International Journal of Geo-Information

MDPI

Article

Identification and Classification of Routine Locations Using Anonymized Mobile Communication Data

Gonçalo Ferreira [1,*], Ana Alves [1,2], Marco Veloso [1,3] and Carlos Bento [1]

[1] Centre of Informatics and Systems (CISUC), University of Coimbra, 3030-290 Coimbra, Portugal; ana@dei.uc.pt (A.A.); mveloso@dei.uc.pt (M.V.); bento@dei.uc.pt (C.B.)
[2] Instituto Superior de Engenharia de Coimbra (ISEC), Polytechnic Institute of Coimbra, 3030-199 Coimbra, Portugal
[3] Escola Superior de Tecnologia e Gestão de Oliveira do Hospital (ESTGOH), Polytechnic Institute of Coimbra, 3030-199 Coimbra, Portugal
* Correspondence: gfferreira@student.dei.uc.pt

Abstract: Digital location traces are a relevant source of insights into how citizens experience their cities. Previous works using call detail records (CDRs) tend to focus on modeling the spatial and temporal patterns of human mobility, not paying much attention to the semantics of places, thus failing to model and enhance the understanding of the motivations behind people's mobility. In this paper, we applied a methodology for identifying individual users' routine locations and propose an approach for attaching semantic meaning to these locations. Specifically, we used circular sectors that correspond to cellular antennas' signal areas. In those areas, we found that all contained points of interest (POIs), extracted their most important attributes (opening hours, check-ins, category) and incorporated them into the classification. We conducted experiments with real-world data from Coimbra, Portugal, and the initial experimental results demonstrate the effectiveness of the proposed methodology to infer activities in the user's routine areas.

Keywords: call detail records; clustering algorithms; human mobility; meaningful places; mobile phone data; points of interest

Citation: Ferreira, G.; Alves, A.; Veloso, M.; Bento, C. Identification and Classification of Routine Locations Using Anonymized Mobile Communication Data. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 228. https://doi.org/10.3390/ijgi11040228

Academic Editors: Luca Pappalardo and Wolfgang Kainz

Received: 31 December 2021
Accepted: 25 March 2022
Published: 29 March 2022

1. Introduction

Human mobility has become a prominent research field in recent years. There is a growing need to understand how people move and use urban space in their daily routines. We know for a fact that human trajectories are characterized by a high degree of temporal and spatial regularity. For each individual, there is frequent time-independent travel distance and a significant probability of returning to a few highly frequented locations, with only minor deviations to visit new destinations [1]. These hidden patterns of motion have importance in applications such as urban planning, traffic forecasting and the spread of biological and mobile viruses. Ubiquitous computing has in fact unlocked the potential to determine the personal movements of the masses that previously were only modeled using household surveys and national or regional census.

In today's society, nothing could be more ubiquitous than mobile phones, each of which is a potential sensor providing a constant data stream. On this basis, specialized spatio-temporal datasets such as GPS records have shown enormous potential as a knowledge base about human mobility patterns. In spite of that, due to the overhead of collecting and analyzing such detailed and high frequency location logs at a large scale, many researchers have been urged to explore other data sources as potential proxies for human mobility.

With this in mind, we take special interest in call detail records. A call detail record (CDR) is a form of data that documents how a user interacts in detail with the cellular network. These records contain information fields such as origin/destination tower ID, user ID, time of start and duration on several types of communication. Collecting the

cellular tower ID that the user is connected to and corresponding Cartesian coordinates means that an accurate location is not possible, only an approximation. In the case of CDRs, records do not maintain regular time intervals, unlike GPS, only when a network or user event occurs (e.g., a call is made) the position of the tower used is known. This leads to data that is both spatially sparse and temporally irregular.

There are some compelling reasons for using call detail records (CDRs) versus more accurate location sources. Compared to other location log data types, the overhead for collection and analysis is inferior. Furthermore, no application has to be running, no additional battery life is consumed and data collection cannot be turned off by the user, it is generated by the regular usage of a mobile communication device and stored by the mobile network provider. Potentially every phone in one provider's network can be used as a data source, resulting in enormous amounts of information regarding a substantial percentage of the population that can be used for research purposes.

Despite the significant benefits of scale on this information type, the detail in the locations recorded is a challenge for research efforts focused on individual user analysis. Since we only have the position of a cellular tower, whose range of action spans between a few hundred meters to several kilometers, it is very difficult to accurately pinpoint a user's location. This uncertainty in location makes it very challenging to identify and classify each user's routine places. Although this question of routine places is far more stabilized with detailed information types, such as GPS [2,3], we feel that there is still room for improvement and innovation working with call detail records.

The main objective of this paper is to present an approach for attaching semantic meaning to a user's routine locations, inferring the motivations behind day-to-day mobility. The focus is on classifying activities (e.g., shopping, dining, outdoor recreation) outside of the normal home–work commute. This study uses call detail records from Portuguese citizens. Data were provided for this research by one of the largest telecommunication service providers in Portugal and records are from a four-month period between July and October 2020 by users who had a majority of mobile events in our study area, the district of Coimbra.

Using several types of complementary records, including those captured without human intervention (network-driven events), we hope to surpass some of the issues related to the sparse and irregular time intervals of events. Filling the gaps of mobility traces eases trajectory reconstruction and, as such, facilitates inferring people's routine places. Additionally, by using points of interest (POIs) datasets to classify routine places, we gather a wealth of information that can be used by other studies and applications. Analyzing each user's habits and tastes creates valuable parameters for recommendation systems, adapting marketing campaigns and improving the quality of service by taking into account clients profiling.

This paper is structured as follows: Section 2 comprises the state of the art, where we look over the current best practices and implemented methods. Section 3 provides an outline of the methodology including the research approach, data analysis techniques and description of the used algorithms. In Section 4, we report experiments conducted to test the behavior of the proposed methodology as well as describe and discuss the obtained results. Validation and evaluation are also presented since ground-truth data were made available from the mobile network provider. Finally, Section 5 addresses the conclusions and future work.

2. State of the Art

With the increasing popularity of personal mobile devices and location-based applications, large-scale trajectories of individuals are being recorded and accumulated at a faster rate than ever. Thus, it is possible to understand human mobility from a data-driven perspective. As a consequence of this continuously increasing availability of data, works based on spatial-temporal data have received a lot of interest, with a large spectrum of methods developed.

2.1. Call Detail Records

Despite the listed benefits, the use of this type of data raises questions regarding the validity of previous works and the obtained conclusions, seeing that CDRs provide limited accuracy along the spatial dimension. In fact, studies such as that in [4] were conducted with the objective of proving that identifying users' most significant locations, such as home and work, is possible with a high degree of success. Another research work, around the same theme, Ref. [5] compared CDR-based individual trajectories with reference information from GPS logs. They found these two types of information to match with a good enough accuracy for extracting the user's movements.

The analysis of CDRs has already revealed the spatial recurrence and temporal periodicity of the movement patterns of people, who show a strong tendency to return to previously visited locations [1]. This entails a high predictability potential for human mobility. Similarly, important places in our lives (e.g., home and workplace) can be inferred from CDRs [6,7]. Routine or hobby-related locations have also been detected with success [8–10]. The use of CDR data in travel and tourism is exemplified in [11] where tourism transportation demand in Shanghai is inferred by mobile phone data and a system to propose new routes is developed. Other examples of previous works making use of CDR analyses include the detection and modeling of aggregate mobility flows at large scales [12], the characterization of individual movement patterns [13], or the computation of origin–destination matrices in urban areas [14].

Preserving privacy and data protection is a concern working with what can be considered sensitive personal information. To that effect, the work presented in [15] demonstrated that, with certain protection techniques applied, the re-identification of an individual via frequently visited locations, co-location pairs or spatial temporal data points with a high probability is not possible. Proper storage and retrieval techniques were also discussed in [16]. Some new attempts are now proposing privacy preserving methods for trajectory data based on CDR information, ranging from building recommendation systems locally on a user device [3] or a novel stay-region based anonymization technique that caters to important locations of a user [17].

2.2. Points of Interest

A point of interest (POI) is an entity of interest with a well-defined location. Points of interest can range from famous landmarks (e.g., museums, churches, towers), natural attractions (e.g., bays, coasts, waterfalls) to commonplace spots (e.g., coffee shops, taverns) [18].

The spatial interactions and distributions of POIs reveal different urban functions which can be associated with activities. For different types of activities (e.g., sports, eating, shopping), people can usually go to specific areas. For the same reason, many scientific studies (e.g., [2,3,10,19–21]) have focused on extracting features for area classification. POIs can be collected from various online sources and are frequently available for free through application programming interfaces (APIs) and online map service providers. For example, Qihang et al. [2] proposed to match the visit of a user to a specific registered location using POIs extracted via Foursquare's API. In contrast, Proux et al. [3] mapped information from four different geographic databases (HERE, Foursquare, Grand-Lyon, IGN) and nine different social and cultural databases (PreditHQ, International Showtime, Evenbrite, Songkick, Allevents.in, Meetup, Sportradar, 10 times) to perform user profiling by detecting significant places and their semantic meaning with external sources of information.

2.3. Semantic Disambiguation of CDRs

The process of semantic enrichment and the disambiguation of places has mostly been seen with the use of GPS trajectories. Studies such as [2,3] center their effort on venue check-ins, where the goal is to match a user to a visit of a registered location (e.g., restaurant, hotel, etc.). These works perfectly showcase the main difficulty with trying to disambiguate stop locations without user input. Using CDRs further augments the issues, as the mobility traces are much less precise. Some authors classify geographic areas by categories and

match these areas with the user's routines, instead of trying to match sparse locations with a specific venue.

For example, in [19], a virtual grid with cells of 500 by 500 m was constructed. Fixed-size cells are a common approach for the classification of geographical areas [20–22]. Each grid cell was classified according to four main categories (eating, shopping, entertainment and recreational) and matched with the user's CDR locations. To obtain that classification, the number of POIs associated with each activity category was recorded for each cell, creating an activity distribution map. Each cell activity proportion was then normalized to a value between the 0 and 1 and the K-means clustering algorithm was applied to create four distinct groups. The final step was finding the most probable activity category for each of the k clusters with a probability function. No ground truth data and no validation were carried out, as the authors fixated in their conclusion on the difficulty and privacy concerns in obtaining such data.

In [10], the metropolitan area of Milan was divided into regions by using density clustering to aggregate the groups of proximate cell towers. In sequence, the POIs obtained from foursquare were used to classify these areas by the most frequent type of POIs present. Study area subdivisions were classified by the top level categories in the POI dataset (e.g., shop, food, nightlife). The main focus of the work was to categorize explorers, those who are inclined to break out of their daily mobility routine and explore new places, and find city areas associated with this behavior. Results and conclusions were based on exploratory analysis, finding, for example, the areas and categories most related to exploration and attempting to validate with existing knowledge of the study area.

Experiments were also conducted with Voronoi diagrams, road segmentation layers, transportation analysis zones (TAZ) and administrative layers [23]. This work primarily uses the statistics of CDRs to classify geographical areas and POIs as a complement. First, based on CDR data, they calculate the parameters required including weekday and weekend CDR density spikes, number of peak values, the intensity of peak values, and the distribution of peak values; this allows the travel behaviors and public cognition to be understood. Then, POI category density was used to complement the analysis of CDR events in each geographical division and based on that, the identification results are modified. Validation was made through existing knowledge of the study area and comparing the results obtained to that of the known reality.

Compared to the state of the art, the main novelties introduced by our work are the following: (i) the inference of mobility routines outside of tourist and exploration activities, using a dataset more tailored for this analysis (Facebook Places); (ii) the circular sector approach, to create signal areas relative to each antenna is, according to our research, an innovative way to subdivide space for classification; (iii) in contrast to previous work, our approach relies on multiple POI features (e.g., category percentage, popularity and opening hours) to match location data with geographical area classification.

3. Materials

This section presents, discusses and analyzes the data obtained or gathered in this work. Exploratory data analysis, using statistical graphics and other data visualization methods, was conducted to summarize the CDRs' main characteristics.

3.1. CDR Dataset

For the development work carried out in this paper, a dataset comprising 35,676 SIMs from the region of Coimbra, Portugal, was used. Data collection corresponded to a period of 4 months, from 1 July 2020 to 30 October 2020, totaling 41,371,218 unique events. This amount of data was large enough to experiment with and apply several state-of-the-art techniques and obtain representative results while still being acceptable and manageable in terms of size.

Data entries were a mix between event-driven and network-driven entries, which means that some events did not require user participation being generated periodically

without human intervention. Each entry in the data has: a user identification field; the timestamp of the event; a unique identifier for the cellular antenna; its corresponding location coordinates; the antenna's initial and final angle of action in degrees; as well as the estimated range value in meters.

Before being made available for research, the user's identifiers were pseudonymized. This means phone numbers were encrypted with a hash function. This function remained unknown to us, the researchers, preserving the anonymity of user identity.

Analyzing the obtained Call Detail Records with regard to the events made by each user, we can ascertain the values present in Table 1. For the 35,676 individuals, there is an average of approximately 1346 unique events recorded with a standard deviation of approximately 425. The number of users with more than 2000 events is small with the maximum recorded being 7288.

Table 1. Dataset analysis on number of events per user.

User Count	35,676
Mean Events	1346.32
Std	425.81
Min	1.000
25%	1109.00
50%	1351.00
75%	1561.00
Max	7288.00

Figure 1 represents a histogram of the events per user. As can be seen from Figure 1, the events do not follow a normal distribution. In fact, they have an appreciable positive skewness while peaking at around the 1300 events mark.

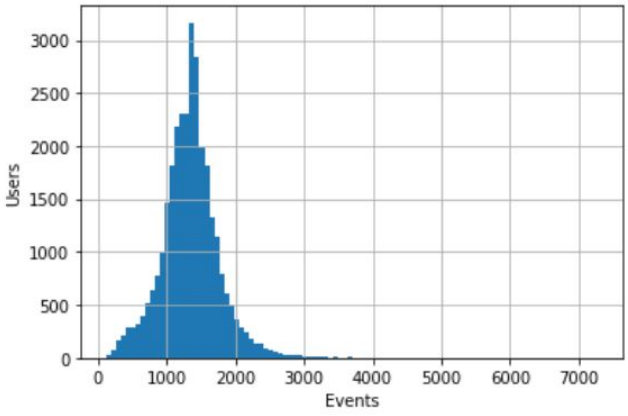

Figure 1. Histogram of the number of events per user.

The period of July/October 2020 represents, at the time of writing, the best possible chance for normal patterns of population movement from the data that can be made available to us since this represents a more relaxed period of COVID-19 confinement in Portugal [24]. Even though we know it will hardly give us an accurate indication of pre-pandemic mobility, it is, however, the closest we obtained since the data collection for this project began.

In search of a better indication of mobility in the data, we conducted an analysis of unique cell towers visited by each user. This would serve as a better indication if this data contains potential for mobility studies, or is compromised by the atypical situation lived throughout the year of 2020.

The information presented in Table 2 and Figure 2 shows that the average user has 25 different locations recorded and the standard deviation is relatively high, which should be expected as there is a big spread of values. This should certify that although time-frame for this type of study is not ideal, the dataset contains a good number of visits for each person.

Table 2. Dataset analysis on unique locations per user.

User Count	35,674
Mean (Unique Locations)	25.309
Std	18.406
Min	1.000
25%	12.000
50%	24.000
75%	34.000
Max	224.000

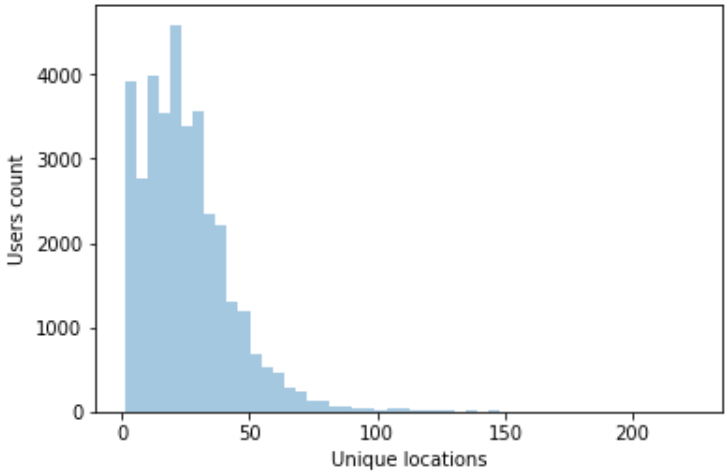

Figure 2. Histogram of unique locations per user.

3.2. POIs Dataset

POIs can be extracted via a single API or obtained and aggregated from several sources. In our particular case, it made more sense to use a single source since places are classified into various categories covering a variety of subcategories, and there are overlapping problems in different datasets, so it would be necessary to reconstruct and reclassify the POI data to join two or more sources.

Facebook is probably the most popular social networking site that makes it easy for people and/or businesses to connect and share with family, friends and clients online. Facebook Places is an associated geolocation service built into Facebook that is designed to help users share their favorite spots and discover new ones. Users can "check in" at various locations, from cities to small stores. Additionally, users are given the ability to create a new POI if the one they intend to 'check-in' or review does not already possess a Facebook Page. Business owners can claim and certify the pages created by a third party by following a verification process.

The main benefit of this data source when compared to Foursquare ([2,10,25]) is the wider reach of the Facebook platform and as such the amount of POIs is increased as expected. There is, in fact, a representation of categories that are not present in Foursquare's database, including organizations, societies, finance and healthcare. These categories, although not as important for tourism or leisure, are important to infer everyday mobility

motivations for the resident population. Furthermore, by observing both datasets, it was perceived that a bigger percentage of Facebook POIs contained information on opening and closing hours.

Furthermore, a dataset had already been constructed for the whole country in a previous work [26]. This allowed access to an extensive offline database using the code provided in the aforementioned work. In total, the dataset has 221,724 unique points spread over hundreds of categories of different hierarchies. An excerpt of the POIs data can be seen in Table 3.

Table 3. Sample of the Facebook Places POI table.

Name	Check-Ins	Hours	Latitude	Longitude
Restaurante Aviz	425	[[8, 0], [9, 0]]	39.82468	−7.4915
AZULMIR	15	[[9, 19], [9, 12]]	40.43211	−8.72678
B-Culture	0	[[9, 19], [9, 13]]	41.45011	−8.33808
...	...	...	...	...
Category	**City**		**Top Category**	
Portuguese Restaurant	Castelo Branco		Food and Beverage	
Wholesale and Supply Store	Mira		Shopping and Retail	
Medical and Health	Guimarães		Medical and Health	
...	...		...	

3.3. User Survey

To carry out validation on predicted user activities, a survey was made by the telecommunications service provider as the information needed was not present and could not be inferred by any data gathered to date. The survey was directed to the existing user pool of the original CDRs dataset in order to compare the knowledge obtained by our methods with reality. Since it was a voluntary questionnaire, it meant that not all users participated. From the total 35,676 users, only 574, or approximately 1.61%, voluntarily gave their answers. This questionnaire included information such as: the professional activity of the client, work schedule, if the client has a second home, where they spend the weekend, main interests/habits, and exercise frequency.

4. Proposed Approach

From this study and the analysis of the state-of-the-art research, we created an initial road map for experimentation and methods. The work can be divided into sections with the final goal being, with CDRs as input, to output a detailed table of routine areas and their classification.

4.1. CDR Pre-Processing

Pre-processing the dataset included retrieving and inferring additional data columns (e.g., the day of the week, workday/weekend) from the existing ones. This was intended to ease the detection of spatio-temporal patterns in the records. An integer for the day of the week (from 0 for Monday to 6 for Sunday) and a Boolean value for the workday or weekend (0 being a workday) were obtained from the timestamp columns. Furthermore, we adopted the time segment division found in [22]. For each entry, taking the timestamp, we verified the corresponding interval. One day is divided into eight time segments to capture the intraday variations in activity participation: early morning (3–6 a.m.); morning—peak hour (6–9 a.m.); morning—work (9 a.m.–12 p.m.); noon (12–2 p.m.); afternoon—work(2–5 p.m.); afternoon—peak hour (5–8 p.m.); night (8 p.m.–12 a.m.); and midnight (12–3 a.m.) [22].

As seen by data exploration in the CDR data description section, there were some cases where users had a lower number of events than average—even those users with less than one event per day. As expected, these will add little to no information for our research purpose, since we seek a higher number of events in order to infer spatial patterns. Thus, we created a simple function that, taking as input an event threshold, filters out all users with a number of events below that threshold. For example, removing users with less than one event per day resulted in a reduction of 0.12% of the dataset or 25,858 unique events.

Another step was the detection of cellular tower reselection in the middle of calls, or in very quick succession, creating impossible trajectories when taking in account the speed of movement. This is due to automatic network load balancing, a phenomenon often called load sharing [7]. With this in mind, distances between the network towers were computed and, consequently, the traveling speeds of users were estimated in consecutive records. For the detection of the load sharing effect, a speed-based method was implemented. A sequence is identified if the tower switching speed exceeds a given threshold. We set the value at 200 km/h inspired by the work of Iovan et al. [27].

After these initial steps and before we could search for routine activity patterns, we needed an accurate identification of each user's home and workplace locations. These are most likely the places where people spend the majority of their time and represent a large portion of their mobile records. Finding these locations first is important because it allows us to focus our attention on relevant records for our research of habits outside of these places. Thankfully this topic has been a subject of many prior studies and there are proven methods with good accuracy.

4.2. Home and Workplace Detection

Motivated by Vanhoof et al.'s work [6], a mixed approach of time filtering and density-based clustering is proposed. Firstly, we selected the temporal intervals of search when someone is not likely to be found in the places we want to identify. In this case, the home time interval was defined as the period from 7 p.m. to 9 a.m. as per [6]. However, because they did not try their method for workplace detection, we defined working hours as the period from 9 a.m. to 5 p.m., a common schedule of 8 h for day workers. Additionally, workplace CDRs were constrained to workdays. Given the state-of-the-art research, we opted for density-based spatial clustering, or DBSCAN, as per the works of [7,8]. DBSCAN is still to this day considered a competent algorithm for grouping CDRs and finding important areas. Its recurrent appearance throughout the literature supported our choice to use it our methodology.

After identifying and excluding homes and workplaces from the individual user's data, we are left with the remaining locations. From these, we then need to understand which are the most relevant to the daily routine, i.e., the most visited ones that account for a substantial time expenditure.

4.3. Other Routine Locations

The chosen method to detect the home and workplace using DBSCAN could also be used to find other routine locations. Without the time restrictions of home/workplace hours and by keeping all the clusters, rather than highlighting the one with the most events, it would be a good candidate solution. The issue found with using this density-based clustering is that we would lose additional precision in pinpointing the exact user position. Antenna locations already have great uncertainty when it comes to matching the user position, and clusters consisting of several antennas would substantially increase the challenge. For routine locations, we want to retain the maximum precision possible. The larger the area, the more difficult it will be to match a specific activity.

Inspired by the work of Quadri et al's [10], which divided users' locations in classes of importance with respect to the number of unique visit days, a similar approach was used. The three classes are: most visited places (MVPs), locations most frequently visited by the user; occasionally visited places (OVP), locations of interest for the user, but only visited

occasionally; exceptionally visited places (EVP): non-routine places. To classify places in these classes, a relevance metric was calculated for each place in the user's records. The initial relevance of a location l for a certain user u: $R(l, u)$ was calculated by the number of unique days that the user visited the location $d_{visit}(l, u)$ over their total number of active days $d_{total}(u)$. As our main goal is not only to detect routine locations but also to infer activities, the relevance metric was modified to accommodate the need for a time window and day type. We separated user places by coordinates, time interval and type of day (workday/weekend). The final metric for calculating the relevance of a location, $R(l_{tm,d}, u)$, is that presented in Equation 1. Instead of counting the unique days that the user visited location l, we counted the unique days that the user visited l in time interval tm and type of day d:

$$R(l_{tm,d}, u) = \frac{d_{visit}(l_{tm,d}, u)}{d_{total}(u)} \quad (1)$$

We used the calculated metric as input to a K-means clustering algorithm, this time with input value k = 3 to obtain the three distinct groups. Figure 3, a 3D scatter plot, shows coordinate points clustering by the relevance metric for one selected person in the data. Note that the Z axis represents the relevance metric while the X and Y are latitude and longitude, respectively. Purple color coded point, with the highest relevance score are MVPs, with orange points being OVPs and blue points EVPs.

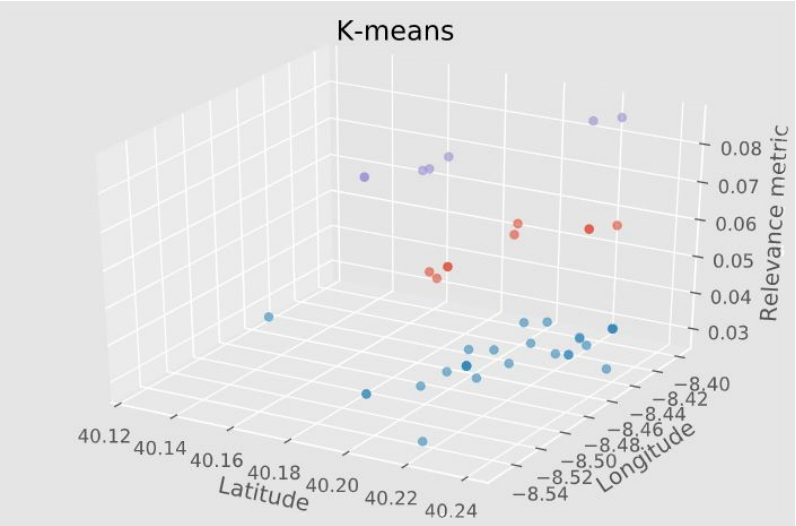

Figure 3. 3D scatter plot of the K-means clustering applied to user locations with K = 3.

Exploration or holiday-related activities (EVPs) do not entail a significant pattern in the data to be considered and are not analyzed further. The idea is that excluding home and work, we find other frequently visited places including MVPs and OVPs, that have significant importance for each user.

4.4. Geographic Regions Classification

To provide better insight into the motivations behind the mobility, at this point, we opted to subdivide the study area and classify the resulting geographic regions with the most likely activity. This is an important step in order to obtain the user's classified routine locations.

The selection of the regions is important as the size and shape can influence the final results. A slightly larger or differently shaped region can encompass more POIs, skewing the activity classification. We needed well-defined region boundaries that represented the search area in order for a function to return all contained POIs. Several approaches were considered, including fixed size ([19–22]) and dynamically sized [23]; however, we proposed a new type of region for activity classification using the antenna's signal attributes.

In our data, we accessed the values of the angle of coverage and the maximum expected range of each telecommunications antenna located in the study area. This resulted in the creation of circular sectors, corresponding to the antenna signal. In our understanding, these regions represent the user position with good estimation, as the user has to be within the boundaries of the signal range in order to connect to the corresponding antenna. Additionally, matching with the routine locations will be easier as these locations are also identified at the antenna level. The identified routine locations are associated with the antenna to which the user is connected, so we can infer that they must be within antenna's signal. An example of an antenna signal area and contained POIs can be seen in Figure 4.

It was necessary to use a specialized points of interest (POIs) dataset as the base data for the region POI mapping. We chose to use Facebook Places, a location-based social network (LSBN) that offers detailed information on 221,724 unique points spread over hundreds of categories of different hierarchies in Portugal [26]. The main attributes of each POI are: the name, Cartesian coordinates, city, opening and closing hours for both workdays and weekends as well as bottom-level category, top-level category and number of check-ins. POIs positioning is defined by the Cartesian coordinates present in the Facebook Places dataset. Although large POIs could theoretically encompass more than one antenna's signal, we only use a single pair of coordinates to infer its presence in those areas. An excerpt of the POI dataset can be seen in Table 3.

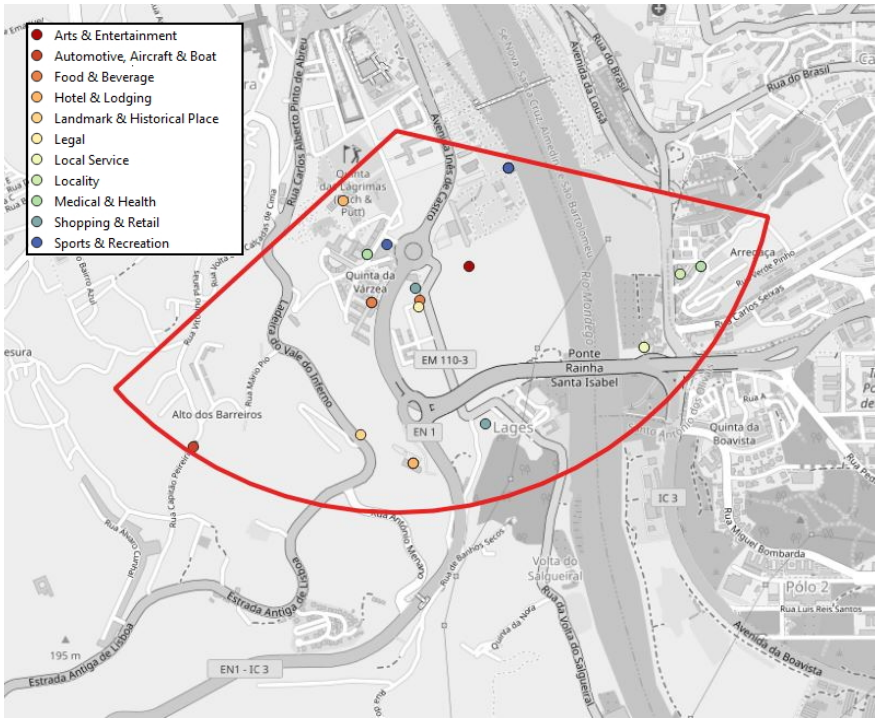

Figure 4. Example of an antenna's signal area.

Running a search function on the POI dataset, we managed to obtain a table with all contained points and respective attributes for each region. To obtain the final output of the classified regions, we need to filter and extract some information from the POIs.

As previously mentioned, in the pre-processing phase, a field was created to divide the CDRs into eight time intervals, as the time of day was taken into account and having a limited number of possible times instead of a continuous timeline facilitates classification. This means that we could classify all regions, at each specific time interval, for each day type (workday/weekend). To filter the POIs, we created a function to check for an intersection between each time interval and the opening hours of each POI. Let (a, b) symbolize the POI opening and closing hours and (x, y) the CDRs time interval; the base rule is to find the cases in which x is in-between (a, b) and the cases where a is in-between (x, y).

In our approach, we do not try to match the sparse user positioning with a single POI or a specific visit. We instead indicate the activity that the user is more exposed to in these locations, depending on the time of visitation and day of the week. Thus, we take into account the percentage of POIs of each category inside the area we are trying to classify. As such, the number of POIs is relevant but only when taking into account the relative percentage of a category when compared with the others. Our confidence measure for classification takes additional features such as opening hours and the popularity of each given POI. This also means that overlapping antennas should have the same activity for the same time periods and not affect the activity prediction for a user.

As points were already assigned to a category by the POI data source, we started by using those as our class labels. There is a problem, however. Lower level categories are sometimes over-specific and need to be grouped into broader classes to increase our chance of an accurate classification. For example, several types of restaurants (e.g., fast-food, Portuguese, Asian) can be grouped under one class label, food and beverage. This is the main reason why we decided to reclassify each POI according to the higher-level Facebook Places categories.

There was still one more value present in the POIs dataset that we could use in the hope of improving results, namely the number of check-ins. A check-in is a user registration of their presence in the location via a social network. As this value is related to the popularity of a given location, it could be inserted into the classification as a weight applied to that class.

From the region POI table, we counted the number of POIs from each label l. Those individual values were then divided by the total number of POIs in the region R, giving us a new area table with the label's percentage. In addition, for each class label l in region R, we sum the number of check-ins and then divided the individual results by the total number of check-ins in the region R. A new column was then added with this percentage to the existing table as can be seen in Table 4. To use both values in the calculation, we multiply the label percentage by the label check-ins to obtain our metric for classification. The resulting equation is written in Equation (2):

$$C(l,R) = \frac{Count_{POIS}(l,R)}{Count_{POIS}(R)} * \frac{Sum_{check-ins}(l,R)}{Sum_{check-ins}(R)} \quad (2)$$

Table 4. Example of a region classification calculation.

Facebook's Top Category	POIs Percentage	Check-Ins Percentage	Result
Food and Beverage	0.125	0.009709	0.001214
Beauty, Cosmetic and Personal Care	0.125	0.019417	0.002427
Religious Organization	0.125	0.064725	0.008091
Medical and Health	0.375	0.284790	0.106796
Shopping and Retail	0.250	0.621359	0.155340

Once we had the regions classified, including all time intervals and day possibilities, we reached the final step where we merged this information with the previous step. The matching key was the corresponding antenna identifier, the *cell id*. This identifier is present in both the region classification table, because regions are associated with an antenna, as well as the identified routine locations. From this merge, we can have a good idea of the user's routine patterns throughout the day, during workdays and weekends.

5. Results and Discussion

In this section, experimental results are summarized and discussed. Some results were validated with ground-truth data and qualitative evaluations were made in cases in which data were not obtained.

5.1. Experimental Results on Regions

Starting with the results in region classification, because our approach for geographic area subdivision does not match any existing administrative or area segmentation diagrams, obtaining ground-truth data for comparison and subsequent validation was not possible. We instead used a discussion approach found in other works with similar objectives [20,23] using knowledge from the study area to make a qualitative analysis of the results.

We take the antenna visualized in Figure 4 with the POI distribution present in Figure 5 as an example. The achieved classification for the antenna is present in Table 5. The predominant activity classification throughout workdays is *Medical and Health* and *Sports and Recreation*. This is due to a high number of POIs pertaining to these categories open in the outlined map region, including care centers, health clinics, a physical fitness center and some outdoor Padel fields. This region also intersects with the locations of some popular restaurants, but is only classified as *Food and Beverage* in the time interval past midnight until 3 a.m., when it is already past closing hours for most other POIs. This is consistent with our knowledge of the region, since POIs related to *Food and Beverage*, which include bars, cafes and restaurants, tend to close at a later hour than health- or sports-related businesses. Furthermore, consistent with known reality is the fact that gyms and sport activities open earlier than other types of POIs, for people that prefer to take part in these activities early in the morning, giving strength to the 6 a.m.–9 a.m. classification. The last point to focus in this area is the change in activity on the weekends between 2 p.m. and 8 p.m. The explanation could be related to the presence of the Hotel Quinta das Lágrimas inside the outlined region, whose gardens are a popular tourist location. As it might be a more frequent choice for a visit during the weekends, the check-ins count could increase its relevance at this particular schedule.

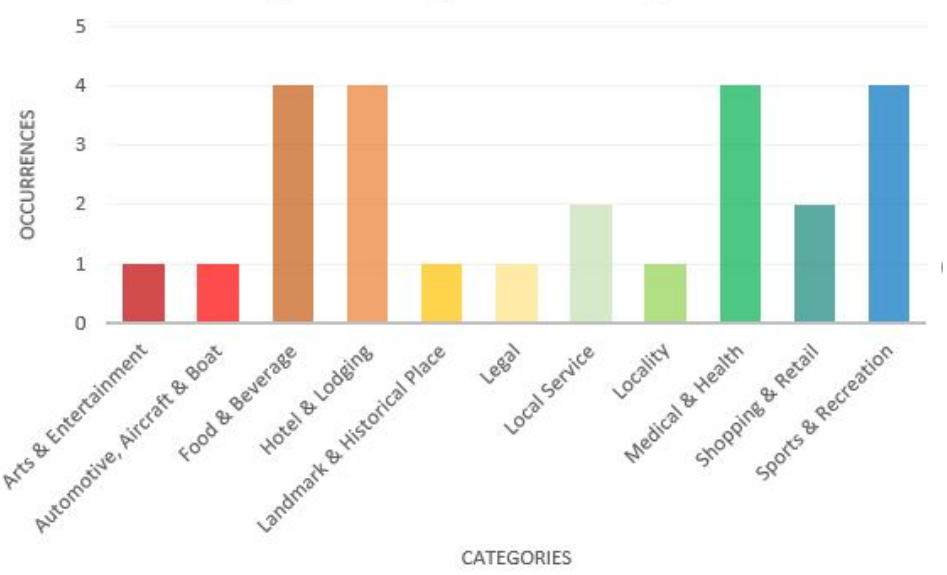

Figure 5. Histogram of high-level categories present in the example antenna of Figure 4.

As this process was repeated with other antennas, subjecting them to the same evaluation, we find that the classifications obtained generally matched our expectations for that particular region. Although the Facebook Places dataset does not contain every existing POI, it contains enough relevant ones to capture the main activities or motivations that might lead an individual to a visit.

Table 5. Region classification by type of day and time interval.

Temporal Interval	Classification	Temporal Interval	Classification
(0, 3)	Food & Beverage	(0, 3)	Food & Beverage
(3, 6)	None	(3, 6)	None
(6, 9)	Sports and Recreation	(6, 9)	Sports and Recreation
(9, 12)	Medical and Health	(9, 12)	Medical and Health
(12, 14)	Medical and Health	(12, 14)	Medical and Health
(14, 17)	Medical and Health	(14, 17)	Hotel and Lodging
(17, 20)	Medical and Health	(17, 20)	Hotel and Lodging
(20, 24)	Sports and Recreation	(20, 24)	Sports and Recreation

5.2. User Routines Validation

To date, excluding home and work, we find other frequently visited places, most visited places (MVPs) and occasionally visited places (OVPs), that have significant importance for each user. In these locations, the time of visitation and the greater offer/popularity of services available in that area may indicate that the user is more exposed to one activity. Even if the user did not visit any of the POIs in the area, the frequency of visitation (not being your home) shows that that place is important to the user, and depending on the time of day, this same place will have different services available.

It is important to reiterate that region classification is done before matching with the detected routine locations. This is the final step in the process, giving us the predicted user activities. This merging of both data tables allows us, for each user, to obtain a visualization similar to the one present in Figure 6, where routine locations are separated and labeled according to the prevalent activity. Some locations might repeat for different time intervals having a different classification. These routines are inferred from 4 months of activity, between July and October of 2020.

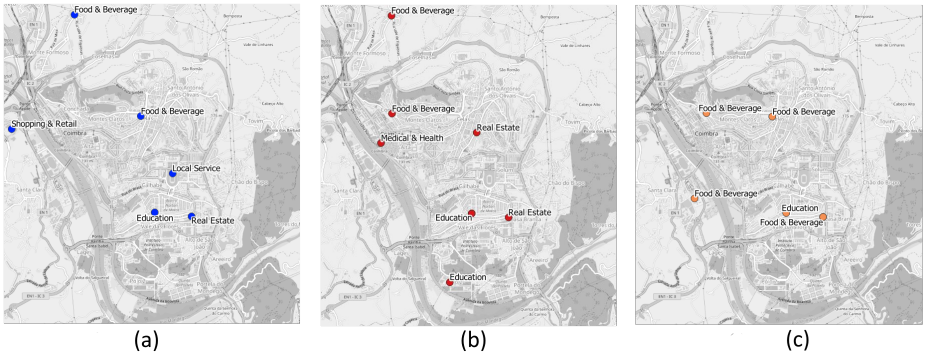

Figure 6. Routine locations classified for one user: (**a**) time interval from 9 a.m.to 12 a.m.; (**b**) time interval from 12 a.m.to 2 p.m.; and (**c**) time interval from 2 p.m. to 5 p.m.

Another way to visualize the activity patterns of users is by grouping the results by the type of day and time interval, as demonstrated in Table 6. This table shows, in each time interval, for both workdays and weekends, what were the activities identified according to the routine places. The user used for this example was the same as that in Figure 6.

Table 6. Example of a user's predicted activities.

Workdays		Weekends	
Temporal Interval	**Activities**	**Temporal Interval**	**Activities**
(9, 12)	[Local Service, Real Estate, Education, Food and Beverage]	(9, 12)	[Real Estate, Shopping and Retail, Food and Beverage]
(12, 14)	[Education, Real Estate, Medical and Health, Food and Beverage]	(12, 14)	[Food and Beverage]
(14, 17)	[Beauty, Cosmetic and Personal Care, Real Estate, Education, Food and Beverage]	(14, 17)	[Food and Beverage]
(17, 20)	[Real Estate, Education, Food and Beverage, Medical and Health, Local Service]	(17, 20)	[Food and Beverage]
(20, 24)	[Food and Beverage, Education, Medical and Health, Shopping and Retail]	(20, 24)	[Food and Beverage]

The main question of interest in the survey had users choose from a list of habits which ones were part of their weekly routine. Results could then be compared between our predictions of the user activities (as present in Table 6) and their given answers. However, the habits present in the question did not exactly match with our classification labels, so to that end, a match had to be made between the two. For example, the Facebook Places category 'arts and entertainment' contains the POIs associated with movie theaters and theatrical plays so we connected it to the survey interest of 'theatre and cinema'. The process was repeated for all six main survey interests: sports; automotive; bars and restaurants; beauty and fashion; theatre and cinema; and travel.

To validate our methods, we transformed the survey interest columns from the several possible answers (yes; no; do not know; blank) into true or false Boolean values. Accordingly, using the matching table, we can verify whether the real interests in an assigned true value are present in our predictions. If we find a match between the user answers and our predictions, we consider it a correct prediction, meaning that accuracy values can be calculated. Figure 7 shows the accuracy results for all six categories. Some categories' performances are worse than others, such as 'bars and restaurants'. This can mostly be explained by the fact that there are many POIs pertaining to this category which lead to an overestimation in the prevalence of this activity.

As concluded in the work of [25], one of the main factors that negatively impact user activity prediction is spatial uncertainty. An uncertain user position means a larger area of search for POIs, giving less importance to those actually visited, which could result in incorrect predictions. Consequently, we conducted experiments in order to reduce the spatial uncertainty by removing regions above a certain threshold of radius. Our x axis in Figure 7 is the average region radius for several radius thresholds. As revealed by the observed values, accuracy tends to increase for certain categories when the average radius is lower. In other cases, accuracy maintains or fluctuates its value very slightly. The correlation between accuracy and spatial granularity resemble that found in [25]. We can infer that if we were to obtain even more precise and relevant regions, results could improve further.

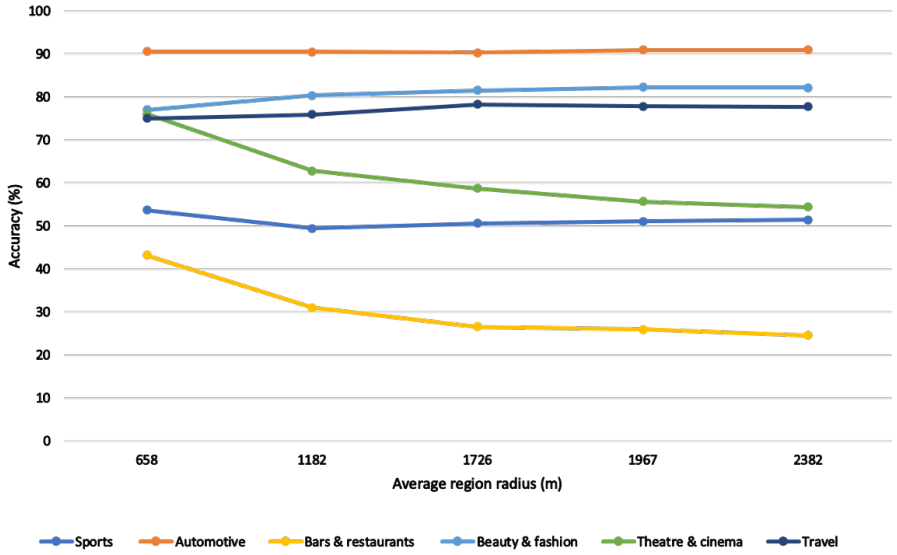

Figure 7. Accuracy of predictions in relation to average antenna's radius.

6. Conclusions

In this work, we addressed the objective of identifying and classifying routine locations using data from anonymized mobile communications events. We presented a group of interacting methods and developed a system to analyze individual mobility behavior. Such an analysis relies on the locations present in call detail records (CDRs) and intends to improve the understanding of user's regular places and their associated routine activities. Throughout the work, we focused on several main steps, data pre-processing, the detection of home and workplace, the detection of other routine places and finally, classification of geographic regions.

The review of the state of the art was essential to understanding the challenges faced by others and potentially explore parts that could be improved in those works. The circular sector approach, to create signal areas relative to each antenna is, according to our research, an innovative way to subdivide space for classification. This allows, in our understanding, for an improved match of the area where a user is when a mobile telecommunication event is made.

Comparing our results with those present in the state of the art, we conclude that for similar categories and taking into account an approximate spatial granularity, our work is on par with state-of-the-art methods being slightly above or below depending on the activity we are trying to identify.

However, we also faced some challenges. Due to the fact that data collection by the telecommunication service provider started in July 2020, the data provided originated from a time period during a global pandemic. Even though the months between July and October of 2020 had a higher user mobility than the stricter confinement period that followed, we know that they do not possess an accurate representation of pre-pandemic mobility.

It is important to reiterate some limitations of this work. For one, we assume that, during the period of study, we are not dealing with phones shared by more than one user and that no particular user changes or has more than one SIM card. Furthermore, there is the possibility that users, for any reason, did not make or receive calls in their workplace or home, precluding the correct detection of these places. The pandemic situation creates further possibilities that users mostly worked from home during the period of data collection.

Future Work

Some possible improvements were found by conducting the analysis of area's activity classification. Manually giving a weight to POIs of certain types to increase their importance depending on the time of day, e.g., for restaurants at regular meal hours, would possibly change the activities to better mirror population tendencies. The same effect could also be achieved with a popularity/check-in value that was hour dependent, but as far as we know, no POI dataset contains this information. There is still the question of points missing from the used dataset, as they might not be registered in the used data provider. One possible solution would be the combination of several POI datasets with the added difficulty of merging completely different category hierarchies into one.

The telecommunications service provider data currently contain cells from 2G to 4G; however, we do not differentiate between these cells. We understand that the signal areas of antennas of different technologies overlap. However, in our approach, these cells will have the same classification, and should not affect the predicted user routines we would like to explore a way to merge overlapping antennas and their records.

In the future, in addition to considering the frequency of visitation, its temporal distribution (every day, weekly, biweekly) could be a factor to take into account regarding the type of activity.

The areas created for classification, although closer to the reality of where the user might be, still remain too large to have a good percentage of certainty in terms of user activity. Newer information sources that have been discussed with the telecommunication service provider for future work have the potential to improve the user's location even further. Doing so allows for a smaller search area and generally more accurate methods. The arrival of 5G networks, with more precise smaller radius antennas, could be the next evolution step in mobility analysis using call detail records. All methods and created and implemented algorithms have the foresight of easy adaptation for future technologies allowing continuation work to be carried out.

Author Contributions: Conceptualization, Ana Alves, Marco Veloso and Carlos Bento; formal analysis, Gonçalo Ferreira; investigation, Gonçalo Ferreira; methodology, Gonçalo Ferreira and Ana Alves; project administration, Carlos Bento; resources, Carlos Bento; software, Gonçalo Ferreira; supervision, Ana Alves, Marco Veloso and Carlos Bento; validation, Gonçalo Ferreira; visualization, Gonçalo Ferreira; writing—original draft, Gonçalo Ferreira; writing—review and editing, Ana Alves and Marco Veloso. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of the data. Data were obtained from a third party and are available from the authors with the permission of said third party.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
CDRs	Call Detail Records
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
EVPs	Exceptionally Visited Places
GPS	Global Positioning System
MVPs	Most Visited Places
OVPs	Occasionally Visited Places
POIs	Points of Interest

References

1. Gonzalez, M.C.; Hidalgo, C.; Barabasi, A.L. Understanding Individual Human Mobility Patterns. *Nature* **2008**, *453*, 779–782. [CrossRef] [PubMed]
2. Gu, Q.; Sacharidis, D.; Mathioudakis, M.; Wang, G. Inferring Venue Visits from GPS Trajectories. In Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA, 7–10 November 2017; Association for Computing Machinery: New York, NY, USA, 2017. [CrossRef]
3. Proux, D.; Roulland, F. Mobile Recommendation Challenges within a Strong Privacy Oriented Paradigm. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Recommendations, Geosocial Networks and Geoadvertising, Chicago, IL, USA, 2019; Association for Computing Machinery: New York, NY, USA, 5–8 November 2019. [CrossRef]
4. Zhang, D.; Huang, J.; Li, Y.; Zhang, F.; Xu, C.; He, T. Exploring Human Mobility with Multi-Source Data at Extremely Large Metropolitan Scales. In Proceedings of the 20th Annual International Conference on Mobile Computing and Networking, Maui, HI, USA, 7–11 September 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 201–212. [CrossRef]
5. Ranjan, G.; Zang, H.; Zhang, Z.L.; Bolot, J. Are Call Detail Records Biased for Sampling Human Mobility? *Mob. Comput. Commun. Rev.* **2012**, *16*, 33–44. [CrossRef]
6. Vanhoof, M.; Reis, F.; Smoreda, Z.; Ploetz, T. Detecting home locations from CDR data: Introducing spatial uncertainty to the state-of-the-art. *arXiv* **2018**, arXiv:1808.06398.
7. Ayesha, B.; Jeewanthi, B.; Chitraranjan, C.; Perera, A.S.; Kumarage, A.S. User Localization Based on Call Detail Record. In *Intelligent Data Engineering and Automated Learning–IDEAL 2019*; Yin, H., Camacho, D., Tino, P., Tallón-Ballesteros, A.J., Menezes, R., Allmendinger, R., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 411–423.
8. Yang, P.; Zhu, T.; Wan, X.; Wang, X. Identifying Significant Places Using Multi-Day Call Detail Records. In Proceedings of the 2014 IEEE 26th International Conference on Tools with Artificial Intelligence, Washington, DC, USA, 10–12 November 2014; pp. 360–366. [CrossRef]
9. Isaacman, S.; Becker, R.; Cáceres, R.; Kobourov, S.; Martonosi, M.; Rowland, J.; Varshavsky, A. Identifying Important Places in People's Lives from Cellular Network Data. In *Pervasive Computing*; Lyons, K., Hightower, J., Huang, E.M., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 133–151.
10. Quadri, C.; Zignani, M.; Gaito, S.; Rossi, G.P. On Non-Routine Places in Urban Human Mobility. In Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 1–3 October 2018; pp. 584–593. [CrossRef]
11. Qian, C.; Li, W.; Duan, Z.; Yang, D.; Ran, B. Using mobile phone data to determine spatial correlations between tourism facilities. *J. Transp. Geogr.* **2021**, *92*, 103018. [CrossRef]
12. Csáji, B.C.; Browet, A.; Traag, V.; Delvenne, J.C.; Huens, E.; Van Dooren, P.; Smoreda, Z.; Blondel, V.D. Exploring the mobility of mobile phone users. *Phys. A Stat. Mech. Appl.* **2013**, *392*, 1459–1473. [CrossRef]
13. Hess, A.; Marsh, I.; Gillblad, D. Exploring communication and mobility behavior of 3G network users and its temporal consistency. In Proceedings of the 2015 IEEE International Conference on Communications (ICC), London, UK, 8–12 June 2015; pp. 5916–5921. [CrossRef]
14. Lenormand, M.; Picornell, M.; Cantú-Ros, O.G.; Tugores, A.; Louail, T.; Herranz, R.; Barthelemy, M.; Frías-Martínez, E.; Ramasco, J.J. Cross-Checking Different Sources of Mobility Information. *PLoS ONE* **2014**, *9*, e105184. [CrossRef] [PubMed]
15. Ding, J.; Ni, C.C.; Gao, J. Fighting Statistical Re-Identification in Human Trajectory Publication. In Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA, 7–10 November 2017; Association for Computing Machinery: New York, NY, USA, 2017. [CrossRef]
16. Vancea, F.; Vancea, C.; Popescu, D.; Zmaranda, D.; Gabor, G. Secure Data Retention of Call Detail Records. *Int. J. Comput.* **2010**, *5*, 961–967. [CrossRef]
17. Yang, J.; Dash, M.; Teo, S. PPTPF: Privacy-Preserving Trajectory Publication Framework for CDR Mobile Trajectories. *ISPRS Int. J. Geo Inf.* **2021**, *10*, 224. [CrossRef]
18. Khosrow-Pour, D.B.A. *Encyclopedia of Information Science and Technology*, 3rd ed.; IGI Global: Hershey, PA, USA, 2015. [CrossRef]
19. Phithakkitnukoon, S.; Horanont, T.; Di Lorenzo, G.; Shibasaki, R.; Ratti, C. Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data. In *Human Behavior Understanding*; Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 14–25.
20. Zhou, X.; Liu, J.; Yeh, A.G.O.; Yue, Y.; Li, W. The Uncertain Geographic Context Problem in Identifying Activity Centers Using Mobile Phone Positioning Data and Point of Interest Data. In *Advances in Spatial Data Handling and Analysis: Select Papers from the 16th IGU Spatial Data Handling Symposium*; Springer International Publishing: Cham, Switzerland, 2015; pp. 107–119. [CrossRef]
21. Wang, F.; Chen, C. On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transp. Res. Part C Emerg. Technol.* **2018**, *87*, 58–74. [CrossRef] [PubMed]
22. Diao, M.; Zhu, Y.; Ferreira, J.; Ratti, C. Inferring individual daily activities from mobile phone traces: A Boston example. *Environ. Plan. Plan. Des.* **2015**, *43*, 920–940. [CrossRef]
23. Yuan, G.; Chen, Y.; Sun, L.; Lai, J.; Li, T.; Zhuo, L. Recognition of Functional Areas Based on Call Detail Records and Point of Interest Data. *J. Adv. Transp.* **2020**, *2020*, 1–16. [CrossRef]
24. Available online: https://www.portugal.gov.pt/pt/gc22/governo/comunicados-do-conselho-de-ministros?p=13 (accessed on 13 February 2021).

25. He, R.; Cao, J.; Zhang, L.; Lee, D. Statistical Enrichment Models for Activity Inference from Imprecise Location Data. In Proceedings of the IEEE INFOCOM 2019-IEEE Conference on Computer Communications, Paris, France, 29 April–2 May 2019; pp. 946–954. [CrossRef]
26. Andrade, R.; Alves, A.; Bento, C. POI Mining for Land Use Classification: A Case Study. *Int. J. Geo Inf.* **2020**, *9*, 493. [CrossRef]
27. Iovan, C.; Olteanu-Raimond, A.M.; Couronné, T.; Smoreda, Z. Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies. In *Geographic Information Science at the Heart of Europe*; Springer International Publishing: Cham, Switzerland, 2013; pp. 247–265. [CrossRef]