A Density-Based approach to identify Home location for College Communities from Big Data

Iván Mendoza[1,*], Andrés Baquero[1], Gustavo Álvarez[1]

Abstract

In the era of big data, the abundance of information presents unprecedented opportunities to gain valuable insights into human behavior and activities. This study focuses on harnessing the power of data collected from mobile applications used by students at a local college, with the aim of identifying the location of their homes. By filtering these locations, the study seeks to improve the understanding of mobility patterns and the most frequent points of interest among students, to obtain useful information for decision-making in mobility-related issues and transportation planning in the surrounding areas. In this paper, a density-based heuristic is proposed to detect home location from big data, which are collected in real-time by a mobile application. The results were satisfactory and their precision seems to increase as more data is acquired, which allows providing insights about potential future applications.

Keywords: Big data, data mining, home detection, mobility patterns, points of interest

1. Introduction

Gaining insights into the mobility patterns and points of interest among college students is crucial in developing effective strategies to optimize transportation for both students and the entire city. In the scope of this study, our objective is to identify and analyze the precise locations of student residences from a college community, utilizing mobile app data.

Mobile phone data can be a useful source for official statistics, but there are challenges and uncertainties that need to be addressed before they can be used. Detecting home locations from mobile phone data and analyzes the performance of five home detection algorithms, offering recommendations for more reliable use of mobile phone data in official statistics.

Some studies like [5] demonstrates the potential of using social media data, specifically Twitter, for studying urban mobility patterns. However, the authors

*Corresponding author

Email addresses: imendoza@uazuay.edu.ec (Iván Mendoza), obaquero@uazuay.edu.ec (Andrés Baquero), galvarezc@uazuay.edu.ec (Gustavo Álvarez)

note the limitations of the data, such as biases in social media usage and the need for ongoing validation and improvement of the methods used. In other study [6] about Home Detection Algorithms (HDAs) using mobile phone data for official statistics. The authors found that the type of data stream used, and the algorithm choice significantly influenced the accuracy of home detection. Algorithms based on weekdays and/or nighttime records were the most accurate and using Extended Detail Records (XDRs) with specific algorithms yielded the best results. Daytime records and spatial perimeter-based algorithms had low accuracy and should be avoided. XDRs and Cellular Positioning Records (CPRs) were more resilient to data reduction compared to Call Detail Records (CDRs). The study had limitations in sample size and lack of demographic information. In [1] concludes that mobile positioning data can be used to monitor population geography and mobility, particularly in socially unstable or rapidly developing regions. However, further work is needed to standardize the data and model for different sources and conditions. The methodology is promising for geographical research and offers opportunities for real-time monitoring tools, geographical applications, tourism development, traffic management, urban planning, and optimizing network services. Detecting the location of households from mobile phone data is a key challenge for the use of this data source in official statistics. The authors argue that current address detection methods suffer from a lack of consensus on criteria and limited validation capabilities. The article presents an analysis of five address detection algorithms applied to a large French Call Detail Record (CDR) dataset, showing that the choice of criteria in Home Detection Algorithms (HDAs) influences address location detection in up to 40% of users, and that HDAs perform poorly when compared to a validation dataset. In [3] the authors used mobile phone data and metadata, including call detail records (CDRs) enriched with gender, socioeconomic segment, and number of phone lines registered under that number. then analyzed the data to reveal a gender gap in mobility and mapped this mobility gap over administrative divisions to observe the association with lower income and lack of public and private transportation options. The paper concludes that there is a gender gap in urban mobility, where women visit fewer unique locations than men and distribute their time less equally among such locations. This mobility gap is associated with lower income and lack of public and private transportation options. In [8] the paper investigates the performance and capabilities of five popular criteria for home detection based on a very large mobile phone dataset from France. The study shows that most Home Detection Algorithms (HDAs) suffer from "blind" deployment of criteria to define homes and from limited possibilities for validation. The paper introduces a data-driven framework to assess the spatial uncertainty related to the application of HDAs. The findings appropriate spatial uncertainty in HDA and, in extension, for detection of meaningful places. The study shows how spatial uncertainties on the individuals' level can be assessed in absence of ground truth annotation, how they relate to traditional, high-level validation practices and how they can be used to improve results for, e.g., nation-wide population estimation. Therefore, the paper concludes that the proposed framework can be used to improve the

accuracy of home detection algorithms and population estimation. In [2] the authors discuss the importance of knowing user location in the digital world, not only in real-time but also predicting future locations. It is necessary to semantically label a place, particularly detecting the most probable home location for a given user. The paper aims to provide insights on the differences among the ways how different types of human digital trails represent actual mobility patterns and the differences between the approaches interpreting those trails for inferring said patterns. The paper starts with an example showing how human mobility patterns described by means of radius of gyration are different for Flickr social network and dataset of bank card transactions. The paper considers several methods for home location definition known from the literature and demonstrates that although for bank card transactions they provide highly consistent results, home location definition detection methods applied to Flickr dataset happen to be way more sensitive to the method selected, stressing the paramount importance of adjusting the method to the specific dataset being used.In universities, there is a need to enhance student mobility, and various strategies have been suggested to achieve this goal. To implement effective policies, it is crucial to have dynamic origin-destination matrices (OD) that represent different scenarios. The widespread use of mobile devices has made it possible to automatically extract daily travel data through tracking apps, eliminating the need for traditional trip surveys. Mendoza et. al [4] presents a new methodology for extracting multi-day mobility demand in universities using logs obtained from dedicated apps regularly used by students. The proposed approach was evaluated using real-life logs from a representative group of students over a five-month period. The results showed that this approach is effective in obtaining average demand data, which can be utilized in planning mobility strategies, as long as continuous tracking of mobility data is feasible through mobile devices.

The primary objective of this study is to filter out student homes from the collected mobile app data and categorize them. By accurately identifying these locations, we can lay the foundation for a more comprehensive analysis of student mobility patterns and points of interest.

This research is part of a larger investigation focused on understanding the unique needs and preferences of a college community. By identifying their homes, we gain crucial insights into their daily commute patterns, transportation modes, and the places they frequent. Such information enables the college authorities and municipal stakeholders, to make data-driven decisions and develop tailored mobility plans that cater to the specific needs of the student population.

In this paper, we present the methodology used to collect and preprocess the mobile app data, with a specific focus on identifying student homes. We discuss the techniques employed to filter and categorize these locations accurately. Additionally, we delve into the analysis of student mobility patterns and points of interest derived from the identified residences. The results and their implications for decision-making are presented and discussed in detail. The findings from this study will provide valuable to the understanding their

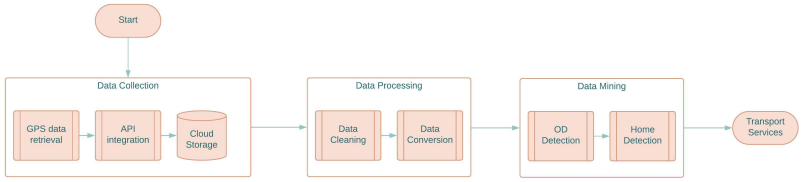


Figure 1: Multi-step procedure methodology for Big Data Home Detection.

mobility patterns, decision-makers can develop targeted interventions, optimize transportation systems, and create a supportive environment for the student community. This research underscores the significance of leveraging mobile app data to gain a comprehensive understanding of student mobility patterns and points of interest, facilitating evidence-based decision-making and the development of effective mobility plans.

For this approach, a density-based heuristic is proposed to detect home location from big data, which are collected in real-time by a dedicated mobile application developed for this purpose. This document is organized in the following way: a literature review of related works about origin-destination (OD) and home detection; then a detailed explanation of the methodology. At the end, conclusions, then a discussion of results and future works are proposed to provide insights about potential future applications and research.

2. Methodology

The Resulting locations of the student homes have to be mined from mobility data after these are processed following a multi-step procedure. The required steps are summarized through the flow chart found in Figure 1, then a deeper explanation follows in the next sections.

In a few words, data are collected by a mobile app which uses the Global Positioning System (GPS) sensors of the mobile device, so that they can be stored in a remote server via an Application Programming Interface (API). Then, these unprocessed data are cleaned to filter out outliers and transformed to a suitable format for further calculations. At last, origins and destinations (OD) are detected by some segmentation algorithm and later classified as potential home locations. Some potential transport services after acquiring this information are planned out to complete the flow.

2.1. Mobile Data Collection

The first step, assuming a mobile app exists so that it can collect GPS data, is storing mobility data together with temporal and identity information in a remote database. Each observation i (a labeled spatiotemporal data point) will have the following structure:

$$p_i = (x_i, y_i, t_i)$$

that is, in its simplest form, it consists on the location's coordinates (x, y) of the tracked user at a specific time t. A more complete version of the observation is:

$$p_i = (uid_i, lat_i, lon_i, alt_i, date_i, t_i, dow_i, acc_i, timestamp_i)$$

- *uid*, is a user's unique identifier, normally a MD5 hash string which guarantees data is collected anonymously, but it still makes possible to know the user in the database.
- *lat. lon, alt*, the point's latitude, longitude and altitude coordinates.
- *date*, a day-month-year string
- *t*, a 24-hour local time format for a continuous variable after applying this formula:

$$t = hours + \frac{minutes}{60} + \frac{seconds}{3600}$$

- *dow*, the day of the week the observation was captured, where 1 is Sunday
- *acc*, accuracy in meters of the measurement as reported by the sensor API, lower values produce more accurate measurements.
- *timestamp*, A Unix based time stamp, allowing to treat dates as continuous variables.

The set P_a of data point observations collected in real-time by a user's mobile device, is stored in this format in a remote server for further off-line procedures. However, this format is not yet suitable for most calculations, so that further processing is required as described in the next steps.

2.2. Data Processing

In order to avoid bias in the calculation of travel destinations, outliers are removed considering position accuracy and temporal constraints. A subset P_a consists of more refined "valid" observations, selected through the following criteria.

$$P_b = \{p_i \in P_a | acc_i < \alpha\}$$

where α is a parameter that denotes the maximum allowed accuracy error in meters. Moreover, users can travel to destinations found in other regions, countries and even continents. For the purpose of this research, home locations must be found inside a study region at a medium-sized city level, so a bounding box defined by $[lon_{min}, lon_{max}, lat_{min}, lat_{max}]$ constraints the valid observation previously found.

$$P_c = \{p_i \in P_b | (lat_{min} < lat_i < lat_{max}) \wedge (lon_{min} < lon_i < lon_{max})\}$$

Because this research is intended to provide insights about transport services for students in a College Community, a time period where users are expected to regularly visit the campus must be selected. Then, a cutoff date defined by $[date_{min}, date_{max}]$ is used to select the final subset P.

$$P = \{p_i \in P_c | (date_{min} < date_i < date_{max})\}$$

At last, since OD detection requires computation of distances, a Cartesian projection is a better approach. For the selected study region the UTM-17S is chosen, so that longitude, latitude and altitude measures are transformed to euclidean space coordinates (x, y, z). Since dates and accuracies are not needed anymore, each observation p_i in the resulting processed set P becomes:

$$p_i = (uid_i, x_i, y_i, z_i, t_i, dow_i)$$

2.2.1. Dataset Description

The full dataset used in this study consists of spatiotemporal data, collected by a dedicated tracking mobile app for a College Community. It involves data from *728* users during one moth period, namely from May 20 to June 20, 2023. The results of the further analysis have proven 30 days of mobility data per user to be sufficient to identify home locations. The number of observations after the data cleaning process is above *11 million*, so that good infrastructure of the cloud storage is required to handle this volume of big data.

A sample from the spatiotemporal dataset is presented below in Table 1 for the five attributes mentioned earlier:

Table 1: Sample observations from the Spatiotemporal Dataset

uid	x	y	z	t	dow	timestamp
614820f9e6	717002.0	9677062	2616.9	23.29667	2	1685402268
614820f9e6	717002.0	9677062	2616.9	23.29667	2	1685402268
614820f9e6	717002.0	9677061	2616.9	23.29917	2	1685402277
614820f9e6	717002.3	9677061	2616.9	23.30222	2	1685402288
614820f9e6	717002.3	9677061	2616.9	23.30222	2	1685402288
614820f9e6	717002.8	9677062	2616.9	23.30444	2	1685402296
614820f9e6	717007.8	9677053	2616.9	23.30500	2	1685402298
614820f9e6	717007.8	9677053	2616.9	23.30500	2	1685402298
614820f9e6	716998.5	9677059	2616.9	23.33444	2	1685402404
614820f9e6	716998.5	9677059	2616.9	23.33444	2	1685402404

The bounding box for the region of Cuenca, Ecuador contains longitudes between -79,084789233 and -78,933588295, and latitudes between -2,938030323 and -2,865347073. A picture of the sample of collected points on a map at scale 1:50000 is given in Figure 2, where it is shown that complete travel trajectories can be retrieved from data. The sampling frequency of the GPS, that is the time difference between measures was not fixed but around 2 seconds on average.

Figure 2: Sample of collected points for the region of Cuenca, Ecuador.

2.2.2. Computing new attributes

Some additional features must be added to the existing dataset so that travel behavior can be evaluated.

Let p_i be the i^{th} observation in a spatiotemporal dataset P. Then, the cumulative distance $D_{a,n}$ for a trajectory starting at point a and consisting of the next n observations is defined as:

$$D_{a,n} = \sum_{i=a+1}^{a+n} d_{i,i-1}$$

where $d_{i,j}$ is the euclidean distance between two observations.

$$d_{i,j} = ((x_i - x_j)^2 + (y_i - y_j)^2)^{1/2}$$

that is, the sum of distances between proximate points; then, instant speed at observation i^{th} is computed by:

$$s_i = \frac{d_{i,i-1}}{t_i - t_{i-1}}$$

The resulting extended dataset has the following structure shown in Table 2, after making the mentioned computations, and then filtering out consecutive measures taken at the same time stamp (possibly duplicates); also the very first observation has to be removed in order to avoid division by zero in the speed calculation. Distances have been transformed to *km* so that speed unit is *km/h*.

Table 2: Extended dataset with distances and speeds.

	x	y	z	t	dow	distance	dt	speed
4	717002.3	9677061	2616.9	23.30222	2	0.000855963	8.003055555	0.28013360
6	717002.8	9677062	2616.9	23.30444	2	0.000643842	8.002222222	0.28972928

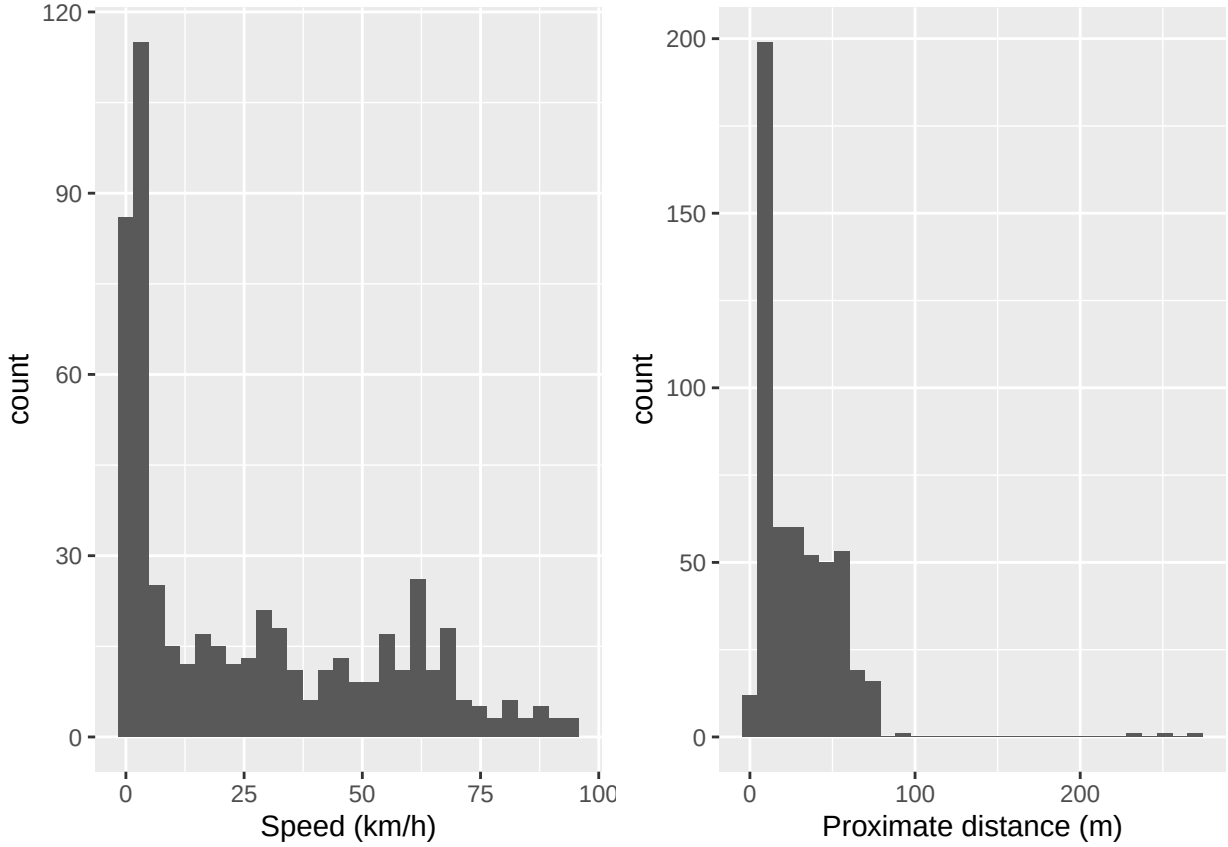
	x	y	z	t	dow	distance	dt	speed
7	717007.8	9677053	2616.9	23.30500	2	0.010478652	0.000555555	68.86157527
9	716998.5	9677059	2616.9	23.33444	2	0.011207170	0.029444444	0.38062089
11	716998.2	9677059	2616.9	23.33500	2	0.000243651	0.000555555	0.43857315
12	717175.5	9676911	2617.3	6.65667	3	0.231104112	2.321666666	7.03156441
15	717228.9	9676883	2617.3	6.65778	3	0.060226605	0.001111111	54.20394450
16	717271.2	9676865	2617.3	6.65861	3	0.045628467	0.000833333	54.75416151
20	717304.9	9676851	2617.3	6.65944	3	0.036599562	0.000833333	43.91947537
21	717338.8	9676843	2617.3	6.66056	3	0.035000220	0.001111111	31.50019820

Finally, a last filtering procedure removes observations with extreme speeds, as values beyond *130km/h* are very improbable (by checking speed limits within the studied region). These data are often related to extreme distances between proximate points, which can occur due to GPS bad measures.

2.3. Data Mining

Data traces of each user must be segmented into individual travels (geometries) so than OD's can be found at the start and end points. In order to aggregate data into trajectories, an heuristic considering low-speed hot spots as destinations is now described.

Taken a single day displacements for a single user, the speeds and distances between proximate points variations that take place when traveling, staying on destination, changing to a different travel mode may allow to detect a trip's endpoints. The speed and distance distributions are shown in Figure **??**.



As seen in the previous figure, most of the speeds and distances are closer to zero, probably because users spend most of their time walking or staying in one destination before starting the next trip. The changes in speed during the day can be seen in Figure 3.

This means that, data points can be segmented into individual travels by detecting those locations when moving at very low speeds (staying still), with respect to a given speed tolerance ϵ.

It can be assumed that actual destinations are found in intervals where users are not moving for a minimum amount of time t_{min}, that is in the "valleys" shown in last figure; in contrast to traffic lights that will also produce zero speeds but will last only a few seconds. The following plot allows appreciating points merged into trips (clusters) for ϵ=2km/h and t_{min}= 10 minutes, see Figure 4. Increasing t_{min} will merge nearby trips into larger ones

At last, the fist (FP) and last (LP) point in each cluster allows extracting coordinates of the origin and destination, as well as the departure and arrival times (from the time stamps of these points). The travel time is simply the difference between the arrival and departure times; also, the cumulative distance of all points in a cluster indicates the trip travel distance. The statistics of the

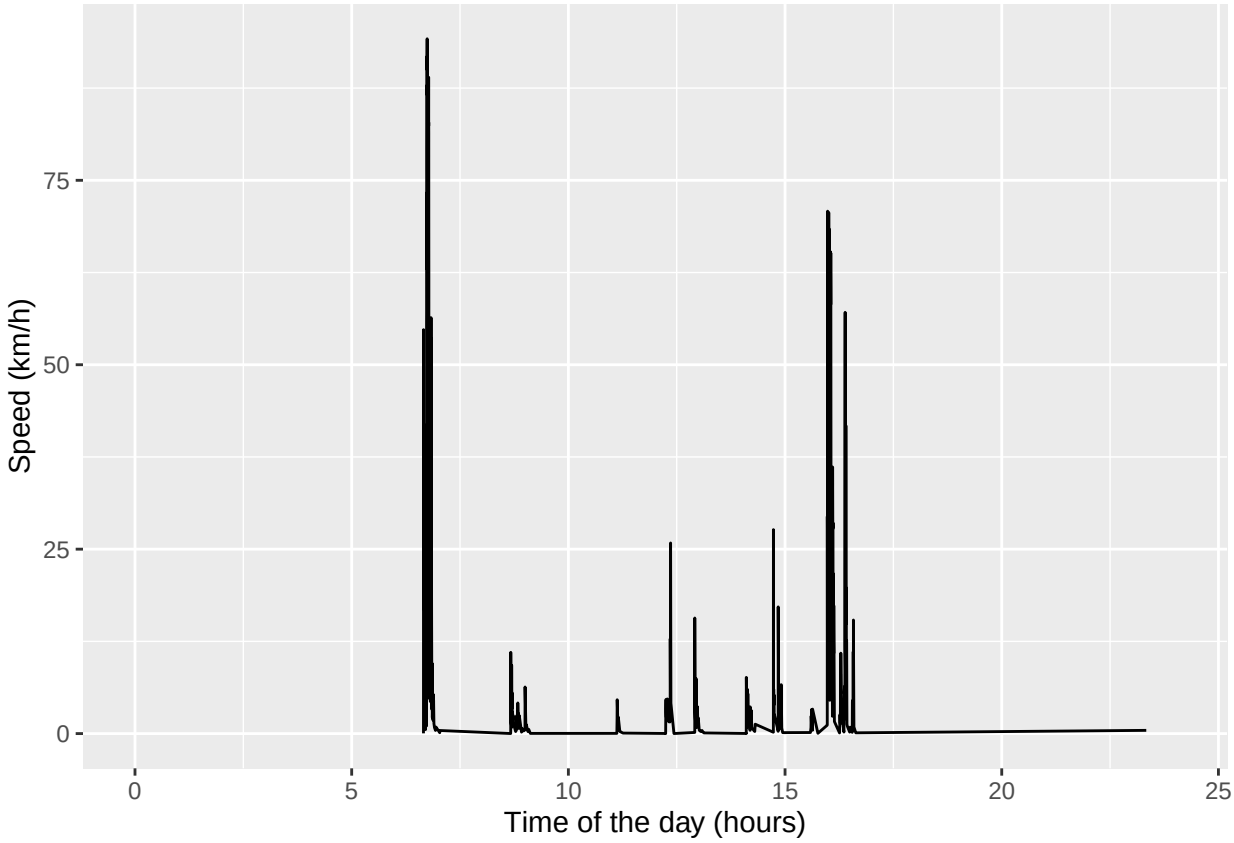


Figure 3: Speed variations along a 24-h single day.

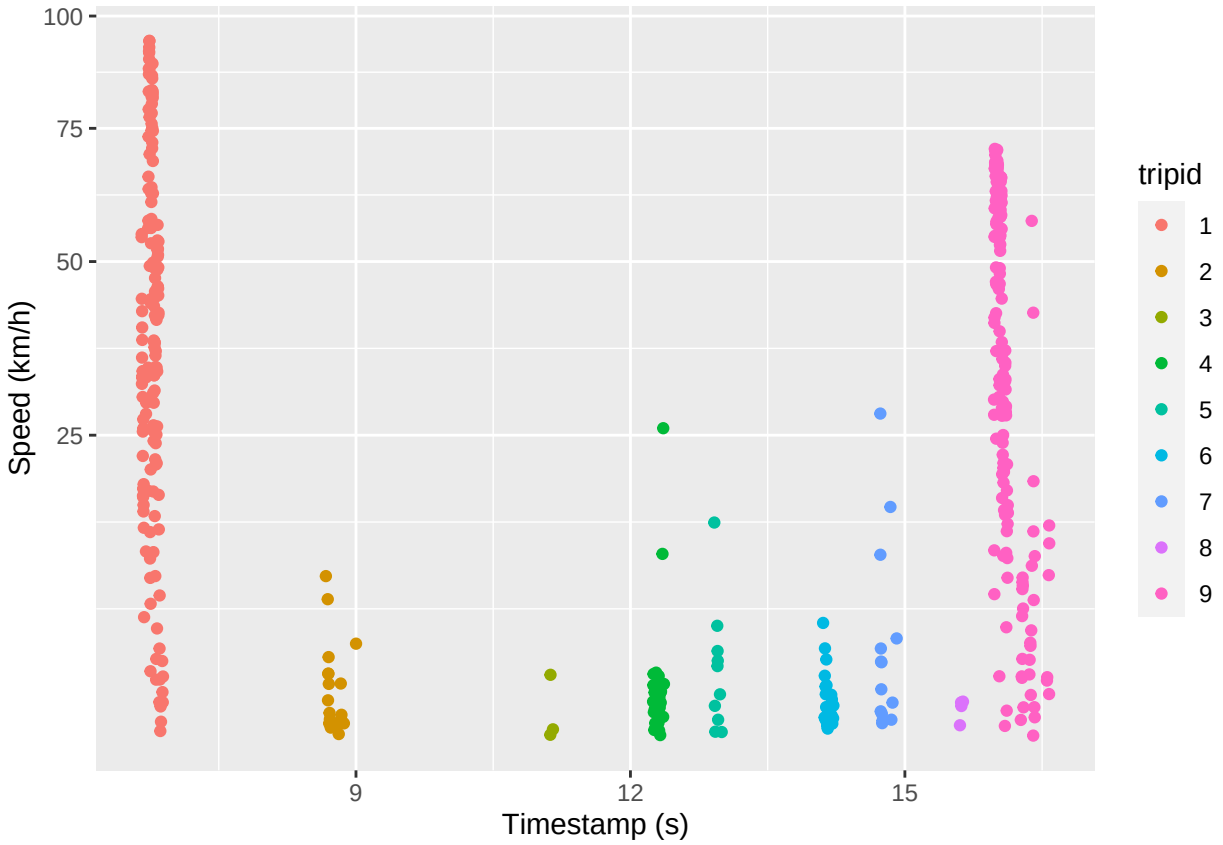


Figure 4: Speed variations along a single day.

extreme points of each cluster are given in table 3.

Table 3: Extreme points in clusters.

Cluster ID	FP time	LP time	FP x	FP y	LP x	LP y
1	6.65778	6.89056	717228.9	9676883	722215.7	9677159
2	8.67111	9.00083	722228.4	9677163	722206.4	9677232
3	11.12417	11.15306	722217.5	9677177	722219.0	9677174
4	12.24778	12.36639	722221.9	9677168	722402.4	9677279
5	12.91694	12.99778	722327.9	9677236	722198.1	9677170
6	14.10639	14.21889	722219.7	9677171	722319.8	9677195
7	14.73167	14.91028	722284.2	9677244	722319.8	9677189
8	15.60222	15.63500	722230.8	9677218	722186.1	9677200
9	15.97472	16.57806	722167.9	9676934	717010.1	9677058

A glimpse of the aggregate data into individual travels (one per cluster) is presented in table 4.

Table 4: Sample trips for one single user and day

tripid	ox	oy	dx	dy	departure	arrival	tdistance	ttime
1	717228.9	9676883	722215.7	9677159	6.65778	6.89056	6.11272980	0.23278
2	722228.4	9677163	722206.4	9677232	8.67111	9.00083	0.20374773	0.32972
3	722217.5	9677177	722219.0	9677174	11.12417	11.15306	0.03148881	0.02889
4	722221.9	9677168	722402.4	9677279	12.24778	12.36639	0.38327980	0.11861
5	722327.9	9677236	722198.1	9677170	12.91694	12.99778	0.11186781	0.08084
6	722219.7	9677171	722319.8	9677195	14.10639	14.21889	0.21750674	0.11250
7	722284.2	9677244	722319.8	9677189	14.73167	14.91028	0.19505546	0.17861
8	722230.8	9677218	722186.1	9677200	15.60222	15.63500	0.04114902	0.03278
9	722167.9	9676934	717010.1	9677058	15.97472	16.57806	6.52144871	0.60334

The resulting segmentation allows OD's to be detected. Their coordinates are given in the table as attributes "dx" and "dy" for destinations locations, and "ox" and "oy" for the origins. Figure 5 presents on a map at scale 1:25000 the resulting user's destinations (as red spots). It can be noticed that as locations are repeatedly visited, some points could be merged into a single destination as possibly they are short displacements around the same location; this can be done by density-based clustering techniques; however this is not necessary for the upcoming analysis.

By applying the algorithm to the full dataset of tracked users, segmented trajectories exhibit the statistics shown in Figure 6.

Then, according to this report, the majority of trips occur on weekdays. Moreover, they are "short trips" below 30 minutes and 10km.

Figure 5: Sample destinations for one single user and day.

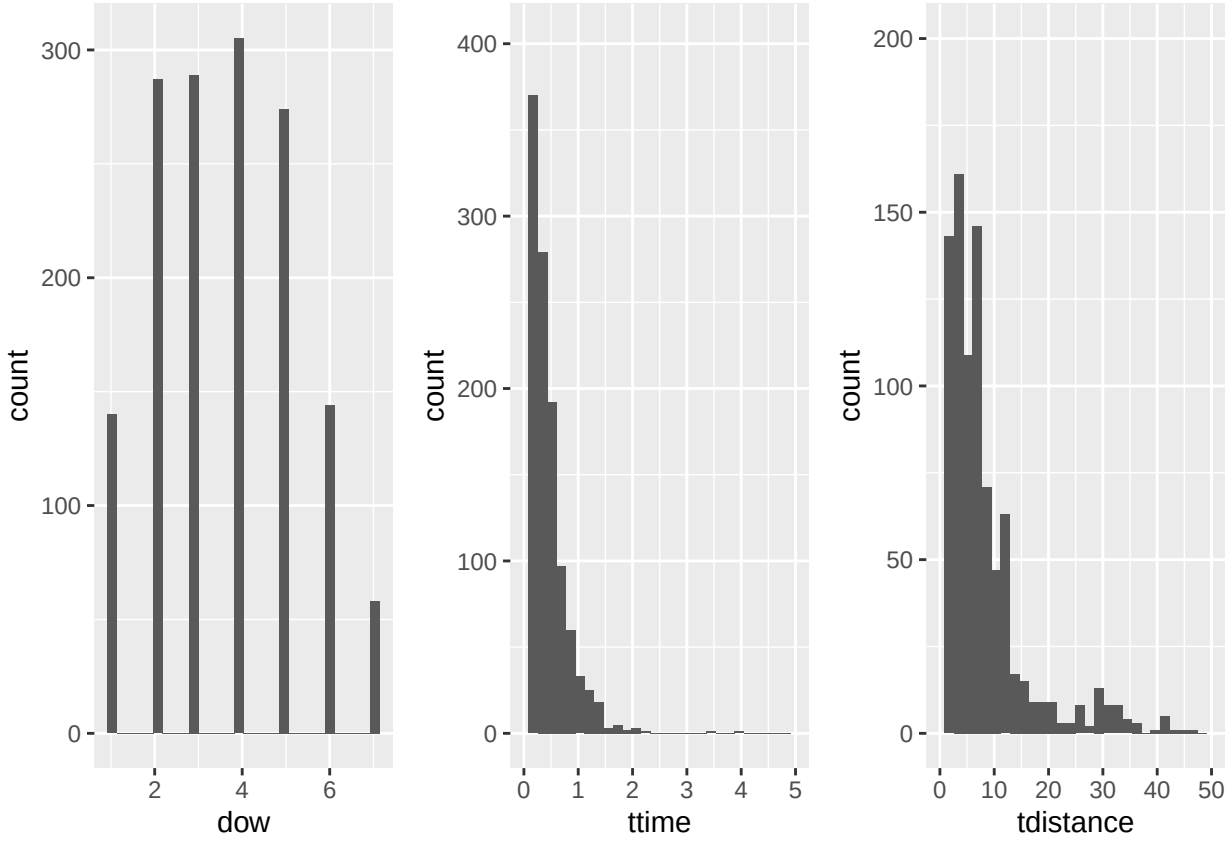


Figure 6: Trip statistics for the full dataset.

2.3.1. Home Detection

After destinations have been detected, another heuristic can be used to identify a user's home. Another concepts must be first defined.

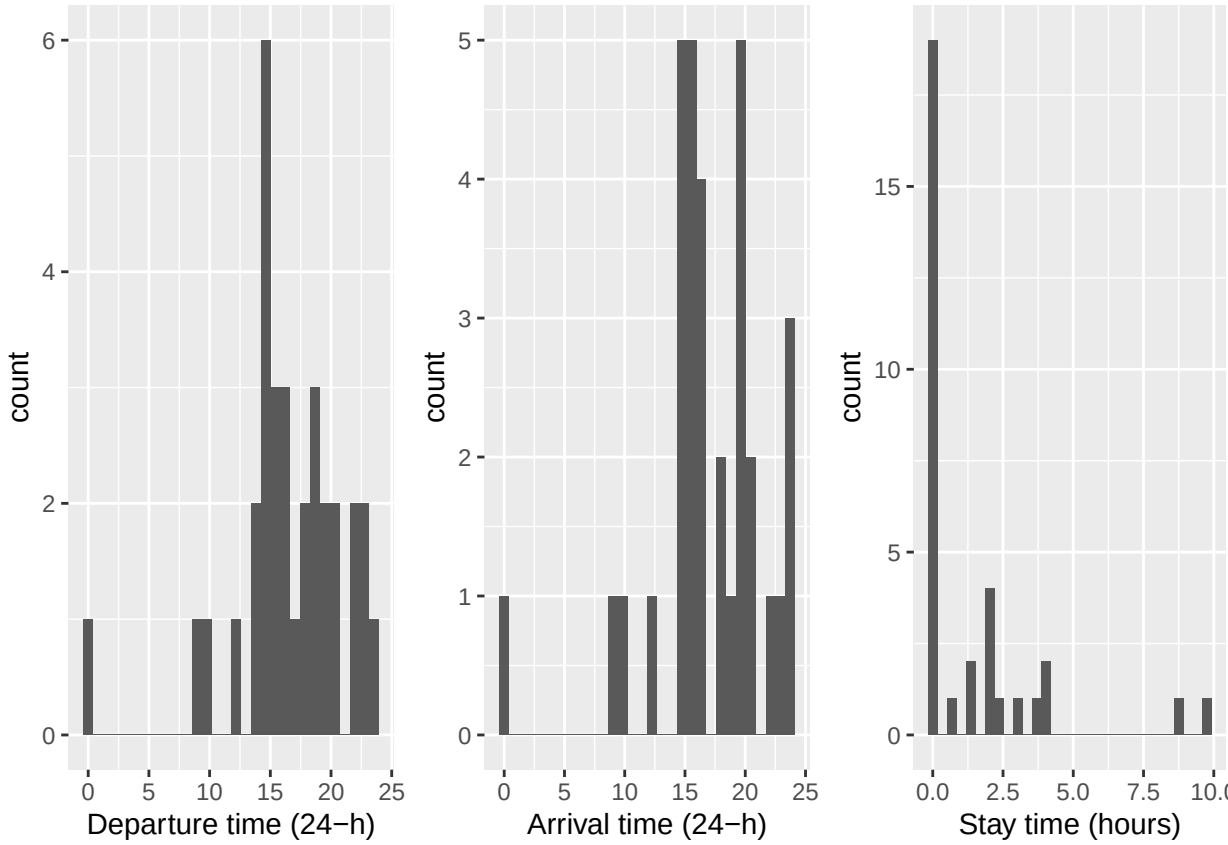
Let T_k be a trip displacement identified by k, the following characteristics are known:

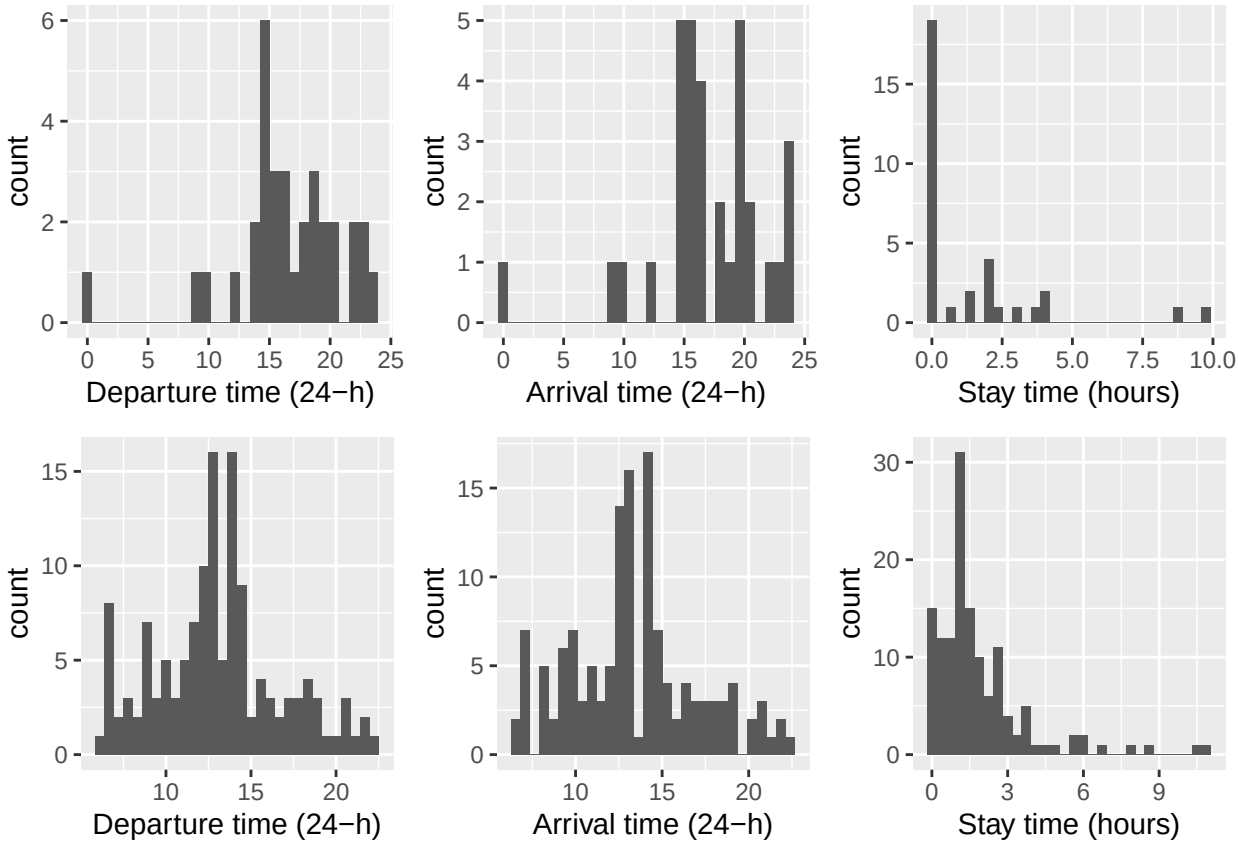
$$T_k = (o_k, d_k, dt_k, at_k, st_k)$$

- o_i, the origin data point with its own coordinates and time stamp
- d_i, the destination data point with its own coordinates and time stamp
- dt_i, the departure time (time at origin data point)
- at_i, the arrival time (time at destination data point)
- st_i, stay time at destination of this trip

$$st_i = dt_{i+1} - at_i$$

That is, the stay time is the time spent on destination before the next trip starts. A first exploratory data analysis must be carried out with 31 users who have voluntarily provided an approximation of the location of their residence. Some statistics of these trips are presented in Figure **??** (top), where it can be noticed that arrival times of home trips occur mostly in the evening, in contrast to no-home trips that do not exhibit a clear pattern as seen in the same figure (bottom).


count
Departure time (24−h)
Arrival time (24−h)
Stay time (hours)



Taking locations of those destinations found on each last day trip (avoiding inter-day trips), the results for the same one-user of the sample used in previous analysis but now for multiple days is shown in 7.

It must be noticed that more than one location could appear as possible last day trip destination; the most frequent will be labeled as the "home location". To find it not only visually, a density based clustering such as DBSCAN [7] must be carried out on this destination points. The dissimilarity measure of the algorithm is the euclidean distance to agglomerate single trip destinations into destinations of interest.

By using a clustering radius of 50 meters with a minimum cluster size of 5 points (which means at least 5 home trips are required to detect it), 5 different possible candidates were found but only the most frequent (the biggest cluster) has been assumed to be home as can be seen in the same Figure 7, where this location has been highlighted as a green spot.

2.4. Discussion and Conclussions

This paper states an approach to detect home locations based on data mining techniques such as clustering, exploratory data analysis and data segmentation.

Figure 7: Locations of last day trip destinations for a single user.

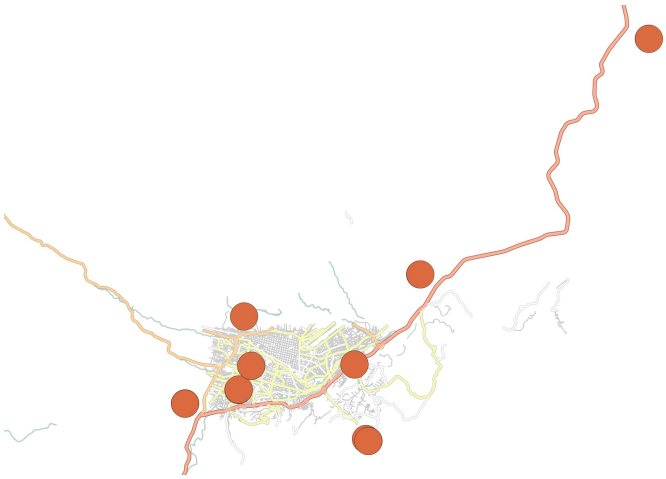


Figure 8: Assumed home locations for a sample of users within the studied region of Cuenca, Ecuador.

Home locations have been validated by users who have voluntarily provided an approximation of the location of their residence. The heuristic considers the last known destination per day, so that the more data are collected the more plausible the chance for the algorithm to capture the correct location. This approach creates clusters of recurrent destinations, after mobility data is aggregated into trips by segmentation techniques. It has been shown that one month of data is sufficient for users to exhibit patterns, which are required to identify those regions of interest.

The segmentation of data points must be done per user, since OD detection expects instant speeds to decrease below a reference value per trip, so that stay time when "not moving" reaches a minimum time threshold. The calibration of these two parameters make trips longer or shorter, and it could increase the chances to find locations of interest where the user stays for long periods. As the dataset contains track points from users, constraining the trips arrival times to an interval (morning, evening) could detect the college campus as well as home locations. In the next section some example use cases are provided.

2.4.1. Transport Service Applications

When taking into consideration the home location of several users within studied area, a list of insights for different applications to provide transport services arise. A sample of assumed home locations are shown in 8 for our region of interest.

Some possible applications are:

- A dedicated transport service for students at the beginning (end) of the day consisting of few bus lines. A current trend is to provide college communities with a electric bus service, and this approach could led to the design of those bus lines by retrieving the demand spatial spots.
- If home locations and the college campus are removed from the trips set, then a subset of regions of interest for the students is retrieved, providing a list of activities the students perform at different times of the day when they are not studying. Private or public transport companies could benefit from this information to provide services according to the expected departure times.
- At last, carpooling and ride-sharing campaigns could use this information to plan groups of students that live nearby, for sharing cars and rides to the college campus or other known regions of interest.

References

[1] Rein Ahas, Siiri Silm, Olle Järv, Erki Saluveer, and Margus Tiru. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of urban technology*, 17(1):3–27, 2010.

[2] Iva Bojic, Emanuele Massaro, Alexander Belyi, Stanislav Sobolevsky, and Carlo Ratti. Choosing the right home location definition method for the given dataset. In *Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings 7*, pages 194–208. Springer, 2015.

[3] Laetitia Gauvin, Michele Tizzoni, Simone Piaggesi, Andrew Young, Natalia Adler, Stefaan Verhulst, Leo Ferres, and Ciro Cattuto. Gender gaps in urban mobility. *Humanities and Social Sciences Communications*, 7(1):1–13, 2020.

[4] Ivan Mendoza, Gustavo Alvarez, Mateo Coello, Joaquín López, and Pablo Carvallo. Automatic estimation of demand matrices for universities through mobile devices. In *2020 IEEE ANDESCON*, pages 1–6. IEEE, 2020.

[5] Joaquín Osorio-Arjona and Juan Carlos García-Palomares. Social media and urban mobility: Using twitter to calculate home-work travel matrices. *Cities*, 89:268–280, 2019.

[6] Luca Pappalardo, Leo Ferres, Manuel Sacasa, Ciro Cattuto, and Loreto Bravo. Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. *EPJ data science*, 10(1):29, 2021.

[7] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

[8] Maarten Vanhoof, Fernando Reis, Thomas Ploetz, and Zbigniew Smoreda. Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics*, 34(4):935–960, 2018.