADRETÉ, CUARTE DE HUERVA, EL BURGO DE Ebro, MARÍA DE HUERVA, PASTRIZ, PUEBLA DE ALFINDÉN (LA), SAN MATEO DE GÁLLEGO, UTEBO, VILLAMAYOR DE GÁLLEGO, VILLANUEVA DE GÁLLEGO, ZARAGOZA, ZUERA, PINSEQUE, SOBRADIEL, TORRES DE BERRELLÉN, MUEL (LA), LECÍNENA</td>
</tr>
<tr>
<td>Palma de Mallorca</td>
<td>552,848</td>
<td>1,016.36</td>
<td>543.95</td>
<td>ANDRATX, BUNYOLA, CALVIÀ, ESPORLES, LLUCMAJOR, MARRATXÍ, PALMA DE MALLORCA, PUIGPUNYENT, SANTA MARIA DEL CAMI</td>
</tr>
</tbody>
</table>

(1) Source: Instituto Nacional de Estadística (), 2009 data.
(2) Sources: Madrid metropolitan area as defined by Comunidad de Madrid (see ‘Atlas de la Comunidad de Madrid en el Siglo XXI’); Barcelona metropolitan area as defined by Área Metropolitana de Barcelona (); Valencia metropolitan area as defined by Ayuntamiento de Valencia; Bilbao metropolitan area as defined by Gobierno Vasco (); Zaragoza and Palma de Mallorca metropolitan areas as defined by the AUDES project ().
Annex V. Methods: mobile phone data description and determination of hotspots

V.1 Phone data - Generalities

The first step of the work has been the extraction of the number of mobile phone users per hour and per Base Transceiver Station (BTS) identified with UTM (WSG84) coordinates. The methodology used to compute these aggregated values is described in Annex 3. We have this information for Spain and for 55 days (from 1 to 15 September 2009, from 1 to 31 October 2009, from 1 to 8 November 2009 and 10 November 2009).

During the first year we first focused on a benchmark composed of six metropolitan areas (Barcelona, Bilbao, Madrid, Palma de Mallorca, Valencia and Zaragoza), those for which aggregated data had been extracted as of 30 September 2013 (see D4.1 intermediary report). These metropolitan areas are very different in terms of population sizes, spatial dimensions and population densities.

<table>
<thead>
<tr>
<th>Metropolitan area</th>
<th>Number of municipalities</th>
<th>Number of inhabitants</th>
<th>Area (km2)</th>
<th>Average number of users per hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Palma</td>
<td>8</td>
<td>552,848</td>
<td>1,016.36</td>
<td>9,861</td>
</tr>
<tr>
<td>Zaragoza</td>
<td>18</td>
<td>747,377</td>
<td>2,289.64</td>
<td>13,983</td>
</tr>
<tr>
<td>Bilbao</td>
<td>34</td>
<td>908,916</td>
<td>500.2</td>
<td>14,584</td>
</tr>
<tr>
<td>Valencia</td>
<td>43</td>
<td>1,549,855</td>
<td>628.81</td>
<td>26,561</td>
</tr>
<tr>
<td>Barcelona</td>
<td>36</td>
<td>3,218,071</td>
<td>634</td>
<td>62,766</td>
</tr>
<tr>
<td>Madrid</td>
<td>27</td>
<td>5,512,495</td>
<td>1,935.97</td>
<td>128,604</td>
</tr>
</tbody>
</table>

The number of users (per hour) represents on average 2% of the total population and at most 5% of the total population. These percentages are almost the same for all the metropolitan areas. An important issue regarding the spatial analysis of these cities lies in the delimitation of their spatial extension: what do we consider as the relevant boundaries to analyse and compare the morphological and mobility patterns of cities which are very different in population sizes, spatial dimensions, density, and age? This is a crucial point: we know that, even if administrative/political boundaries are chosen in compliance with what is perceived as coherent territorial units, they often mismatch the reality of the functioning of territories [1]. As a starting point we have relied on the metropolitan areas delimitations, which correspond to regions comprised of a densely populated urban core (the ancient and main municipality which give its name to the metropolitan area) and its less-populated and dense surrounding territories, sharing industry, infrastructure, and housing. As we will see in section 3, given the nature of the data, relying upon these spatial boundaries may raise some issues in the comparison of urban forms. For some cities, the lower-density municipalities included might bias the values of the global descriptive measures of the spatial structure of the cities, since these are zones for which the intensity of phone activity in the dataset is very low. In addition in these zones antennas are very sparse, which increases the noise in the density estimates.
V.2 Identification of hotspots

The data gives access to the spatial density \( \rho(i, t) \) of users for different moments. The full density is a complex object and we have to extract relevant and useful information. In particular, we will focus on particular locations that display a density much larger than the others. These locations - the hotspots - give a good picture of the city by showing where most of the people are. The hotspots thus contain important information about points of interest and activities in the city. At this point hotspots should not be confused with polycentres, even if the nuance at this point may be subtle and is not clear in the literature. We will keep the term “polycentres” to describe places of attraction of daily travels (important starting and ending points of journeys). These polycentres will be discussed through the analysis of OD matrices.

The determination of centres and subcentres from density data is a problem which has been broadly tackled in urban economics during the last twenty years (Giuliano and Small 1991, McMillen 2001, McMillen and Smith 2003). Starting from a distribution of spatial densities, we have to identify the local maxima. This is in principle a simple problem solved by the choice of a threshold \( \delta \) for the density \( \rho \): a cell \( i \) is a hotspot at a time \( t \) if the instantaneous density of users \( \rho(i, t) > \delta \). This is for example what is done by Giuliano and Small to determine employment centres in Los Angeles (1991). It is however clear that this method introduces some arbitrariness due to the choice of \( \delta \), and also requires prior knowledge of the city to which it is applied. Other authors have proposed nonparametric methods based on the regression of the natural logarithm of employment density on distance from the centre (McMillen 2001). The exponent of the negative exponential fit of the density distribution has also been used to determine the number of centres in polycentric cities (Griffith 1981). Potential limits of these methods lie in the fact that they return a unique number of centres that could reveal to be biased when the actual density distribution is not properly fitted by a negative exponential law. Here we will propose an alternative method that allows us to control the impact of this choice.

A first simple criterion is to choose the point that corresponds to the average \( \mu(t) = \overline{\rho(i, t)} \) of the distribution at time \( t \): all the cells whose density is larger than \( \mu \) are hotspots. This is a weak definition of what can be considered as a hotspot, and we propose here to use it as a 'lower' bound. In order to understand how the various properties of hotspots will depend on this definition, we introduce a stronger and more restrictive definition that will be considered as an upper bound of what can be considered as a hotspot. In the following we discuss how to find this upper bound.

In order to characterize the disparity of the activity in the city and to isolate the dominant places, we first plot the Lorenz curve of the density distribution in the city at each hour. The Lorenz curve is a standard object in economics, it is a graphical representation of the cumulative distribution function of an empirical probability distribution. For a given hour, we have the distribution of densities \( \rho(i, t) \) of the spatial units composing the metropolitan area, and we sort them in increasing rank and denote them by \( \rho(1, t) < \rho(2, t) < ... < \rho(n, t) \) where \( n \) is the number of cells. The Lorenz curve is constructed by plotting on the x-axis the proportion of cells \( F = i/n \) and on the y-axis the corresponding proportion of users density \( L \) with:

\[
L(i, t) = \frac{\sum_{j=1}^{i} \rho(j, t)}{\sum_{j=1}^{n} \rho(j, t)}
\]
If all the densities were of the same order the Lorenz curve would be the diagonal from (0,0) to (1,1) and in general we observe a concave curve, with a more or less strong curvature. The area between the diagonal and the actual curve is related to the Gini coefficient, an important indicator of inequality used in economics.

Regarding the Lorenz curve, the stronger is the curvature the stronger is the inequality and, intuitively, the smaller is the number of hotspots. This remark allows us to construct a new criterion by relating the number of dominant hotspots (i.e. those that have a very dominant value compared to the other cells) to the slope of the Lorenz curve at point $F = 1$: the larger the slope, the smaller the number of dominant individuals in the statistical distribution. The natural way to identify the typical scale of the number of hotspots is to take the intersection point $F$ between the tangent of $L(F)$ at point $F = 1$ and the horizontal axis $L = 0$ (see Figure 19).

![Figure 19. Illustration of the thresholds selection on the Lorenz curve.](image)

This method is inspired from the classical scale determination for an exponential decay: if the decay from $F = 1$ was an exponential of the form $e^{-(1-F)/\alpha}$ where $\alpha$ is the typical scale we want to extract, this method would give $1 - F = \alpha$.

Another criterium is given by the point of the Lorenz curve with tangent slope equal to 1. The general expression of the Lorenz curve for the set of densities $\rho(i, t)$ whose cumulative function is $F(\rho)$ is:

$$L(F) = \frac{1}{\mu} \int_0^F \rho(F)dF$$

where $\mu$ is the average of $\rho(i, t)$ and $\rho(F)$ is the inverse function of the cumulative. This point thus satisfies:

$$\frac{dL}{dF} = 1$$
which gives \( \mu = \rho(F_H) \) or in other words, the hotspots will be those with densities larger than the average. Our criterium gives:

\[
F = 1 - \frac{\mu}{\rho_M}
\]

where \( \rho_M \) is the maximum value of \( \rho(t) \) (for a given time \( t \)). We thus see that in general \( F_{Avg} < F \) and that our new criteria, more restrictive, does not only depend on the average value of the density but also on the dispersion: the larger \( \rho_M \) and the larger the value of \( F \), and consequently the smaller is the number of hotspots detected.

All other possible methods will then essentially give a value comprised in the interval \([F_{Avg}, F] \) between the average criterion and our criterion (denoted in the following by `LouBar`). Instead of choosing a particular point, we will thus study most of the properties computed for hotspots with the two methods, giving us both lower and upper bounds. In particular, we will be able to test the robustness of our results against the arbitrariness of the hotspot identification method.

Figure 20 shows the location of the hotspots selected according to the two methods/criteria according at different moments of the day, in the metropolitan area of Barcelona.

These maps should be watched as the boundaries of the set of hotspots maps that reasonable delimitation methods could produce, that is to say that any reasonable method would select a number of hotspots comprised between \( F_{Avg} \) and \( F \).
V.3 Influence of the spatial scale of aggregation

Another arbitrary parameter in the hotspots identification process is the size of the cells on which we aggregate the number (resp. densities) of users associated to each BTS. Figure 21 illustrates how the ratio $H/N$ (#hotspots / #cells) evolves with the size of the grid for the two boundary methods.

![Graph](image)

**Figure 21.** $H/N = f(a)$. When $a$ varies from 100m to ~ 2000m, $H/N$ grows in a monotonous way, and then fluctuations appear, due to boundary effects when the grid cell become too large. In all cases, it would not be relevant to analyse the spatial structure of cities with base spatial units bigger than about 4km$^2$. For sizes of cells comprised between 500m and 2000m the ratio between the $H/N$ value obtained with the Average criteria and the one obtained with our criteria (LouBar) is approximately constant.

We can reasonably make this cell size $a$ vary from 500 meters to 2 km. Figures in the annex 4 give an idea of how much the proportion of hotspots and their location patterns change from one cell size to another. Also, the cell size $a$ should primary be chosen with regards of what sizes could be considered as reasonable sizes for an urban hotspot. From the pedestrian point of view, every size between 500 metres and 2 kilometres seems *a priori* acceptable. Below 500m, it would clearly be necessary to aggregate contiguous hotspots: for example, for $a=100m$ ($10^{-2}km^2$ cells), two contiguous hotspots could not be considered as different from a pedestrian point of view. In contrast, a size of 2000m can be considered as an upper bound for the same reasons: if two contiguous cells are classified as hotspots, it is reasonable to consider them as two distinct hotspots. It is however a question of perception and should be discussed carefully. In the hypothesis of $a=1000m$ ($1km^2$ cells), we chose to consider that *two adjacent cells are two different hotspots*. 
Figure 22. Time evolution of the ratio (#hotspots/#cells) for two hotspots definitions and different sizes of grid cells. Every reasonable method for defining hotspots would give a value between the two lines of each plot.

For reasonable sizes of grid (i.e. reasonable values of $\alpha$), values of indicators should be robust with a change of aggregation grid. We will then have to test the sensitivity of various results with respect to different resolutions in order to check for the existence of significant aggregation effects.