Verma et al. *EPJ Data Science* (2024) 13:6
https://doi.org/10.1140/epjds/s13688-023-00447-w

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Comparison of home detection algorithms using smartphone GPS data

Rajat Verma[1*], Shagun Mittal[1], Zengxiang Lei[1], Xiaowei Chen[1] and Satish V. Ukkusuri[1]

*Correspondence:
verma99@purdue.edu
[1]Lyles School of Civil Engineering, Purdue University, 550 Stadium Mall Avenue, West Lafayette, 47907, Indiana, USA

Abstract

Estimation of people's home locations using location-based services data from smartphones is a common task in human mobility assessment. However, commonly used home detection algorithms (HDAs) are often arbitrary and unexamined. In this study, we review existing HDAs and examine five HDAs using eight high-quality mobile phone geolocation datasets. These include four commonly used HDAs as well as an HDA proposed in this work. To make quantitative comparisons, we propose three novel metrics to assess the quality of detected home locations and test them on eight datasets across four U.S. cities. We find that all three metrics show a consistent rank of HDAs' performances, with the proposed HDA outperforming the others. We infer that the temporal and spatial continuity of the geolocation data points matters more than the overall size of the data for accurate home detection. We also find that HDAs with high (and similar) performance metrics tend to create results with better consistency and closer to common expectations. Further, the performance deteriorates with decreasing data quality of the devices, though the patterns of relative performance persist. Finally, we show how the differences in home detection can lead to substantial differences in subsequent inferences using two case studies—(i) hurricane evacuation estimation, and (ii) correlation of mobility patterns with socioeconomic status. Our work contributes to improving the transparency of large-scale human mobility assessment applications.

Keywords: Human mobility; GPS; Cell phone data; Home location; Data inference

1 Introduction

Home location detection is an important step in several fields of human mobility analysis such as transportation planning [1], migration and evacuation studies [2, 3], accessibility analysis [4], and the theory of human mobility [5, 6]. This task involves predicting people's 'home location' based on geolocation data, often collected passively by their devices via location-based services, call detailed records, social media activity, smart-card transactions, and in-vehicle location trackers [7]. Home detection plays an essential role in understanding large-scale human mobility patterns. For instance, in the event of a hurricane, one needs the home locations both before and after the disaster to identify their evacuation status [8]. In urban planning, identifying home locations serves as the foundational data for vital information including home-based trips [9] and human mobility metrics

[10], and this forms the basis for optimizing existing infrastructure [11]. It is consequently important to have a robust understanding of home detection approaches.

Despite its significance, existing studies using home detection algorithms (HDAs) have paid little attention to the effectiveness of their algorithms. Researchers have developed several HDAs for geolocation data of different kinds whose assumptions, methods, and parameters are not necessarily consistent with one another [12]. This raises doubts about the validity of their findings as the error in home detection may propagate to the downstream calculation of home-related metrics such as evacuation counts [3], home-based trip rates [13], and data representativeness figures for accessibility analysis [14].

This issue stems primarily from a lack of ground truth home locations associated with large geolocation datasets. The collection of accurate home location collection on a large scale poses significant risks to privacy [15]. Mobility data vendors provide anonymized device identifiers and modify sensitive trajectories to prevent an accurate tracking of people's trip origins and destinations [3]. In the absence of ground truth data, it becomes difficult to compare the accuracy of different HDAs using supervised learning methods. Researchers have largely relied on unsupervised methods for home detection, such as rule- and clustering-based HDAs. Small-scale studies such as [12] and [16] have sought to compare the effectiveness of HDAs but have focused only on the parameters of a few HDAs. Further, their small experiments do not provide insights about the impact of study region and period and data quality on the performance of the HDAs.

In this study, we tackle this issue of a lack of a systematic comparative assessment of commonly used HDAs. In doing so, we contribute to the literature on the home location detection problem in the following ways:

1. We review the state-of-the-art HDAs that use large-scale mobility data, including their benefits, assumptions, and limitations.
2. We propose three intuitive metrics to quantify the quality of the home-location detection results in the absence of ground truth home location information.
3. We develop a comprehensive experiment where a set of HDAs are quantitatively compared in terms of the introduced performance metrics and their sensitivity to the data quality.
4. We propose a new HDA that overcomes some of the limitations of the above methods and shows superior performance.

The framework and experiment design of this study is shown in Fig. 1 and described in detail in the following sections. The main objective is to compare the performance of different HDAs across different input datasets. On the basis of the review of research literature on HDAs, we have selected four popular and unique HDAs, and additionally proposed an HDA for comparison in this study. The testing is done on eight input samples of passively collected smartphone GPS data covering four U.S. metropolitan areas of different data qualities and different time periods spawned by mobility-influencing events. Once home locations are estimated for each combination of the sample dataset and HDA, they are compared using three approximate accuracy metrics proposed in Sect. 3.2. The performance of the HDAs under the different dataset conditions is discussed in Sect. 4.1. In the subsequent sensitivity analysis Sect. 4.2, the performance metrics are recomputed for different subsamples of the datasets by changing the quality of the users in the input dataset. Finally, the impacts of these HDAs on subsequent applications, such as hurricane

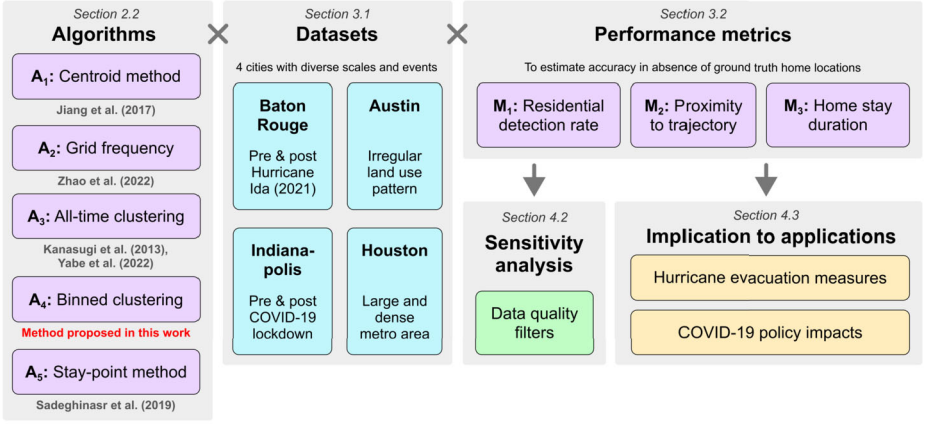

Figure 1 *Framework of the study*. The figure shows the key components of the experiment—HDAs, datasets, and metrics. The cross symbol denotes Cartesian product

evacuation assessment and analysis of mobility change during COVID-19, are shown in Sect. 4.3.

2 Home detection algorithms

2.1 Literature review

HDAs from mobility data can be categorized on the basis of several characteristics, such as the type of input geolocation data (such as social media and passively collected GPS data), the modeling paradigm (supervised vs. unsupervised and rule-based vs. data-driven), and constraints for filtering the input data. Based on these classifications, some prominent HDAs are reviewed and summarized in Table 1.

2.1.1 Supervised methods

Supervised methods predominantly rely on GPS-based travel surveys that involve the subjects carrying GPS-enabled devices that track movements. In addition to individual-level information such as actual (ground truth) home locations, demographic characteristics, and personal preferences, the devices provide detailed travel entries such as the origin and destination, the departure and arrival time, the trip purpose, and the travel mode. Such mixed methods have been used in many pilot studies [12, 18, 19]. In some cases, it is also possible to obtain CDR data of specific groups for whom individual-level data is also available, such as employees of a telephone carrier [12]. In other cases, such as in standalone travel diaries like the National Household Travel Survey, the respondents' street addresses are geocoded to coordinates, though other large-scale information is not obtained for them [38]. With true home locations of the small survey sample, it is possible to create sophisticated supervised machine learning models, such as random forests and AdaBoost [21] or artificial neural networks [29].

Although supervised HDAs are powerful, they suffer from a major limitation of training data availability due to privacy reasons. In the recent past, growing pressure from human rights organizations and the subsequent government regulations has made it difficult to obtain actual home locations of individuals at a large scale [39]. GPS surveys and CDR samples used in supervised HDAs are usually very small, often with fewer than 100 subjects [12, 18]. Samples also typically form specialized volunteer subjects such as students

Table 1 Summary of commonly used HDAs for different data and algorithm types

Kind	Algorithm class	Dataset type	Sources	Definition of home location
Supervised	Clustering	CDR	[17]	Most popular important cluster (Hartigan clustering of cell towers) using a logistic regression model
		GPS Survey / Tracking	[18]	Density-based spatial clustering of points with noise (DBSCAN)
			[19]	Most popular of the clusters based on DJ-cluster algorithm (modified DBSCAN clustering)
			[20]	Most popular of the 'locations' (obtained using modified k-means clustering of "places")
	Clustering and heuristic	CDR	[21]	Binary classification algorithms; logistic regression, random forest, adaboosting and neural network models
	Heuristic	CDR	[12, 16, 22]	Most active tower for several data filter criteria such as nighttime constraints, weekday/weekend, and distinct days
Unsupervised	Clustering	CDR	[23]	Most frequent stay place (determined based on mean-shift clustering of sequenced cell tower locations)
		Passive GPS	[24]	Largest hierarchical cluster of stay points (detected based on Liu et al. (2008))
			[3, 25]	Largest cluster of nighttime records using mean-shift clustering
	Heuristic	CDR	[26]	Location of the more popular of the two cell towers with the most records during non-work time
			[27]	Most frequently communicated tower during nights of weekdays, and weekends over the study period
			[28]	Most frequent location during night time
			[29]	Most common visited locations during night time
			[30]	Anchor point determination model (cell tower location satisfying specific rules of call count)
		Passive GPS	[31]	The centroid of the most visited 20 × 20 m cell during night hours
		Smart card	[32]	Center point-based HDA (iteratively updated centroid between pairs of subway stations)
			[33]	Most visited transit station
			[34]	Most popular transaction place (overall and active days); place with most nighttime activity
		Social media	[35, 36]	Place with the most check-ins on 3 social networks
			[37]	Place with the most check-ins during midnight

[17] and older patients [40], raising concerns about sample representativeness [16]. In addition, GPS travel surveys do not represent longitudinal data. These issues make supervised HDAs much less popular in the research literature.

2.1.2 Unsupervised methods

Basic assumptions

Due to the difficulty in obtaining high-quality home location data at a large scale, researchers have relied heavily on unsupervised HDAs. These methods necessarily depend on a set of assumptions about people's home locations that are found throughout the research literature [3, 29, 37]. These include:

- People are more likely to stay at their homes during the off-work period. This normally includes nighttime, but in some cases, can be extended to weekends or even the office after-hours.
- The most observed place for an individual, especially at night, is usually their home.

These assumptions intuitively make sense, although there are several exceptions, such as people who work from home or who work night shifts. However, since these assumptions are almost always used in unsupervised HDAs, we consider these assumptions to be axiomatic.

Dataset types

The input datasets for unsupervised HDAs are abundantly available on scale, including longitudinal data [31, 41], although they lack the demographics and travel preferences of the subjects [27, 42, 43]. Some of the most prominent dataset kinds include the following:

- *Social media location data* include posts on websites such as Twitter, Foursquare, and Flickr that a user tags with the location of the mentioned place [41]. They are usually available at large scales and in several time periods but are usually spatiotemporally sparse and biased toward certain demographics for effective home location detection, and access to the data can disappear quickly [44–46].
- *Smart card data* include transactions at payment booths such as at subway stations and inside public transit buses [32]. These are usually anonymized and frequent, but they can only be used to infer public transit mobility patterns adequately as opposed to home locations.
- *Call detailed records (CDRs)* provide geolocation data at the cell tower level. Such datasets are characterized by large spatiotemporal density and coverage, but the quality of the detected homes is subject to the spatial distribution of the cell towers rather than the users' activity patterns [26]. Nonetheless, they have been used extensively to understand people's travel and activity patterns during the recording period [26–28].
- *Passively collected GPS data* are usually obtained from mobile devices such as smartphones and tablets and automobiles that have location-based services (LBS) enabled [43]. GPS data overcomes the main problem with CDRs by providing the exact locations and overcomes geotagged posts by providing continuous and high-frequency records. Further, GPS can provide more detailed information about the movements of individuals, including their speed, direction, and stop durations along the way. Therefore, it has seen a substantial increase in availability and use in the last decade. In this study, we use this data kind for our analysis.

Method types

Density-based clustering methods are commonly used in the literature to estimate home locations, such as DBSCAN [47] (used in [48]) and mean shift clustering [49] for home detection (used in [50]).

Mean-shift clustering [49] is a popular density-based clustering method that has been used in several studies [3, 20, 50, 51], probably owing to its simplicity in having just one main parameter—the radius of flat kernel for kernel density estimation (KDE). DBSCAN [47] and its variants (e.g., [48, 52]), on the other hand, have two main parameters – the maximum intra-cluster distance at each iteration and the minimum number of points in an acceptable neighborhood. In both these methods, the results of the clustering can be substantially sensitive to the choice of these parameters [53].

Heuristic algorithms are widely applied to detect home locations, which rely on various decision rules on the time and frequencies of user records in specific areas during observations [12, 16, 22, 43]. The most intuitive assumption is that users have the highest records at home, and their home locations are identified based on the density of the data. Different variants are proposed by shifting the rules, such as determining their home as locations with the highest number of nighttime records or the most distinct days.

Li et al. (2008) [54] developed a rule-based method for detecting 'stay points' which represent spatiotemporal regions of low movement and are thus helpful in trip segmentation. These stay points are computed by identifying the breaks in the time gap and distance between the first and last point of a sequential set of points based on given thresholds of time gap (30 min) and distance (200 m). This method was further modified by Sadeghinasr et al. (2019) [43] who clustered these stay points using hierarchical clustering into stay regions and identified home locations as the most visited stay regions during nighttime. Other methods, such as the center-point algorithm by Zou et al. (2018) [32], which uses one's middle point of the first to-subway trip's origin and the last from-subway trip's destination to represent the home location, are easy to compute but have been shown to perform fairly well up to a large radius of tolerance (e.g., [32, 44]).

Current studies that compared different HDAs already demonstrated that the results are sensitive to criteria choice, such as night time periods. For instance, Vanhoof et al. [22] primarily focused on assessing the effects of different night periods on the home detection results while ignoring the limitations of the HDA. Pappalardo et al. [12] compared five similar HDAs and validated the results with multiple small-scale datasets, yet they neglected to consider factors such as data quality and period. In contrast, this study concerns comparing the HDAs, with a particular emphasis on testing across scenarios spanning different regions, data periods, and data quality.

2.2 Algorithms used in this study

Five HDAs are compared in this study, including a simple baseline algorithm, three algorithms listed in the 'Passive GPS data' section of Table 1, and a derivative of one of those algorithms as proposed in this study. The steps involved in these algorithms, labeled as $A_1, \ldots, A_5$, are illustrated in Fig. 2. The same input dataset is used for each of these HDAs. For clustering-based methods, the implementations of `scikit-learn`, a popular Python-based machine learning library, are used. The common set of users resulting from each of these HDAs is used for subsequent performance assessment.

*2.2.1 A*1: *centroid method*

This is the simplest of all the considered HDAs and is meant to serve as the baseline for comparison with the other algorithms. In this case, a user's home location is simply computed as the centroid (or alternatively the medoid) of all their nighttime ping locations over the entire study period, following the assumption that a person's most probable location during the night is their home. This is similar to, but not exactly the same as, most popular cell tower-based algorithms in the case of CDR data [27–29].

*2.2.2 A*2: *grid frequency method*

This HDA was used in Zhao et al. (2022) [31]. They first divided the study region into a square grid with cells of 20 $\times$ 20 meters. They considered the home location as the

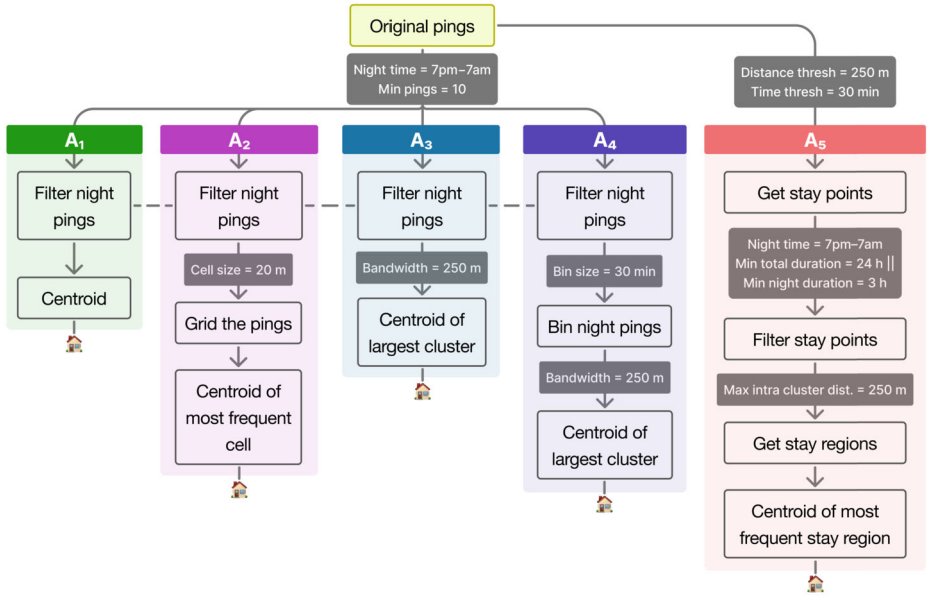

Figure 2 *Flowchart of the steps of the HDAs compared in this study*. The values shaded in grey depict the algorithms' parameters. The dashed lines between two HDAs depict the same or equivalent step between the two HDAs

mean location of the pings of the cell with the most nighttime pings over the study duration.

2.2.3 A3: all-time clustering method

This method involves finding the most popular cluster of all the pings in the nighttime data taken together without distinguishing the temporal variation in locations during this night time. Though several clustering methods exist as explained in Sect. 2.1.2, this method particularly uses mean-shift clustering with the same parameters as in [3, 23, 50]. All these studies use a flat kernel with a radius of 250 m for KDE. In this study, other parameters in this method such as the sampling strategy for the KDE process and the number of iterations in the hill climb process are controlled to prefer accuracy over runtime speed.

2.2.4 A4: binned clustering method

HDA A_3 uses clustering of all the nighttime points at once, meaning it does not distinguish between the following cases: (i) a scenario where most of the nighttime points are concentrated in a small time period (e.g., 10:00–10:20 PM) where the user might possibly be in movement and thus more likely to enable LBS, and (ii) a scenario where the same number of nighttime points as in case (i) are distributed evenly across the night. It can be argued that the latter case provides more confidence in the inferred home location since it relies on better-sampled data.

To overcome this limitation of A_3, we propose an adaptation in the form of A_4 where the nighttime points are collected at fixed time intervals over the study period. The centroids of these locations are computed and used as inputs in mean-shift clustering. Similar to A_3, the centroid of the largest cluster is labeled the home location. This HDA introduces a parameter in addition to those of A_3—the binning period, which is taken as 30 minutes in this study.

2.2.5 A5: stay-point method

This HDA was proposed by Sadeghinasr et al. (2019) [24] where they used the stay point detection algorithm proposed by Li et al. (2008) [54] to first identify stay points and then cluster the stay points using hierarchical clustering into stay regions by setting a threshold of a maximum intra-cluster distance of 250 m. Then, they considered the home locations as the most visited stay regions during nighttime (8 PM–5 AM) which had a visit duration of at least 3 hours during the nighttime or a total duration of at least 24 hours.

3 Experimental setup

3.1 Data description

3.1.1 Smartphone GPS data

This study uses GPS trace data collected using LBS on smartphones and tablets, aggregated and anonymized by a private vendor. The trace table (illustrated in Table 2) comprises events (called 'pings' here) which include a mobile device's (called 'user' here) anonymized identifier, latitude and longitude of the point, an estimate of the radius of GPS recording error for that ping, and the Unix-style timestamp of the event (seconds passed since Jan 1, 1970 UTC+00:00). More details about the data are provided in the Supplementary Sect. 1 in Additional file 1.

LBS data is usually slightly erroneous due to inaccuracies in the GPS logging system and thus needs preprocessing for better results. The preliminary data filtering done to create the dataset samples includes removing pings with an error radius of more than 50 m, those with segment speed of more than 50 m/s (180 km/h), and those with acceleration outside the range of −10 to 10 m/s^2 (based on works like [55, 56]). For reference, for the ith ping in the sequence trace with coordinates $\mathbf{x}_i = (x_i, y_i)$ and timestamp t_i, its speed is given by $v_i = \frac{d(\mathbf{x}_i, \mathbf{x}_{i-1})}{t_i - t_{i-1}}$ and the acceleration by $a_i = \frac{v_i - v_{i-1}}{t_i - t_{i-1}}$, where d is the Haversine distance function. By definition, $v_1 = a_1 = a_2 = 0$.

3.1.2 Study regions and periods

Four U.S. metropolitan statistical areas (MSAs) are assessed in this study—(i) Austin, TX, (ii) Baton Rouge, LA, (iii) Houston, TX, and (iv) Indianapolis. The counties included in these MSAs, their total area, and their total population (as of the 2020 5-year estimates of the American Community Survey (ACS)) are shown in Fig. 3.

These regions are chosen from the cities with available land use and smartphone GPS data so as to cover a diverse set of scales and land use patterns. Baton Rouge has a large but sparsely populated MSA, whereas Houston has a much larger MSA. Austin and Indianapolis lie in between but represent cities with very different land use distributions and urban layouts. Houston is known for its sprawling layout, with significant suburban development extending into multiple counties. The city is characterized by a lack of zoning laws,

Table 2 A sample of the GPS data used in this study. The coordinates have been fuzzed for illustration

Row	Device ID	Longitude	Latitude	Timestamp (s)	Error radius (m)
1	107258c2-c027-41c9-aa4d-166951bd5007	−86.964964	40.064320	1552588288.0	22
2	ad96c788-965d-4074-bf28-306a3cf6cb07	−85.982222	39.848495	1552594937.0	6
…	…	…	…	…	…
…	d3286a43-a68c-42cf-ba71-e838e2276b1a	−86.514245	41.672555	1552579184.0	7

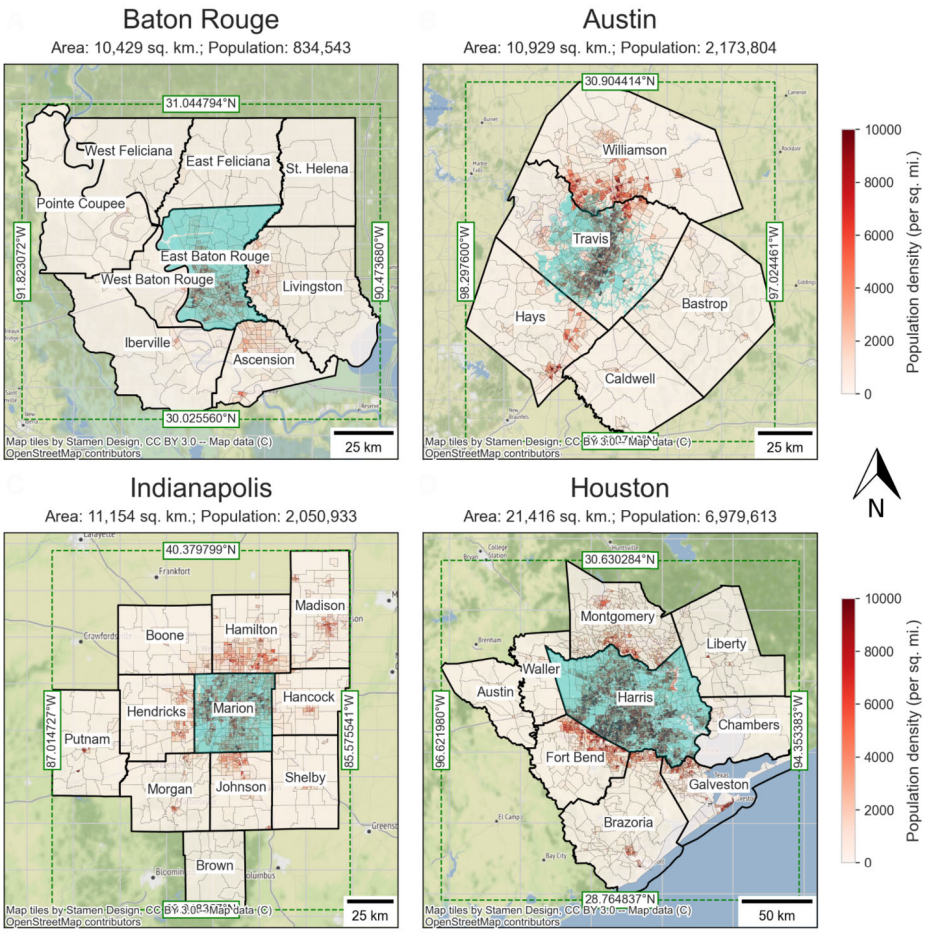

Figure 3 *Study regions showing the metropolitan statistical area (MSA) counties and bounding boxes*. The population density of the census block groups as per ACS 2020 data is colored in red. The regions covered in the land use maps are shaded in cyan

which has led to a unique pattern of residential, commercial, and industrial areas often being interspersed [57]. Austin is characterized by a higher density in the city center, with the urban core being home to a mix of residential, commercial, and cultural facilities (towards mixed-use developments) [58]. By conducting analysis for these four cities with different socio-economic contexts, underlying data characteristics, and scale and complexity, we ensure that our tests are robust and generalizable across various urban settings.

In addition to spatial variation, the datasets used for testing the HDAs are created so as to include temporal variation as well. Particularly, two case studies are chosen to represent the potential temporal difference of HDA outputs before and after two specific events. The first event is Hurricane Ida which caused damage in southeastern Louisiana upon landfall on August 29, 2020, causing waves of evacuation and displacement around the region, including in Baton Rouge. The periods depicting stability before the landfall, during the mobilization period around landfall, and long after the event are considered in this analysis.

The second event is the first government-mandated lockdown in Indiana on March 16, 2020 following the outbreak of COVID-19 in the United States which was known to have

Table 3 Description of the study datasets (combinations of region and analysis period)

ID	Region	Period	#(Days)	#(Users) (k)	% of Popu.	#(Pings) (M)
D_1	Baton Rouge, LA	Aug 1–25, 2021	25	165.5	19.8%	245.8
D_2		Aug 26–Sep 7, 2021	13	87.7	10.5%	65.5
D_3		Sep 8–Nov 30, 2021	84	316.5	37.9%	1120.0
D_4	Indianapolis, IN	Mar 1–15, 2020	15	273.9	13.4%	101.6
D_5		Mar 16–31, 2020	16	251.6	12.3%	127.2
D_6		Mar 1–31, 2020	31	445.1	21.7%	241.6
D_7	Austin, TX	Jul 1–7, 2021	7	166.9	7.7%	97.6
D_8	Houston, TX	Jul 1–7, 2021	7	538.8	7.7%	331.1

drastically reduced mobility. A before-after comparison of the HDAs of these events is deemed useful in explaining the robustness of the HDAs. This is explained in Sect. 4.3.

With these two combinations of study regions and periods, a total of 8 datasets are prepared for testing the HDAs. These are shown in Table 3. The number of unique devices (referred to as 'users') obtained after cleaning the GPS data and their ratio to the regional population are also shown. Similarly, the number of filtered pings is also shown for each dataset, reflecting the scale of variation in the test datasets. Note that the pings are filtered within the regions' bounding boxes (shown in green dashed outlines in Fig. 3) instead of filtering within the MSA counties for the sake of performance speed.

3.2 Performance metrics

In the absence of ground-truth information on device users' home locations, the accuracy of the HDAs is tested using three approximate or pseudo-performance metrics. All these are based on some assumptions that are generally considered valid intuitively and in the literature.

*3.2.1 M*1: *residential detection rate*

This metric makes use of the idea that a good HDA should detect more homes in a city's residential areas as opposed to other land use categories such as commercial, industrial, and forests. This metric is given by the proportion of homes detected by a given HDA in the residential area of the region based on its land use distribution (see Supplementary Sect. 2.1 for more details). To offset some potential mislocation errors due to the nature of the GPS data and the often convoluted land use maps, tolerance buffers of different widths are also considered in the calculation. This results in the following definition of the performance metric:

$$M_1(A) = \sum_{r=0}^{r_{\max}} w(r)\rho_A(r), \quad \text{where } w(r) = \frac{r_{\max} - r}{\sum_{r=0}^{r_{\max}} r_{\max} - r}. \tag{1}$$

Here, for buffers of width r, ranging from zero to $r_{\max}$, $\rho_A(r)$ is the proportion of homes detected in the combined buffered residential area detected by HDA A. For instance, a value of $M_1(A) = 0.4$ can be roughly interpreted as 40% of the users' home locations detected by HDA A lying within a region of the city classified as 'residential'. In the subsequent experiments, the value of $r_{\max}$ is taken as 50 m, with buffer width increments of 5 m, the same as the maximum allowed error in GPS spatial accuracy as explained in Sect. 3.1.1.

3.2.2 *M*2: *proximity to daily data*

This metric uses the idea that a home location should be the origin/destination of one's daily trips. Given a user's home location detected by a given HDA, this metric involves computing its distance to the closest ping in that user's nighttime pings on each day in the study period using Haversine distance. Then, the median of these daily shortest distances is taken for each user. The cumulative density function (CDF) of this median shortest distance is drawn and the normalized area under the curve is computed. This represents the proximity performance metric, given by the following:

$$M_2(A) = \frac{1}{\delta_{\max}} \int_{\delta=0}^{\delta_{\max}} F_A^{\Delta}(\delta) \cdot d\delta, \quad \text{where } \delta_{A,i} = q^{0.50}_{t \in 1:n_T} \Big(\min_{x_{i,t} \in \mathbf{X}_{i,t}} \| h_{A,i} - x_{i,t} \|_2 \Big). \tag{2}$$

Here, F_A^{Δ} is the CDF of the median shortest distance of the users detected by HDA A, $\delta_{A,i}$ is the median shortest distance for the user i whose home location is given by $h_{A,i}$, $\mathbf{X}_{i,t}$ is the set of nighttime pings at night t, $q^{0.50}$ represents the 0.50 quantile (that is, median) over all study days up to n_T, and $\delta_{\max}$ is a reasonable upper limit, taken as 5 km.

3.2.3 *M*3: *home stay duration*

This metric is based on the idea that people typically spend the majority of their nighttime at their homes. For a given user, we first identify the locations they visit during nighttime hours using a stay region detection method similar to Sadeghinasr et al. (2019) [43] but with an adaptive linkage calculation (details of this method are provided in Supplementary Sect. 2.2). The stay region closest to each user's detected home location is assigned as their 'home region'. With the detected stay regions, the performance metric for each user is simply the ratio of time outside the home region to the total time spent in all stay regions. The overall performance metric is given by the area under the curve of the CDF of this value:

$$M_3(A) = \int_0^1 F_A^{\tau}(r) \cdot dr, \quad \text{where } r_{A,i} = \frac{\tau(C_{h_{A,i}})}{\sum_{k=1}^{K_i} \tau(C_{h_{k,i}})}. \tag{3}$$

Here, F_A^{τ} is the CDF of the ratio, $r_{A,i}$, of time (τ) spent in the home stay region, $C_{h_{A,i}}$, to the maximum time spent in any stay region $C_{k,i}$ over all the users i detected by HDA A, and K_i is the total number of stay regions detected for user i. Similar to M_1 and M_2, a higher value of M_3 indicates a better HDA.

4 Results

The HDAs listed in Sect. 2.2 are compared on the basis of their precision, as approximated by the three performance metrics in Sect. 3.2 and their sensitivity to data quality. These are described in the subsequent sections.

4.1 Performance comparison

The visual comparison of the performance metrics M_1, M_2, and M_3 across the HDAs over all the datasets is shown in Fig. 4. The generating curves of these metrics are provided in Supplementary Fig. 2. In Fig. 4, the size of the radar polygons depicts the overall performance of an HDA, while the skewness of the polygons hints at the differences in the behavior of the HDAs across different datasets.

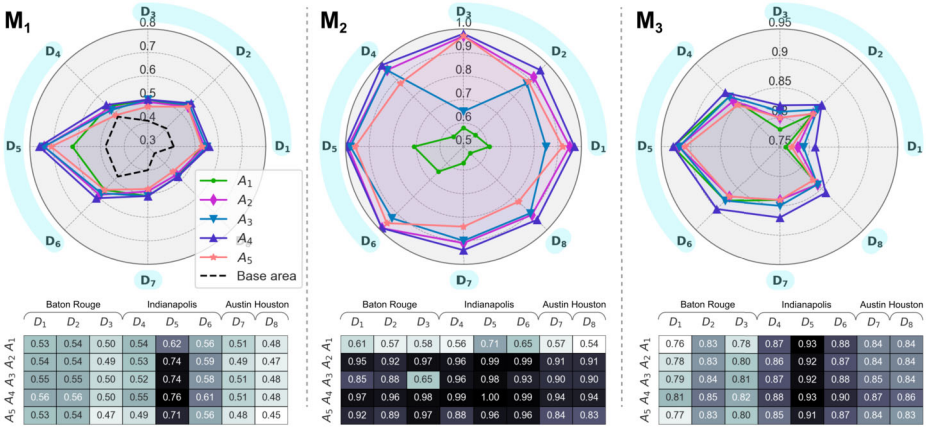

Figure 4 *Performance metrics for the HDAs across the study datasets*. For M_1, the dashed black line represents a uniform random selection algorithm based on the residential area buffers up to 50 m. The datasets of the same city are grouped in cyan

4.1.1 Overall differences by HDA

The findings from the plots in Fig. 4 are diverse and vital. First, A_1 consistently performs the worst in these overall results. This is expected, as A_1 is a very straightforward HDA with several key limitations (i) It is difficult to find the most frequent place among GPS points is not easy due to high data precision; (ii) The centroid may not necessarily be the most probable location; (iii) The results of this method are heavily susceptible to disturbances due to outliers; (iv) This method does not distinguish between spatiotemporal regions of stay and movement. For people with high movement during the night, the mean value of the coordinates can shift the detected home substantially far away from the user's trajectory. This explains why A_1 performs substantially worse in the case of M_2 compared to the other HDAs, since M_2 directly involves computing the distance of the detected home location with the closest nighttime trajectory point.

The performance of the other algorithms is largely similar, with some exceptions. Algorithm A_4 consistently performs better than the others, as is evident from the largest radar polygons corresponding to A_4 in the three metrics. In particular, although A_4 requires a data filtering criterion on its base HDA A_3 and thus operates on fewer data points than A_3, it performs better than that. This might be attributed to the focus on data quality over quantity by discretizing the data temporally, as explained in Sect. 2.2.3. This is important because it is possible for users to have high LBS activity during traveling (e.g., for navigation services) which may overshadow the location data during stay periods such as at home. Since traveling generally occurs far from home, all HDAs other than A_4 are more likely to consider these irrelevant points for the home detection process.

This bias is reduced to a lesser extent in A_3 and A_5 that rely on clustering. This positive impact of discretization is also evident in terms of space. A_2, which is a very simple heuristic that only involves finding the most visited grid cell, i.e., the discretization of space, performs, with metric values finishing close to A_4 in most cases.

The rule-based HDA A_5 is generally found to perform slightly worse than A_3, although this pattern reverses in the case of M_2. Both A_3 and A_5 involve clustering, but the order and kind of clustering are different between the two. It may be argued that the time and distance-based thresholds involved in the stay point detection step of A_5 might hamper

the performance of the algorithm since those thresholds do not take into account the continuity of the data.

4.1.2 Differences by dataset

The radar plot in Fig. 4 also shows the significant differences in the performance of the same HDA in different datasets. Notably, all the metrics are observed to be the highest in the case of D_5. It should be noted that datasets D_4, D_5, and D_6 have the same underlying urban land use and transportation networks. D_5 corresponds to the period of reduced mobility and high stay-at-home rates during the surge of COVID-19 in the Indianapolis region. It includes the date of the first death related to COVID-19 recorded in the region on March 16, 2020, and the imposition of the government-mandated lockdown on March 23 [59]. Since people were more likely to stay at home during the period of D_5, the data quality for the HDAs was substantially better than the other datasets, making it easier for all the HDAs to perform the best. This is made further prominent in the stark difference between D_4 and D_5 in the value of M_3 which depends on the time spent at home.

Moreover, the performance metric values for D_6 are consistently near the corresponding values of D_4 and D_5. This makes sense given that the period of D_6 is the union of the periods of D_4 and D_5 which are of equal length. This indicates that the better data quality of D_5 does not inordinately skew the performance metrics.

In Baton Rouge, the effect of Hurricane Ian is observed to be small yet important. This is evident in the higher values of M_3 for dataset D_2 that corresponds to the period close and immediately after the hurricane landfall compared to the pre-landfall (D_1) and long-term post-landfall (D_3) periods. However, the values of M_1 and M_2 do not vary significantly between D_1, D_2, and D_3.

4.1.3 Differences by metric

The ranges and behaviors of the three performance metrics also shed light on the nature of the analysis of this study. First, M_1 has a large range of 0.45 to 0.76. All the tested HDAs perform substantially better than a random uniform HDA where the residential detection rate curve is plotted by simply computing the proportion of land use region covered by residential areas. This is evident in Fig. 4A where the black dashed curve (denoting this uniform random HDA) is significantly smaller than those of the other HDAs in the plot. It must be noted, however, that M_1 relies on assumptions about home location that might not always hold true and could have skewed the results. For example, some users may stay at places other than their homes (such as a hotel or a relative's residence). Similarly, the home locations of night-shift workers may be overrepresented in the commercial areas of a city and thus reduce the value of M_1.

The case for M_2 is also similar. It has a substantially small range outside of the poorest performing A_1. This could be attributed to the fact that M_2 is unidirectional in its utility. That is, a small shortest distance of trajectory points from home only serves as a necessary condition for a good HDA, not a sufficient condition. Its computation relies on the distance to the closest point to the trajectory. Since the home locations are detected based on the trajectory itself, it is highly probable for an HDA to produce a high value of M_2 for a set of users who do not travel very long distances.

4.2 Sensitivity to data quality

In the previous sections, we observed the difference in the performance of the test HDAs. While it was shown that the continuity of data discretized in space and time substantially influences the goodness of an HDA, there is substantial nuance to the effect of data quality in terms of overall ping density on this goodness. In this section, we particularly ask the question: "if an analyst has geolocation data of a specific ping density, which HDA should they choose for their analysis?"

Building on the notion of ping density, the data quality of a user in this section is defined as the mean number of pings per night in their data points. Users with more pings on average are expected to have higher quality data and yield better home location detection results. At the same time, however, owing to the nature of mobile phone geolocation data, most users have very few data points, making home detection a difficult task (for reference, see SM). This means that a good HDA should strike the balance between good data quantity and quality.

To achieve this, we recomputed the performance of the HDAs on several subsets of the users by dividing them by their data quality, given by their mean nightly ping count. To simplify the decision-making for HDA choice, we further computed the mean value of the three metrics for the subset of users contained in each bin, given by $\bar{M} = \frac{1}{3}(M_1 + M_2 + M_3)$. The results of these aggregate metrics are shown in Fig. 5.

The findings of this figure are aligned with those in the previous section. First, we see here that at nearly all levels of data quality, the order of performance is largely consistent with the overall results shown in Fig. 4. A_4 still consistently performs the best, closely followed by A_3 and A_2, while A_1 and A_5 perform considerably worse. When the data quality is measured in relative terms, i.e., using the ping count distribution of each dataset, the trends are considerably different (see the Supplementary Fig. 3).

Notably, in Fig. 5A, though A_1 performs worse than A_5 in the case of M_1 and M_2, the trend is reversed for M_3. The trends of M_3 are also different from those of the other two metrics in that, unlike them, M_3 decreases with increasing data quality. It is likely because it involves computing the ratio of time spent in the detected stay-at-home region, which is likely to be exactly the same as the only (or one of the only) stay region detected for low-quality data users since they do not have enough data, to begin with. In contrast, M_1

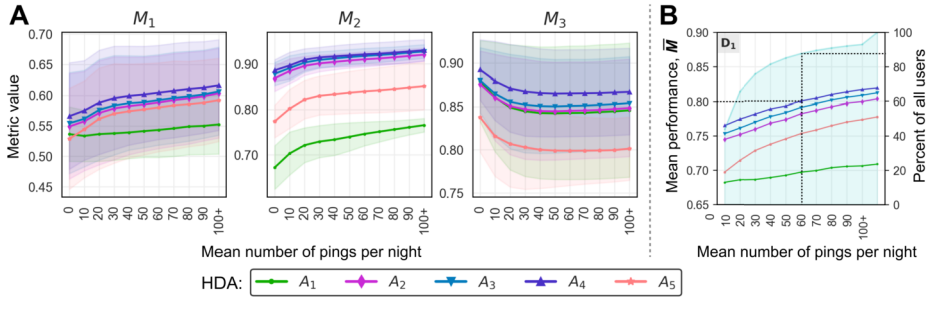

Figure 5 *Impact of data quality on HDA performance.* Each value x on the x-axis represents the subset of users having at least x pings per night on average. (A) Comparison of the mean value ($\bar{x}$) of each metric across all the datasets. The shaded regions correspond to the range $\bar{x} \pm \sigma$, where σ is the standard deviation across the datasets (B) Comparison of the mean of the three metrics for one dataset. For reference, the CDF of the users sorted by the average nightly ping count (x-axis) is shown in the shaded blue curve on the right y-axis

and M_2 rely on the richness of the data in increasing the likelihood of locating a user in a residential region and closeness to the trajectory respectively.

To compare the overall relative performance of the HDAs, we also computed the mean value of each of the three metrics across all the study datasets. The result of one dataset D_1 is shown in Fig. 5B. Similar results for the other datasets are shown in the Supplementary Fig. 4. It can be seen that the opposite trends of M_1 and M_2 with M_3 are balanced to some extent when their values are averaged. There is a steady but small increase variation in the value of $\bar{M}$ as the user quality increases in D_1. This shows that there is merit in choosing these metrics as their values do not show any abrupt behavior over different data quality categories.

This comparison is also helpful in making the choice of data filtering required for any downstream application of home location detection. For example, suppose we decide that a mean performance value of 0.8 is acceptable in a dataset similar in ping count distribution to D_1 and an urban land use similar to Baton Rouge. Then, we can refer to Fig. 5B to see that, for example, for HDA A_4, users with at least 50 pings per night would be required for analysis (dotted vertical line). This corresponds to the 13% best quality users of the dataset since 87% of the users have fewer than 50 pings (right horizontal dotted line).

4.3 Impact on applications

To see how different HDAs would influence applications of human mobility assessment and how our performance metrics could help improve the results, we conduct two experiments on common tasks where smartphone data is considered superior to other sources. These are explained in the subsequent sections.

4.3.1 Hurricane evacuation identification

Large-scale GPS data is used to estimate the evacuation/return patterns during natural disasters [8, 31]. In this task, a crucial factor is the distance between individuals' post-disaster stay locations and original home locations before the disaster.

Here, we calculate this factor based on D_1 (before landfall) and D_3 (aftermath of Hurricane Ida) using the five test HDAs. Then, we estimate the evacuation ratio using the threshold, i.e., if the distance between an individual's pre- and post-disaster home locations exceeds 1 km, we consider them as evacuated.

We observe that among the five HDAs, A_1 and A_5 produce significantly different distributions of the distance between pre- and post-disaster homes (see Fig. 6). Even for HDAs with similar CDF curves, it can be seen from Fig. 6B that they can generate a significant estimation of evacuation ratio in some areas (e.g., the northwestern part and the southern part of the city). When connecting these results with the observations of Sect. 4.1, we notice that the HDAs with good and similar performance metrics (namely A_3, A_4, and A_2) tend to create similar results. In contrast, A_1 and A_5 result in much higher evacuation rates. Since evacuation rates are essential in assessing policies and equity issues related to home evacuation, in-place sheltering, and disaster recovery, it can be imaged that adopting an arbitrary HDA can yield substantial negative impacts on policymaking [60].

4.3.2 COVID-19 policy impacts assessment

GPS-based cell phone location data has been extensively used to evaluate mobility patterns and potential solutions during COVID-19 [61]. These include evaluating alterations

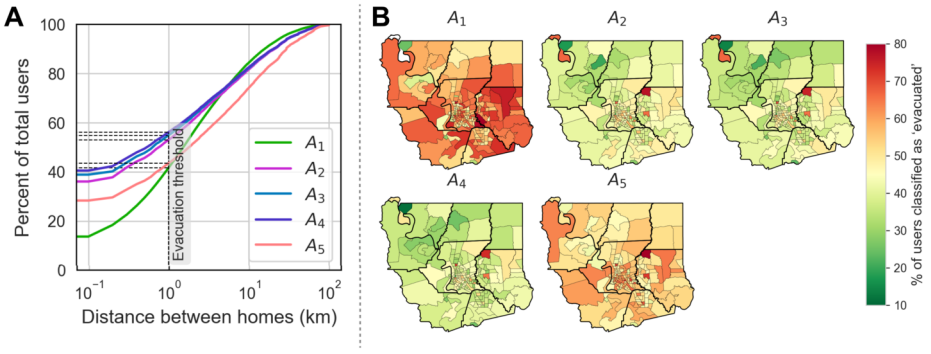

Figure 6 *Evacuation identification results under different HDAs*. (A) CDF of the distance between homes identified before (dataset D_1) and after Hurricane Ida (D_3) in Baton Rouge MSA. The threshold to classify users as displaced (1 km) is highlighted. (B) Identified percentage of users classified as evacuated by census tracts

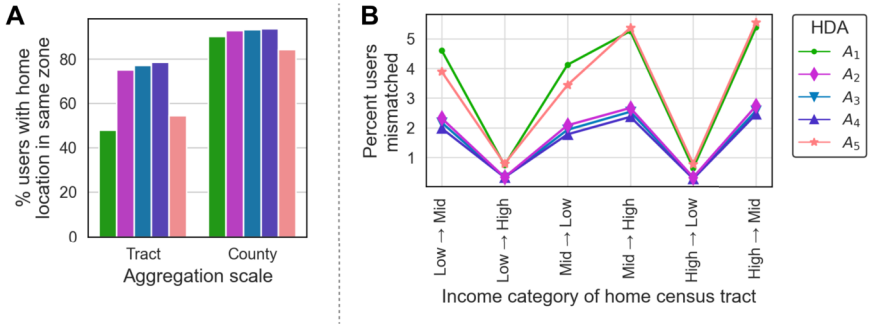

Figure 7 *Consistency of home locations in the Indianapolis region before (D_4) and after (D_5) COVID-19 mobility restrictions*. (A) Percentage of users with consistent home inference in the same zone displayed across different aggregation levels. (B) Percentage of users with income category mismatches in the inferred home tract. 'Low → Mid' indicates a change from 'Low' income category in the restricted mobility period (data D_5) to 'Medium' category in the pre-COVID normalcy mobility period (data D_4)

in population-wide mobility [62], compliance with COVID-19 policies in various demographic groups [63], and the spread and associated risk of disease from different regions [64]. However, erroneous home location inference may lead to inaccurate assessment of mobility changes and policy compliance of regions or demographics, resulting in resource misallocation and ineffective policies.

To test this concern, we compare the locations of homes inferred from each HDA in two periods: the pre-COVID-19 normal mobility period (1–15 March 2020; data D_4) and the post-lockdown mobility period (16–31 March 2020; data D_5). Ideally, high proximity between the homes inferred from both datasets for each HDA should be expected. However, significant inconsistencies are observed in certain HDAs that may lead to inaccurate inferences.

To demonstrate the consistency of inferred homes, we report the percentage of users with home locations within the same zone for each HDA (Fig. 7A). We show the results for both an aggregated administrated boundary (county) and disaggregated one (tract). With HDA A_1, only 47% of the users exhibit a consistent census tract, while with A_5, 56% of such users were observed. For all remaining users, demographic considerations can be inconsistent and imprecise. For every HDA at the spatially aggregated county level, more

than 80% of the users are classified within the same county. At both spatial levels, A_4 shows the highest consistency, with A_3 and A_2 being comparable. Therefore, for reliable analysis, this suggests using HDAs A_2, A_3, and A_4 rather than A_1 and A_5, aligning with the findings presented in Sect. 4.1.

We further investigate its potential impact on realistic applications, and income-based inequality assessment, which rely on demographic information inferred from home location. Income-based inequalities have been extensively examined using cell phone data in aspects such as access to opportunities [65], the well-being of individuals [66], emissions [67], and evacuation [50]. An inadequate HDA may result in the misclassification of users into different income groups, compromising the accuracy of assessing inequalities and characteristics associated with people from specific income groups.

We assess the percentage of users exhibiting income group discrepancies based on the median income of inferred home's census tract for two datasets for an HDA. Income groups are categorized from the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Series (LODES) dataset, comprising three categories based on monthly income: low (less than $1250), mid ($1250–$3333), and high ($3333 and above) [68]. Figure 7B shows the percentage of users with inconsistent income group classification across the two datasets. A minimal proportion of users experienced misclassification between high and low-income groups. However, a significant number of low-income users were incorrectly classified as middle-income and vice versa, resulting in a blending of categories and inaccurate assessment of behavior. Both A_1 and A_5 exhibit the highest percentage of misclassified users. The consistent performance of A_2, A_3, and A_4 suggests their suitability for studies involving demographics. These findings underscore the importance and precision of the inferred metrics, as these findings align with the results from Sect. 4.1.

5 Discussion and conclusion

In this study, we examine several home detection algorithms (HDAs) for mobile phone geolocation data, an important source that opens novel opportunities on several crucial topics. To evaluate the quality of identified home locations, we propose three performance metrics. Each metric corresponds to a feature that the true home location would likely hold: most identified homes should be located in residential areas (metric M_1), the home should be close to one's daily trajectories for every day (M_2), and people typically spend most of the nighttime at their homes (M_3). We test four representative HDAs together with one which we propose and calculate the metrics on eight datasets in four US cities with different urban layouts and population distributions. We also conduct a sensitivity analysis against data density to understand the impact of data quality on the relative quality of the detected home locations.

We find that different HDAs, even well-established in the literature, can lead to significantly different home location results. Among the five HDAs tested in this study, we observe that two of them (A_1 and A_5) consistently perform worse than other algorithms in all eight datasets. More than 20% of the homes detected by these two HDAs fall outside a 2-mile radius from the home locations estimated by the other three HDAs in the eight datasets. A_1 is a simple centroid-based algorithm that is primarily used in call detailed records (CDR) mobile phone data. Its poor performance can be attributed to its sensitivity to outlier records and a lack of consideration for other data filtering criteria

and nuances. A_5 is a more sophisticated algorithm that uses both clustering and a rule-based approach to identify the location of the home. The choice of its many parameters might be attributed to some or all of its poorer performance. The other three HDAs (A_3, A_4, and A_2) perform similarly to each other. In addition to this, it is found that all three metrics agree with each other in terms of the rank of the performance, which supports the strength of their design.

We also propose a new algorithm (A_4), which is based on A_3 with an additional process to bin every 30-minute pings to consider spatial data continuity. Under our metrics, we report that A_4 consistently performs better than other HDAs studied. It is worth noting that by adding the binning process, we also manage to reduce computational time when compared with A_3. Although computational time is not a big concern for this offline task, it becomes important if the size of the samples is substantially large.

We perform a sensitivity analysis of the data quality to provide useful suggestions to researchers who might encounter different data collection frequencies and sample rates. It is found that the order of relative performance remains largely the same even for different subsets of mobile phone devices ranked by their data quality.

Further, we explore the implications of different HDAs and their corresponding metrics subsequent applications in human mobility assessment. We use two tasks to build our experiments: evacuation identification and COVID-19 policy impact assessment. There are two main takeaways: first, the use of different HDAs could significantly influence downstream results; second, by preferring HDAs with high (and similar) performance metrics, the results are more consistent and closer to expected behavior.

We expect our work can provide the following values to researchers and practitioners who are using HDAs. First, we hope that this study can shed light on a previously unexamined issue: the quality of detected homes and their potential influences on findings in subsequent applications in human mobility assessment. These findings could be different for different fields. For example, in evacuation assessment, geolocation data may not be available for a lot of nights. Using a limited amount of data may impact the quality of the estimated home locations. Second, we recommend that researchers use these metrics to compare the performance of the used HDA with others for their use case (geographic location, time period, and data quality) before finalizing that HDA. In urban planning, for example, planners might want to select a different HDA based on the data quality threshold they choose for their planning region to estimate home-based trip rates. Third, we expect our results to establish general ideas about what makes a good HDA. In the literature, we observe different researchers tend to adopt or even design different HDAs based on their available data. In this case, information such as data continuity (across different times of the day) matters more than the data density and can provide useful guidance for their methodology design. Moreover, since HDAs are commonly shared in many applications of passively collected human mobility data, we have created an open-source toolbox [69] to facilitate access to our proposed metrics and different HDAs.

We also recognize some limitations of our study and some related topics that merit further examination. First, due to the absence of large-scale true home locations, our evaluation can only be indirect. Note this is also the motivation for performing home detection, which suggests that this would be a limitation for all HDAs when they are applied in practice. Here, we introduce the COVID-19 scenario to alleviate this issue as the impact of the lockdown influence on human mobility is well studied and accepted. Given that the infor-

mation about people's exact home locations is very sensitive, we expect the restrictions to be unlikely to be fully resolved, but we expect future events to provide opportunities to create more evidence. Second, we recognize that the datasets used in this study may not reflect the nature, quality, and quantity of data available to other researchers. Finally, our proposed metrics are 'necessary' conditions in the sense that the detected homes are good, as they align with our intuition of the features that a real home location would follow. It would be interesting to establish the 'sufficient' conditions for an HDA's results to be acceptable. To establish such standards, we posit the need for more and diverse empirical evidence with our proposed metrics.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-023-00447-w.

Additional file 1. A Supplementary Material document is available for this manuscript. (PDF 7.1 MB)

Acknowledgements
The geolocation data was purchased from a well-established data vendor.

Funding
This study was not sponsored by any funding agency.

Abbreviations
ACS, American Community Survey; CDF, Cumulative Density Function; CDR, Call Detailed Records; COVID-19, Coronavirus Disease 2019; DBSCAN, Density-Based Spatial Clustering of Applications with Noise; GPS, Global Positioning System; HDA, Home Detection Algorithm; KDE, Kernel Density Estimation; LBS, Location-Based Services; LEHD, Longitudinal Employer-Household Dynamics; LODES, LEHD Origin-Destination Employment Statistics; MSA, Metropolitan Statistical Area; UTC, Coordinated Universal Time.

Data availability
The complete smartphone geolocation data analyzed in the current study are not publicly available due to a contract with the private data provider. However, samples of the processed datasets are available from the corresponding author upon reasonable request. All the relevant code used to process the data sets used in this study is available at https://github.com/rvanxer/home_detection.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author contributions
RV, SM, XC, and ZL conceived the study design. XC reviewed the research literature. SU collected the mobile phone geolocation data and supervised the study. RV, ZL, and SM designed the three performance metrics M_1, M_2, and M_3 respectively. RV analyzed the results and prepared the results for the 'Performance comparison' section. ZL and SM prepared the results for the 'Impact on applications' section. All authors read and approved the final manuscript.

Received: 2 August 2023 Accepted: 23 December 2023 Published online: 16 January 2024

References

1. Çolak S, Alexander LP, Alvim BG, Mehndiratta SR, González MC (2015) Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. Transp Res Rec 2526(1):126–135
2. Lai S, Erbach-Schoenberg EZ, Pezzulo C, Ruktanonchai NW, Sorichetta A, Steele J, Li T, Dooley CA, Tatem AJ (2019) Exploring the use of mobile phone data for national migration statistics. Palgrave Commun 5(1):1–10
3. Yabe T, Jones NK, Rao PSC, Gonzalez MC, Ukkusuri SV (2022) Mobile phone location data for disasters: a review from natural hazards and epidemics. Comput Environ Urban Syst 94:101777. https://doi.org/10.1016/j.compenvurbsys.2022.101777
4. Guo Y, Chen Z, Stuart A, Li X, Zhang Y (2020) A systematic overview of transportation equity in terms of accessibility, traffic emissions, and safety outcomes: from conventional to emerging technologies. Transp Res Interdiscip Perspect 4:100091. https://doi.org/10.1016/j.trip.2020.100091
5. Schneider CM, Belik V, Couronné T, Smoreda Z, González MC (2013) Unravelling daily human mobility motifs. J R Soc Interface 10(84):20130246

6. Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782
7. Anda C, Erath A, Fourie PJ (2017) Transport modelling in the age of big data. Int J Urban Sci 21(sup1):19–42
8. Yabe T, Sekimoto Y, Tsubouchi K, Ikemoto S (2019) Cross-comparative analysis of evacuation behavior after earthquakes using mobile phone data. PLoS ONE 14(2):0211375. https://doi.org/10.1371/journal.pone.0211375
9. Demissie MG, Phithakkitnukoon S, Kattan L (2018) Trip distribution modeling using mobile phone data: emphasis on intra-zonal trips. IEEE Trans Intell Transp Syst 20(7):2605–2617
10. Song C, Koren T, Wang P, Barabási A-L (2010) Modelling the scaling properties of human mobility. Nat Phys 6(10):818–823
11. Harrison G, Grant-Muller SM, Hodgson FC (2020) New and emerging data forms in transportation planning and policy: opportunities and challenges for "track and trace" data. Transp Res, Part C, Emerg Technol 117:102672
12. Pappalardo L, Ferres L, Sacasa M, Cattuto C, Bravo L (2021) Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. EPJ Data Sci 10(1):29. https://doi.org/10.1140/epjds/s13688-021-00284-9
13. Dypvik Landmark A, Arnesen P, Södersten C-J, Hjelkrem OA (2021) Mobile phone data in transportation research: methods for benchmarking against other data sources. Transportation 48:2883–2905
14. García-Albertos P, Picornell M, Salas-Olmedo MH, Gutiérrez J (2019) Exploring the potential of mobile phone records and online route planners for dynamic accessibility analysis. Transp Res, Part A, Policy Pract 125:294–307
15. Jiang S, Fiore GA, Yang Y, Ferreira J Jr, Frazzoli E, González MC (2013) A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: Proceedings of the 2nd ACM SIGKDD international workshop on urban computing, pp 1–9
16. Vanhoof M, Reis F, Ploetz T, Smoreda Z (2018) Assessing the quality of home detection from mobile phone data for official statistics. J Off Stat 34(4):935–960. https://doi.org/10.2478/jos-2018-0046
17. Isaacman S, Becker R, Cáceres R, Kobourov S, Martonosi M, Rowland J, Varshavsky A (2011) Identifying important places in people's lives from cellular network data. In: Pervasive computing: 9th international conference, pervasive 2011, San Francisco, USA, June 12–15, 2011. Proceedings 9. Springer, Berlin, pp 133–151. https://doi.org/10.1007/978-3-642-21726-5_9
18. Bayat S, Naglie G, Rapoport MJ, Stasiulis E, Chikhaoui B, Mihailidis A et al (2020) Inferring destinations and activity types of older adults from GPS data: algorithm development and validation. JMIR Aging 3(2):18008. https://doi.org/10.2196/18008
19. Zhou C, Frankowski D, Ludford P, Shekhar S, Terveen L (2007) Discovering personally meaningful places: an interactive clustering approach. ACM Trans Inf Syst 25(3):12. https://doi.org/10.1145/1247715.1247718
20. Ashbrook D, Starner T (2003) Using GPS to learn significant locations and predict movement across multiple users. Pers Ubiquitous Comput 7:275–286. https://doi.org/10.1007/s00779-003-0240-0
21. Oosterlinck D, Baecke P, Benoit DF (2021) Home location prediction with telecom data: benchmarking heuristics with a predictive modelling approach. Expert Syst Appl 170:114507. https://doi.org/10.1016/j.eswa.2020.114507
22. Vanhoof M, Lee C, Smoreda Z (2020) Performance and sensitivities of home detection on mobile phone data. In: Big data meets survey science: a collection of innovative methods, pp 245–271. https://doi.org/10.1002/9781118976357.ch8
23. Kanasugi H, Sekimoto Y, Kurokawa M, Watanabe T, Muramatsu S, Shibasaki R (2013) Spatiotemporal route estimation consistent with human mobility using cellular network data. In: 2013 IEEE international conference on pervasive computing and communications workshops (PERCOM workshops). IEEE, New York, pp 267–272. https://doi.org/10.1109/percomw.2013.6529493
24. Sadeghinasr B, Akhavan A, Wang Q (2019) Estimating commuting patterns from high-resolution phone GPS data. In: Computing in civil engineering 2019: data, sensing, and analytics—selected papers from the ASCE international conference on computing in civil engineering 2019, pp 9–16. https://doi.org/10.1061/9780784482438.002
25. Yabe T, Ukkusuri SV, Rao PSC (2019) Mobile phone data reveals the importance of pre-disaster inter-city social ties for recovery after hurricane Maria. Appl Netw Sci 4(1):1–18. https://doi.org/10.1007/s41109-019-0221-5
26. Xu Y, Shaw S-L, Zhao Z, Yin L, Fang Z, Li Q (2015) Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. Transportation 42:625–646. https://doi.org/10.1007/s11116-015-9597-y
27. Jiang S, Ferreira J, Gonzalez MC (2017) Activity-based human mobility patterns inferred from mobile phone data: a case study of Singapore. IEEE Trans Big Data 3(2):208–219. https://doi.org/10.1109/tbdata.2016.2631141
28. Chen G, Hoteit S, Viana AC, Fiore M, Sarraute C (2018) Enriching sparse mobility information in call detail records. Comput Commun 122:44–58. https://doi.org/10.1016/j.comcom.2018.03.012
29. Leca C-L, Nicolaescu I, Rîncu C-I (2015) Significant location detection & prediction in cellular networks using artificial neural networks. Comput Sci Inf Technol 3:81–89. https://doi.org/10.13189/csit.2015.030305
30. Ahas R, Silm S, Järv O, Saluveer E, Tiru M (2010) Using mobile positioning data to model locations meaningful to users of mobile phones. J Urban Technol 17(1):3–27. https://doi.org/10.1080/10630731003597306
31. Zhao X, Xu Y, Lovreglio R, Kuligowski E, Nilsson D, Cova TJ, Wu A, Yan X (2022) Estimating wildfire evacuation decision and departure timing using large-scale GPS data. Transp Res, Part D, Transp Environ 107:103277. https://doi.org/10.1016/j.trd.2022.103277
32. Zou Q, Yao X, Zhao P, Wei H, Ren H (2018) Detecting home location and trip purposes for cardholders by mining smart card transaction data in Beijing subway. Transportation 45:919–944
33. Hasan S, Schneider CM, Ukkusuri SV, González MC (2013) Spatiotemporal patterns of urban human mobility. J Stat Phys 151:304–318. https://doi.org/10.1007/s10955-012-0645-0
34. Bojic I, Massaro E, Belyi A, Sobolevsky S, Ratti C (2015) 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. Springer, Berlin, pp 194–208. https://doi.org/10.1007/978-3-319-27433-1_14
35. Scellato S, Noulas A, Lambiotte R, Mascolo C (2011) Socio-spatial properties of online location-based social networks. In: Proceedings of the international AAAI conference on web and social media, vol 5, pp 329–336. https://doi.org/10.1609/icwsm.v5i1.14094

36. Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1082–1090. https://doi.org/10.1145/2020408.2020579
37. Sadilek A, Kautz H, Silenzio V (2012) Modeling spread of disease from social interactions. In: Proceedings of the international AAAI conference on web and social media, vol 6, pp 322–329. https://doi.org/10.1609/icwsm.v6i1.14235
38. Jenkins D (2023) Travel survey state of the practice. Technical report, Federal Highway Adminstration. https://nhts.ornl.gov/assets/NextGen%20NHTS_State%20of%20Practice_032423.pdf
39. Stalla-Bourdillon S, Knight A (2016) Anonymous data v. personal data-false debate: an EU perspective on anonymization, pseudonymization and personal data. Wis Int Law J 34:284
40. Wojtusiak J, Nia RM (2021) Location prediction using GPS trackers: can machine learning help locate the missing people with dementia? Internet Things 13:100035. https://doi.org/10.1016/j.iot.2019.01.002
41. Luo F, Cao G, Mulligan K, Li X (2016) Explore spatiotemporal and demographic characteristics of human mobility via Twitter: a case study of Chicago. Appl Geogr 70:11–25. https://doi.org/10.1016/j.apgeog.2016.03.001
42. Shen L, Stopher PR (2014) Review of GPS travel survey and GPS data-processing methods. Transp Rev 34(3):316–334. https://doi.org/10.1080/01441647.2014.903530
43. Sadeghinasr B, Akhavan A, Wang Q (2019) Estimating commuting patterns from high resolution phone GPS data. In: Computing in civil engineering 2019: data, sensing, and analytics. American Society of Civil Engineers, Reston, pp 9–16. https://doi.org/10.1061/9780784482438.002
44. Jurdak R, Zhao K, Liu J, AbouJaoude M, Cameron M, Newth D (2015) Understanding human mobility from Twitter. PLoS ONE 10(7):0131469. https://doi.org/10.1371/journal.pone.0131469
45. Wang Q, Phillips NE, Small ML, Sampson RJ (2018) Urban mobility and neighborhood isolation in America's 50 largest cities. Proc Natl Acad Sci 115(30):7735–7740. https://doi.org/10.1073/pnas.1802537115
46. Phillips NE, Levy BL, Sampson RJ, Small ML, Wang RQ (2021) The social integration of American cities: network measures of connectedness based on everyday mobility across neighborhoods. Sociol Methods Res 50(3):1110–1149
47. Ester M, Kriegel H-P, Sander J, Xu X et al (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, vol 96, pp 226–231
48. Xu Y, Clemente RD, González MC (2021) Understanding vehicular routing behavior with location-based service data. EPJ Data Sci 10(1):12. https://doi.org/10.1140/epjds/s13688-021-00267-w
49. Comaniciu D, Meer P (2002) Mean shift: a robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell 24(5):603–619. https://doi.org/10.1109/34.1000236
50. Yabe T, Ukkusuri SV (2020) Effects of income inequality on evacuation, reentry and segregation after disasters. Transp Res, Part D, Transp Environ 82:102260. https://doi.org/10.1016/j.trd.2020.102260
51. Comito C, Falcone D, Talia D (2016) Mining human mobility patterns from social geo-tagged data. Pervasive Mob Comput 33:91–107. https://doi.org/10.1016/j.pmcj.2016.06.005
52. Shah GH (2012) An improved dbscan, a density based clustering algorithm with parameter selection for high dimensional data sets. In: 2012 nirma university international conference on engineering (NUiCONE). IEEE, New York, pp 1–6. https://doi.org/10.1109/nuicone.2012.6493211
53. Fahad A, Alshatri N, Tari Z, Alamri A, Khalil I, Zomaya AY, Foufou S, Bouras A (2014) A survey of clustering algorithms for big data: taxonomy and empirical analysis. IEEE Trans Emerg Top Comput 2(3):267–279. https://doi.org/10.1109/tetc.2014.2330519
54. Li Q, Zheng Y, Xie X, Chen Y, Liu W, Ma W-Y (2008) Mining user similarity based on location history. In: Proceedings of the 16th ACM SIGSPATIAL international conference on advances in geographic information systems. ACM, New York, pp 1–10. https://doi.org/10.1145/1463434.1463477
55. Nyhan M, Sobolevsky S, Kang C, Robinson P, Corti A, Szell M, Streets D, Lu Z, Britter R, Barrett SR et al (2016) Predicting vehicular emissions in high spatial resolution using pervasively measured transportation data and microscopic emissions model. Atmos Environ 140:352–363. https://doi.org/10.1016/j.atmosenv.2016.06.018
56. Böhm M, Nanni M, Pappalardo L (2022) Gross polluters and vehicle emissions reduction. Nat Sustain 5(8):699–707. https://doi.org/10.1038/s41893-022-00903-x
57. Qian Z (2010) Without zoning: urban development and land use controls in Houston. Cities 27(1):31–41
58. Handy SL, Clifton K, Fisher J (1998) The effectiveness of land use policies as a strategy for reducing automobile dependence: a study of Austin neighborhoods. Technical report
59. King C (2021) Our year of COVID: key dates in Indiana's fight against the coronavirus. https://www.indystar.com/in-depth/news/2021/03/18/indiana-covid-timeline-key-dates-states-fight-vs-pandemic/6813412002. Accessed 2023-05-18
60. Martín Y, Li Z, Cutter SL (2017) Leveraging Twitter to gauge evacuation compliance: spatiotemporal analysis of hurricane Matthew. PLoS ONE 12(7):0181701
61. Grantz KH, Meredith HR, Cummings DA, Metcalf CJE, Grenfell BT, Giles JR, Mehta S, Solomon S, Labrique A, Kishore N et al (2020) The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. Nat Commun 11(1):4961. https://doi.org/10.1038/s41467-020-18190-5
62. Xiong C, Hu S, Yang M, Luo W, Zhang L (2020) Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. Proc Natl Acad Sci 117(44):27087–27089. https://doi.org/10.1073/pnas.2010836117
63. Bargain O, Aminjonov U (2020) Trust and compliance to public health policies in times of COVID-19. J Public Econ 192:104316
64. Verma R, Yabe T, Ukkusuri SV (2021) Spatiotemporal contact density explains the disparity of COVID-19 spread in urban neighborhoods. Sci Rep 11(1):10952. https://doi.org/10.1038/s41598-021-90483-1
65. Mittal S, Yabe T, Arroyo Arroyo F, Ukkusuri S (2023) Linking poverty-based inequalities with transportation and accessibility using mobility data: a case study of greater maputo. Transp Res Rec 2677(3):668–682. https://doi.org/10.1177/03611981221116623
66. Pappalardo L, Vanhoof M, Gabrielli L, Smoreda Z, Pedreschi D, Giannotti F (2016) An analytical framework to nowcast well-being using mobile phone data. Int J Data Sci Anal 2:75–92. https://doi.org/10.1007/s41060-016-0013-2

67. Guo H, Li W, Yao F, Wu J, Zhou X, Yue Y, Yeh AG (2020) Who are more exposed to PM2.5 pollution: a mobile phone data approach. Environ Int 143:105821. https://doi.org/10.1016/j.envint.2020.105821
68. Bureau UC (2020) Lehd origin-destination employment statistics (2002–2017). Longitudinal-Employer Household Dynamics Program
69. Verma R, Mittal S, Ukkusuri SV (2023) Mobilkit Plus. https://github.com/rvanxer/mk

Publisher's Note