Received: 24 October 2023 | Revised: 24 July 2024 | Accepted: 27 August 2024

IET Intelligent Transport Systems

DOI: 10.1049/itr2.12563

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Investigating the relative accuracy of GPS, GSM and CDR data for inferring spatiotemporal travel trajectories

Khatun E. Zannat | **Charisma F. Choudhury** | **Stephane Hess** | **David Watling**

Institute for Transport Studies, University of Leeds, Leeds, UK

Correspondence
Khatun E Zannat, Institute for Transport Studies, University of Leeds, Leeds, LS2 9JT, UK.
Email: K.E.Zannat@leeds.ac.uk

Funding information
Schlumberger Foundation: Faculty for the Future Fellowship; European Research Council, Grant/Award Number: 101020940-SYNERGY; UK Research and Innovation, Grant/Award Number: MR/T020423/1-NEXUS

Abstract

The potential of passively generated big data sources in transport modelling is well-recognised. However, assessing their accuracy and suitability for policymaking remains challenging due to the lack of ground-truth (GT) data for validation. This study evaluates the accuracy of inferring human mobility patterns from global positioning system (GPS), call detail records (CDR), and global system for mobile communication (GSM) data. Using outputs from an agent-based simulation platform (MATSim) as 'synthetic GT' (SGT), synthetic GPS, CDR, and GSM data were generated, considering their positional disturbances and conventional spatiotemporal resolutions. Mobility information, including activity location, departure time, and trajectory distance, derived from the synthetic data, was compared with SGT to evaluate the accuracy of passive trajectory data at both disaggregate and aggregate levels. The results indicated a higher accuracy of GPS data in identifying stay locations at high resolution. But, GSM data at a lower resolution effectively accounted for over 80% of the variability in stay locations. Comparisons of departure time distribution and travel distance revealed higher measurement errors in GSM and CDR data than in GPS data. The proposed simulation-based accuracy assessment framework will aid transport planners select the most suitable data for specific analyses and understand the potential margin of error involved.

1 | INTRODUCTION

Over the past decade, passively generated spatiotemporal data have emerged as popular sources for extracting mobility information, such as activity location, departure time, and mode of transportation. Among various types of passive data, the most commonly used data for understanding travel behaviour and travel demand modelling include anonymous global positioning system (GPS), call detail records (CDR), global system for mobile communication (GSM) data, social media data, and public transport smart card records [1–4]. These datasets are typically characterised by their large size and the provision of updated, near-real-time spatial and temporal information from a substantial sample size over an extended observation period [5, 6]. However, utilising such data for transport planning poses challenges primarily due to their varying degree of spatiotemporal density, restrictions imposed by the General Data Protection Regulation (GDPR)[1], and the presence of various types of noise that impact the accuracy and precision of extracted mobility information [9].

Errors in data can arise from different sources, including the devices and technology used for data collection, data processing skills, software, and algorithms employed [10]. For example, multipath interference, sampling bias and satellite errors are sources of error in GPS data or influence of tower location density and cell size effect in GSM and CDR data [11–13]. Additionally, noise may appear due to aspects related to users,

[1] The implementation of GDPR varies significantly across different regions globally, and its presence can impact data quality and availability [7, 8]. For example, it might result in the exclusion of brief and infrequent travel patterns, which could potentially expose personally identifiable information. In this particular context, the GDPR did not affect the results as it is not in place in Dhaka (and in many other parts of the developing world).

such as randomness[2] in users' mobility behaviour, access to the service (which could be influenced by network coverage and socio-demographic characteristics), communication, and technology usage patterns [14–16]. For example, the frequency of phone calls and the number of users using a single device influence the quality of CDR data [17, 18]. Other contextual factors such as land use, topography, vegetation, and urban density can introduce additional inaccuracies in the data [19–24]. Therefore, it is essential to assess the reliability and precision of trajectory information extracted from various passive sources before employing mobility information for transport planning and modelling. Such an accuracy assessment is necessary to validate the passive trajectory data as an alternative to conventional survey data and is particularly important in the context of the Global South where the collection of survey data is often labour and resource intensive.

Many studies have focused on the accuracy and precision of passive trajectory data due to different types of noise and its determinants, defining three main perspectives: data source (technology), spatial context, and user attributes. Evaluating accuracy from the standpoint of data source and spatial context underscores the crucial significance of temporal and spatial precision within trajectory data, as these factors directly impact the validity of research on human mobility [25, 26]. For instance, when utilising passive data for trip-based demand modelling as the starting point for generating origin-destination (O-D) matrices, precise identification of the geo-referenced location and trip timing is a prerequisite [27]. Similarly, in order to use passive data in activity- and agent-based demand modelling, precise information regarding activity locations is necessary to infer trip purposes and accurately reflect corresponding travel patterns [28]. While methodologies for assessing the accuracy of trajectory data collected by surveys (e.g. manual survey and smartphone app-based survey) and validating them against ground truth (GT) are well-established [29–35], there remains a significant research gap in addressing the specific challenges of accurately assessing and validating mobility information derived from anonymous passive data sources [32]. One reason for the scarcity of accuracy assessment (both positional and temporal) studies on passive trajectory data is the limited availability of a suitable GT that is compatible with different passive data types and their corresponding spatial and temporal resolutions.

In the contemporary literature, GT data used to validate passive trajectory data can be broadly categorised into two classes: (1) externally collected data, including census data, travel surveys and traffic counts, and (2) internally collected data, which includes information concurrently recorded with passive data through parallel surveys involving subsets of individuals using GPS or app-based location updates [36–40].

Different studies have attempted to validate passive trajectory data against externally collected GT data. For instance, Vanhoof, Lee [41] and Vanhoof, Reis [42] compared home location information extracted from mobile phone data with the cell tower level aggregated population counts. This aggregated approach avoids the translation error that may emerge from converting census grid data to cell tower network grid data. However, the CDR data and geolocated home information were not from the same time period. While using census data along with passive trajectory data, such a time lag can also be found in other similar studies [43, 44]. In all cases, an ad-hoc assumption was made that the population distribution or growth does not change drastically throughout the course of the time gap. This assumption may hold true for certain developed countries where transition and growth are not rapid; however, it may not be appropriate for rapidly growing developing countries [45]. Gordon, Koutsopoulos [46] manually collected trace counts to compare boarding and alighting times, locations, and interchanges inferred from automatic fare collection (AFC) and automatic vehicle location (AVL) data. Similarly, manual travel surveys, conducted independently from the individuals represented in passive data, are often used to validate the trajectory information derived from other passive trajectory data sources such as GPS data loggers [40]. Such validation often requires spatial and temporal adjustments to make meaningful comparisons between travel surveys and passive data. For instance, conventional travel surveys often provide location and time information at the traffic analysis zone level (TAZ), while passive trajectory data (e.g., mobile phone data, smart card data, etc.) offer individual-level location updates throughout a journey [47]. Moreover, validation using manual traffic counts or travel surveys poses a risk of overestimating or underestimating predicted travel demand. Short trips are frequently omitted from travel surveys, hence, models built using GPS data may be labelled as overestimating travel demand because GPS data can capture information about short stays and visits [48].

To overcome the issues associated with validation against externally collected data, a few studies have validated passive trajectory data using internally collected data, where the actual location (e.g. home or activity location) of a subset of individuals was known in advance due to their voluntary participation [26, 49, 50]. However, more often than not parallel surveys involving GPS or apps for individuals can be more complex, costly (particularly for developing countries), and less feasible for extensive passive datasets like GSM and CDR data [39]. This is particularly because passive data acquired through third-party entities, such as mobile network operators or service providers, are subject to privacy concerns and data protection regulations. Consequently, access to passive data is typically granted to transport modellers in an aggregated and anonymised format. This anonymisation process complicates the attribution of devices or services to specific individuals, making it challenging to collect true travel trajectories associated with a given device or service.

Some studies have also used both externally and internally collected GT data simultaneously to validate passive trajectory data. For instance, Toole, Colak [51] leveraged mobile phone call data records to extract both the origin-destination (O-D) matrix and routes, incorporating census data and travel diary survey information. Other studies comparatively analysed different

[2] The randomness in users' mobility is the intrinsic unpredictability and variability in each person's movement pattern. This can be caused by a variety of factors, including impulsive choices, shifting environmental circumstances, or personal preferences. Even with attempts to create mobility patterns using transition probabilities, time periods, and association rules, these models are frequently unable to adequately represent the unpredictable and dynamic character of user movements [14].

data sources to assess the relative accuracy of different passive data sources. Bwambale, Choudhury [18] compared travel time sensitivity, schedule delay, and stoppage number extracted using two different data types with different temporal resolutions (GPS and GSM) in the context of departure time choice. In an experimental study, Forghani, Karimipour [12] compared trajectories generated from CDR data with a GPS logger. This approach combines and comparatively assesses the strengths of different data sources to enhance validation accuracy. Nevertheless, integrating diverse data sets can be challenging due to differences in data resolution, collection times, and the potential for inconsistencies.

Evaluating the precision of passive trajectory data against real-world ground truth data, which accurately reflects the true trajectories of respondents, is a formidable challenge and a timely topic that requires detail investigation. Many challenges related to externally and internally collected GT data such as temporal and spatial mismatches, high costs, limited sample sizes, and privacy concerns could be addressed by generating synthetic ground truth (SGT) data compatible with different passive data types and their corresponding spatial and temporal resolutions. SGT data offers distinct advantages by allowing the formation of a comprehensive range of potential scenarios, including different activity patterns, mode use, and activity locations, spanning a range of spatial and temporal resolutions depending on the type of passive data under consideration. This capability is lacking in both externally and internally collected GT data. Synthetic data is often used for accuracy assessment when it is difficult to access real-world data or when true data is unavailable [52, 53]. Nevertheless, a similar examination of the relative precision of spatiotemporal data extracted from passive data sources with appropriate (S)GT is still in its early stages. Further investigation of positional and temporal accuracy assessment with reliable GT is needed to make informed decisions based on passive data [54].

In this study, we proposed a framework for assessing the accuracy of passive trajectory data with appropriate GT that matches the resolution of passive data. We considered three mainstream passive data types—anonymous data from GPS[3] devices, GSM[4], and CDR[5]—for accuracy assessment. In the absence of real-world disaggregate ground truth data for benchmarking, we treated outputs generated from an agent-based simulation platform (MATSim simulation) as the 'SGT' for each traveller. Synthetic GPS, GSM and CDR data corresponding to this SGT were generated based on the standard spatial and temporal resolutions typical for each type of data. We compared the statistical and spatial characteristics of individual mobility information extracted from the GPS, GSM, and CDR data with the SGT data to assess the accuracy of passive trajectory data at both the disaggregate and aggregate levels. Also, this work generated synthetic datasets, allowing us to quantify the magnitude of errors (due to the added noise) in different travel information extracted from passive data compared to GT. Therefore, our proposed framework can be employed to introduce any inherent noise (caused by sensor technology, sampling bias, data loss, or environmental conditions [9]) associated with the data under consideration, and assess the error level in the travel attributes compared to the generated GT. It is important to note that the current study is limited only to examining a straightforward type of noise that is, spatial-temporal disturbance that affects the positional and temporal accuracy of trajectory data. However, our proposed framework can be employed recursively to incorporate any inherent noise associated with the data under consideration or arising from different geographical landscapes which will enable the evaluation of positional and temporal accuracy and reliability in comparison to the ground truth generated.

The remainder of the paper is organised as follows: The following section discusses the methodology employed, as well as the data sources used in this study. Subsequent sections present the analysis results, followed by a discussion of the findings. The paper concludes with a summary and directions for future research.

2 | METHODOLOGY

The methodology employed in this article can be categorised into four main steps: (1) generating synthetic ground truth (SGT) data; (2) generating corresponding passive mobility data; (3) extracting mobility attributes from the synthetic trajectory data and (4) accuracy assessment. The overall methodology of this study is illustrated in Figure 1.

2.1 | Generating synthetic ground truth (SGT) data

The agent-based simulation tool MATSim (Multi-agent Transport Simulation) was used to generate the SGT data. At its core, MATSim operates by allowing a group of agents to interact within a virtual environment. The inputs required for MATSim simulation include activity plans, transport networks, and configuration files. An activity plan serves as a sequence of actions that agents are required to perform within this artificial setting. Typically, such plans are generated using microdata from a representative sample or synthetic population. In this study, we employed household-level travel diary data derived from the subway study by TYPSA (https://www.typsa.com/en/) in Dhaka. Detailed demographic information about the sample can be found in supplementary Table S1. We selected this data from 2019 for two primary reasons: (1) the 2019 dataset was the most recent available that included trip information across a wide range of transportation modes, such as ride-hailing

[3] GPS data is timestamp location data generated by global positioning system satellites. These satellites emit signals that are received by GPS receivers in devices like smartphones, GPS trackers, or navigation devices.

[4] GSM is a type of mobile phone data that provides triangulated location information when the mobile phone with a valid sim card is turned on. It offers higher temporal and spatial resolution than CDR data (another type of mobile phone data) as one single call will generate multiple sightings in GSM data.

[5] CDR data is a type of mobile phone data that records locations at the tower level when the mobile phone is in use (such as during calling and texting). In CDR data, each phone call generates a single record, representing the location associated with the tower used during that call.

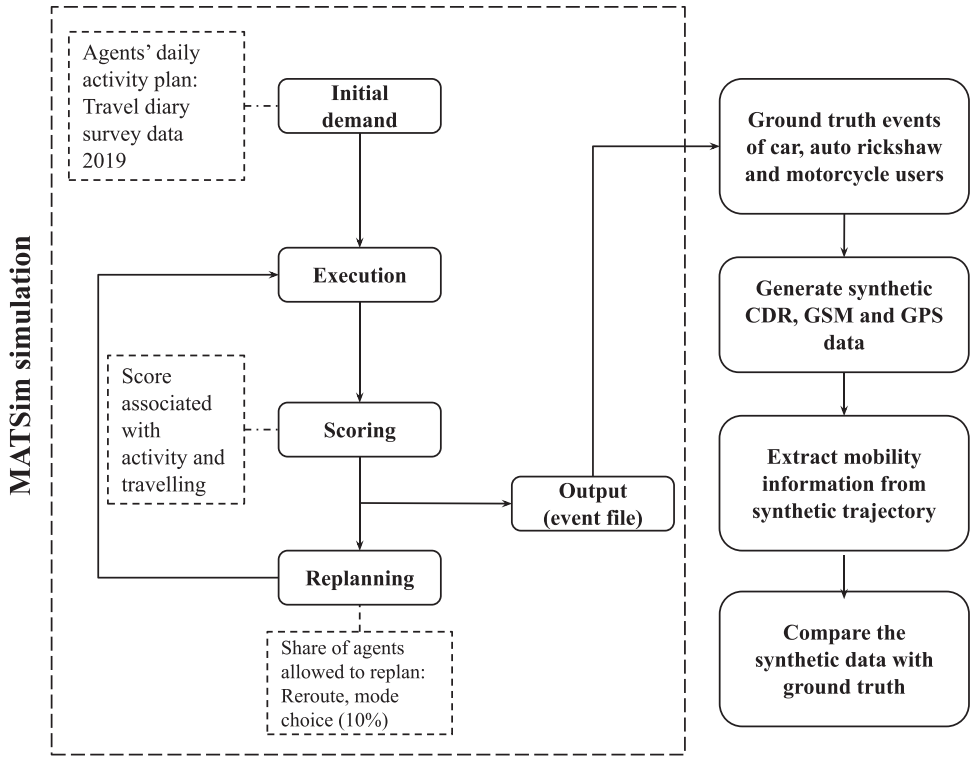


FIGURE 1 Framework to check the accuracy of trajectory data. CDR, call detail records; GPS, global positioning system; GSM, global system for mobile communication.

services and motorbikes, which have been widely used in Dhaka, alongside traditional modes of transport and (2) this data represented a period before the COVID-19 pandemic, ensuring that it was not influenced by the mobility restrictions that affected travel patterns during and after the pandemic. This allowed for an accurate representation of typical working day travel behaviour in Dhaka.

However, the survey data could not be directly employed as ground truth because it lacked specific route information. In contrast, passive data typically contains location records within the trajectory while agents are traversing. Furthermore, the trip diary data only offered activity locations at the TAZ level and provided detailed geographic information solely for each participant's home location. To address this, ArcGIS 10.8 was employed to randomly allocate activity locations within TAZ boundaries, ensuring compliance with the user-stated travel time [55]. Following the assignment of activity locations, an activity profile was generated for each agent based on trip information extracted from the travel diary survey data. Each activity plan included information regarding the activity's location (x–y-coordinate), the end time of the first activity, the chosen 'leg' mode, and the maximum duration allocated for that activity.

To represent a virtual urban transport landscape mirroring Dhaka, this study integrated a road network and available information on existing transport services. The road network for the study area was obtained from the Open Street Map (OSM). In MATSim, the available modes include car, public transport (PT), bike, and walking. While Dhaka also features modes such as auto-rickshaws and motorcycles, these are not predefined options in MATSim. To account for them, they were modelled using special *vehicular specifications* within the existing framework of MATSim. It is noteworthy to mention that the car category encompassed various types of private vehicles, including personal cars, ride-hailing cars, and taxis. Similarly, the motorcycle category included both personal and ride-hailing motorcycles. The default configuration settings of the MATSim simulation were used to generate the SGT data.

In the MATSim simulation, each agent strives to optimise their actions based on a utility function. During the iterative process, a specific proportion of agents are allowed to modify their typical choices in an attempt to identify strategies with higher utility. This iterative process continues until the overall score of the population reaches equilibrium within the simulation. The strategy adopted by each agent at this equilibrium is intended to be a realistic approximation of their actual behaviour.

To generate the SGT data in this study, the marginal utility of travel time and cost derived from a joint RP-SP mode choice model was employed. A study by Zannat, K. E., et al. [56] described in detail the MATSim model of Dhaka scenario. Supplementary Tables S1 and S2 provide further details on the RP-SP data and the mode choice model used in the simulation respectively. A predetermined proportion of agents were allowed to change their trip mode (randomly selecting a leg mode) and route during the iteration process in an effort to find strategies with higher utility. All agents made attempts to adjust their plans to increase their utility by tracking each action chain.

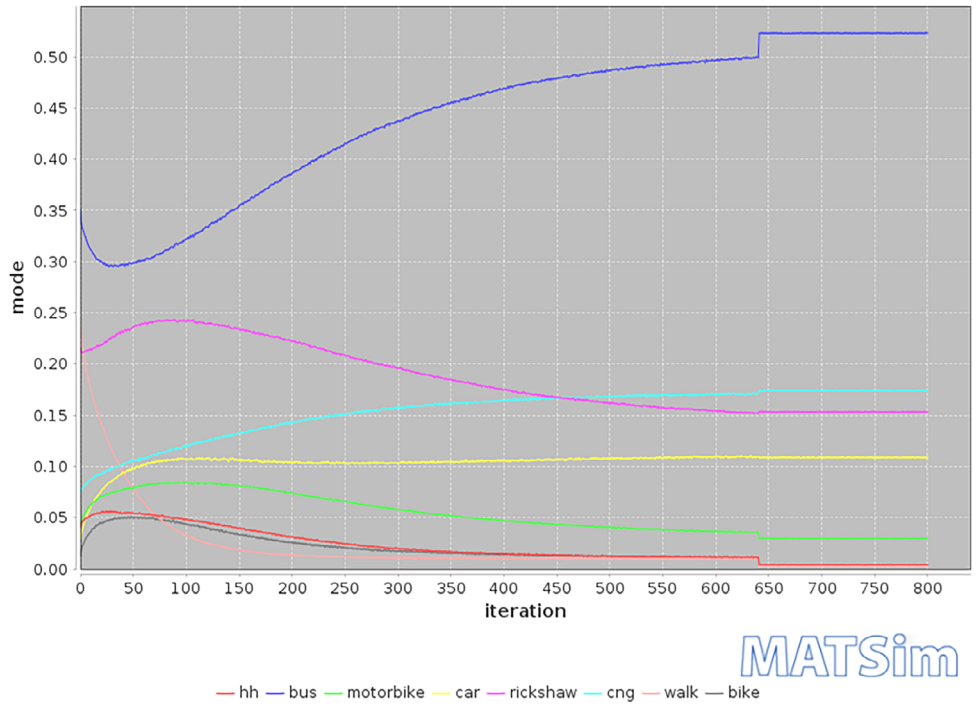


FIGURE 2 Mode statistics from the MATSim simulation.

The simulation was run until the population achieved an equilibrium state, requiring a total of 800 iterations. Figure 2 presents the mode statistics of the simulation model. We compared the predicted modal share and preferred departure time with actual data collected from the 2019 travel diary survey. Additionally, the modal share of hypothetical passenger trips was compared with the modal share of the passenger trips collected by the Japan International Cooperation Agency (JICA) research team using the inner cordon line survey in Dhaka in 2014, details of which can be found in supplementary Table S3.

The traffic simulation component was used to ascertain the "real" travel schedules for cars, motorcycles, and autorickshaws. As for the bus and human hauler routes, they were simulated utilising the "teleportation" feature, but with network-derived journey distances since precise information about these routes and schedules was lacking. Similarly, non-motorised modes were also simulated using the "teleportation" feature. Therefore, the simulation produced event files that comprehensively documented every specific action taken by agents who exclusively used motorcycles, autorickshaws, and cars throughout the simulation period. These actions included activity start times, activity end times, and link-level interactions.

We extracted the complete, unprocessed list of actions carried out by agents who opted for a car, autorickshaw, or motorcycle during the simulation period from the final event file. This extraction detailed the entire trajectory and activities of each agent. In this study, the extracted event file served as the SGT and the baseline for generating synthetic CDR, GSM, and GPS data. Notably, since PT users were simulated using the teleportation feature of MATSim, only the origin and destination, along with the activity start and finish times, were recorded in their event files, as opposed to their full trajectories. As a result, we limited our analyses to private mode users and excluded the PT users from the comparison.

2.2 | Generating detailed trajectory data

Accessing real-world mobility-related passive data can be a challenging endeavour for transport modellers due to a myriad of concerns including privacy issues, the potential for re-identification, legal restrictions, data ownership, and data availability. To address these challenges, significant research efforts have focused on the generation of synthetic data. Such efforts aim to obfuscate or mask real-world location data derived from mobile phones and navigation devices for privacy and security purposes [52, 57]. In this study, we adopted the fundamental principles established in previous research within this domain. These principles underpinned the following assumptions:

- It was assumed that the mobile device remained active throughout the entire journey and was carried by the travellers for the duration of their trips.
- All agents were presumed to choose the shortest route during their travels.
- The sociodemographic characteristics of the chosen agents were consistent.
- The mobile phone tower network strength within the study area was assumed to be uniform, with no differentiation

TABLE 1 Overview of global positioning system (GPS) and mobile phone data.

Data	Noise source	Positional disturbance range (approx. values)	Spatial data density	Temporal data density	Sources
GPS device	Orbital error	<5 cm to 4 m	High	High	See [58–61] for further details of each noise type and algorithm to filter that noise
	Satellite clock error	3 cm to 150 cm			
	Ionospheric error	≈5 m			
	Tropospheric error	1 cm to 2.3 m			
	Multipath error	10 m to 15 m			
	Receiver error	0.8 mm to 15 m			
	Satellite geometry and availability	1 m to 20 m (when pseudo-range error[a] is 1 m)			
	Interference and signal noise	1 m to 40 m			
Mobile phone data	Signal oscillation	Few metres to kilometres	Medium to low	Medium to low	See [20, 62–64] for further details
	Cell tower density				
	Cellular network size and coverage				

[a]Inaccuracies in the measured distance between a GPS satellite and a receiver.

between 2G, 3G, and 4G networks. Additionally, there were no gaps or incompleteness in the data due to topographic features or 'urban tunnelling' effects.

Based on these assumptions, we generated synthetic GPS, GSM and CDR data, taking into account their spatiotemporal resolutions according to the source type. Table 1 provides an overview of typical GPS and mobile phone data, including their temporal and spatial data resolution, as well as the observed positional disturbances due to the existence of different noises. It is noteworthy that all positional disturbances of GPS data mentioned in Table 1 are subject to weather, satellite elevation, time of day, and the surroundings of the receiver.

2.2.1 | Synthetic GPS data

GPS data is generated by smartphones or GPS devices, which may have varying degrees of measurement noise. These data typically consist of anonymous timestamped latitudes and longitudes. For this research, we generated synthetic data from GPS devices. In Dhaka city, where there is no underground or tunnel infrastructure, GPS devices can provide relatively accurate location data when used aboveground. These devices record positions both outdoors and inside various structures such as buildings, buses, elevated trains, bridges, and urban canyons [65]. However, it is important to note that the accuracy of the recorded data may vary. The accuracy of spatiotemporal information from GPS devices is influenced by three primary factors related to the satellite, the GPS receiver, and the surrounding environment (details of errors are shown in Table 1). Aspects related to these factors are associated with different spatiotemporal disturbances in GPS data, which inevitably result in measurement errors, positioning jumps, and irregular sampling biases, all of which can lead to poor-quality measurements. Although filtering techniques such as Kalman filters, carrier smoothing, and moving average filters can mitigate some of these errors, errors in GPS data can persist despite their application. Therefore, to generate synthetic GPS data, we followed a systematic procedure designed to account for these various sources of error and ensure the generation of high-quality data. The MATSim simulation yielded temporal data whenever an agent interacted with a specific event. In contrast, depending on the type of GPS receiver and signal strength in the target region, GPS devices provide more frequent location and time information. Therefore, to produce intermediary points that bridge the gap between the SGT point feature and create more realistic GPS data, we used linear interpolation. GPS points were generated at 10 s intervals. Typically, the positioning error for raw GPS-enabled systems falls within the range of 1–20 m [66–68]. After applying filtering techniques, the amount of residual error can still range from 3 to 5 m [69]. Therefore, to account for the geo-positioning noise, we introduced two types of simulated noise within this range as horizontal positioning disturbance—random shift and random drift, following the method proposed by Bösche and Sellam [70].

First, we applied Gaussian noise to the spatial data with a mean of 0 and a standard deviation of 3 m. A well-designed GPS receiver generally offers horizontal accuracy of at least 3 m [71]. However, this could lead to abrupt changes in direction between successive GPS points. To address this issue, we applied shifts with a probability of 0.05 per second and otherwise, added the deviation from the previous point to the current point. Second, we simulated random drift with a probability of 0.03 per second, representing shifts perpendicular to the current driving direction. The maximum drift distance was determined from a Gaussian distribution with a mean of 0 and a standard deviation of 10 m, as GPS in moving vehicles can achieve dynamic accuracy of up to 10 m in urban areas [72]. It took precisely 30 s to reach the maximum drift distance, after which the drift

TABLE 2 Summary of parameters used for adding noise while generating synthetic global positioning system (GPS) and global system for mobile communication (GSM) data.

	Parameter	Value
GPS	Shift parameter (mean)	0
	Shift parameter (standard deviation)	3 m
	Shift probability	0.05
	Drift parameter (mean)	0
	Drift parameter (standard deviation)	10 m
	Drift probability	0.03
GSM	Shift parameter (mean)	0
	Shift parameter (standard deviation)	5 m
	Drift parameter (mean)	0
	Drift parameter (standard deviation)	15 m
	Drift probability	0.03

distance gradually decreased to zero over another 30 s. Table 2 summarises all the parameters selected for adding noise to the GPS data.

2.2.2 | Synthetic GSM data

GSM data comprises the identifiers of all GSM cells that a mobile phone passes through at regular intervals during its use [73]. Notably, GSM data offers finer spatial and temporal resolution compared to CDR data because it becomes accessible as soon as the phone, equipped with a valid SIM card, is activated. In our study, we leveraged the location data from the SGT to generate synthetic GSM data. The temporal resolution of the GSM data in our study was set at 60 s, aligning with the temporal resolution of real GSM data [18]. For each sighting time, we selected locations from the respondent's SGT data while preserving the timestamped location sequence. If the time interval between two successive SGT points exceeded 60 s, we applied the linear interpolation method to generate intermediary cell points. However, it is crucial to acknowledge that the positioning accuracy of GSM data can widely vary, ranging from 1 to 600 m, depending on factors such as the location type (indoor/outdoor), cell size and data collection techniques (triangulation, radio camera, signalling messages, and GPS, etc.) [73–76]. Similar to GPS data, various filtering approaches such as recursive naïve filters, look-ahead filters, and Kalman filters have been used to remove noise from mobile phone data. Empirical investigations have shown that while filtered GSM data contain significantly less noise than raw mobile phone data, the noise levels remain higher compared to filtered GPS data. [77]. Therefore, to generate synthetic GSM data for the densely urbanised and heavily populated area of Dhaka city, we added noise to the location points generated from SGT, which was relatively larger than the noise added with GPS data (GSM data tends to be noisier compared to GPS data, as per Bwambale, Choudhury [18]). The summary of parameters used for generating synthetic GSM is outlined in Table 2. Gaussian noise (with a mean of 0 and stand deviation of 5 m) was included as positioning disturbance. Given that successive GSM points can exhibit abrupt jumps from one side of the road to the other, we did not apply corrections for directional deviations in GSM points. However, similar to GPS data, we introduced random drift for moving vehicles at a probability rate of 0.03 per second in the case of GSM data.

2.2.3 | Synthetic CDR data

Real-world CDR data includes time-stamped tower locations whenever a user initiates a phone call, sends a text message, or accesses mobile Wi-Fi. Table 3 provides an example of CDR data from various hypothetical users. In Dhaka, the most recent available CDR data was collected between 19 June, 2012, and 18 July, 2012. For detailed statistics regarding the available CDR data, refer to [27, 44].

To construct the CDR trajectory for a group of agents, we leveraged the set of SGT location and time data generated via the MATSim simulation. Using the distribution of actual call rate (the average number of calls per hour) observed in real CDR data, we produced one-day trajectory data, with the hourly call rate during a typical working day depicted in Figure 3. For each 10-s interval within this timeframe, we randomly generated call rates from a normal distribution centred around the population's median call rate (0.053 per hour). By using a Poisson distribution[6], we determined the number of calls for each 10-s interval based on the call rate. Subsequently, we selected a location for each call from the agent's SGT data and generated a call duration ranging from 1 to 60 min, following a uniform distribution. Call start and end times were then generated based on these durations. This procedure was repeated for all 10-s intervals within the specified time period. It is noteworthy that to generate the synthetic CDR data, SGT locations were updated to correspond to the nearest mobile phone tower location (tower locations were extracted from the real CDR data from 2012).

2.3 | Extracting mobility information

In order to assess the relative accuracy of CDR, GPS and GSM mobility data, we extracted mobility information from the synthetic trajectory. The following paragraphs summarise extracted mobility information and the methods used.

2.3.1 | Stay location

To extract stay locations or points of interest (POIs) such as congestion stay points and potential activity locations from the

[6] Poisson distribution is a discrete probability distribution that can be used to simulate the number of calls when it is known how many calls are made on average per hour during that time [78].

TABLE 3 An excerpt from call detail records data in a typical working day.

Unique caller ID	Date	Time	Call duration	Latitude	Longitude
AAH03JABkAAHvEkAQE	20120622	13:32:38	530	23.7186	90.4494
AAH03JACKAAAgfBALW	20120622	13:41:25	15	23.9139	90.2931
AAH03JAC8AAAbZfAHB	20120622	13:41:49	73	23.7911	90.2603
AAH03JAC5AAAdAkAJZ	20120622	13:45:40	16	23.7172	90.3556
AAH03JAC3AAAdDZAEe	20120622	13:46:22	17	23.1581	90.4119

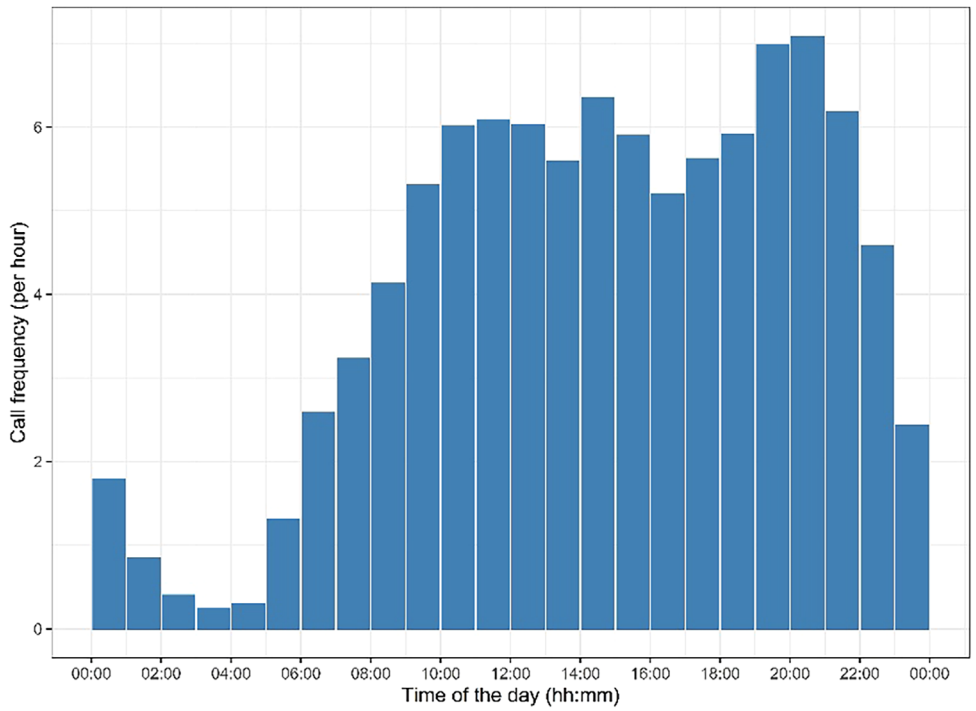


FIGURE 3 Hourly call rate on a typical working day.

trajectory data, different clustering algorithms have been used in the literature. These algorithms can be broadly classified into five groups of methods — partitioning based, hierarchical based, density based, grid based, and model based [79]. A summary of these methods is outlined in Table 4. Among these methods, partition-based algorithms (e.g., K-mean, FCM) are very straightforward to implement, however, they require the number of clusters as an input which is difficult to fix beforehand. As the cluster number affects the granularity of cluster analysis, arbitrarily fixing this input could affect the clustering result. Hierarchical clustering builds a tree (dendrogram) of clusters. It does not require the number of clusters to be specified in advance. Model-based approaches optimise the fit between the data and pre-defined mathematical models. MCLUST or EM are widely used model-based algorithms. These algorithms assume that a collection of observed objects consists of instances drawn from several probabilistic clusters. Therefore, this method has the potential to evaluate the likelihood that a given observation belongs to any of the existing clusters. However, it also necessitates specifying a maximum number of clusters as input. On the other hand, density- or grid-based methods do not require the number of clusters to be known in advance. Nevertheless, grid-based techniques frequently suffer from the "sharp-edge" problem, in which two closely comparable places could be split into two distinct zones. Hence, we utilised a geographic clustering method based on density for GPS and GSM data. Density-Based Spatial Clustering of Applications with Noise (DBSCAN), a classic density-based algorithm, was selected due to its ability to identify clusters of various shapes without the need to specify the number of clusters in advance [80]. Figure 4 shows the conversion process of point density into stop points. The effectiveness of DBSCAN heavily relies on the appropriate selection of two key parameters: eps (epsilon), which defines the maximum distance between two points to be considered as neighbours, and the minPts (minimum points), the minimum number of points required to form a stay location cluster. For the GPS dataset, we selected an eps value of 50 m and the minPts value was set to 5. This choice reflected the precision

TABLE 4 Summary of clustering algorithms.

Clustering method	Overview of the method	Algorithms	Strength	Weakness	Literature on stay locations
Partition based	Divide a dataset into a discrete, non-overlapping cluster subset (of spherical shape), with each data point belonging to a single cluster	K-means K-modes K-medoids PAM CLARA CLARANS FCM	- Efficient in clustering small- to medium-size data sets based on distance division - Easy to interpret the results	- Require prior knowledge on number of clusters - Less flexible for complex datasets with varying cluster shapes and densities	[81, 82]
Hierarchical based	Data organisation follows a hierarchical approach based on the medium of proximity	BIRCH CURE ROCK ECHIDNA	- The number of clusters is not required a priori - Effective to reveal the hierarchical structure of the data	- Less flexible approach - The assumption of nested and hierarchical cluster may not be appropriate for all data distribution	[83, 84]
Density based	Data objects are divided according to their boundary, connectivity, and density	DBSCAN OPTICS DBCLASD DENCLUE	- The number of clusters is not required a priori - Effective at identifying clusters of arbitrary shapes and handling noise	- Points on the boundary between clusters can be difficult to classify accurately - Computationally intensive	[85, 86]
Grid based	Divide the data into finite number of cells that form a grid to separate the dense grid area from the less dense ones	Wave-Cluster STING CLIQUE OptiGrid	- Determine clusters of arbitrary shapes - Can handle different types of datasets and remove noise elements	- Difficult to apply the method in dimensional space size - Results are sensitive to the choice of the grid size	[87]
Model based	Using either a statistical or neural network approach, this method automatically determines the number of clusters while taking into account noise	EM COBWEB CLASSIT SOMs	- Have the flexibility to model clusters of various shapes and sizes - Provides a probabilistic membership for each data point to each cluster	- Results are sensitivity to parameter initialisation - Model selection criteria may not provide a clear indication of the optimal clumber of clusters	[88, 89]

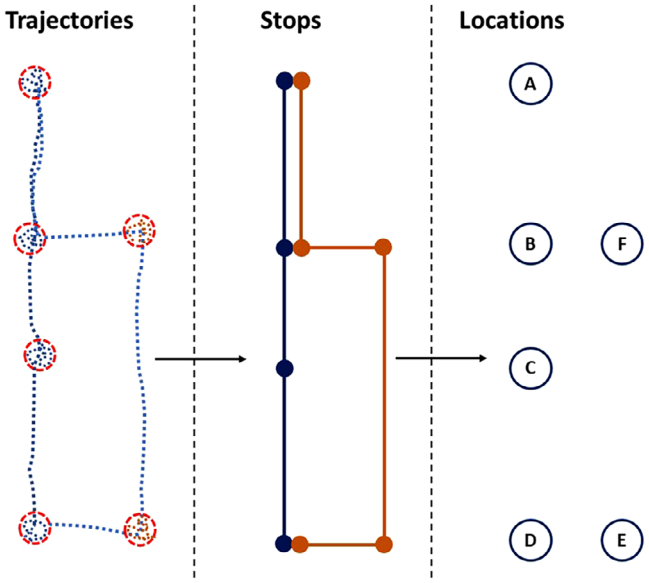


FIGURE 4 Location extraction by density clustering.

and typically close spacing of GPS data points. For the GSM dataset, we chose a larger eps value of 100 m and the minPts value was also set to 5. This decision considered the less precise and more spread-out nature of GSM data.

Then, we applied a temporal rule (>10 min) within each cluster to remove potential congestion stays before extracting potential activity locations from the cluster data. For the synthetic CDR trajectory generated for a single day, call locations were assumed as potential stay locations. However, to distinguish these potential stay locations from those recorded during travel, we examined the driving distance between the two locations and the time lag between call times. If the time lag between call times exceeded the time required to travel between the two locations by car, the locations were considered as stay locations.

2.3.2 | Home and activity location identification

With GPS and GSM data, which offer records with a high level of temporal precision, we were able to deduce likely home locations and other activity places from their trajectories. Home locations were identified as places where agents revisited multiple times, with sightings predominantly in the early morning (before 6:00) and late evening (after 20:00). In contrast, CDR data recorded very few locations, primarily while calls or texts were made. Distinguishing home and other activity location information from CDR data, especially for agents with a single day of call data following the median call rate distribution, posed challenges. In the absence of location data, we made the

assumption that each stay location in the CDR data represented a potential activity location, aligning with the rule proposed by Zilske and Nagel [52] for generating synthetic CDR data.

2.3.3 | Trip attributes

Following the identification of the stay locations, we extracted various trip-related attributes—departure time and trajectory distance, for each trajectory. To extract departure time from each stay location cluster in GPS and GSM trajectories, we sequenced the clustered points based on their observed times. We then examined the first observation and duration within each stay location cluster point, sequentially capturing different start times for repeated activities and their corresponding stay durations. The first observation and corresponding duration in each cluster were used to find the departure time for each stay location to travel for the succeeding stay location cluster. CDR data, however, did not allow for the capture of departure times, as the call and text times recorded in this data reflected the sighting times at activity locations. To calculate the trajectory distance from GPS and GSM data, we sequenced the stay location clusters and calculated the Euclidean distance of trip segments, considering the centroid of each stay location cluster points. The total trajectory distance for each agent was the sum of these individual trip segment distances. For CDR data, we sequenced activity locations based on their sighting times and calculated the total trajectory distance by summing the individual Euclidean trip segment lengths.

2.4 | Accuracy assessment

To assess the accuracy of passive mobility data, we compared the statistical and spatial characteristics of individual mobility information extracted from GPS, GSM, and CDR data with the SGT data. We assessed the accuracy of the stay locations (potential activity locations). To achieve this, we created grid cells of varying resolutions (e.g. 50 m × 50 m, 100 m × 100 m, 200 m × 200 m, 500 m × 500 m) within the study region. For each synthetic trajectory, we counted the number of stay locations found within each grid cell. The precision of stay location estimation was evaluated using bivariate analysis. If x_k is the count of stay location from passive data for grid cell k and y_k is the count of stay locations from SGT for grid cell k, the fitted linear regression model can be expressed as:

$$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k \tag{1}$$

The predicted $\hat{y}_k$ and mean counts from passive data and SGT were used to calculate R^2 value, which indicates how well the counts of stay locations from passive data explain the variation in the counts from SGT data. Here, the total number of cell ($\sum_{k=1}^{K} k$) is dependent on the resolution of grid under consideration.

Furthermore, to understand the spatial distribution pattern of stay location, we calculated Global Moran's I index [90], a measure of spatial autocorrelation, to assess the degree of clustering or dispersion of stay locations. Using the following Equation (2), an index value bounded by −1 and 1 for each dataset was calculated. The positive index value indicates a tendency toward clustering, zero suggests a random distribution with no autocorrelation, and the negative value indicates dispersion.

$$I = \frac{n}{S_o} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} \tag{2}$$

where z_i is the deviation of an attribute for feature i from its mean $(x_i - \bar{X})$, w_{ij} is the spatial weight between feature i and j, n is equal to the total number of features, and S_o is the aggregate of all the spatial weights:

$$S_o = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \tag{3}$$

Furthermore, local Moran's I index [91] values were calculated to identify spatial clusters (positive significant values indicate cluster and negative values indicate outlier) using the following equation:

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1, j \neq i}^{n} w_{i,j} \left(x_j - \bar{X}\right) \tag{4}$$

where x_i is an attribute for feature i, $\bar{X}$ is the mean of the corresponding attribute, $w_{i,j}$ is the spatial weight between feature i and j, n is the total number of features, and:

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^{n} \left(x_j - \bar{X}\right)^2}{n - 1} \tag{5}$$

Moreover, the cumulative distribution of local Moran's I index of identified clusters were further compared using Kolmogorov–Smirnov (K-S) test. This non-parametric test was employed to assess the differences between each passive data source and the ground truth. Unlike other tests, the K-S test does not assume any specific distribution for the data, making it widely applicable [92]. The test statistic D can be derived by:

$$D_{n,m} = \max \left| F_1(x) - F_2(x) \right| \tag{6}$$

Here, n is the number of observations on passive data and m is the number of observations in the SGT. $F_1(x)$ is the empirical cumulative distribution function of the passive data and $F_2(x)$ is the empirical cumulative distribution function of the SGT. Like the cluster of stay location, departure time distribution from GPS, GSM, and sightings from CDR data were compared with the cumulative distribution of SGT using the K-S test value.

We employed Pearson's correlation (r) statistic to determine the relationship between the synthetic trajectory's distance and the SGT data. The correlation coefficient was derived by the following equation:

$$r = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum \left(X_i - \bar{X}\right)^2 \sum \left(Y_i - \bar{Y}\right)^2}} \quad (7)$$

X_i are the individual trajectory distances derived from the passive data and Y_i are the individual trajectory distances derived from the SGT. $\bar{X}$ and $\bar{Y}$ are the means of the X and Y datasets, respectively. All statistical analysis was conducted using R (version 4.2.3) programming language. Moreover, for spatial analysis, ArcGIS 10.8 was used in this study.

3 | RESULTS

3.1 | Synthetic ground truth (SGT)

The MATSim simulation output mimics real-world agents' activity, departure times, routes, and mode choices. The output is represented in an 'events' file, which comprehensively documents the movements and activities of each agent throughout the simulated day. The various event types encompassed in the MATSim output include *"Activity End Event"*, *"Person Departure Event"*, *"Person Enters Vehicle Event"*, *"Vehicle Enters Traffic Events"*, *"Link Leaves Event"*, *"Link Enters Event"*, *"Vehicle Leaves Traffic Event"*, *"Person Leaves Vehicle Event"*,*"Person Arrival Event"*, and *"Activity Start Event"*. A schematic diagram illustrating the full range of events stored in the MATSim simulation output is provided in Figure 5a. MATSim output offers comprehensive situational information about the agents' actions. For this study, we extracted activity start and end times, and time-stamped link IDs, for 9704 agents. An example of the extracted event information is demonstrated in Figure 5b. It should be noted that precise activity location information (exact latitude and longitude) was not saved in the event file; instead, it was derived from network file using the time-stamped link IDs corresponding to each event. The coordinates of activity start and end points provided insight into the potential home and activity locations of agents, while link coordinates depicted locations during the trips. In total, 20661 unique locations were identified as potential activity locations, while 269429 locations were identified as en-route point locations, collectively constituting the SGT. Figure 6a illustrates SGT data generated from the MATSim simulation for a single agent.

3.2 | Synthetic trajectory data

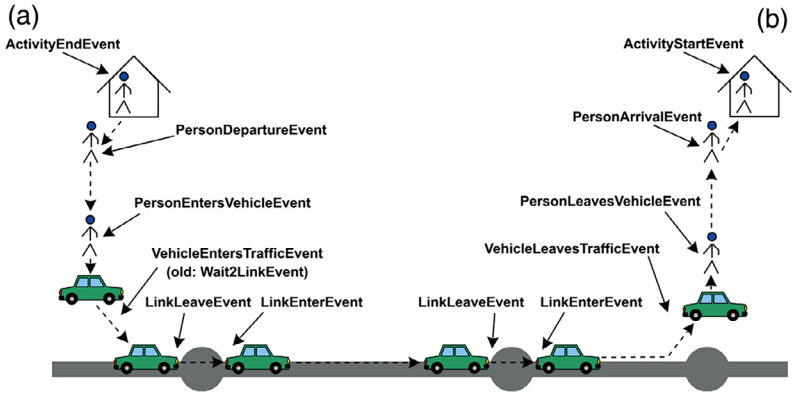
Three distinct sets of synthetic trajectory data were generated, each incorporating different levels of noise (detail in Section 2.2). For 9704 agents, a total of 19333 CDR trajectory points, 464253 GSM points, and 30728597 GPS traces were generated. These trajectory data sets exhibited varying levels of spatial and temporal granularity for an equivalent number of agents. Synthetic GPS, GSM, and CDR data for a representative agent are shown in Figure 6b–d. Figure 6b exhibits the accuracy and precision of GPS data, providing precise location information along the entire route of the trip (assuming the device and location apps were active during the journey). This accuracy was achieved because the synthetic GPS data in this study did not account for urban canyons and topographic effects (Dhaka city's topography is generally flat with fewer concentrations of skyscrapers). Therefore, the temporal and spatial resolution of GPS data from GPS-enabled devices was solely influenced by the GPS receiver's type in the vehicle and the satellite network's availability. Figure 6b,d demonstrates that GPS and GSM data offered reasonably accurate information at coarser spatial resolutions, aligning with the SGT. These datasets not only illustrated the travel route but also depicted stopover locations or congested points through point density. Therefore, low-precision GPS data could provide trip-related location updates similar to GSM data. Conversely, CDR data only exclusively presented records at the tower level based on call times (Figure 6c), resulting in coarser temporal and spatial precision compared to GPS and GSM data types.

3.3 | Analysing the accuracy of different trajectory data

In order to evaluate how well GPS, GSM, and CDR could be useful in extracting mobility information, we compared the information extracted from synthetic data with the MATSim simulation-generated SGT data. This comparison encompassed statistical properties of trip-related information and the spatial distribution of location information between synthetic trajectory data and SGT data.

3.3.1 | Stay location accuracy

To evaluate the accuracy of the spatial distribution of stay locations (includes both home and other activity locations), we compared the stay locations extracted from SGT data and synthetic passive trajectory data. Figure 7 illustrates the distribution of stay locations (at a 500 m resolution) obtained from the three types of passive data in the Dhaka City Corporation (DCC) region. It is visually evident that the distribution of GPS and GSM data was closely aligned with the distribution of observed activity/stay locations within 500 m × 500 m grid cells. In contrast, the location data from the CDR dataset conformed to the SGT in central Dhaka, where mobile phone towers were densely concentrated, but exhibited notable discrepancies in other areas, such as the outskirts (e.g. the eastern fringe region). Such discrepancies were more evident when mapping the stay location distributions in the RAJUK area due to variations in tower location density between the DCC area and the surrounding regions. To measure these differences through quantitative analysis, the coefficient of determination from the bivariate analysis of SGT and the stay locations from synthetic passive



Event type	Time	Person ID	Link ID
actend	23404	134	5108733800005f
left link	23406	134	5108733800005f
left link	23444	134	5108733800005r
left link	23505	134	2965586160006f
left link	23518	134	2638027510020f
left link	23536	134	266421830010f
left link	23555	134	5066951920002f
left link	23570	134	353238520029f
left link	23572	134	353238520030f
left link	23576	134	353238520032f

FIGURE 5 (a) MATSim events by Axhausen, Horni [93]. (b) Extracted synthetic ground truth information from the event file.

TABLE 5 Summary of bivariate statistical analysis of stay location capturing accuracy between synthetic ground truth and synthetic trajectory.

Data type	Bivariate component	Cell size resolution			
		50 m	100 m	200 m	500 m
CDR	R^2	0.0007	0.0071	0.0583	0.305
GSM	R^2	0.2653	0.478	0.6393	**0.8101**
GPS	R^2	**0.5502**	**0.6476**	**0.7671**	**0.8979**

Abbreviations: CDR, call detail records; GPS, global positioning system; GSM, global system for mobile communication.

TABLE 6 Global Moran's *I* value for different dataset.

	Global Moran's *I*	Remarks	*p*-value
SGT	0.596018	+ values indicate a tendency toward clustering	1% likelihood that this clustered pattern could be the result of random chance
CDR	0.370308		
GSM	0.560002		
GPS	0.617444		

Abbreviations: CDR, call detail records; GPS, global positioning system; GSM, global system for mobile communication.

data was compared to assess stay location accuracy, particularly in Dhaka city's RAJUK area.

The results of this analysis at different spatial granularities (i.e. 50, 100, 200, and 500 m) are summarised in Table 5. At the finest spatial resolution (50 m × 50 m grid), only 0.07% of the variance in the stay locations from the SGT could be explained by the stay locations extracted from the one-day CDR data. In comparison, GPS data demonstrated the highest explanatory power for stay location distribution at this finer resolution. Indeed, at the finest spatial resolution (50 m × 50 m grid), the explanatory power of GPS data for stay location distribution was twice than that of GSM data (R^2 for SGT vs GPS was 0.55, whereas R^2 for SGT vs GSM was 0.27). The explicability of GSM stay locations noticeably improved at 100 m resolution, roughly doubling from the 50 m resolution (R^2 for SGT vs GSM = 0.478). As the grid size increased, GPS and GSM showed comparable explicability in stay position. However, even at a 500 m resolution, CDR data could only explain around 30% of the variation in SGT stay locations, significantly less than the over 80% achieved by GPS and GSM data.

After finding 500 m resolution as a better cell size to capture activity location from GPS, GSM and CDR data, we investigated their spatial distribution patterns at this scale. The estimated Global Moran's *I* index values are presented in Table 6. For both SGT and passive data, the estimated values suggested a positive spatial autocorrelation. In another term, the activity locations in all datasets tend to be spatially clustered rather than randomly

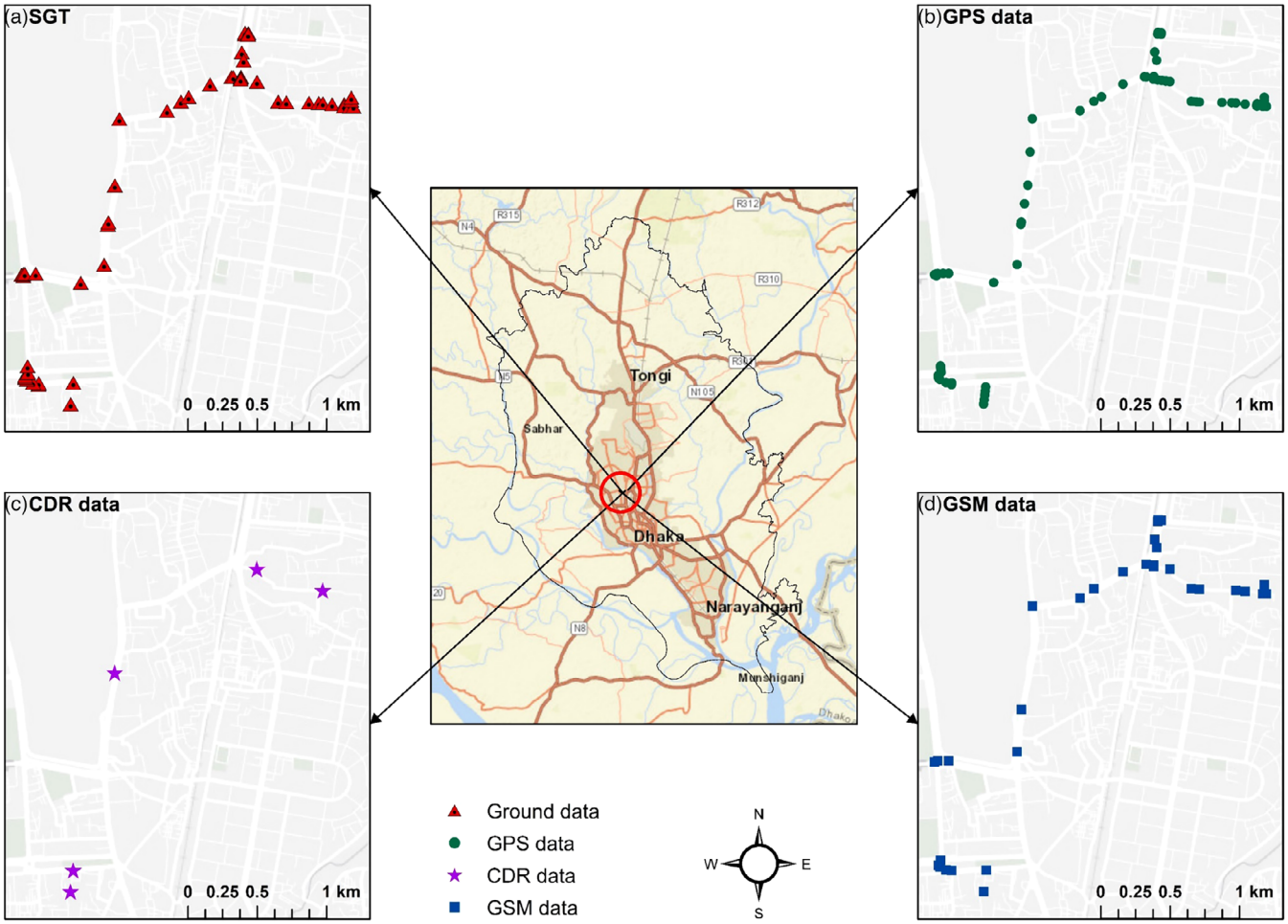


FIGURE 6 Synthetic ground truth (SGT) and trajectory data (one individual's trajectory). CDR, call detail records; GPS, global positioning system; GSM, global system for mobile communication.

distributed or dispersed. Among the three synthetic passive data, Moran's *I* value of GPS data (0.617444) suggested the highest level of positive spatial autocorrelation among the four datasets. Like the GPS, the GSM dataset exhibited a moderately high level of positive spatial autocorrelation.

While there was a possibility for stay locations to be clustered together in the CDR data, this clustering was not as strong as in the SGT dataset. To further analyse the clustering pattern of stay location, local Moran's *I* values were estimated, and a map of distribution has been included in supplementary Figure S1. The estimated local Moran's *I* values of clustered stay locations were compared using the K-S test. Results from the K-S tests are summarised in Table 7. The results of the K-S tests demonstrated that there were significant differences in local Moran's *I* distributions between SGT and passive data compared, with the most substantial differences observed between SGT and CDR, followed by SGT and GSM, and the least but still significant difference between SGT and GPS. The lowest *D* value (0.23437) from the K-S test between SGT and GPS data highlighted relatively less difference between the CDFs of the two samples.

TABLE 7 Comparison of local Moran's *I* index of clustered stay location between synthetic ground truth (SGT) and synthetic passive data using Kolmogorov–Smirnov (K-S) test.

	K-S test (*D*)	*p*-value
SGT vs GPS	0.23437	2.2e-16
SGT vs GSM	0.66632	2.2e-16
SGT vs CDR	0.86657	2.2e-16

Abbreviations: CDR, call detail records; GPS, global positioning system; GSM, global system for mobile communication.

3.3.2 | Trip-related statistics

Departure time information was extracted from the SGT, GPS, and GSM trajectories. The *end time* of each activity served as the departure time for the subsequent activity in the SGT. For instance, the beginning of a trip from home to work involved leaving the house and the end of home-related activity. Therefore, the end of home activity was the time for departure for work activity. The method followed to extract departure time

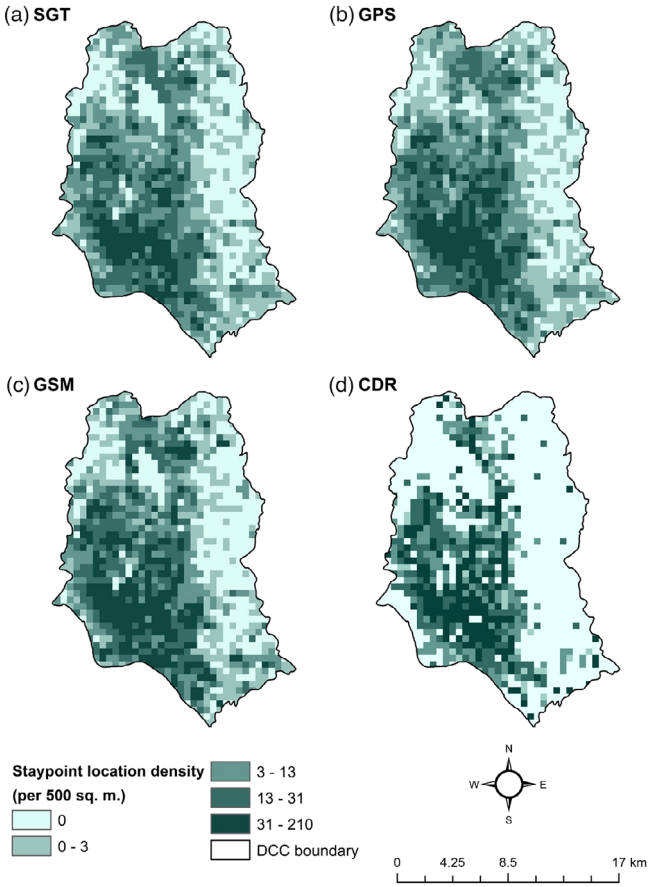


FIGURE 7 Stay-location distribution at the Dhaka City Corporation (DCC) area. CDR, call detail records; GPS, global positioning system; GSM, global system for mobile communication; SGT, synthetic ground truth.

TABLE 8 The Kolmogorov–Smirnov (KS) test results from departure time distribution.

	K-S test (*D*)	*p*-value
SGT vs GPS	0.067797	2.2×10^{-16}
SGT vs GSM	0.10054	2.2×10^{-16}
SGT vs CDR	0.13309	2.2×10^{-16}

Abbreviations: CDR, call detail records; GPS, global positioning system; GSM, global system for mobile communication; SGT, synthetic ground truth.

from GPS and GSM trajectory is explained in Section 2.3. Figure 8a,b illustrates respectively the departure time frequency distribution and CDFs of GPS, GSM, and SGT trajectories. In both figures, it is evident that GPS data was able to more accurately capture the variance in departure times during peak hours and better represented agents' departure times for different activities compared to GSM data. The K-S test results are shown in Table 8; the K-S test compared the distributions of a sample i.e., the passive data and the reference probability distribution of SGT. The K-S test statistics (*D* value) indicated that the maximum difference between the CDFs of the SGT and GPS datasets was approximately 0.068 suggesting a relatively small difference between the distributions of SGT and GPS. The SGT versus CDR comparison had the largest *D* value (0.133), indicating that the distributions of departure time in SGT and CDR were the most different (significant at 95% confidence level). However, in all cases the *p*-values suggested that the differences between the distributions were statistically significant.

Moreover, as shown in Figure 8a, both GPS and GSM trajectories substantially understated early morning departure times compared to the SGT. Conversely, GPS and GSM trajectories overrepresented late morning and late afternoon departure times. Consequently, the disparity between the SGT and GPS/GSM trajectories was more pronounced during the morning peak and late afternoon hours (In Dhaka, the morning peak starts from 8:00 to 10:00 and afternoon peak from 16:00 to 18:00 [94]) (Figure 8b). This variation may be attributed to the merging of activity or stay clusters near the journey's origin with the nearby congestion clusters, leading to over or underrepresentation during peak hours. Extracting agents' departure times from the one-day synthetic CDR data was challenging. Therefore, the distribution of sighting times (call/text times) extracted from the CDR data is presented in Figure 8. The sighting distribution from the one-day CDR data markedly differed from the SGT departure time distribution which was also supported by the K-S test results.

In addition to departure time, we compared trajectory distances derived from SGT and synthetic passive data. Figure 9 depicts the bivariate relationships and correlation coefficients (r) between the SGT trajectory and passive trajectories (GPS, GSM, and CDR). GPS trajectories exhibited the highest correlation ($r = 0.77$, p-value 2.2×10^{-16}), followed by the GSM trajectories ($r = 0.55$, p-value 2.2×10^{-16}), and CDR trajectories ($r = 0.47$, p-value 2.2×10^{-16}). GPS devices provided frequent (10 s intervals), accurate, and precise location and time information, likely contributing to the stronger association between the GPS and the SGT trajectory. In contrast, GSM provided triangulated approximate location data (60 s intervals), influenced by the network strength and mobile phone tower density, resulting in a moderate degree of correlation between synthetic GSM trajectories and the SGT trajectory. The lower spatial resolution of the CDR data (attributable to the lack of location records when the mobile phone was not in use) likely accounted for the low correlation coefficient between the CDR trajectory and the SGT. Furthermore, identifying home and possible activity locations using one-day CDR data proved challenging. This difficulty arose because sighting locations in the CDR data could represent locations observed during trips rather than end points. Additionally, capturing stay locations from CDR data was influenced by variations in mobile phone usage frequencies among users. Conversely, GPS and GSM data, along with their corresponding trajectories, featured finer temporal and spatial granularity, simplifying the identification of starting, intermediate, and ending locations within the trajectory.

4 | DISCUSSION

In this study, we introduced a comprehensive framework for assessing the relative accuracy of mobility information extracted from passive trajectory data, taking into account the limitations

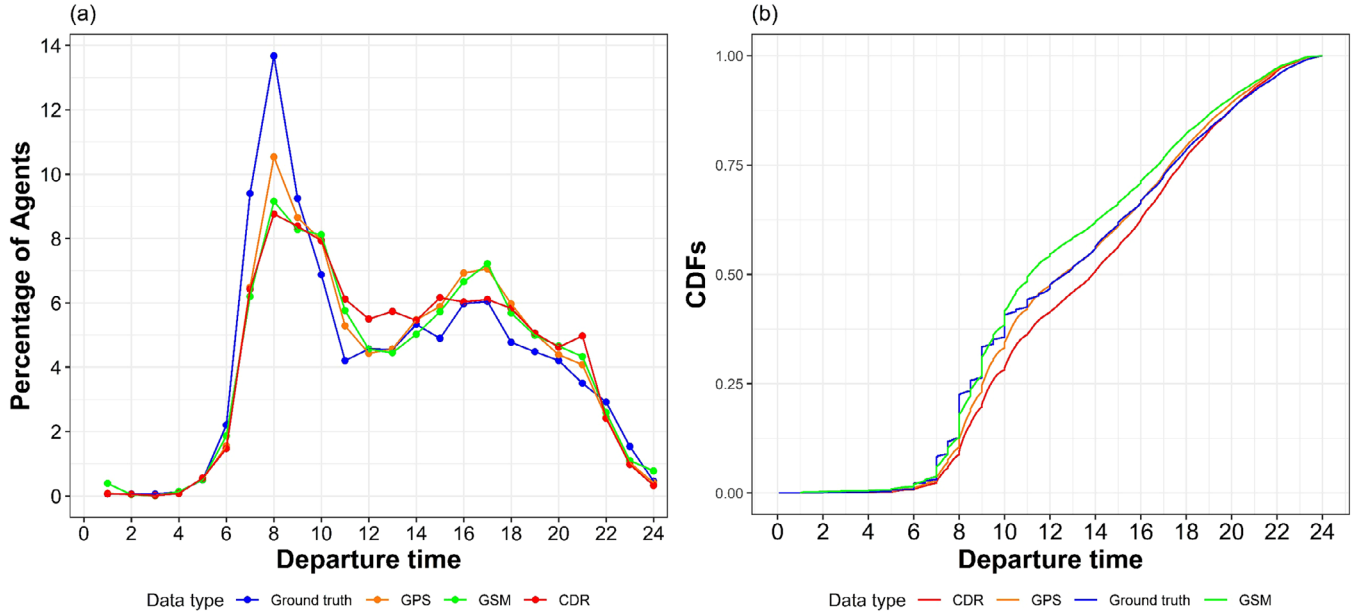


FIGURE 8 Time dimension of global positioning system (GPS), global system for mobile communication (GSM), call detail records (CDR) and synthetic ground truth (SGT) trajectory. (a) Frequency distribution of agents' departure time; (b) cumulative distributions of departure time (for GPS, GSM and SGT the time dimension shown in the figure are departure time while for CDR is the sighting time).

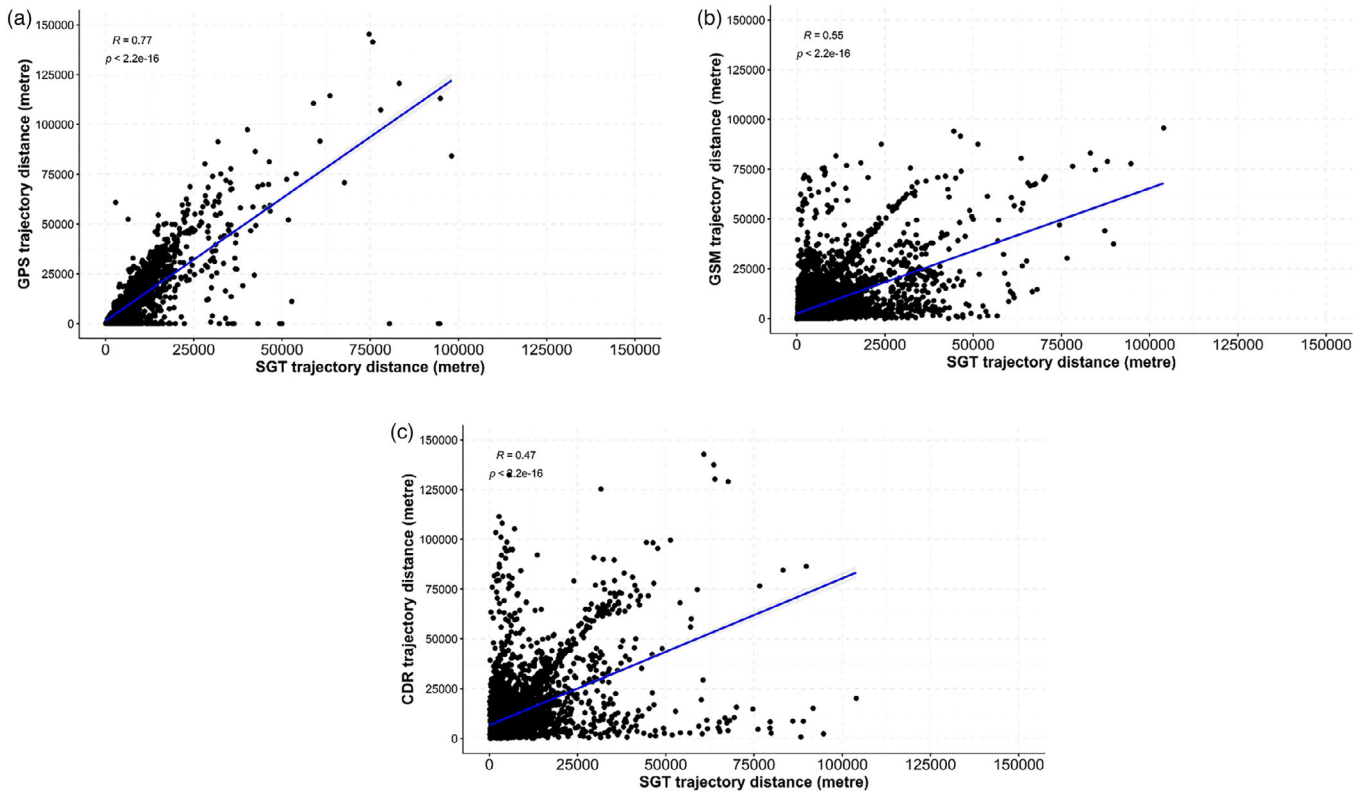


FIGURE 9 Correlation between SGT and synthetic traversed trajectory distance (Euclidean distance).

of conventional ground-truth data. We utilised MATSim simulation to generate synthetic ground truth (SGT) data. The mobility profile obtained from the MATSim simulation provided the essential inputs, such as information of individual geocoded activity location, and potential traversed route information, necessary for generating synthetic trajectories using a conventional trajectory generation module. The generation of synthetic data is a crucial step in the field of transport planning, as it allows for the testing of the accuracy and applicability of developed models with reproducible data [95]. Additionally, the generated SGT enabled us to evaluate the accuracy of mobility information from three different passive data sources (GPS, GSM, and CDR) at both disaggregate and aggregate levels.

Our assessment of stay location accuracy using GPS, GSM, and CDR data revealed that higher accuracy was associated with lower spatial resolution. Patrick [96] also emphasised the impact of CDR data resolution on the accuracy of mobility information derived from CDR trajectories. Bwambale and Choudhury [18] noted that when GSM cell sizes decrease, GSM time lags also reduce. We found at the finest resolution of 50 m, both CDR and GSM data exhibited poor performance in explaining the variability of SGT stay locations compared to GPS data. However, at a lower resolution of 500 m, both GPS and GSM accounted for over 80% of the variability in stay locations. These findings reinforced the notion that GSM data can effectively serve as a source of mobility information for evaluating activity profiles at coarse spatial resolution. A compromise in the resolution of the scale of analysis has the potential to reduce the measurement error of different passive data.

Moreover, the spatial autocorrelation test results highlighted the potential clustering of stay locations both in the SGT and synthetic passive data. The aggregated index values (Global Moran's *I*) confirmed the presence of clustering in stay location both in SGT and passive data which is supported by the empirical study related to identifying temporal and spatial regularity in travel trajectory [97]. On the other hand, the distribution of local Moran's *I* values for different passive datasets was significantly different from those of the SGT data. For the GSM and CDR data, the method used to derive stay locations—whether through triangulation or at the cell tower level [20, 63]—could result in different local area clustering values compared to the SGT data. On the other hand, the high resolution of GPS data and the extraction of longer congestion points as potential stay locations could also account for the differences in cluster indices.

Furthermore, the findings of this study shed light on the reliability of passive data in describing mobility profiles during various times of the day, including peak and off-peak hours. Notably, we observed that GPS and GSM data exhibited discrepancies (e.g. either understated or overstated) in the departure time distribution during peak hours (both in the early morning and late afternoon), potentially attributed to congestion near activity locations. However, it is important to note that GSM data exhibited a significantly higher measurement error compared to GPS data. The largest difference in departure time distribution between SGT and GPS data was about 3% (with a standard deviation of ± 0.78), while the difference between GSM and SGT departure time distribution was about 4.5% (with a standard deviation of ± 0.98). The K-S test results also supported these findings. Interestingly, Bwambale and Choudhury [18] reported that GSM data exhibited greater accuracy than GPS data when studying departure time choices in southwestern Switzerland. This discrepancy can be mainly attributed to observed GPS data in areas where topographical factors, dense foliage, and human factors (e.g. deactivated location service, and battery power loss) resulted in larger time gaps. Additionally, when compared to GPS and GSM, the time distribution generated from CDR data showed a larger deviation from the SGT time distribution, with a maximum deviation of approximately 5% (with a standard deviation of ± 1.5). This difference can be explained by the fact that one-day CDR data reflects sighting distribution rather than the departure time distribution when a text message or phone call is made [20]. However, it is worth noting that as mobile phone internet usage (e.g. calls, texts, and browsing) increases, CDR data is evolving towards continuous data, which could eventually achieve a temporal precision comparable to that of GSM data. Additionally, the statistically significant difference between the departure time distributions of the SGT and GPS data (particularly in the Dhaka case) can be attributed to the long-duration congestion detected as stay locations. The departure times extracted from these stay locations likely influenced the K-S test results. Such differences were also highlighted in the local Moran's *I* distribution of stay location cluster in SGT and GPS data.

The results of the one-to-one comparison between passive data and SGT also revealed that GPS data exhibited the highest level of agreement with the SGT when estimating travelled trajectory distance. Conversely, CDR data demonstrated the lowest level of agreement with the actual trajectory distance. Saarik [98] further emphasised the error in constructing mobility patterns due to the tower-level resolution of CDR data. Since we calculated the trajectory distance in this study as the sequenced Euclidean distance, accuracy depended on the sequencing and retrieval of the activity or stay locations from passive trajectory data. The sequence of activities in CDR data might not match the actual trajectory because locations are only recorded when the mobile phone is in use. While increased mobile phone usage frequency could potentially improve the sequencing of missing activity locations, the trajectory distance would still be limited to tower positions. Similarly, a weaker association between GSM trajectory distance and SGT, compared to GPS and SGT, can be attributed to the triangulated approximation of stay location information in GSM data.

In addition to device performance, the results showed that location errors in GPS and GSM data also depended on external factors such as road congestion and the relative distance between activity locations and congested roadways. As a result, deriving activity and travel locations from GPS and GSM data can introduce additional errors beyond positional shift, drift, and discontinuity noise. This may have a significant effect on the methods for location extraction [80]. Thus, it can be difficult to distinguish short stays (e.g., pick-up or drop-off locations) and activity locations from congestion stays when using GPS and GSM data. This has been particularly problematic in the

context of this case study due to the mixed land use and on-street parking facilities on most of the roads in Dhaka. Combining passive trajectory data with secondary data on traffic, land use, weather, and parking could help differentiate various types of stay locations, a potential avenue for future research. Additionally, conducting further research with multi-day panels of GPS and GSM data could aid in better distinguishing activity locations from long stay points resulting from traffic congestion.

Eventually, the proposed simulation-based framework effectively demonstrated the impact of various errors that cause positional disturbances and affect the accuracy of spatiotemporal mobility attributes from passive trajectory data. While this study specifically highlighted the effects of positional disturbances on mobility attributes, the use of different passive data sources with varying temporal resolutions also underscores the framework's effectiveness in assessing the impact of sampling bias on extracting mobility information from passive trajectory data.

5 | CONCLUSION

This article introduced a comprehensive four-step framework for assessing the accuracy of mobility information extracted from trajectory data. We employed this framework to generate finely detailed SGT data and synthetic trajectory data (GPS, GSM, and CDR) infused with realistic noise using MATSim simulation. The use of simulation-based GT allowed us to precisely evaluate multiple passive trajectory data sources and their accuracy in depicting mobility information. Through both visual and statistical analysis, we compared the statistical attributes and accuracy of trip-related factors (e.g. stay location, departure time, and travel distance) extracted from synthetic GPS, GSM, and CDR data with those from SGT. Our findings highlighted that the generated synthetic data had the potential to closely resemble real-world GPS, GSM, and CDR data. Furthermore, when considering the additional positional disturbance, GPS data outperformed GSM and CDR data in terms of deriving departure time, trajectory distance, and activity location information.

We also demonstrated empirically that the accuracy of passive data depends on various assumptions made during their evaluation, such as assuming the mobile device remained active throughout the journey. External factors, such as congestion and the relative distance from the road to the activity locations, also influenced accuracy. Additionally, the proposed framework offers several key advantages:

- Importance of SGT: This study underscored the significance of having SGT that closely matches the spatial and temporal granularity of passive data sources. We tested its significance with three mainstream passive data. Also, SGT facilitated precise and rigorous comparisons between passive data (from different sources) and a reliable reference in a controlled environment, enabling researchers to assess how accurately these sources capture human mobility patterns and choose the most suitable data source for their specific needs.
- Addressing challenges with real-life data: While real-life datasets are essential for accuracy assessment, they pose challenges in distinguishing relative inaccuracies caused by different noise levels and their impact on model uncertainties. The proposed framework would allow researchers to isolate and assess the influence of various types of noise, beyond positional disturbance, validating the stability of model outputs derived from passive data and their sensitivity to error size and extraction assumptions.
- Alignment with other accuracy assessments: The accuracy assessment results aligned closely with those from other assessments using travel diary surveys, census, or other GT sources. This alignment underscores the suitability and significance of the proposed SGT for future accuracy assessment studies. Additionally, the framework's versatility enables the assessment of passive trajectory data accuracy at both the individual and aggregate levels.

Finally, this framework provides a foundation for benchmarking models developed with passive data, aiding in the evaluation of various data management solutions. The results from the comparative analyses can help identify data requirements for different scales of transport planning and modelling. Additionally, accuracy assessments of passive data with appropriate GT have the potential to address challenges related to data collection, processing, and model specification complexities. They can reveal the most suitable passive data source for specific concerns, such as using GSM data instead of GPS data [25].

It is important to note that the findings deduced from the MATSim-based study using the Dhaka network may not be applicable universally. The accuracy level of passively generated data is affected by local factors like the topology of the transport network, topographic characteristics of the area (e.g. if the area is flat or mountainous), the presence of clusters of high-rise buildings, and coverage of the mobile phone network. For instance, in Dhaka, the mobile phone tower location density was notably seen to influence the accuracy of CDR data in the eastern part of the city. However, this effect and its magnitude may not hold true for cities in developed countries. Considering such contextual factors is imperative when utilising this framework for accuracy assessment. Furthermore, it is essential to have a comprehensive MATSim model realistically representing agents' activity and travel behaviour prevailing in the selected context for conducting such accuracy assessments, which entails substantial data requirements (e.g. MATSim model inputs such as network, activity plans, behavioural model), and complex technology-dependent simulations (e.g. high-performance computer to run the scenarios). These factors can introduce additional noise during the event file generation stage. For example, if there is no potential route between the specified origin and destination, agents will end up teleporting without traversing the network. This will generate misleading trajectories in the event file.

Moreover, this study investigated the accuracy of three mainstream passive data sources with a standard positioning (random shift and drift) and temporal resolution. Future research should explore the sensitivity of these data sources to different noise

levels, including sampling bias, variations in shift and drift parameters, tower density, call rate, and location update frequency. The accuracy of mobile phone CDR data was assessed only using the distribution of calls. CDR data may include a variety of records beyond just calls, such as text messages, mobile data usage, and location updates, which collectively contribute to a richer dataset. Incorporating these diverse types of CDR records can significantly enhance the resolution of the data and the accuracy of the extracted mobility information. For example, by combining records of calls, texts, and data usage, we can obtain a more comprehensive view of both specific activities and continuous movement patterns. While this approach can resemble low-resolution GPS data, it is important to note that the locations provided by CDR data are still at the tower level. Therefore, as our findings highlighted, the density of mobile tower locations remains a critical factor in determining the accuracy of the data.

Additionally, the impact of noise, arising from external sources such as land use, built environment, and topographic conditions etc., on trajectory data can be explored using the proposed framework. This can be achieved by generating synthetic passive data with added noise from external sources through integrating MATSim event files with external data sources, such as land use maps or weather data. The impact assessment of different types of noise (both internal and external) on trajectory data will also enable the assessment of the effectiveness of different data processing algorithms for extracting trajectory information from passive data sources. Similar investigations can assess the accuracy of other passive data sources, such as smart card data and automatic vehicle location information. Expanding this research to compare model outputs derived from different passive data sources, such as the value of travel time (VTT), is also a valuable avenue for exploration. Additionally, conducting a comparative analysis between outputs obtained from MATSim simulation and real GPS, GSM, and CDR data collected from the same sample used for calibrating MATSim can also provide insights into the influence of other factors like traffic density and congestion on assessing the accuracy of mobility information obtained from passive data. While this work focused on checking stay location accuracy at various spatial resolutions (50–500 m), future research can delve into comparing the relative accuracy of real-time passive trajectory positions and related link/lane level locations by testing various map-matching algorithms [67, 99]. Ultimately, the proposed framework holds the promise for generating trajectory data in data-scarce cities and validating them with appropriate GT information to make informed decisions based on validated models developed using passive data. The research community can further enhance and develop new datasets according to their specific requirements.

AUTHOR CONTRIBUTIONS

Khatun E. Zannat: Conceptualization; data curation; formal analysis; investigation; methodology; software; validation; visualization; writing—original draft; writing—review and editing. **Charisma F. Choudhury**: Data curation; investigation; methodology; resources; supervision; validation; writing—review and editing. **Stephane Hess**: Investigation; methodology; resources; supervision; validation; writing—review and editing. **David Watling**: Investigation; methodology; resources; validation; writing—review and editing.

ACKNOWLEDGEMENTS

This research is funded by the Faculty for the Future Program of the Schlumberger Foundation. Charisma Choudhury's time has been partially supported by the UKRI Future Leader Fellowship [MR/T020423/1-NEXUS]. Stephane Hess acknowledges support by the European Research Council through advanced Grant 101020940-SYNERGY. The data used for the study has been made available by the Dhaka Transport Coordination Agency (DTCA). The authors acknowledge the support from Mr Anisur Rahman and Mr Dhrubo Alam of DTCA for providing clarifications regarding the data. The authors also acknowledge the training and logistic support from the Choice Modelling Centre, University of Leeds, UK, and Transport Systems Planning and Transport Telematics, Technical University of Berlin, Germany. Also, this simulation work was undertaken on ARC4, part of the High-Performance Computing facilities at the University of Leeds, UK.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data will be made available on request due to privacy/ethical restrictions.

ORCID

Khatun E. Zannat https://orcid.org/0000-0003-3108-5732
Charisma F. Choudhury https://orcid.org/0000-0002-8886-8976
Stephane Hess https://orcid.org/0000-0002-3650-2518
David Watling https://orcid.org/0000-0002-6193-9121

REFERENCES

1. Huang, H., Cheng, Y., Weibel, R.: Transport mode detection based on mobile phone network data: A systematic review. Transp. Res. Part C Emerging Technol. 101, 297–312 (2019)
2. Pelletier, M.-P., Trépanier, M., Morency, C.: Smart card data use in public transit: A literature review. Transp. Res. Part C Emerging Technol. 19(4), 557–568 (2011)
3. Wang, Z., He, S.Y., Leung, Y.: Applying mobile phone data to travel behaviour research: A literature review. Travel Behav. Soc. 11, 141–155 (2018)
4. Shen, L., Stopher, P.R.: Review of GPS travel survey and GPS data-processing methods. Transp. Rev. 34(3), 316–334 (2014)
5. Harrison, G., Grant-Muller, S.M., Hodgson, F.C.: New and emerging data forms in transportation planning and policy: Opportunities and challenges for "Track and Trace" data. Transp. Res. Part C Emerging Technol. 117, 102672 (2020)
6. Zannat, K.E., Choudhury, C.F.: Emerging big data sources for public transport planning: A systematic review on current state of art and future research directions. J. Indian Inst. Sci. 99(4), 601–619 (2019)
7. von Mörner, M.: Application of call detail records-chances and obstacles. Transp. Res. Procedia 25, 2233–2241 (2017)
8. Jansen, R., Kovacs, K., Esko, S., Saluveer, E., Sõstra, K., Bengtsson, L., et al.: Guiding principles to maintain public trust in the use of mobile operator data for policy purposes. Data Policy 3, e24 (2021)

9. Liu, J., Li, J., Li, W., Wu, J.: Rethinking big data: A review on the data quality and usage issues. ISPRS J. Photogramm. Remote Sens. 115, 134–142 (2016)
10. Eagle, N., Pentland, A.: Reality mining: Sensing complex social systems. Pers. Ubiquitous Comput. 10, 255–268 (2006)
11. Hendawi, A., Shen, J., Sabbineni, S.S., Song, Y., Cao, P., Zhang, Z., et al., editors: Noise patterns in GPS trajectories. In: Proceedings of the 2020 21st IEEE International Conference on Mobile Data Management (MDM). IEEE, Piscataway, NJ (2020)
12. Forghani, M., Karimipour, F., Claramunt, C.: From cellular positioning data to trajectories: Steps towards a more accurate mobility exploration. Transp. Res. Part C: Emerging Technol. 117, 102666 (2020)
13. Su, R., Dodge, S., Goulias, K.G.: Understanding the impact of temporal scale on human movement analytics. J. Geograph. Syst. 24(3), 353–388 (2022)
14. Wu, R., Luo, G., Shao, J., Tian, L., Peng, C.: Location prediction on trajectory data: A review. Big Data Min. Anal. 1(2), 108–127 (2018)
15. Etter, V., Kafsi, M., Kazemi, E., Been there, done that: What your mobility traces reveal about your behavior. Paper presented at the Mobile Data Challenge by Nokia Workshop, in Conjunction with International Conference on Pervasive Computing, Newcastle, UK, 18–19 June 2012
16. Dong, Y.: Disaggregated Short-Term Travel Location Prediction. United Kingdom Imperial College London, UK (2022)
17. Li, Y., Ran, Z., Tsai, L., Williams, S.: Using call detail records to determine mobility patterns of different socio-demographic groups in the western area of Sierra Leone during early COVID-19 crisis. Environ. Plann. B: Urban Anal. City Sci. 50(5), 1298–1312 (2023)
18. Bwambale, A., Choudhury, C.F., Hess, S.: Modelling departure time choice using mobile phone data. Transp. Res. Part A Policy Pract. 130, 424–439 (2019)
19. Kang, X., Liu, L., Zhao, D., Ma, H.: TraG: A trajectory generation technique for simulating urban crowd mobility. IEEE Trans. Ind. Inf. 17(2), 820–829 (2020)
20. Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M.: The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp. Res. Part C Emerging Technol. 68, 285–299 (2016)
21. Anda, C., Erath, A., Fourie, P.J.: Transport modelling in the age of big data. Int. J. Urban Sci. 21(1), 19–42 (2017)
22. Ahas, R., Laineste, J., Aasa, A., Mark, Ü.: The spatial accuracy of mobile positioning: some experiences with geographical studies in Estonia. In: Location Based Services and Telecartography, pp. 445–460. Springer, New York (2007)
23. Fund, F., Lin, R., Korakis, T., Panwar, S.S.: How bad is the flat earth assumption? Effect of topography on wireless systems. In: Proceedings of the 2016 14th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), pp. 1–5. IEEE, Piscataway, NJ (2016)
24. Paul, B.S., Rimer, S., editors: Wireless sensor node placement due to power loss effects from surrounding vegetation. In: Proceedings of the 9th International Conference on Quality, Reliability, Security and Robustness in Heterogeneous Networks, pp. 915–927. Springer, Berlin, Heidelberg (2013)
25. Ahas, R., Aasa, A., Silm, S., Aunap, R., Kalle, H., Mark, Ü.: Mobile positioning in space–time behaviour studies: Social positioning method experiments in Estonia. Cartogr. Geogr. Inf. Sci. 34(4), 259–273 (2007)
26. Song, X., Long, Y., Zhang, L., Rossiter, D.G., Liu, F., Jiang, W.: Spatial accuracy evaluation for mobile phone location data with consideration of geographical context. IEEE Access 8, 221176–221190 (2020)
27. Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C.: Development of origin–destination matrices using mobile phone call data. Transp. Res. Part C Emerging Technol. 40, 63–74 (2014)
28. Liao, L., Fox, D., Kautz, H.: Extracting places and activities from GPS traces using hierarchical conditional random fields. Int. J. Rob. Res. 26(1), 119–134 (2007)
29. Jones, P., Stopher, P.R.: Transport Survey Quality and Innovation. Emerald Group Publishing Limited, Bingley, UK (2003)
30. Wolf, J., Bricka, S., Ashby, T., Gorugantua, C.: Advances in the application of GPS to household travel surveys. Paper presented at the National Household Travel Survey Conference, Washington DC, 1–2 November 2004
31. Stopher, P., FitzGerald, C., Xu, M.: Assessing the accuracy of the Sydney household travel survey with GPS. Transportation 34, 723–741 (2007)
32. Harding, C., Faghih Imani, A., Srikukenthiran, S., Miller, E.J., Nurul Habib, K.: Are we there yet? Assessing smartphone apps as full-fledged tools for activity-travel surveys. Transportation 48, 2433–2460 (2021)
33. Janzen, M., Vanhoof, M., Smoreda, Z., Axhausen, K.W.: Closer to the total? Long-distance travel of French mobile phone users. Travel Behav. Soc. 11, 31–42 (2018)
34. Kohla, B., Meschik, M.: Comparing trip diaries with GPS tracking: Results of a comprehensive Austrian study. In: Transport Survey Methods: Best Practice for Decision Making, pp. 305–320. Emerald Group Publishing Limited, Bingley, UK (2013)
35. Zmud, J., Lee-Gosselin, M., Munizaga, M., Carrasco, J.A.: Transport Survey Methods: Best Practice for Decision Making. Emerald Group Publishing Limited, Bingley, UK (2013)
36. Bar-Gera, H.: Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. Transp. Res. Part C Emerging Technol. 15(6), 380–391 (2007)
37. Sternfeld, B., Jiang, S.F., Picchi, T., Chasan-Taber, L., Ainsworth, B., Quesenberry, C.P.: Evaluation of a cell phone-based physical activity diary. Med. Sci. Sports Exercise 44(3), 487–495 (2012)
38. Tsoleridis, P., Choudhury, C.F., Hess, S.: Deriving transport appraisal values from emerging revealed preference data. Transp. Res. Part A Policy Pract. 165, 225–245 (2022)
39. Pan, Y., Sun, Q., Yang, M., Darzi, A., Zhao, G., Kabiri, A., et al.: Residency and worker status identification based on mobile device location data. Transp. Res. Part C Emerging Technol. 146, 103956 (2023)
40. Bohte, W., Maat, K.: Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. Transp. Res. Part C Emerging Technol. 17(3), 285–297 (2009)
41. Vanhoof, M., Lee, C., Smoreda, Z.: Performance and sensitivities of home detection on mobile phone data. In: Big Data Meets Survey Science: A Collection of Innovative Methods, pp. 245–271. Wiley, Hoboken, NJ (2020)
42. Vanhoof, M., Reis, F., Ploetz, T., Smoreda, Z.: Assessing the quality of home detection from mobile phone data for official statistics. J. Off. Stat. 34(4), 935–960 (2018)
43. Sadeghinasr, B., Akhavan, A., Wang, Q.: Estimating commuting patterns from high resolution phone GPS data. In: Proceedings of the ASCE International Conference on Computing in Civil Engineering 2019. American Society of Civil Engineers, Reston, VA (2019)
44. Bwambale, A., Choudhury, C.F., Hess, S., Iqbal, M.S.: Getting the best of both worlds—a framework for combining disaggregate travel survey data and aggregate mobile phone data for trip generation modelling. Transportation 48, 2287–2314 (2019)
45. Shackman, G., Wang, X., Liu, Y.-L.: Brief review of world demographic trends: Explaining population trends: Birth, death and migration. A Report from the Global Social Change Research Project (2012). https://doi.org/10.2139/ssrn.2163196
46. Gordon, J.B., Koutsopoulos, H.N., Wilson, N.H., Attanucci, J.P.: Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transp. Res. Record 2343(1), 17–24 (2013)
47. Li, Z., Yu, L., Gao, Y., Wu, Y., Song, G., Gong, D.: Identifying temporal and spatial characteristics of residents' trips from cellular signaling data: Case study of Beijing. Transp. Res. Record 2672(42), 81–90 (2018)
48. Wolf, J., Loechl, M., Thompson, M., Arce, C.: Trip rate analysis in GPS-enhanced personal travel surveys. In: Transport Survey Quality and Innovation, pp. 483–498. Emerald Group Publishing Limited, Bingley, UK (2003)
49. Pappalardo, L., Ferres, L., Sacasa, M., Cattuto, C., Bravo, L.: An individual-level ground truth dataset for home location detection. arXiv:201008814 (2020)
50. Yang, Y., Xiong, C., Zhuo, J., Cai, M.: Detecting home and work locations from mobile phone cellular signaling data. Mob. Inf. Syst. 2021, 1–13 (2021)

51. Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C.: The path most traveled: Travel demand estimation using big data resources. Transp. Res. Part C Emerging Technol. 58, 162–177 (2015)
52. Zilske, M., Nagel, K.: Studying the accuracy of demand generation from mobile phone trajectories with synthetic data. Procedia Comput. Sci. 32, 802–807 (2014)
53. Zhang, G., Rui, X., Poslad, S., Song, X., Fan, Y., Wu, B.: A method for the estimation of finely-grained temporal spatial human population density distributions based on cell phone call detail records. Remote Sens. 12(16), 2572 (2020)
54. Hemmings, T., Goves, C.: Utilising Mobile Network Data for Transport Modelling: Recommendations Paper. Department for Transport, United Kingdom (2017)
55. Bekhor, S., Dobler, C., Axhausen, K.W.: Integration of activity-based and agent-based models: Case of Tel Aviv, Israel. Transp. Res. Rec. 2255(1), 38–47 (2011). https://doi.org/10.3141/2255-05
56. Zannat, K.E., Laudan, J., Choudhury, C.F., Hess, S.: Developing an agent-based microsimulation for predicting the Bus Rapid Transit (BRT) demand in developing countries: A case study of Dhaka, Bangladesh. Transp. Policy 148, 92–106 (2024)
57. Smith, D.M., Clarke, G.P., Harland, K.: Improving the synthetic data generation process in spatial microsimulation models. Environ. Plan. A 41(5), 1251–1268 (2009)
58. Olynik, M.: Temporal Characteristics of GPS Error Sources and Their Impact on Relative Positioning. University of Calgary, Calgary, Canada (2002)
59. Tahsin, M., Sultana, S., Reza, T., Hossam-E-Haider, M.: Analysis of DOP and its preciseness in GNSS position estimation. In: Proceedings of the 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), pp. 1–6. IEEE, Piscataway, NJ (2015)
60. Enge, P.K.: The global positioning system: Signals, measurements, and performance. Int. J. Wireless Inf. Networks 1, 83–105 (1994)
61. Dinesh, S., Faudzi, M., Fitry, M.Z.: Evaluation of the effect of radio frequency interference on global positioning system (GPS) accuracy via GPS simulation. Def. Sci. J. 62(5), 338–347 (2012)
62. Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J., Yin, L.: Understanding the bias of call detail records in human mobility research. Int. J. Geogr. Inf. Sci. 30(9), 1738–1762 (2016)
63. Jiang, S., Fiore, G.A., Yang, Y., Ferreira, J. Jr., Frazzoli, E., González, M.C.: 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. Association for Computing Machinery, New York, NY (2013)
64. Zhao, Z., Shaw, S.-L., Yin, L., Fang, Z., Yang, X., Zhang, F., et al.: The effect of temporal sampling intervals on typical human mobility indicators obtained from mobile phone location data. Int. J. Geograph. Inf. Sci. 33(7), 1471–1495 (2019)
65. Gong, H., Chen, C., Bialostozky, E., Lawson, C.T.: A GPS/GIS method for travel mode detection in New York City. Comput. Environ. Urban Syst. 36(2), 131–139 (2012)
66. Merry, K., Bettinger, P.: Smartphone GPS accuracy study in an urban environment. PLoS One 14(7), e0219890 (2019)
67. Quddus, M.A., Noland, R.B., Ochieng, W.Y.: Validation of map matching algorithms using high precision positioning with GPS. J. Navig. 58(2), 257–271 (2005)
68. Hasan, A.M., Samsudin, K., Ramli, A.R., Azmir, R., Ismaeel, S.: A review of navigation systems (integration and algorithms). Aust. J. Basic Appl. Sci. 3(2), 943–959 (2009)
69. Zhu, W., Hou, J., Liu, Z., Ding, Z.: GPS positioning error compensation based on kalman filtering. J. Phys. Conf. Ser. 1920, 012088 (2021)
70. Bösche, K., Sellam, T., Pirk, H., Beier, R., Mieth, P., Manegold, S.: Scalable generation of synthetic GPS traces with real-life data characteristics. In: Proceedings of the 4th TPC Technology Conference; Selected Topics in Performance Evaluation and Benchmarking, pp. 140–155. Springer, Berlin, Heidelberg (2013)
71. Renfro, B.A., Stein, M., Reed, E.B., Villalba, E.J.: An Analysis of Global Positioning System Standard Positioning Service performance for 2020. Space and Geophysics Laboratory Applied Research Laboratories, University of Texas at Austin, Austin, Texas (2022)
72. Schipperijn, J., Kerr, J., Duncan, S., Madsen, T., Klinker, C.D., Troelsen, J.: Dynamic accuracy of GPS receivers for use in health research: A novel method to assess GPS accuracy in real-world settings. Front. Public Health 2, 21 (2014)
73. Steenbruggen, J., Borzacchiello, M.T., Nijkamp, P., Scholten, H.: Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: A review of applications and opportunities. GeoJournal 78, 223–243 (2013)
74. Ratti, C., Frenchman, D., Pulselli, R.M., Williams, S.: Mobile landscapes: Using location data from cell phones for urban analysis. Environ. Plan. B: Plan. Des. 33(5), 727–748 (2006)
75. Chen, M.Y., Sohn, T., Chmelev, D., Haehnel, D., Hightower, J., Hughes, J., et al., editors: Practical metropolitan-scale positioning for gsm phones. In: Proceedings of the UbiComp 2006: Ubiquitous Computing: 8th International Conference, UbiComp 2006 Orange County, CA, USA, September 17–21, 2006 Proceedings 8, pp. 225–242. Springer, Berlin, Heidelberg (2006)
76. Varshavsky, A., De Lara, E., Hightower, J., LaMarca, A., Otsason, V.: GSM indoor localization. Pervasive Mob. Comput. 3(6), 698–720 (2007)
77. Horn, C., Klampfl, S., Cik, M., Reiter, T.: Detecting outliers in cell phone data: Correcting trajectories to improve traffic modeling. Transp. Res. Rec. 2405(1), 49–56 (2014)
78. Letkowski, J.: Applications of the Poisson probability distribution. Academic and Business Research Institute Conference - San Antonio, 22–24 March 2012
79. Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A.Y., et al.: A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE Trans. Emerging Top. Comput. 2(3), 267–279 (2014)
80. Fu, Z., Tian, Z., Xu, Y., Qiao, C.: A two-step clustering approach to extract locations from individual GPS trajectory data. ISPRS Int. J. Geo-Inf. 5(10), 166 (2016)
81. Hafezi, M.H., Liu, L., Millward, H.: Identification of representative patterns of time use activity through fuzzy C-means clustering. Transp. Res. Rec. 2668(1), 38–50 (2017)
82. Sun, H., Chen, Y., Lai, J., Wang, Y., Liu, X.: Identifying tourists and locals by K-means clustering method from mobile phone signaling data. J. Transp. Eng. A: Syst. 147(10), 04021070 (2021)
83. Wang, Y., Qin, K., Chen, Y., Zhao, P.: Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi GPS data. ISPRS Int. J. Geo-Inf. 7(1), 25 (2018)
84. Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Fang, Z., Li, Q.: Understanding aggregate human mobility patterns using passive mobile phone location data: A home-based approach. Transportation 42, 625–646 (2015)
85. Gong, L., Sato, H., Yamamoto, T., Miwa, T., Morikawa, T.: Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines. J. Mod. Transp. 23, 202–213 (2015)
86. Zimmermann, M., Kirste, T., Spiliopoulou, M.: Finding stops in error-prone trajectories of moving objects with time-based clustering. In: Proceedings of the International Conference on Intelligent Interactive Assistance and Mobile Multimedia Computing, pp. 275–286. Springer, Berlin, Heidelberg (2009)
87. Hazan, I., Shabtai, A.: Improving grid-based location prediction algorithms by speed and direction based boosting. IEEE Access 7, 21211–21219 (2019)
88. An, N.T., Phuong, T.M.: A gaussian mixture model for mobile location prediction. In: Proceedings of the 2007 IEEE International Conference on Research, Innovation and Vision for the Future, pp. 152–157. IEEE, Piscataway, NJ (2007)
89. Braham, H., Jemaa, S.B., Fort, G., Moulines, E., Sayrac, B.: Spatial prediction under location uncertainty in cellular networks. IEEE Trans. Wireless Commun. 15(11), 7633–7643 (2016)
90. Getis, A., Ord, J.K.: The analysis of spatial association by use of distance statistics. Geog. Anal. 24(3), 189–206 (1992)
91. Anselin, L.: Local indicators of spatial association—LISA. Geog. Anal. 27(2), 93–115 (1995)

92. Massey, F.J. Jr.: The Kolmogorov-Smirnov test for goodness of fit. J. Am. Stat. Assoc. 46(253), 68–78 (1951)
93. Axhausen, K.W., Horni, A., Nagel, K.: The Multi-Agent Transport Simulation MATSim. Ubiquity Press, London (2016)
94. Zannat, K.E., Choudhury, C.F., Hess, S.: Modeling departure time choice of car commuters in Dhaka, Bangladesh. Transp. Res. Rec. 2676(2), 247–262 (2022)
95. Chatterjee, S., Byun, Y.-C.: A synthetic data generation technique for enhancement of prediction accuracy of electric vehicles demand. Sensors 23(2), 594 (2023)
96. Patrick, A.J.: Using Simulation to Investigate the Accuracy of Mobility Data Derived from Call Detail Records. University of Leeds, Leeds, England (2016)
97. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.-L.: Understanding individual human mobility patterns. Nature 453(7196), 779–782 (2008)
98. Saarik, A.: Trajectory Reconstruction and Mobility Pattern Analysis Based on Call Detail Record Data. University of Tartu, Tartu, Estonia (2017)
99. Bierlaire, M., Chen, J., Newman, J.: A probabilistic map matching method for smartphone GPS data. Transp. Res. Part C Emerging Technol. 26, 78–98 (2013)

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zannat, K.E., Choudhury, C.F., Hess, S., Watling, D.: Investigating the relative accuracy of GPS, GSM and CDR data for inferring spatiotemporal travel trajectories. IET Intell. Transp. Syst. 1–21 (2024). https://doi.org/10.1049/itr2.12563