This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3366088

Using App Usage Data from Mobile Devices to Improve Activity-based Travel Demand Models

Ana Belén Rodríguez González, Javier Burrieza-Galán, Juan José Vinagre Díaz, Inés Peirats de Castro, Mark Richard Wilby, and Oliva G. Cantú-Ros

Abstract—In the last years we have seen several studies showing the potential of mobile network data to reconstruct activity and mobility patterns of the population. These data sources allow continuous monitoring of the population with a higher degree of spatial and temporal resolution and at a lower cost compared with traditional methods. However, for certain applications, the spatial resolution of these data sources is still not enough since it typically provides a spatial resolution of hundreds of meters in urban areas and of few kilometers in rural areas. In this work, we fill this gap by proposing a methodology that utilises GPS data from the usage of different applications in mobile devices. This approach improves the spatial precision in the location of activities, previously identified with the mobile network data.

Index Terms—Application usage data, travel demand models, mobile phone data, location of activities, big data analytics.

◆

1 INTRODUCTION

DETAILED knowledge of human activity and in particular of population's distribution and dynamics is key for public policy planning and services provision in domains like transport, health and urban planning, among others. Traditionally, the analysis of population's distribution and mobility has been based on data from surveys (e.g., census, travel surveys, etc.). This approach has the disadvantage of being expensive, thus providing small datasets and, in most cases, a static picture of the population distribution.

Advances in information and communication technologies and data analysis techniques have opened new possibilities for the study of population's activity dynamics [1] and for the detection of conflicts in urban areas [2]. In the last ten years, we have seen different examples of population's activity-mobility analysis leveraging on geographically located big data sources. These enable the continuous collection of activity and mobility data with high spatio-temporal resolution, opening the door to longitudinal studies that monitor short and long-term changes in citizens' behaviour [3]. A variety of studies have demonstrated the potential of: Bluetooth sensors for traffic monitoring [4], [5], estimating presence at mass events [6], [7] and analysing urban structure [8]; ticketing and smart card data for monitoring mobility patterns in public transport [9], [10] and estimating exposure to advertising [11]; mobile crowdsensing for resilient parking search [12]; and mobile network data (MND) for analysing urban structure [13], [14], [15], traffic and mobility monitoring [16], [17], [18], monitoring urban dynamics [19] and estimating exposure to pollutants [1], [20] among other examples of data sources and applications.

From the mentioned data sources MND are particularly interesting for the analysis of population's dynamics thanks to their large samples sets for most population segments and their constant data generation along one or a sequence of days [1], [20]. However, these data present some limitations. Their spatial and temporal resolution is heterogeneous among areas and users and, in some cases, can be considered coarse for some types of analyses [21]. The temporal resolution of the records depends on the frequency of use of the mobile device; most users typically generate a register every 15–20 min at least. On the other hand, the spatial information depends on the network structure, defined by the positions of the antennas. They are typically spaced from dozens to hundreds of meters in urban environments, and up to a few kilometers in rural areas, where the mobile network is less dense. In most cases, it is assumed that the device is connected to the closest antenna and, hence, it can be located within its coverage area. This coverage area corresponds to a complex geographical region, which overlaps with adjacent regions by design. Consequently, they are commonly approximated by Voronoi polygons for simplicity, which result from dividing the space with a tessellation of Voronoi, whose seeds are the antennas.

On the contrary, the spatial information obtained from mobile apps is much more precise, since it corresponds to GPS data [22]. However, this information provides small sample sets and is discontinuous in time. GPS tracking shows a high battery consumption and, consequently, apps tend to use it only when they are active. For this reason, we can find users with intermittent traces.

In this paper we propose and demonstrate a methodology to refine the location of users performing an activity (extracted from MND), using mobile apps records as ancillary

---

- *A. B. Rodríguez González, J. J. Vinagre Díaz, and Mark Richard Wilby are with the Department of Mathematics Applied to Information and Communication Technologies, Universidad Politécnica de Madrid, Avda. Complutense, 30, 28040 Madrid, Spain*
E-mail: anabelen.rodriguez@upm.es; juanjose.vinagre@upm.es; mark.wilby@upm.es.
- *Javier Burrieza-Galán, Inés Peirats de Castro, and Oliva G. Cantú-Ros are with NOMMON, Pl. Carlos Trías Bertrán, 4, 28020 Madrid, Spain*
E-mail: javier.burrieza@nommon.es; ines.peirats@nommon.es; oliva.garcia-cantu@nommon.es.

Manuscript received MMM DD, 20YY; revised MMM DD, 20YY.

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3366088

data. This methodology takes advantage of the strengths found in each data source: the continuous longitudinal information of large data sources and the detailed spatial-temporal information of small sample sets, available for specific moments of the day.

2 RELATED WORK

In this section we review previous research devoted to the analysis of the population's distribution from the analysis of MND. More precisely, based on the analysis of mobile phone records. Mobile phone records are produced every time the mobile phone's user interacts with the network. These records contain information about the position and the time at which the interaction took place.

Before proceeding with the review, it is important to clarify a distinction between user location and activity location. The user location refers to the position of the user (specifically, the mobile device) at the moment the record occurs, independently of whether the user is in transit or performing an activity. We understand an activity as the action that motivates the users' displacement. The activity location refers to that position where the users are located while performing an activity. The identification of activities requires a longitudinal processing these records. There are different methodologies for identifying activities from MND [23]. This section will concentrate on the location estimation of both single registers or activities and will not focus on the methodologies followed for activities' identification.

The majority of studies that use MND to analyse population dynamics estimate the users' position at a Voronoi polygon level and, most commonly, in a subsequent step, they assign the users in each polygon to one of the zones of a predefined zoning system (e.g., census tracks, transport zones, regular grids, etc.) that intersect the polygon. Depending on the scope of the study it may be the case that the spatial accuracy of the Voronoi polygon level may be enough. [1] compare the dynamic population densities obtained from call detail records (CDRs) from Portugal and France with the dynamic densities obtained from the census data applying a dasymetric model with ancillary data from land use, OpenStreetMap-derived infrastructure, satellite nightlights and slope among others. In this work, the authors conclude that, even considering the Voronoi polygons as the minimum spatial granularity, the results obtained with CDRs are superior in accuracy compared with those obtained with the dasymetric models applied to the census.

Different methodologies are adopted when a higher, than a Voronoi polygon, level of spatial accuracy or the adoption of a specific zoning system is needed for the analysis. These may be based on distance, land use or densities criteria. In [24], activity patterns were inferred from CDRs of one million users in San Francisco (USA) based on hidden Markov models. This work discriminates between primary and secondary activities and assign them to Transportation Analysis Zones (TAZs). In [25], a methodology is proposed to generate origin-destination (OD) matrices using MND from 2.87 million users in Dhaka (Bangladesh). Spatially, the OD matrices are determined from tower-to-tower transitions in a certain time window and then associated to nodes of the traffic network, by geographical proximity.

[17] reconstructed the activity-travel patterns of one day of 10 % of the population of the metropolitan area of Barcelona (Spain) from MND of one month in autumn, 2019. In this work, the location of all activities, except for Home activity, is randomly distributed inside the identified Voronoi area. Residence location (home activity) is assigned probabilistically to one of the census tracks intersecting the Voronoi area identified as home. The probability to be assigned to each track is a function of the socio-demographic characteristics (age, gender) of the user and of the population in the different intersecting census tracks. [19] studied the population dynamics of Madrid (Spain) during a pre-COVID-19, COVID-19 and post COVID-19 period based on the longitudinal analysis of MND. In this work, the assignation of activities from Voronoi polygon to the specific zone was made using a probabilistic function based on land use information and activity type (e.g., users performing a work activity will have a higher probability to be assigned to those zones inside the polygon with a predominant business and offices land use). This required not only to identify the different activities the user performs but also the type of activity. The exploitation of land use data as ancillary data to refine location of activities obtained from MND is a commonly observed practice [23], [26].

Other works, estimate the users' position applying triangulation algorithms to consecutive records connected to different antennas. This methodology tries to increase the precision of the spatial location of the records. [27] analyzed more than 8 billion mobile phone records of 2 million users in Boston (USA), whose position was estimated by triangulation, to identify users' activities, incorporating surveys as ancillary data. A location precision of 300 meters was stated in this work. [28] also used mobile records which position was approximated by triangulation to monitor mobility in the state of Massachusetts (USA). In this work, no explicit validation of location's accuracy is presented. [11] analysed data from CDR in Singapore registered on 5 000 towers for 14 days in March/April, 2011. In the study, mobility patterns are extracted and types of activities are inferred, with transport planning purposes. The activities' location is estimated at area level.

To the best of our knowledge, none of the previous works used GPS data from mobile apps in order to refine the spatial accuracy of activities extracted from MND. On the other hand, we can consider the work of Blasco et al. [29] as a predecessor of the work we present here, given that activity-mobility diaries extracted from mobile phone data are enhanced with information coming from mobile phone apps data for the detailed reconstruction of the users' itinerary inside the Palma de Mallorca Airport (Spain).

3 DATA SETS

This study is based on the use of multiple big data sources. Each data source provides partial information for the analysis we want to perform. What we need to do is fuse the data in a systematic and unbiased way. In this work, we will concentrate on the integration of geographical information.

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3366088

Primarily from a geographical perspective, we use a MND database, provided by one of the largest mobile network operator in Spain, which serves as a basis to identify activities. The geographical information is augmented using a database of application usage on mobile devices, provided by *Pickwell*, a company that records location and use on each mobile device subscribed. It is important to note the relative scales of these data sources. The MND data provides a massive sampling of users, whilst the Pickwell database, which is much smaller, has a much more detailed representation of the geographical distribution of users, albeit intermittently.

3.1 Mobile Network Data

These data consist of a set of anonymised mobile phone records generated by the users in Madrid (Spain) in August, 2019. This data was obtained through a collaboration agreement with one of the three main Mobile Network Operators (MNOs) in Spain, with a market share of more than $20\,\%$. The homogeneous penetration of the MNO in virtually all socioeconomic groups of the population, together with the size of the sample set, grants a good representativeness of the whole Spanish population.

The records include call detailed records (CDRs) produced every time the user interacts with the network, which include making or receiving a call, a message or an Internet data connection, as well as passive events coming from network probes. Among other information, each record contains an anonymised identifier of the user, a timestamp and the ID of the cell or tower to which the device is connected at that particular moment.

In addition to the CDRs, the data provided by the MNO includes the position of the different cells. This produces an indication of the geographical position of the user at certain moments within the day. The records do not provide the exact location of the users. Users could be located anywhere inside the coverage area of the cell to which they are connected.

Ancillary data of land use and census information is used for the identification of activities performed in a single day by the users of the network which have at least one stay in Madrid the studied day. The ancillary data has the following characteristics. *Land use* data was obtained from the Directorate General for Cadastre in Spain. The databases define the surface area m^2 of each type of land use. These data are updated every 6 months and the data set we used corresponds to the update of January 24, 2020. For exploitation purposes, this data was discretised in the following way: the Spanish territory was divided in a regular square mesh (125 meters side). For each square, the predominant land use is assigned to it. *Census data* for 2019 was obtained from the National Institute of Statistics. This data has been used as the sampling frame for expanding the sample of the MNO customers.

3.2 App Usage Data

The Pickwell database contains $346\,\mathrm{GB}$ of information on the use of more than 500 applications across all the corresponding mobile devices. It was collected in Spain during the month of August, 2019. In total, there are $1\,591\,954\,031$ records, each of which consists of 16 fields. Among these, the relevant ones for this study are: mobile device identifier, timestamp, and longitude and latitude coordinates. $2\,371\,218$ devices were monitored during the study period. Below, we show a typical record as an illustrative example:

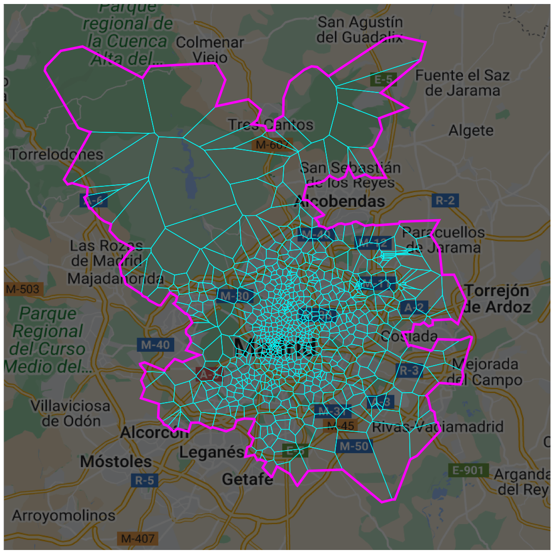


Fig. 1. Voronoi polygons in Madrid's metropolitan area.

id: `80352571-ca26-4685-bcd9-b48940d592a9`
ts: `03/08/2019 09:49:15`
lon: `-3.8109977`
lat: `40.358947`

As it provides detailed information about the location of the devices, it can be used as a model of the geographical distribution of those devices. We consider that the devices in this database and those in the MND set are typically the same. In this respect, for each hour in the day, we have studied the number of devices that show activity in the MND set and app usage in the Pickwell database, obtaining a correlation coefficient of 0.86. However, we do not have information connecting specific devices between these two databases.

3.3 Study Area

This study focuses on the metropolitan area of Madrid, Spain (see Fig. 1). This region has a surface area of $904\,\mathrm{km}^2$ in which 787 Voronoi polygons are defined, representing the mobile cell areas in the MNO network.

In this scenario, $42\,097\,627$ activities were identified at a Voronoi polygon level, and $114\,527\,692$ application uses are available to locate those activities inside the polygons, which amount to $25\,\mathrm{GB}$ of the original database.

3.4 Ethical issues

The use of MND and location data from mobile apps inherently opens a set of ethical issues to be considered. For this reason, the authors representing their corresponding institutions signed an agreement that protects the users data privacy and restricts the use of the data provided by Pickwell to scientific research. Specifically, (i) the data set is restricted to August 2019; and (ii) the data have only been

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3366088

used within this particular work and accessed by the research team. In addition, MND data has been anonymized. The activity registers contain no personal information and cannot be tracked longitudinally; therefore, individuals cannot be identified.

4 METHODOLOGY

4.1 Identification of Activities

The extraction of activity information from mobile phone records followed previous works reported in [18], [19], [20]. This is summarised in the sequence of steps: i) Data preprocessing and cleaning: filtering errors in the raw data in order to ensure the quality of the results. ii) Sample selection: selecting those users with such level of mobile phone activity that makes it possible to reconstruct their activity patterns in an accurate and reliable way; after this selection phase, we are left with a sample set that represents approximately $15\,\%$ of the population of Spain. iii) Identification of activities, by the longitudinal analysis of CDRs: we define an "activity" as an interaction or set of interactions with the environment that takes place in the same location and motivates an individual to move there; and a "trip" as a sequence of one or more journeys ("stages" or "legs") between two consecutive activities. This way, a trip has a main purpose determined by the activity at origin and/or destination. Different criteria applied to the analysis of several days are used to identify activities and distinguish them from stops between two legs of a single trip based on frequency of appearance, time of appearance and length of stay in the observed locations. The information associated to each activity includes its location, the start and end times, and the type of activity: home, work, study, other frequent activities and non-frequent activities. This classification is based on the analysis of the user's longitudinal behavioral patterns during several weeks/months (e.g. the place of residence of each user is identified as the place where the user more often sleeps). iv) Activities location: once an activity is extracted at an antenna level, a layer of land use information is used to refine the estimation of the user position inside the antenna coverage areas, approximated by the Voronoi tessellation. Users are assigned to different areas served by the same antenna through a probabilistic method that takes into account the type of land use (residential, commercial, industrial, etc.) and the type of activity. The assignment is made in two steps: first, the identified activities are associated to one of the regular squares (125 meters side) intersecting the Voronoi polygon; second, an actual longitude and latitude are assigned at random within the square element. This method of assigning spatial locations will be enhanced by the algorithms we propose in this paper. v) Expansion of the sample to the total population: in order to extract meaningful indicators, the sample is expanded to the total population of Spain. The expansion factor is calculated at a district level as the ratio between the number of residents of the district, according to the census information, and the sample of users assigned to the given district. vi) Finally, the sample is filtered once again to keep only those users performing at least one activity in the study region.

4.2 Location of Activities

We have developed a two-stage algorithm to merge the statistically relevant geographical data with the imprecise location definition of the CDR's. In the first step, algorithm 1, we build a distribution function of the geographical extent of users. In the second, algorithm 2, we use this distribution function to assign a statistically likely detailed position to each activity identified with the mobile phone data.

The size and representativeness of the two data sets provide sufficient ground to reasonably affirm the existence of correlations between them. This fact allows us to relate both data sets and fuse them in order to generate a statistically valid result for the location of activities.

Our starting point is the Pickwell data set, which takes the form of a set of tuples $(i, \vec{p}, t, \ldots)$. Only the identifier, i, the position, $\vec{p}$ and the time, t, are important for this study. The identifier i is defined as the index $i \in I$ where I is the set of all devices and m_i is a label associated with the device i. The raw data, as always, has inconsistencies due to incorrectly registered data and missing records, so we must process them to eliminate these problems. In addition, as outlined in Section 4.1, there are certain restrictions we need to apply to the population distribution model. Specifically, we need to restrict the model to users that confine themselves to a restricted region for the duration of the activity.

4.2.1 Filter Algorithm

To process the data, we define three separate filters that are applied to the data once it has been partitioned into ordered subsets. Each subset is defined by the time-ordered set of all the records corresponding to identifier i. So for each device i, we have the set of events for this particular device

$$\mathcal{E}_i = \{(m_i, \vec{p}_e, t_e) \mid 0 < e \leq |\mathcal{E}_i|\} \tag{1}$$

where $|\mathcal{E}_i|$ is the cardinality of the set and it is time ordered by the constraint $t(e) < t(e+1)$, $\forall e$. This simply defines the trajectory followed by the device.

We now apply the following set of filters to these subsets.

- *minimum size filter:* Eliminates small subsets.
- *maximum speed filter:* Eliminates physically unfeasible movements.
- *stationarity filter:* Restricts the records to cases where the device is reasonably stationary.

The minimum size filter imposes a minimum size to subsets, $|\mathcal{E}_i| \geq N$; if the subset does not meet this criterion, it is discarded.

As the set of events corresponds to the trajectory followed by the device, we know that subsequent records must correspond to geographical positions showing a physical separation corresponding to some feasible velocity. Thus, we define a maximum speed, V, to cover the straight line separation between two subsequent records. If a new record does not match this criterion, it is directly removed and the next event is checked. If the result is valid, the algorithm moves on to the next element in the list and is tested against the next record. If, at any point, the condition $|\mathcal{E}_i| < N$ is met, the entire subset is removed.

The stationarity filter ensures we select only events where devices are reasonably stationary, i.e., their motion is restricted to a few meters, for example, inside a working place or residence. The records are grouped into sub-clusters, based on their location. Each sub-cluster is built by starting with the next available record, and then checking subsequent records in order. If the record's coordinates lie inside a bounding box defined by a maximum diagonal distance, D, it is added to the sub-cluster. Otherwise, the sub-cluster is considered complete and the test value becomes the starting point for the next sub-cluster. Any sub-cluster with a size below N is removed from the data set. This effectively eliminates points where the device is in transit.

For each subset $\mathcal{E}_i$, this processing results in a set of time stamped locations, $\mathcal{E}_i'$, which represent the population distribution, satisfying our stationarity condition as well as reflecting the real underlying population distribution. We can reconstruct the entire data source by taking the union of the filtered event sub sets, $\mathcal{E}' = \cup_{i \in I} \mathcal{E}_i'$.

As a whole, the procedure defines a 3-parameter algorithm, whose pseudo-code is presented in Algorithm 1.

4.2.2 *Allocation Algorithm*

The second stage, algorithm 2, constructs a function, based on $\mathcal{E}'$, that can generate a statistically accurate position vector $\vec{p}_*$ for a given activity identified in the CDR data, given the Voronoi polygon and the corresponding time window.

The CDR data segregates the study area into a group of Voronoi polygons, each one denoted by v_j, with $j = 1, 2, \ldots, J$, where J is the total number of polygons.

Let Γ_j represent the border of the polygon v_j. We now subdivide $\mathcal{E}'$ into the time ordered sub sets of the Voronoi polygon *profile* $\mathcal{U}_j$

$$\mathcal{U}_j = \{(m_q, \vec{p}_q, t_q) \mid \vec{p}_q \subseteq \Gamma_j\} \tag{2}$$

where all symbols maintain their usual meaning and q is defined as an integer index in the range $0 < q \leq |\mathcal{U}_j|$ and has the property $t(q) < t(q+1), \forall q$. This effectively associates all records that occur in the region Γ_j with the set $\mathcal{U}_j$. The time ordering allows further subdivisions into time windows. The union of the subsets reconstructs the original set, so we have $\mathcal{U} = \cup_j^J \mathcal{U}_j$ and $\mathcal{E}' = \mathcal{U}$

These subsets effectively define the population density functional in the region Γ_j. In theory, we would have to discretize the space, calculate the density of measurements, at each discrete point, then construct a functional approximation. Once we had a complete functional, we would then need to invert it to be able to generate coordinates on demand. Fortunately, this complexity is not necessary. We can use the measurements themselves to generate approximate coordinates on demand.

The subsets $\mathcal{U}_j$ contain a collection of positions that quite naturally form a distribution that matches the one we wish to emulate. So, to generate a statistical significant coordinate we just need to select one of the contained points at random in an unbiased way. Consequently, the next step is to define this selection process.

We define two position selection processes: (1) *Random position*, which randomly selects one position among the total set of locations. (2) *Random device*, which randomly selects a device among the total set of devices.

Algorithm 1 Filter algorithm

```
Input:
  $\mathcal{E}$: raw app usage data
  $N$: minimum number of records per location
  $V$: instantaneous speed threshold
  $D$: diagonal distance threshold
Output:
  $\mathcal{E}'$: cleaned app usage data
procedure FILTER( $\mathcal{E}, N, V, D$ )
    $\mathcal{E}' \leftarrow \varnothing$
    for each $\mathcal{E}_i \in \mathcal{E}$ do
        $\mathcal{W} \leftarrow \varnothing$                              ▷ speed filter
        $\mathcal{Z} \leftarrow \varnothing$
        $K_i \leftarrow |\mathcal{E}_i|$
        if $K_i < N$ then
            continue
        else
            $\mathcal{W} \leftarrow \{(m_i, \vec{p}_1, t_1)\}$
            $W \leftarrow |\mathcal{W}|$
            for $k = 2$ to $K_i$ step $1$ do
                $\delta = \| \vec{p}_k - \vec{p}_W \|$
                $\tau = t_k - t_W$
                $\nu = \frac{\delta}{\tau}$
                if $\nu \leq V$ then
                    $\mathcal{W} \leftarrow \mathcal{W} \cup \{(m_i, \vec{p}_k, t_k)\}$
                    $W \leftarrow |\mathcal{W}|$
                end if
            end for

            For simplicity, let us denote here:
                $\mathcal{W} = \{(m_i, \vec{p}_w, t_w) \mid 0 < w \leq W\}$,
                $\mathcal{W}_{x,y} = \{(m_i, \vec{p}_w, t_w) \mid x < w \leq y\}$, and
                $\Delta_{x,y}$ being the diagonal distance of the minimum
                bounding box that contains all the coordinates of
                the set $\mathcal{W}_{x,y}$.

            $\mathcal{Z} \leftarrow \varnothing$                          ▷ stationarity filter
            $a \leftarrow 1$
            $b \leftarrow 1$
            for $w = 2$ to $W$ step $1$ do
                if $\Delta_{a,w} \leq D$ then
                    $b \leftarrow w$
                else
                    if $|\mathcal{W}_{a,b}| \geq N$ then
                        $\mathcal{Z} \leftarrow \mathcal{Z} \cup \mathcal{W}_{a,b}$
                    end if
                    $a \leftarrow w$
                    $b \leftarrow w$
                end if
            end for
            if $|\mathcal{W}_{a,b}| \geq N$ then
                $\mathcal{Z} \leftarrow \mathcal{Z} \cup \mathcal{W}_{a,b}$
            end if
        end if
        $\mathcal{E}' \leftarrow \mathcal{E}' \cup \mathcal{Z}$
    end for
    return $\mathcal{E}'$
end procedure
```

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3366088

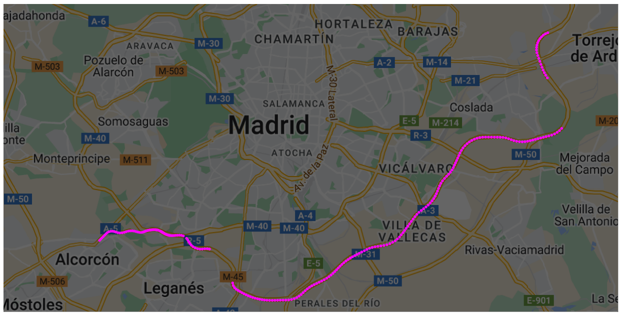


Fig. 2. Example of app usage records removed by the filter algorithm.

The activity data taken from the CDR's creates a set $\mathcal{A}$ of records of the form (v_j, a_z, t_z^I, t_z^F), where z is an index between $0 < z \leq |\mathcal{A}|$, a_z is the activity type, and t^I and t^F are the start and end times of the activity, respectively. This quite naturally forms a collection of subsets

$$\mathcal{A}_j = \left\{(v_j, a_z, t_z^I, t_z^F) \mid 0 < z \leq |\mathcal{A}_j|\right\} \tag{3}$$

each subset containing all the activities in a given Voronoi polygon, v_j. Again it can be time ordered by imposing the condition $t_z^I < t_{z+1}^I, \forall z$.

Now, to assign a location to the activities, we take a specific activity (v_j, a_z, t_z^I, t_z^F), then create the subset $\mathcal{H}$ from $\mathcal{U}_j$ by taking only the events that fall within the time window $[t_z^I, t_z^F]$. If $\mathcal{H} = \emptyset$, it is replaced by the Voronoi polygon *profile* (set of records associated with each Vononoi polygon in the application usage database). Either position selection algorithms can be used to assign a position to the activity record.

The pseudo-code for the allocation algorithm is presented in Algorithm 2.

5 RESULTS

In this section, we show some illustrative results of the performance of the developed algorithms.

5.1 Filter Algorithm

First, some results are shown at the individual level; subsequently, global results are included.

The device with identifier:

`eb40a292-764b-40b7-ae62-cb1d100dce84,`

generated 264 records on $08/01/2019$, in approximately half an hour (between 05:40:07 and 06:05:13). The spatial location of those records is shown in Fig. 2. Clearly, the device is moving and, after applying the filtering algorithm, there is no record associated with that device on that day in the clean database.

At the other end, the device:

`6fbf0f24-cae5-49ef-9bb2-0dce3a008048,`

generated 257 records on $08/01/2019$, in slightly less than 40 minutes (between 14:48:51 and 15:26:06). All those records share the same longitude and latitude coordinates $(-3.7047205, 40.3786122)$ and are kept in the clean database, once the filtering algorithm is applied.

Globally, we can explore the records associated with all devices for a full day, and compare the locations before and after applying the filtering algorithm. We will use as an example the day $08/05/2019$. That day, $4\,296\,798$ records appear in the original database, of which $2\,632\,268$ (61.26%) remain in the clean database. Both cases are drawn in Fig. 3. As we can observe, the majority of records are located on roads, clearly shown in Fig. 3a, correspond to moving devices. Once filtered, those records are no longer found in the database, Fig. 3b.

Algorithm 2 Allocation algorithm

```
Input:
  $\mathcal{U}$: cleaned app usage data
  $\mathcal{A}$: not located activity data
  $\psi$: allocation method
Output:
  $\mathcal{A}'$: located activity data
procedure ALLOCATE( $\mathcal{U}, \mathcal{A}, \psi$ )
    $\mathcal{A}' \leftarrow \varnothing$
    for $j = 1$ to $J$ step $1$ do                              ▷ each polygon
        for $z = 1$ to $Z_j$ step $1$ do          ▷ each activity in polygon
            $\mathcal{H} \leftarrow \{u_j(q) \in \mathcal{U}_j \mid t_I(z) \leq t(q) \leq t_F(z)\}$
            if $\mathcal{H} = \varnothing$ then
                $\mathcal{H} \leftarrow \mathcal{U}_j$                          ▷ replace by profile
            end if

  Let us denote here:
      $\mathcal{H} = \{h(s) \mid 0 < s \leq S\}$.
      with $h(s) = (m(s), \vec{p}_s, t_s)$ and $S = |\mathcal{H}|$
  Selecting only one usage per device (the first), let us denote:
      $\mathcal{H}_{\mathrm{U}} = \{h_{\mathrm{U}}(r) \mid 0 < r \leq R\}$,
      with $h_{\mathrm{U}}(r) = (m(r), \vec{p}(r), t(r))$ and $R = |\mathcal{H}_{\mathrm{U}}|$
      and $m(r) \neq m(r')\ \forall r, r'$.

            switch $\psi$ do
                case "random record"
                    $s^* \leftarrow \mathrm{randi}(S)^\dagger$
                    $\mathcal{A}' \leftarrow \mathcal{A}' \cup \{(v_j, a_z, t_z^I, t_z^F, \vec{p}(s^*))\}$
                                            ▷ where $\vec{p}(s^*) \subset h(s) \in \mathcal{H}$
                end case
                case "random device"
                    $r^* \leftarrow \mathrm{randi}(R)^\dagger$
                    $\mathcal{A}' \leftarrow \mathcal{A}' \cup \{(v_j, a_z, t_z^I, t_z^F, \vec{p}(r^*))\}$
                                       ▷ where $\vec{p}(r^*) \subset h_{\mathrm{U}}(r) \in \mathcal{H}_{\mathrm{U}}$
                end case
            end switch
        end for
    end for
    return $\mathcal{A}'$
end procedure
```

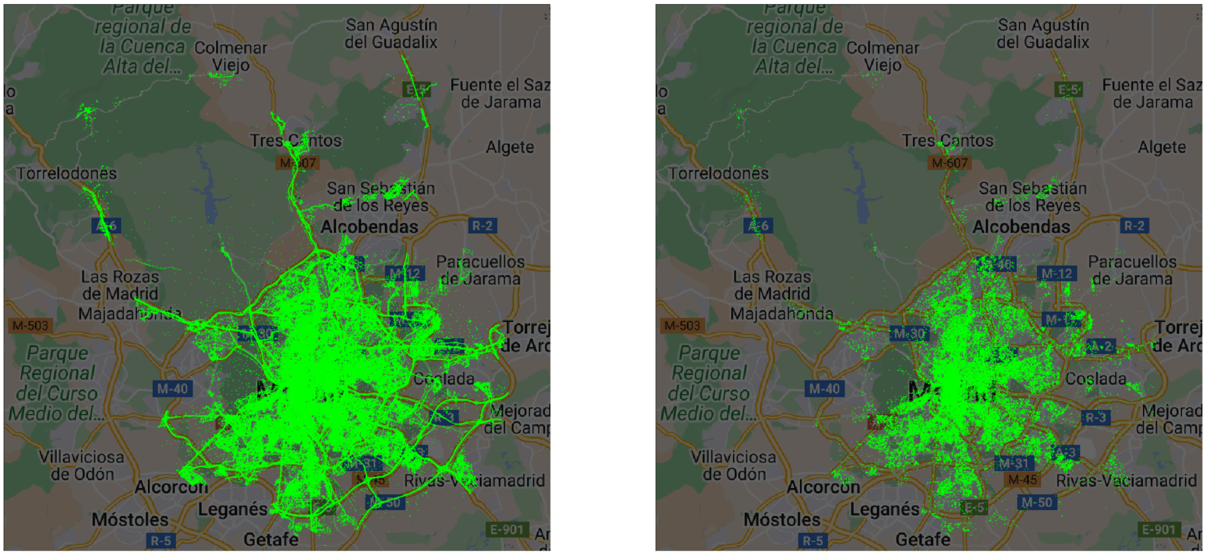
---

† $\mathrm{randi}(X)$ denotes a random integer between 1 and X.

5.1.1 Allocation Algorithm

The results of the allocation algorithm are illustrated with several examples. In Fig. 4, we show the location of activities in a certain Voronoi polygon. As a gold standard to compare against, in Fig. 4a activities are located considering land use data, which is the approach used [19] in the absence of any other data source. In this case, the assignment of

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3366088



(a) *Raw* app usage data.

(b) *Cleaned* app usage data.

Fig. 3. App usages recorded on 05/08/2019: (a) before and (b) after applying the filter algorithm.

specific coordinates (longitude and latitude) to each activity detected by the mobile phone base station uses the following weighted algorithm: the Voronoi polygon associated to the base station is divided into squares of 125 meters. An a priory weight for each type of use is assigned to each square. In reality, activities tend to be located in regions associated with the highest intensity of that type of activity, for example residential, commercial, industrial, etc. The assignment is then made in two steps: first, a square element of 125 meters is selected at random based on the weight value from the associated activity. Second, an actual longitude and latitude is assigned at random within the square element.

The Voronoi polygon included in the example in Fig. 4 has uniform land uses. As a consequence, the coordinates assigned considering land uses are distributed at random (see Fig. 4a). On the other hand, when the application usage data is employed (Fig. 4b and Fig. 4c) the activities are concentrated in certain areas, which match those showing the greatest presence of mobile devices. In addition, the two proposed allocation methods (register-based, in Fig. 4b; device-based, in Fig. 4c) generate similar results.

Finally, Fig. 5 includes the results of locating activities within a Voronoi polygon with *non-uniform* land uses. The conclusions are similar to those of the previous example. In this case, the difference in land use concentrates activities in certain zones but, within those zones, the distribution is completely random (see Fig. 5a). On the contrary, we avoid this randomness employing application uses to locate activities (Fig. 5b and Fig. 5c).

6 VALIDATION

To the best of our knowledge, there are no previous works in the state-of-the-art that propose refining the spatial accuracy of activities extracted from MND using GPS data from mobile apps, which blocks the possibility of performing a quantitative assessment of the proposed methodology. Considering this issue and in the absence of a ground-truth to check the validity of the results obtained, we opted to carry out an event-based validation, following an approach similar to that used in [30] and [31]. We will use two case studies: a special calendar event and a specific region with atypical activities.

6.1 Case Study 1: Fiestas de la Paloma

The *Fiestas de la Paloma* is a summer celebration located in the neighborhood of *La Latina*, between August 14 and 18, 2019, with different cultural activities carried out in public spaces, from eight in the afternoon till dawn.

In the first place, we select the activities carried out during the days of the festivities, between eight in the afternoon and six in the morning of the following day (908 activities). Then, we locate these activities using two approaches considering (i) land use, and (ii) usage data of mobile applications. Fig. 6 shows the results of each case. As we can observe, employing land use results in an even distribution of activities throughout the Voronoi polygon. However, applying the proposed algorithm, most activities are concentrated around three leisure areas: the *Jardín de las Vistillas*, the *Plaza de la Paja*, and the *Mercado de la Cebada*, which correspond to the three main settings where the festive events took place.

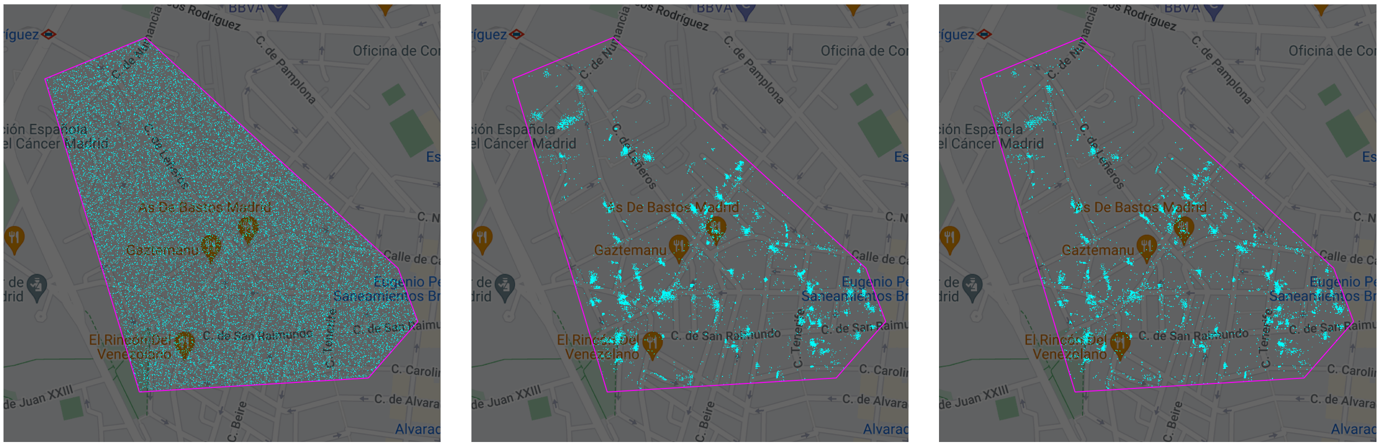
Next, we analyze the activities carried out on those same days, but at other times; specifically, between 09:00 and 15:00 (103 activities). The location results for the two methods (land use and app usage) are shown in Fig. 7. As expected, the proposed algorithm locates the activities in certain areas, without a uniform spatial distribution, showing no bias corresponding to nightlife activities.

Lastly, we analyze the same night time slot, but in the previous week (see Fig. 8). As we can observe, the concentration around the leisure areas related to the festivities is no longer present.

6.2 Case Study 2: University Campus

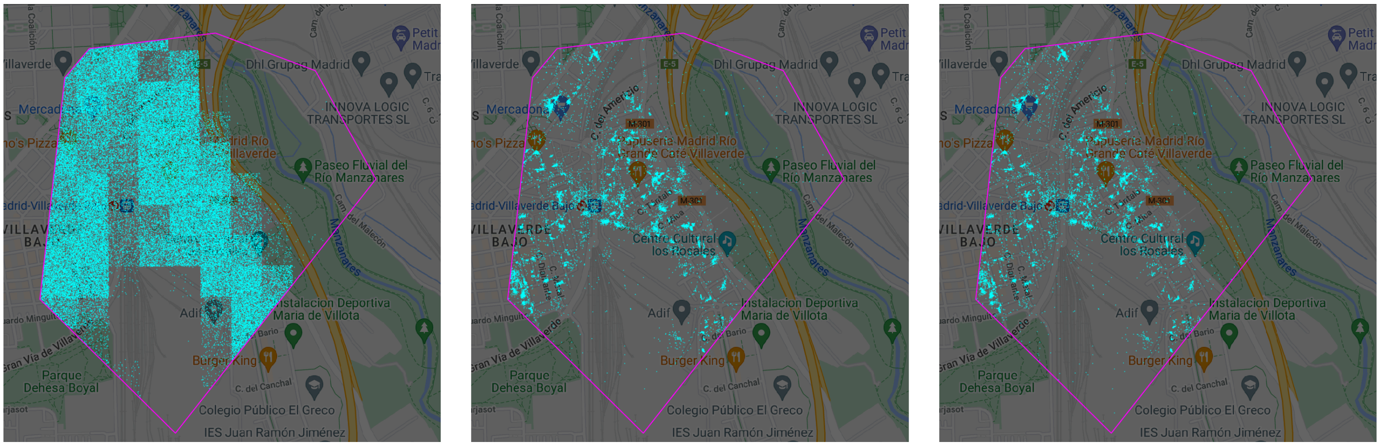
In the second case study, we analyze activities in a university area. Since the month of August is a non-teaching

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3366088



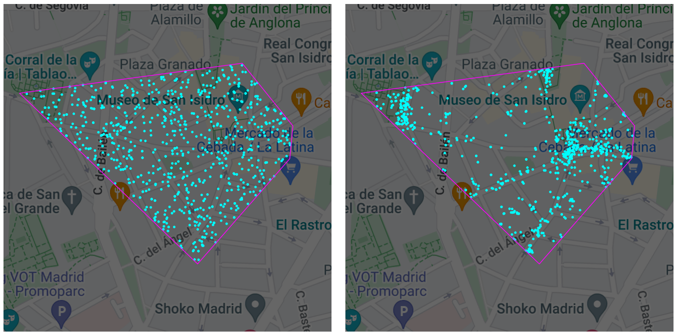
(a) Location based on land use. (b) Location based on app usage (method 1). (c) Location based on app usage (method 2).

Fig. 4. Example of location of activities for a Voronoi polygon with *homogeneous* land use.



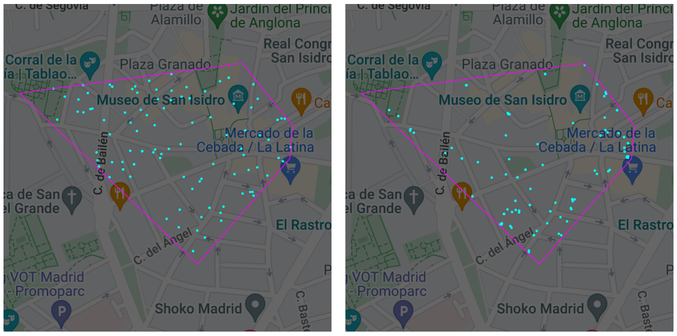
(a) Location based on land use. (b) Location based on app usage (method 1). (c) Location based on app usage (method 2).

Fig. 5. Example of location of activities for a Voronoi polygon with *non homogeneous* land use.



(a) Location based on land use. (b) Location based on app usage.

Fig. 6. Case Study 1 (I): from 14/08/2019 (Wed.) to 18/08/2019 (Sun.); at night, from 20:00 to 06:00 next day.



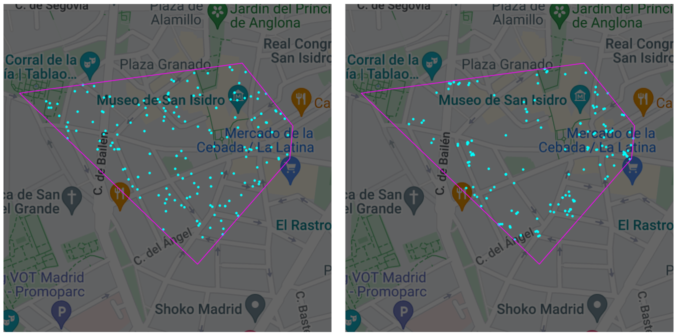
(a) Location based on land use. (b) Location based on app usage.

Fig. 7. Case Study 1 (II): from 14/08/2019 (Wed.) to 18/08/2019 (Sun.); during daytime, from 09:00 to 15:00.

period, we can a priori expect little or no presence of people in the educational centers, along with greater activity in the recreational areas of the campus.

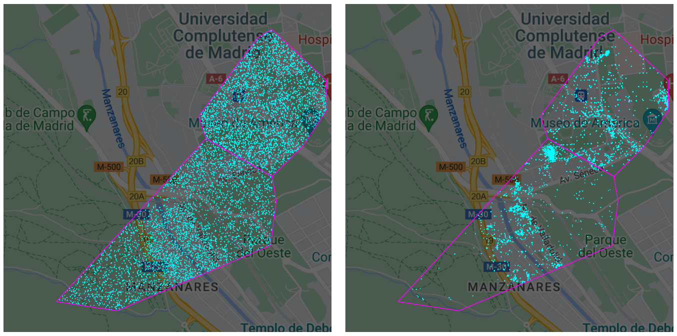
First, we select the activities carried out throughout the month during the day, from 10:00 to 20:00 (18 700 activities). As we can observe in Fig. 9, the algorithm concentrates most of the activities in the swimming pools of the Complutense University and only a few in the educational centers. The other two accumulation points are the San Carlos Clinical Hospital and the access to the subway (Ciudad Universi-

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3366088



(a) Location based on land use. (b) Location based on app usage.

Fig. 8. Case Study 1 (III): from 07/08/2019 (Wed.) to 11/08/2019 (Sun.); at night, from 20:00 to 06:00 next day.

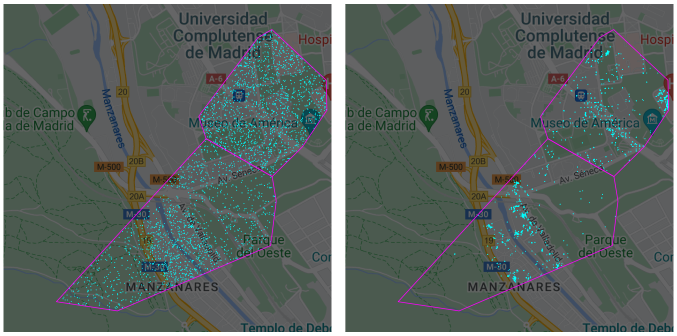


(a) Location based on land use. (b) Location based on app usage.

Fig. 9. Case Study 2 (I): August 2019; during daytime hours, from 10:00 to 20:00.

taria).

Second, we look at the activities carried out throughout the month at night, from 22:00 to 08:00 in the following day (3 854 activities). Now (see Fig. 10), the concentration around the pools of the Complutense University disappears, but those corresponding to the San Carlos Clinical Hospital and the access to the subway (Ciudad Universitaria) remain, at a lower level.



(a) Location based on land use. (b) Location based on app usage.

Fig. 10. Case Study 2 (II): August 2019; at night, from 22:00 to 08:00 next day.

7 CONCLUSION

In this paper we have presented a methodology to refine the location of activities detected with mobile phone records, using GPS data from mobile apps as ancillary information. This methodology can be used to improve the spatial accuracy of the location of activities generated by any data source with some degree of uncertainty using whichever data source that can provide precise geographical location, thus extending the applicability of this work.

Future research lines in this field will focus on applying the proposed methodology to new data sets in order to prove its applicability and scalability. In addition, we will refine previous developments in the transportation area using the enhanced information about users' locations to improve the accuracy in the extraction of origin-destination matrices in both urban and interurban scenarios.

In the absence of a ground truth, we have validated the proposed methods using two case studies in which a higher activity level was expected to happen in specific locations, but not observed using the data set directly obtained from the mobile network data. The performance of the proposed algorithms shown through this validation demonstrates that, despite the reduced sample size and the discontinuous temporal granularity of the data from mobile apps, they have a great potential to improve the spatial granularity of the information obtained from other larger but less spatially accurate data sources, like mobile network data.

This opens the door for the reconstruction of highly detailed activity-travel diaries, through the fusion of mobile network and mobile apps data. Among the wide variety of sectors that would potentially benefit from these findings, transportation and urban planning would enrich the existing knowledge about citizen's mobility in order to optimize the services they provide, at a minimal cost.

ACKNOWLEDGMENTS

This work was partially supported by Comunidad de Madrid and the European Regional Development Fund, under the Research Grant S-2020/L3-736 (SHAPEMOV).

REFERENCES

[1] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem, "Dynamic population mapping using mobile phone data," *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 45, pp. 15 888–15 893, Nov. 2014.

[2] M. Zhang, T. Li, Y. Yu, Y. Li, P. Hui, and Y. Zheng, "Urban anomaly analytics: Description, detection, and prediction," *IEEE Trans. Big Data*, vol. 8, no. 3, pp. 809–826, May/Jun. 2022.

[3] B. Klein, D. Lazer, T. Eliassi Rad, S. V. Scarpino, M. Chinazzi, and A. Vespignani, "Assessing changes in commuting and individual mobility in major metropolitan areas in the United States during the COVID-19 outbreak," Northeastern University, Network Science Institute, Mar. 2020. [Online]. Available: https://www.networkscienceinstitute.org/publications/assessing-changes-in-commuting-and-individual-mobility-in-major-metropolitan-areas-in-the-united-states-during-the-covid-19-outbreak

[4] J. J. Vinagre Díaz, A. B. Rodríguez González, and M. R. Wilby, "Bluetooth traffic monitoring systems for travel time estimation on freeways," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 123–132, Jan. 2016.

[5] M. R. Wilby, A. B. Rodríguez González, R. Fernández Pozo, and J. J. Vinagre Díaz, "Short-term prediction of level of service in highways based on Bluetooth identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 142–151, Jan. 2022.

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3366088

[6] F. M. Naini, O. Dousse, P. Thiran, and M. Vetterli, "Population size estimation using a few individuals as agents," in *Proc. of the 2011 IEEE International Symposium on Information Theory*, St. Petersburg, Russia, 31 Jul. – 5 Aug. 2011, pp. 2499–2503.

[7] M. Versichele, T. Neutens, M. Delafontaine, and N. Van de Weghe, "The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities," *Appl. Geogr.*, vol. 32, no. 2, pp. 208–220, Mar. 2012.

[8] S. Sarkar, S. Chawla, S. Ahmad, J. Srivastava, H. Hammady, F. Filali, W. Znaidi, and J. Borge-Holthoefer, "Effective urban structure inference from traffic flow dynamics," *IEEE Trans. Big Data*, vol. 3, no. 2, pp. 181–193, Apr./Jun. 2017.

[9] R. Fernández Pozo, M. R. Wilby, J. J. Vinagre Díaz, and A. B. Rodríguez González, "Data-driven analysis of the impact of COVID-19 on Madrid's public transport during each phase of the pandemic," *Cities*, vol. 127, pp. 1–12, Aug. 2022.

[10] A. B. Rodríguez González, M. R. Wilby, J. J. Vinagre Díaz, and R. Fernández Pozo, "Characterization of COVID-19's impact on mobility and short-term prediction of public transport demand in a mid-size city in Spain," *Sensors*, vol. 21, no. 19, pp. 1–19, Sep. 2021.

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

[12] D. Wu, Z. Zeng, F. Shi, W. Yu, T. Wu, and Q. Liu, "Human as a service: Towards resilient parking search system with sensorless sensing," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 13 863–13 877, 2022.

[13] M. Lenormand, M. Picornell, O. G. Cantú Ros, T. Louail, R. Herranz, M. Barthelemy, E. Frías Martínez, M. San Miguel, and J. J. Ramasco, "Comparing and modelling land use organization in cities," *Royal Society open science*, vol. 2, no. 12, Aug. 2015.

[14] T. Louail, M. Lenormand, O. G. Cantú Ros, M. Picornell, R. Herranz, E. Frías Martínez, J. J. Ramasco, and M. Barthelemy, "From mobile phone data to the spatial structure of cities," *Scientific reports*, vol. 4, no. 1, pp. 1–12, Jun. 2014.

[15] T. Louail, M. Lenormand, M. Picornell, O. G. Cantú Ros, R. Herranz, E. Frías Martínez, J. J. Ramasco, and M. Barthelemy, "Uncovering the spatial structure of mobility networks," *Nature communications*, vol. 6, no. 1, Jan. 2015.

[16] A. Janecek, D. Valerio, K. A. Hummel, F. Ricciato, and H. Hlavacs, "The cellular network as a sensor: From mobile mhone data to real-time road traffic monitoring," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2551–2572, Oct. 2015.

[17] A. Bassolas, J. J. Ramasco, R. Herranz, and O. G. Cantú Ros, "Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona," *Transp. Res. Part A Policy and Pract.*, vol. 121, pp. 56–74, Mar. 2019.

[18] "Estudio de movilidad con Big Data durante la pandemia," Ministerio de Transportes, Movilidad y Agenda Urbana, 2020. [Online]. Available: https://www.mitma.gob.es/ministerio/covid-19/evolucion-movilidad-big-data

[19] G. Romanillos, J. C. García Palomares, B. Moya Gómez, J. Gutiérrez, J. Torres, M. López, O. G. Cantú Ros, and R. Herranz, "The city turned off: Urban dynamics during the COVID-19 pandemic based on mobile phone data," *Appl. Geogr.*, vol. 134, pp. 1–14, Sep. 2021.

[20] M. Picornell, T. Ruiz, R. Borge, P. García Albertos, D. de la Paz, and J. Lumbreras, "Population dynamics based on mobile phone data to improve air pollution exposure assessments," *Journal of Exposure Science & Environmental Epidemiology*, vol. 29, pp. 278–291, Mar. 2019.

[21] R. A. Becker, R. Cáceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "A tale of one city: Using cellular network data for urban planning," *IEEE Pervasive Comput.*, vol. 10, no. 4, pp. 18–26, Apr. 2011.

[22] Z. Tu, R. Li, Y. Li, G. Wang, D. Wu, P. Hui, L. Su, and D. Jin, "Your apps give you away: Distinguishing mobile users by their app usage fingerprints," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–23, sep 2018.

[23] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, "The promises of big data and small data for travel behavior (aka human mobility) analysis," *Transp. Res. Part C Emerg. Technol.*, vol. 68, pp. 285–299, Jul. 2016.

[24] M. Yin, M. Sheehan, S. Feygin, J.-F. Paiement, and A. Pozdnoukhov, "A generative model of urban activities from cellular data," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 6, pp. 1682–1696, Jun. 2018.

[25] M. S. Iqbal, C. F. Choudhury, P. Wang, and M. C. González, "Development of origin-destination matrices using mobile phone call data," *Transp. Res. Part C Emerg. Technol.*, vol. 40, pp. 63–74, Mar. 2014.

[26] D. Bachir, V. Gauthier, M. El Yacoubi, and G. Khodabandelou, "Using mobile phone data analysis for the estimation of daily urban dynamics," in *Proc. of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, Yokohama, Japan, Oct. 16–19, 2017, pp. 626–632.

[27] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin-destination trips by purpose and time of day inferred from mobile phone data," *Transp. Res. Part C Emerg. Technol.*, vol. 58, pp. 240–250, Sep. 2015.

[28] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Comput. Netw.*, vol. 64, pp. 296–307, May 2014.

[29] J. Blasco Puyuelo, C. Blasco, R. Jordá Muñoz, J. B. nad O. G. Cantú Ros, and D. Mocholí, "Data fusion for the analysis of air travel behavior: Application to Palma de Mallorca Airport," in *Proc. of the 12th SESAR Innovation Days*, Budapest, Hungary, Dec. 5–8, 2022, pp. 1–9.

[30] V. A. Traag, A. Browet, F. Calabrese, and F. Morlot, "Social event detection in massive mobile phone data using probabilistic location inference," in *Proc. of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, Boston, MA, USA, Oct. 9–11, 2011, pp. 625–628.

[31] Y. Dong, F. Pinelli, Y. Gkoufas, Z. Nabi, F. Calabrese, and N. V. Chawla, "Inferring unusual crowd events from mobile phone call detail records," in *Proc. of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Porto, Portugal, Sep. 7–11, 2015, pp. 474–492.

Ana Belén Rodríguez González was born in Plasencia, Cáceres, Spain. She respectively received the B.S. and Ph.D. degrees in Telecommunication Engineering from the Universidad de Valladolid (2000) and Universidad Carlos III de Madrid (UC3M, 2008), and the B.A. degree in Economics from the Universidad Nacional de Educación a Distancia (2002). She was a Professor with UC3M and with Universidad Rey Juan Carlos. In 2013, she joined the Department of Mathematics Applied to Information and Communication Technologies, Universidad Politécnica de Madrid where her research focuses on data analytics, smart cities, energy efficiency in buildings, and intelligent transportation systems.

Javier Burrieza Galán , mobility analytics consultant at Nommon, graduated as Civil and Territorial Engineer from Universidad Politécnica de Madrid (UPM). He holds a double master's degree in Sustainable Urban and Transport Planning from the UPM and the Royal Institute of Technology (KTH). He carried out his Master Thesis on public participation and cycling planning initiatives in Stockholm and Madrid. In January 2018, he joined Nommon, where he conducts transport planning and mobility management projects based on big data solutions.

Juan José Vinagre Díaz received the B.S. degree from the Universidad Politécnica de Madrid (UPM) in 1998 and the Ph.D. degree from the Universidad Carlos III de Madrid (UC3M) in 2005, both in Telecommunications Engineering. He is currently with the Department of Mathematics Applied to Information and Communication Technologies, UPM. He was a Professor with Universidad Rey Juan Carlos and with UC3M. His research focuses on smart cities, intelligent transportation systems, and energy efficiency in buildings. He has led 10 research projects.

Inés Peirats de Castro graduated as Industrial Engineer at the Universidad Carlos III de Madrid. She holds a master's degree in Industrial Engineering from the Universidad Politécnica de Madrid. She started as a business consultant in PricewaterhouseCoopers and joined Goal Systems in 2019. In March 2021, she joined Nommon as a product manager, where she leads the development of decision support platforms for shared mobility operators and transport planning authorities.

Mark Richard Wilby received the PhD in Physics from Imperial College London. He was a Lecturer with the Department of Electronic and Electrical Engineering, University College London; Universidad Carlos III de Madrid; and Universidad Rey Juan Carlos. He was CTO and CEO of several companies. He is currently with the Department of Mathematics Applied to Information and Communication Technologies, Universidad Politécnica de Madrid. His interests are concerned with the processing, understanding, and control of distributed data and sensor systems.

Oliva Garcia Cantu Ros, Chief Research and Development Officer at Nommon, graduated in Physics from the Universidad Nacional Autónoma de México. She holds a Master in Advanced Mathematics from the University of Cambridge and a PhD in Theoretical Physics from Imperial College London. From 2008 to 2012 she worked as a postdoctoral researcher at Universidad Complutense de Madrid and Universidad Carlos III. In October 2012 she joined Nommon, where she applies complex systems theory and AI to the study of social and economic systems. She has participated in national and European research projects.