REDIN Revista Facultad de Ingeniería Universidad de Antioquia

No.114, pp. 51-62, Jan-Mar 2025

A density-based heuristic for household detection in college communities through Big Data analysis

Una heurística de densidad para detectar ubicaciones de hogares en comunidades universitarias mediante Big Data

Iván Mendoza [1*], **Andrés Baquero-Larriva** [2], **Gustavo Álvarez-Coello** [3]

[1]Centro de Estadística, Universidad del Azuay. Avenida 24 de Mayo 7-77 y Hernán Malo. Ecuador, Cuenca

CITE THIS ARTICLE AS:

I. Mendoza, A. Baquero-Larriva and G. A. Álvarez-Coello "A Density-based heuristic for household detection in college communities through big data analysis", *Revista Facultad de Ingeniería, Universidad de Antioquia*, no. 114, pp. 51-62, Jan-Mar 2025. [Online]. Available: https://www.doi.org/10.17533/udea.redin.20240938

ARTICLE INFO:

Received: January 29, 2024
Accepted: August 30, 2024
Available online: September 02, 2024

KEYWORDS:

Algorithms; data analysis; data visualization; data processing; pattern recognition

Algoritmos; análisis de datos; visualización de datos; procesamiento de datos; reconocimiento de patrones

ABSTRACT: In the age of big data, the wealth of information offers unprecedented opportunities to glean valuable insights into human behavior and activities. This study focuses on leveraging data collected from mobile applications used by students at a local college to identify their home locations and other shared points of interest. Through this research, we aim to enhance understanding of mobility patterns within student communities, providing valuable information for decision-making in transportation planning and mobility-related issues in surrounding areas. This paper introduces a heuristic based on density-related clustering to detect home locations from real-time big data collected by a mobile application. The results demonstrate satisfactory precision, with potential for further improvement as additional data is acquired, thus offering insights into potential future applications and services.

RESUMEN: En la era del Big Data, la abundancia de datos ofrece oportunidades sin precedentes para obtener información valiosa sobre el comportamiento y las actividades humanas. Este estudio se centra en el aprovechamiento de los datos recogidos de las aplicaciones móviles utilizadas por los estudiantes de una universidad local para identificar la ubicación de sus hogares y otros puntos de interés compartidos. A través de esta investigación, se pretende mejorar la comprensión de los patrones de movilidad dentro de las comunidades de estudiantes, proporcionando información valiosa para la toma de decisiones en la planificación del transporte y las cuestiones relacionadas con la movilidad en las zonas circundantes. Este artículo presenta una heurística a partir de clustering basado en densidades para detectar las ubicaciones de los hogares a partir de grandes volúmenes de datos, recopilados en tiempo real por una aplicación móvil. Los resultados demuestran una precisión satisfactoria, con potencial de mejora a medida que se adquieren datos adicionales, evidenciando un potencial de posibles aplicaciones y servicios futuros.

1. Introduction

Home location detection plays a pivotal role in various domains, from urban planning to personalized services, and is essential for understanding human mobility patterns and informing policy decisions [1–3]. Social media platforms, particularly Twitter, have emerged as valuable sources of data for inferring travel patterns and home-work travel matrices [4]. Meanwhile, mobile phone data analytics offer insights into population geography and mobility patterns, aiding in the development of real-time monitoring tools and urban planning strategies [5, 6].

Studies have underscored the importance of selecting appropriate methodologies for home location detection, considering the characteristics of the dataset, and addressing inherent biases [7, 8]. Various techniques have been proposed for inferring home locations, including clustering algorithms, trajectory analysis, and machine learning models [9, 10]. Ensemble approaches that combine multiple algorithms have shown promise in enhancing detection accuracy [11, 12]. However, privacy concerns related to location data persist, necessitating the

* Corresponding author: Iván Mendoza
E-mail: imendoza@uazuay.edu.ec
ISSN 0120-6230
e-ISSN 2422-2844

implementation of robust privacy protection measures, especially in the context of location-based services [13, 14]. Furthermore, recent research has focused on the development of novel methodologies for home location detection [15, 16]. These advancements highlight the interdisciplinary nature of home location detection research, drawing insights from fields such as computer science, geography, and social sciences.

In fact, accurate home location detection is crucial for various applications, including urban planning, personalized services, and ensuring user privacy. Addressing the methodological challenges and privacy concerns associated with home location detection requires ongoing research and collaboration across disciplines.

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 methodology proposed in the following section employs a density-based approach, a type of agglomerative clustering, to address the unique challenges presented by spatiotemporal data, which often form large clusters around user points of interest. This approach is particularly effective for detecting these hotspots while simultaneously discarding outliers, such as occasional stops during a trip. In our case study, the parameters of the heuristic approach were calibrated using home data provided by volunteers. Unlike previous studies, which broadly address various populations, our research focuses specifically on the residential patterns of college students. Our study offers a more targeted analysis by incorporating characteristics such as the last-day trip destination identified through a data mining process on raw data. Additionally, home locations identified by our methodology were validated against a set of known student data, ensuring the accuracy and relevance of our findings.

2. Methodology

In order to obtain locations of the student homes, data have to be mined from a typical tracking dataset, which contains primarily coordinates and timestamps by following a multi-step procedure described in Figure 1. In summary, the methodology follows this sequence:

Data are collected by a mobile app that uses the location services of the mobile device, so that they can be temporarily kept in the local storage; then, data are securely stored in a remote server via an Application Programming Interface (API). In the second stage, these unprocessed data are filtered to detect outliers and possible errors and then transformed to a suitable format for further processing. Finally, data are aggregated into origins and destinations (OD) by a segmentation algorithm and classified as the inferred home locations of students. These locations are validated for a selected sample of known locations. To complete the flow, some potential transport services and strategies, including routes among these locations are planned out. The steps followed in the methodology are described below.

2.1 Mobile Data Collection

The first step is storing mobility data together with temporal and user's identity information in a remote database. Each observation *i* (a labeled spatiotemporal data point) will have the following structure in Equation 1. The simplest form consists of the location´s coordinates (x, y) of the tracked user at a specific time, *t*.

$$p_i = (x_i, y_i, t_i) \tag{1}$$

A more complete version of the observation is shown in Equation 2.1.

$\mathrm{p}_i = (uid_i, lat_i, lon_i, alt_i, date_i, t_i, dow_i, acc_i, timestamp_i)$

- *uid*, is a unique identifier of the user, normally a MD5 hash string which guarantees data is collected anonymously, but it still makes it possible to know the user in the database
- *lat. lon, alt*, the points of latitude, longitude, and altitude coordinates
- *date*, a day-month-year string
- *t*, a 24-hour local time format for a continuous variable after applying the formula shown in Equation 2

$$t = hours + \frac{minutes}{60} + \frac{seconds}{3600} \tag{2}$$

- *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 timestamp, 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 offline treatment. However, this format is not yet suitable for most calculations, so that further processing is required as described in the following steps.

2.2 Data Processing

In order to avoid bias in the calculation of travel destinations, outliers are removed, considering the position accuracy and temporal constraints. A subset P_a consists of more refined "valid" observations, selected through the following criteria shown in Equation 3.

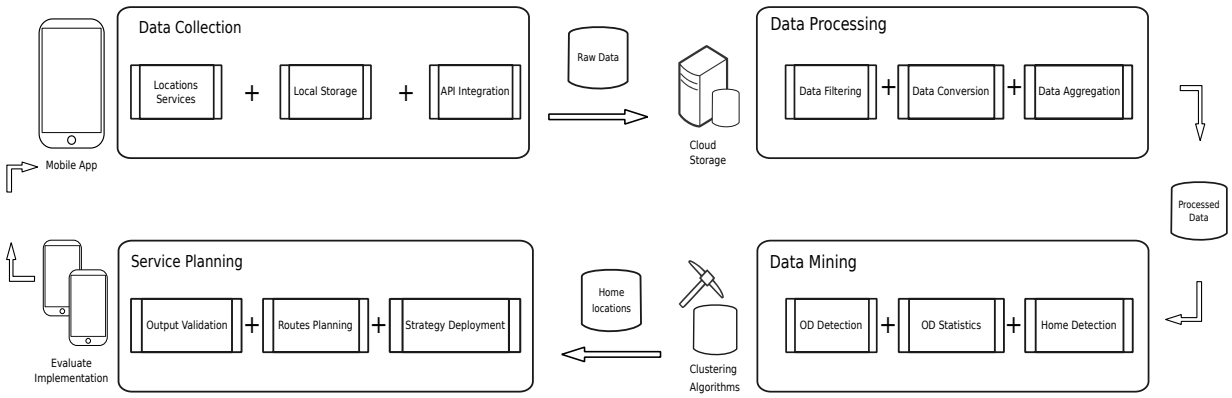

Figure 1 Proposed Multi-step procedure methodology for Home Detection.

$$P_b = \{p_i \in P_a | acc_i < \alpha\} \tag{3}$$

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; this is shown in Equation 2.2.

$$\begin{aligned} \mathrm{P}_c = \{&p_i \in P_b \mid \\ &(lat_{\min} < lat_i < lat_{\max}) \\ &\wedge\ (lon_{\min} < lon_i < lon_{\max})\} \end{aligned}$$

Because this research is intended to provide insights into transport services for students in a College Community, a time period when 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, shown in Equation 4.

$$P = \{p_i \in P_c | (date_{min} < date_i < date_{max}) \tag{4}$$

Lastly, since OD detection requires the 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 is shown in Equation 5.

$$p_i = (uid_i, x_i, y_i, z_i, t_i, dow_i) \tag{5}$$

Computing new attributes

Additional features must be incorporated into the existing dataset to comprehensively evaluate travel behavior.

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 following n observations is defined in Equation 6.

$$D_{a,n} = \sum_{i=a+1}^{a+n} d_{i,i-1} \tag{6}$$

where $d_{i,j}$ is the Euclidean distance between two observations, defined in Equation 7.

$$d_{i,j} = ((x_i - x_j)^2 + (y_i - y_j)^2)^{1/2} \tag{7}$$

that is, the sum of distances between proximate points; then, the instant speed at observation i^{th} is computed by Equation 8.

$$s_i = \frac{d_{i,i-1}}{t_i - t_{i-1}} \tag{8}$$

The resulting extended dataset has the following structure shown in Table 1, 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 the speed unit is *km/h*. The field "dt" in Table 1 corresponds to the difference in hours between consecutive points.

Finally, a last filtering procedure removes observations with extreme speeds, as values beyond *130 km/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 bad GPS measures.

x	y	z	t	dow	distance	dt	speed
717002.3	9677061	2616.9	23.3022	2	0.0009	0.0031	0.2801
717002.8	9677062	2616.9	23.3044	2	0.0006	0.0022	0.2897
717007.8	9677053	2616.9	23.3050	2	0.0105	0.0006	18.8616
716998.5	9677059	2616.9	23.3344	2	0.0112	0.0294	0.3806
716998.2	9677059	2616.9	23.3350	2	0.0002	0.0006	0.4386
717175.5	9676911	2617.3	6.6567	3	0.2311	7.3217	0.0316
717228.9	9676883	2617.3	6.6578	3	0.0602	0.0011	54.2040
717271.2	9676865	2617.3	6.6586	3	0.0456	0.0008	54.7542
717304.9	9676851	2617.3	6.6594	3	0.0366	0.0008	43.9195
717338.8	9676843	2617.3	6.6606	3	0.0350	0.0011	31.5002

Table 1 Extended dataset with distances and speeds.

2.3 Data Mining

A heuristic is proposed to segment the data traces of each user into individual travels (geometries) and OD pairs at the start and end points. This heuristic focuses on identifying low-speed hot spots as potential destinations to aggregate the data into trajectories.

Low-Speed Detection: Data are traversed to look for stay hot spots, by comparing instant speeds with a threshold parameter ϵ so that data points are segmented into a sequence of displacements.

False positives filtering: An acceptable displacement by a set of known travel modes should involve a minimum stay time t_{min} in location, as well as minimum time and travel distance. After this process, the dataset size will be reduced (approximately up to a 0.5 factor).

GPS Traces Clustering: Finally, the first (FP) and last (LP) point in each cluster allow 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.

Home Detection: After detecting destinations, additional heuristics must be applied to find out a user's place of residence. It is, therefore, essential to define certain concepts beforehand. Let T_k be a trip displacement identified by k, the following characteristics are known in Equation 9.

$$T_k = (o_k, d_k, dt_k, at_k, st_k) \tag{9}$$

- o_i, the origin data point with its own coordinates and timestamp
- d_i, the destination data point with its own coordinates and timestamp
- 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 is defined in Equation 10

$$st_i = dt_{i+1} - at_i \tag{10}$$

That is, the stay time is the time spent on the destination before the next trip starts. Some points of interest will be found with this process, including, for sure, the college campus and the expected home location. The heuristic assumes that the most frequent last trip destination of the day must be a student's home.

It must be noted that every new visit to this place will yield a slightly different pair of coordinates. To find their centroid, a density-based clustering such as DBSCAN [**?**] must be carried out on these points. The dissimilarity measure is the Euclidean distance and a set of points or destinations of interest is found.

In the following section, the results after applying this methodology to the collected data set are presented and discussed; an approach to validate these results is also provided.

3. Tests and results

3.1 Dataset Description

The dataset used in this study comprises spatiotemporal data collected through a specialized tracking mobile application designed for a university community over five months. It encompasses data from 728 users, chosen over a one-month period, specifically from May 20 to June 20, 2023, with data totaling about five million monthly records. Subsequent analysis results have shown that 30 days of mobility data per user is sufficient to identify household locations domicile. The number of observations after the data cleaning process is above *11 million*, so that good infrastructure for cloud storage is required to handle this volume of big data. 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, showing 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 it was around 2 seconds on average.

3.2 Heuristic Results

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, and changing to a different travel mode may allow a trip's endpoints to be detected. The speed and distance distributions are shown in Figure 3.

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 4.

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 the last figure; in contrast to traffic lights that will also produce zero speeds but will last only a few seconds. Figure 5 shows the points merged into trips (clusters) for ϵ=2km/h and t_{min}= 10 minutes. Increasing t_{min} will merge nearby trips into larger ones.

Following the steps mentioned in the methodology, the data points have been clustered into trips that fit the time and distance constraints. The statistical data for the extreme (border) points of each cluster are given in Table 2.

A glimpse of the aggregate data on individual travels (one per cluster) is presented in Table 3.

The resulting segmentation allows ODs to be detected. Their coordinates are given in the table as attributes "dx" and "dy" for destination locations, and "ox" and "oy" for the origins. Figure 6 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 they are possibly 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 7.

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

Home Detection

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 Figures 8 and 9, 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 Figure 9 .

Taking the locations of those destinations found in each last trip of each day (avoiding trips between days), the results for the same one-user of the sample used in previous analysis, but now for multiple days, is shown in Figure 10.

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 10, where this location has been highlighted as a green spot, and in Figure 11, the assumed home locations for a sample of users are shown.

Data Validation

The last stage intends to test the approach's accuracy, and for that, a set of 30 volunteers who provided actual home locations is used. In Figure 12, a map with these actual locations and the detected ones by the algorithm is presented.

In Table 4, the corresponding coordinates are shown together with the measured distance. A histogram of these errors can be seen in Figure 13, suggesting an approximated normal distribution, with a mean of 27.9 meters and a margin of 5.62 meters for a 95% confidence interval.

4. Conclusions

This paper presents an approach to detecting home locations through density-based clustering, exploratory data analysis and data segmentation.. 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 the more data collected, the more plausible it is that the algorithm will pick up 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.

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

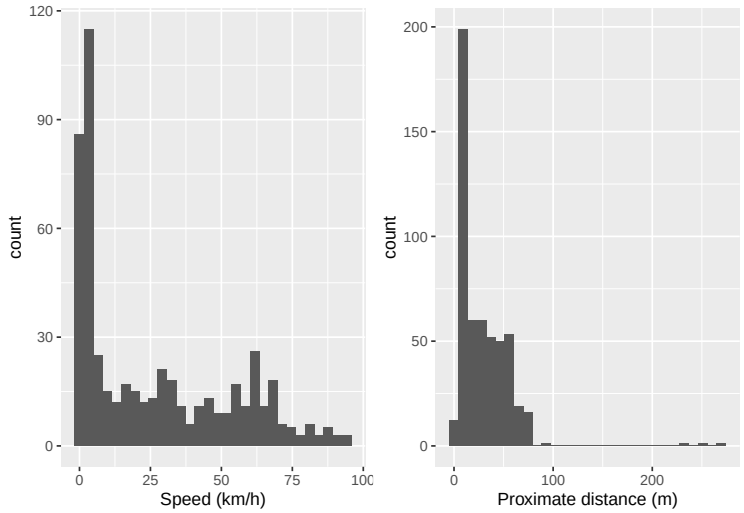

Figure 3 Sample of collected points for the region of Cuenca, Ecuador.

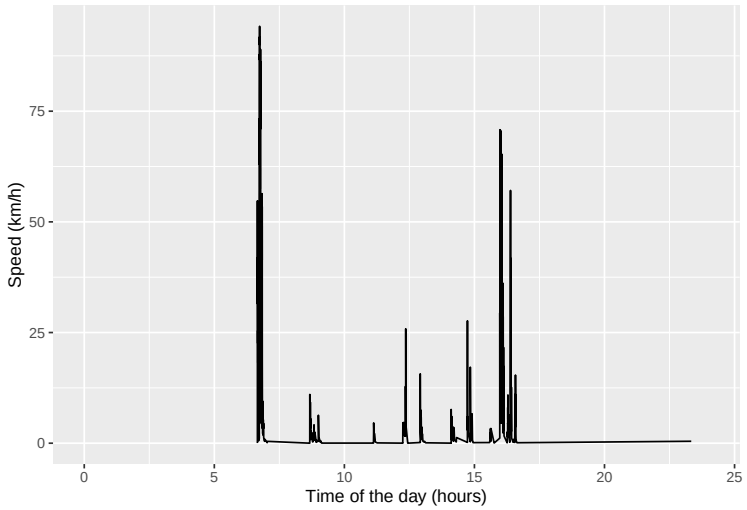

Figure 4 Speed variations along a 24-h single day.

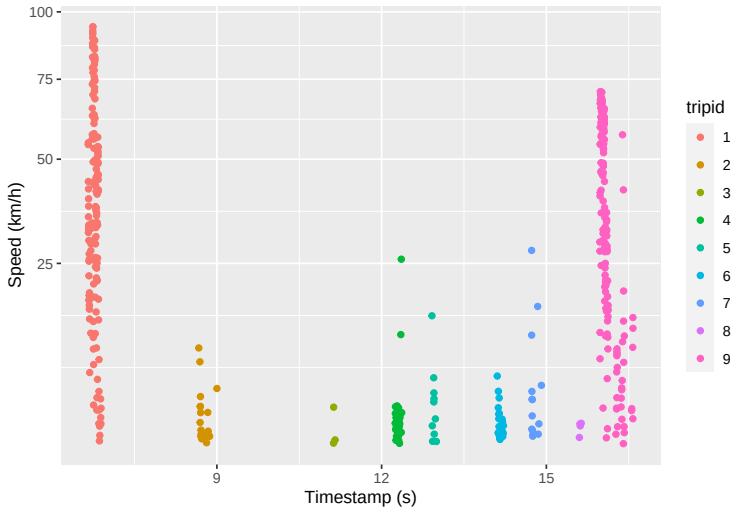

Figure 5 Speed variations along a single day, where gaps between clusters exhibit stops.

Table 2 Extreme points in clusters.

Cluster ID	FP time	LP time	FP x	FP y	LP x	LP y
1	6.66	6.89	717228.90	9676883	722215.70	9677159
2	8.67	9.00	722228.40	9677163	722206.40	9677232
3	11.12	11.15	722217.50	9677177	722219.00	9677174
4	12.25	12.37	722221.90	9677168	722402.40	9677279
5	12.92	13.00	722327.90	9677236	722198.10	9677170
6	14.11	14.22	722219.70	9677171	722319.80	9677195
7	14.73	14.91	722284.20	9677244	722319.80	9677189
8	15.60	15.64	722230.80	9677218	722186.10	9677200
9	15.97	16.58	722167.90	9676934	717010.10	9677058

Table 3 Sample trips for one single user and day

tripid	ox	oy	dx	dy	departure	arrival	tdistance	ttime
1	717228.9	9676883	722215.7	9677159	6.66	6.89	6.11	0.23
2	722228.4	9677163	722206.4	9677232	8.67	9.00	0.20	0.33
3	722217.5	9677177	722219.0	9677174	11.12	11.15	0.03	0.03
4	722221.9	9677168	722402.4	9677279	12.25	12.37	0.38	0.12
5	722327.9	9677236	722198.1	9677170	12.92	13.00	0.11	0.08
6	722219.7	9677171	722319.8	9677195	14.11	14.22	0.22	0.11
7	722284.2	9677244	722319.8	9677189	14.73	14.91	0.20	0.18
8	722230.8	9677218	722186.1	9677200	15.60	15.64	0.04	0.03
9	722167.9	9676934	717010.1	9677058	15.97	16.58	6.52	0.60

Figure 6 Sample destinations for one single user and day.

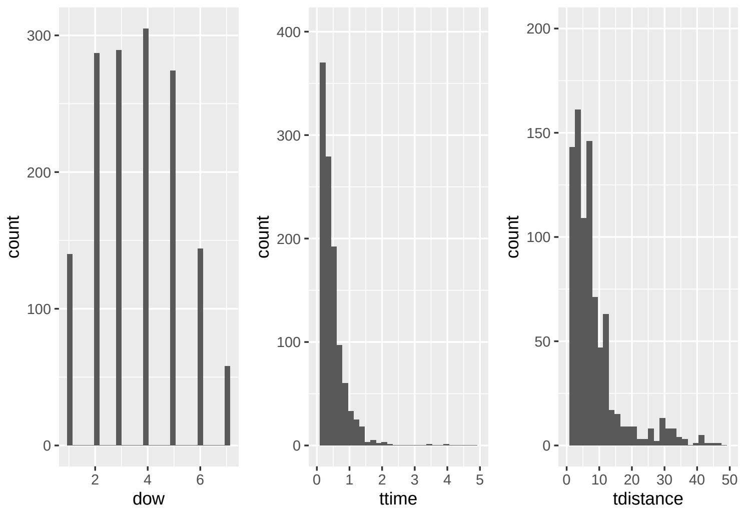

Figure 7 Main trip statistics for the entire dataset.

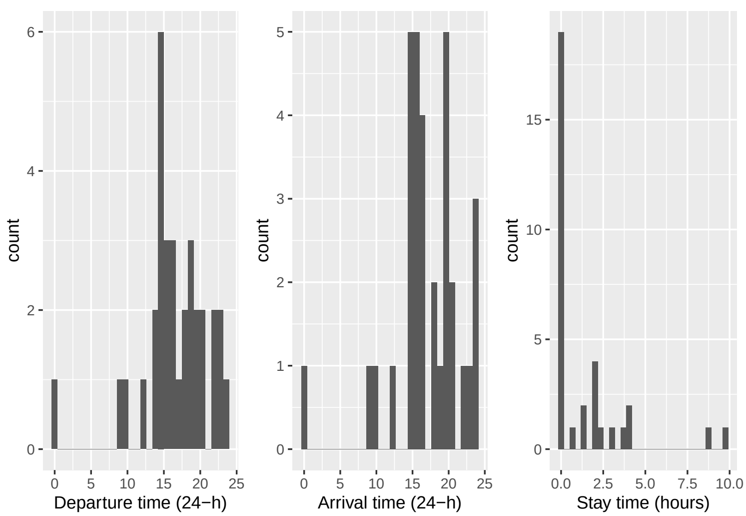

Figure 8 Temporal statistics for trips in the entire dataset.

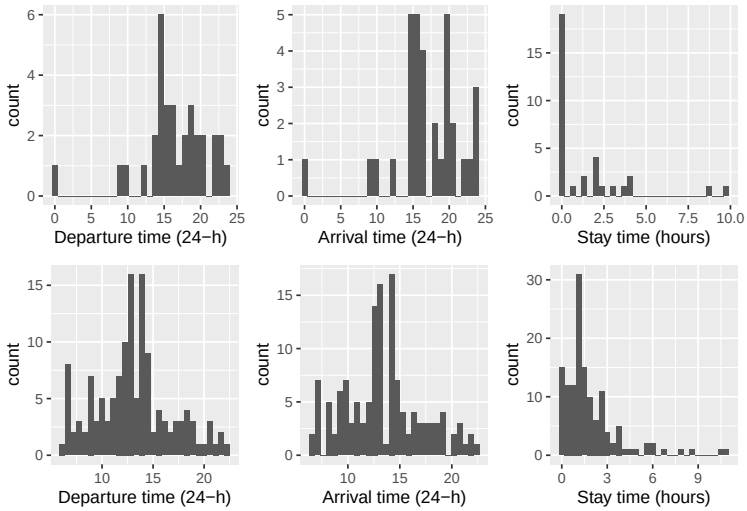

Figure 9 (top) Statistics of home trips and (bottom) for the rest of trips in the entire dataset.

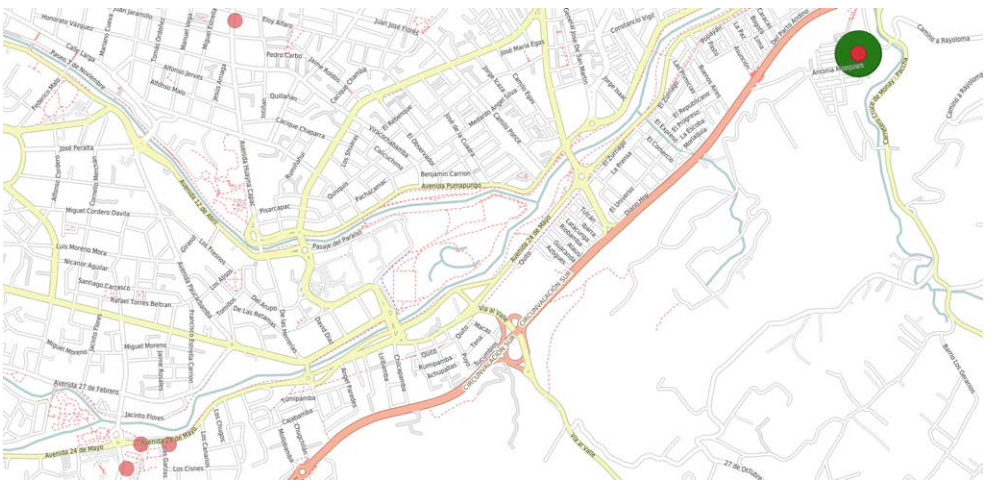

Figure 10 Locations of last-day trip destinations for a single user.

Table 4 Distances between actual and inferred home locations for the sample of volunteers. *DOlon* and *DOlat* correspond to the coordinates detected by the algorithm, while *ROlon* and *ROlat* correspond to the real coordinates

HomeID	DOlon	DOlat	ROlon	ROlat	distance (m)
1	-78.87934	-2.845382	-78.87921	-2.845159	28.65025
2	-79.00991	-2.910327	-79.01020	-2.909943	53.42883
3	-79.01243	-2.904658	-79.01269	-2.904569	30.23294
4	-78.98280	-2.891793	-78.98267	-2.891945	22.24010
5	-79.00700	-2.912737	-79.00692	-2.912505	26.93325
6	-78.96580	-2.877730	-78.96603	-2.877840	28.54877
7	-78.98228	-2.903040	-78.98266	-2.902823	48.80563
8	-78.99913	-2.918980	-78.99965	-2.919037	58.56616
9	-78.96086	-2.894924	-78.96078	-2.895052	17.03338
10	-79.00684	-2.905486	-79.00688	-2.905770	31.57670

Figure 11 Assumed home locations for a sample of users within the studied region of Cuenca, Ecuador.

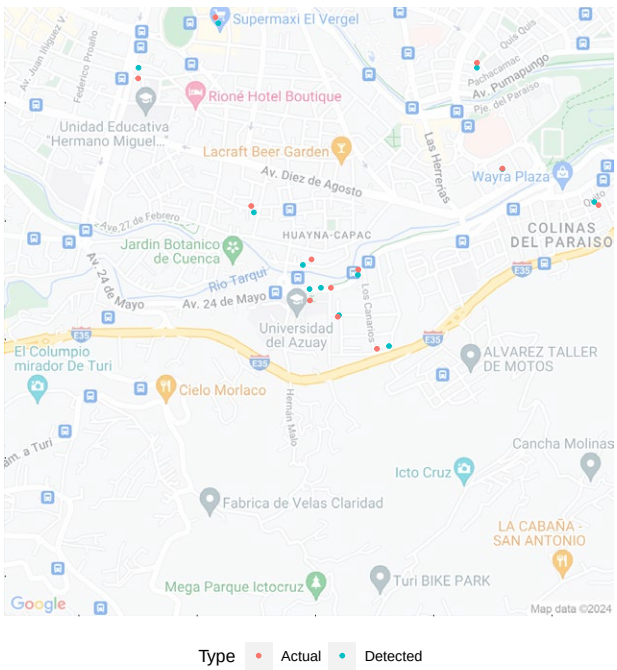

Figure 12 Home locations of student volunteers (red circles) and their assumed locations by the algorithm (blue circles).

5. Discussion and Future Works

Some key aspects must be considered in order to increase the chance of success of this approach. For instance, 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 the max speed threshold makes 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 all tracking data from users, constraining a trip's arrival time to a given interval (morning, evening) and detecting its travel mode helps discriminate destinations among college campuses, home locations, and other relevant places a student visits. In the following section, some examples of use cases are provided.

5.1 Transport Service Applications

When considering the home location of several users within the studied area, a list of insights for different applications to provide transport services arise. A sample of assumed home locations is shown in Figure 11 for our region of interest.
Some possible applications are:

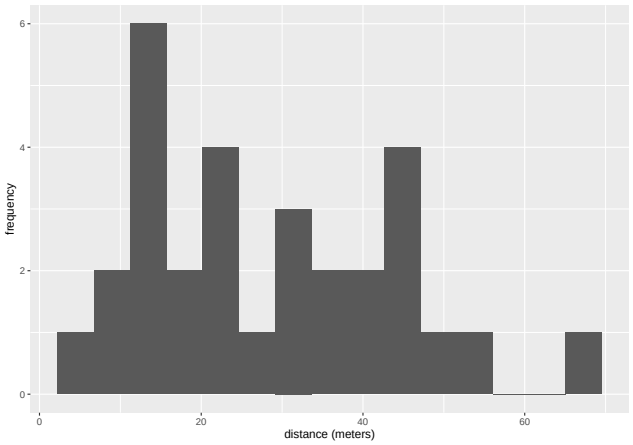

Figure 13 Histogram of distances between actual and inferred home locations.

- A dedicated transport service for students at the beginning or end of the day consisting of a few bus lines. A current trend is to provide college communities with an electric bus service, and this approach could lead 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 by providing services according to the expected departure times.
- Finally, 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.

Declaration of competing interest

We declare that we have no significant competing interests, including financial or non-financial, professional, or personal interests interfering with the full and objective presentation of the work described in this manuscript.

Acknowledgement

We extend our heartfelt appreciation to the Vice Rector for Research at the University of Azuay, the Faculty of Science and Technology at the University of Azuay, and the Latin American Research Network on Energy and Vehicles (RELIEVE).

Author contributions

I. Mendoza: Data acquisition, data analysis, algorithm implementation. A. Baquero-Larriva: Data acquisition. G. Álvarez-Coelo: Data acquisition, data sample selection, data analysis, algorithm implementation. All authors have made significant contributions across various domains, encompassing the design, construction, maintenance, and operation of the instrument(s) utilized for data acquisition. Their combined expertise and dedication have enriched the comprehensive scope of our research endeavor.

Data availability statement

The processed and raw datasets used in this work, as well as the implementation scripts for the algorithms and analysis are available in this repository: https://github.com/ivanmendozav/patronesbigdata

References

[1] J. Osorio-Arjona and J. C. García-Palomares, "Social media and urban mobility: Using twitter to calculate home-work travel matrices," *Cities*, vol. 89, pp. 268–280, 2019. [Online]. Available: https://doi.org/10.1016/j.cities.2019.03.006

[2] R. Ahas, S. Silm, O. Jarv, E. Saluveer, and M. Tiru, "Using mobile positioning data to model locations meaningful to users of mobile phones," *Journal of Urban Technology*, vol. 17, no. 1, pp. 3–27, 2010. [Online]. Available: https://doi.org/10.1080/10630731003597306

[3] I. Mendoza, G. Alvarez, M. Coello, J. López, and P. Carvallo, "Automatic estimation of demand matrices for universities through mobile devices," in *2020 IEEE ANDESCON*. IEEE, 2020, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ANDESCON50619.2020.9272071

[4] S. Bayat, G. Naglie, M. J. Rapoport, E. Stasiulis, B. Chikhaoui, and A. Mihailidis, "Inferring destinations and activity types of older adults from gps data: Algorithm development and validation (preprint)," *JMIR Aging*, 2020. [Online]. Available: https://doi.org/10.2196/18008

[5] L. Pappalardo, L. Ferres, M. Sacasa, C. Cattuto, and L. Bravo, "Evaluation of home detection algorithms on mobile phone data

using individual-level ground truth," *EPJ Data Science*, vol. 10, no. 1, p. 29, 2021. [Online]. Available: https://doi.org/10.1140/epjds/s13688-021-00284-9

[6] M. Vanhoof, F. Reis, T. Ploetz, and Z. Smoreda, "Assessing the quality of home detection from mobile phone data for official statistics," *Journal of Official Statistics*, vol. 34, no. 4, pp. 935–960, 2018. [Online]. Available: https://doi.org/10.2478/jos-2018-00

[7] I. Bojic, E. Massaroa, A. Belyi, S. Sobolevsky, and C. 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*, 2015, pp. 194–208. [Online]. Available: https://doi.org/10.1007/978-3-319-27433-1_14

[8] L. Gauvin, M. Tizzoni, S. Piaggesi, A. Young, N. Adler, and *et al.*, "Gender gaps in urban mobility," *Humanities and Social Sciences Communications*, vol. 7, no. 1, pp. 1–13, 2020. [Online]. Available: https://doi.org/10.1057/s41599-020-0500-x

[9] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, 2017. [Online]. Available: https://doi.org/10.1145/3068335

[10] Q. Chen and A. Poorthuis, "Inferring home locations in human mobility data: An open-source R package for comparison and reproducibility," 2021. [Online]. Available: https://doi.org/10.1080/13658816.2021.1887489

[11] T. Hua, J. Luo, H. Kautz, and A. Sadilek, "Home location inference from sparse and noisy data: Models and applications," *Frontiers of Information Technology & Electronic Engineering*, vol. 17, pp. 389–502, 2016. [Online]. Available: https://doi.org/10.1631/FITEE.1500385

[12] X. Pan, C. Weizhang, and L. Wu, "Mobile user location inference attacks fusing with multiple background knowledge in location-based social networks," *Mathematics*, vol. 8, no. 2, 2020. [Online]. Available: https://doi.org/10.3390/math8020262

[13] P. Golle and K. Partridge, "On the anonymity of home/work location pairs," in *Lecture Notes in Computer Science*, vol. 5538, 2009. [Online]. Available: https://doi.org/10.1007/978-3-642-01516-8_26

[14] D. Zheng, T. Hu, Q. You, H. Kautz, and J. Luo, "Inferring home location from user's photo collections based on visual content and mobility patterns." [Online]. Available: https://doi.org/10.1145/2661118.266112

[15] J. Lin and R. G. Cromley, "Inferring the home locations of twitter users based on the spatiotemporal clustering of twitter data," *Transactions in GIS*, vol. 22, no. 1, pp. 82–97, 2018. [Online]. Available: https://doi.org/10.1111/tgis.12297

[16] I. Chen and A. Poorthuis, "Identifying home locations in human mobility data: An open-source R package for comparison and reproducibility," 2021. [Online]. Available: https://doi.org/10.1080/13658816.2021.1887489