Comparison of GPS imputation methods in environmental health research

Sungsoon Hwang,[1] Kashica J. Webber-Ritchey,[2] Elizabeth Moxley[3]

[1]Department of Geography, DePaul University, Chicago, IL; [2]School of Nursing, DePaul University, Chicago, IL; [3]College of Health and Human Sciences, Northern Illinois University, DeKalb, IL, USA

Abstract

Assessment of personal exposure in the external environment commonly relies on global positioning system (GPS) measurements. However, it has been challenging to determine exposures accurately due to missing data in GPS trajectories. In environmental health research using GPS, missing data are often discarded or are typically imputed based on the last known location or linear interpolation. Imputation is said to mitigate bias in exposure measures, but methods used are hardly evaluated against ground truth. Widely used imputation methods assume that a person is either stationary or constantly moving during the missing interval. Relaxing this assumption, we propose a method for imputing locations as a function of a person's likely movement state (stop, move) during the missing interval. We then evaluate the proposed method in terms of the accuracy of imputed location, movement state, and daily mobility measures such as the number of trips and time spent on places visited. Experiments based on real data collected by participants (n=59) show that the proposed approach outperforms existing methods. Imputation to the last known location can lead to large deviation from the actual location when gap distance is large. Linear interpolation is shown to result in large errors in mobility measures. Researchers should be aware that the different treatment of missing data can affect the spatiotemporal accuracy of GPS-based exposure assessments.

Correspondence: Sungsoon Hwang, Department of Geography, DePaul University, 990 W Fullerton Ave., Chicago, IL 60614, USA.
Tel.: +1(773)325.8668 - Fax: +1(773)325.4590.
E-mail: shwang9@depaul.edu

Key words: GPS trajectory data; geographic imputation; exposure assessment; daily mobility measures; data quality.

Conflict of interest: The Authors declare no conflict of interest.

Acknowledgements: This research was funded by DePaul University, grant number 601317.

Received for publication: 15 February 2022.
Revision received: 1 July 2022.
Accepted for publication: 4 July 2022.

Geospatial Health 2022; 17:1081
doi:10.4081/gh.2022.1081

Publisher's note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Introduction

GPS allows measuring a person's movement and activity patterns, which is a key element in assessing personal exposure in the external environment (Loh *et al*., 2017). GPS trajectory data are increasingly used to monitor community mobility (participation) of people with limited mobility (Evans *et al*., 2012; Hordacre *et al*., 2014; Jayaraman *et al*., 2014; Hanke *et al*., 2019) and determine environmental correlates of physical activity (PA) (Jones *et al*., 2009; Maddison *et al*., 2010; Wheeler *et al*., 2010; Quigg *et al*., 2010; Almanza *et al*., 2012; Kerr *et al*., 2012; McGrath *et al*., 2015; Rundle *et al*., 2016; Jansen *et al*., 2018). Despite the growing use of GPS, there is no good consensus on GPS data processing procedures for environmental health research (Kerr *et al*., 2011; Krenn *et al*., 2011; Klinker *et al*., 2014; McCrorie *et al*., 2014; Fillekes *et al*., 2019). This is partly because human trajectory data are sparsely sampled where built structure interferes with GPS signals. In many health studies using GPS, missing intervals (or time gaps) in GPS trajectories are often excluded for subsequent analysis although those gaps represent significant portions of important places visited. Their exclusion can generate biased exposure measures and thus affect inference with respect to environment-health links (Mennis *et al*., 2018). The spatial buffer of raw GPS data is of limited use as a measure of activity spaces (Hirsch *et al*., 2014) because temporal aspects of visited places (*e.g*., arrival time, duration of stay) can be miscalculated due to missing data (Christensen *et al*., 2021).

Not surprisingly, researchers have attempted to impute locations over missing intervals in trajectories (geographic imputation in short) to improve the assessment of spatiotemporal exposure of persons to the external environment. It has been shown that there can be a difference of several minutes for PA at various locations before and after imputation (Meseck *et al*., 2016). Dwell time at home was substantially underestimated from raw data (*i.e*. without imputation) and imputation rendered this measure less biased (Yoo *et al*., 2020). Imputation is increasingly seen as a viable strategy for dealing with missing data in trajectories. Imputation is even more necessary as a large volume of low-resolution GPS data is generated by smartphones. Given the prevalence of missing location data in GPS trajectories, it is important to consider effects of treating missing GPS data when assessing exposures.

Although making GPS data complete is integral to accurate and reliable analysis, imputation is often conducted without a good knowledge of how it works. An existing imputation method typically assumes that a person is either stationary or constantly moving during the missing interval. We propose a geographic imputation method that considers a person's potential stop and move (SAM) characteristics during time gaps. To achieve this, we compared how accurately different geographic imputation methods infer locations and movement states during time gaps. We also compared the accuracy of daily mobility measures (*e.g.*, the number of trips and time spent on places visited) as mentioned by Fillekes *et al*. (2019) resulting from different imputation methods. This article is intended to shed light on what makes effective imputation strategies and how imputation methods affect exposure assessment.

Materials and methods

How missing GPS data are treated in health research

To understand the geographic context of health behaviour, researchers have resorted to GPS, accelerometers and geographic information systems (GIS). A GPS-equipped device measures an individual's movement (or location) at given intervals, generating spatial trajectory data; an accelerometer measures the level of PA (or body posture); while GIS puts together the geospatial data (*e.g.*, walkability, access to food, air pollution, green space, *etc.*) in a geographic context. The combination of these technologies enables objective and contextual measurement of health behaviours at a high level of specificity and accuracy (Chaix *et al*., 2012, 2013; Jankowska *et al*., 2015). To determine environmental correlates of PA, researchers synchronize GPS data with accelerometer data based on timestamp and overlay the synchronized data with contextual data in GIS (Klinker *et al*., 2014; Oreskovic *et al*., 2015). Spatiotemporal accuracy of imputed GPS trajectories (*i.e.* where a person was at a particular time) is not only crucial to the integration of these data but can also affect the validity of conclusions based on such data.

To examine how missing data in GPS trajectories are treated in this type of research, we selected reputable articles in environmental health using GPS trajectory data with high specificity (*e.g.*, actual locations rather than indoor/outdoor or within/outside of neighbourhoods), where accelerometers or GIS are used at varying degrees. Table 1 summarizes approaches to treating missing GPS data alongside findings in these articles. The review reveals that gaps (signal loss) are prevalent. Missing data are discarded for analysis (*i.e.* not imputed) in many studies. It appears that the imputation of missing location data has recently become commonplace. Below we describe existing imputation methods used in studies listed in Table 1. An unknown location is typically imputed to the last known location of a GPS track point before signal loss ('last fix' onwards). The unknown location is less commonly imputed to the first known location after signal loss ('first fix' onwards). Location data are typically missing when persons are in an area where the GPS signal does not penetrate (*e.g.*, buildings, subways, tunnels and the like). In other words, 'missingness' in GPS trajectories is not a random event. This type of imputation method is denoted as stop-based (ST) in Table 1. This method can impute location well if a person is stationary during signal loss. However, most studies using this method fail to check whether the gap is part of a stop.

Another common method is to impute an unknown location with the assumption that persons are constantly moving during signal loss. It is possible to delineate a space-time prism (STP) or a maximal possible boundary of a person's whereabouts with spatiotemporal coordinates of two anchor points (last fix, first fix) (Miller, 1991; Pfoser and Jensen, 1999; Hornsby and Egenhofer, 2002). In a typical form, unknown locations are linearly interpolated along the path between two anchor points with a constant velocity as a representative location of STP. The path can be represented as the straight line, the least-cost route (Chaix *et al*., 2019) using external data (*e.g.*, transportation network), or popular routes mined from a person's historical trajectories (Wei *et al*., 2012; Zhao *et al*., 2021). This type of imputation method is denoted as linear interpolation (LI) in Table 1. It performs well if the assumption of constant motion is valid.

Some recent studies listed in Table 1 use *personal activity and location measurement system* (PALMS) (Carlson *et al*., 2015). This approach allows researchers to clean GPS trajectory data, identify trips (moves) and locations (stops), and integrate GPS data with PA data. With default parameter values, PALMS detects signal losses longer than 10 min (LOS) in GPS data to mark last fix and first fix. PALMS marks a trip start if it detects movement (speed >0.57 m/sec) and marks a trip end if the device is stationary for 3 min. It marks a stop if first fixes, last fixes, trip starts and trip ends are spatially clustered within 30 m for at least 5 min. This means that PALMS imputes the unknown location of missing data to the centre of the spatial cluster if the gap exceeds 10 min and forms part of a stop. Otherwise, missing data is imputed as part of a move; thus imputing location is highly contingent on LOS.

A statistical approach to imputation has also been proposed (Barnett and Onnela, 2018). In this approach, an unknown sub-trajectory over the missing interval is modelled as a sequence of events, in which an event is either a flight (move) or a pause (stop). It determines whether a move occurs based on probability involving similarities between the current data point and observed events and then models the event displacement using density functions for events conditional on the time and location at last fix. The method is designed to account for GPS data that are missing completely at random (MCAR) with short duration (less than 10 min). However, many GPS trajectories, such as being inside a building, are missing not at random (MNAR).

In synthesis, a geographic imputation method for GPS trajectory data can be classified into different types depending on: i) which movement state missing data is imputed as; ii) whether it models random or non-random gaps; and iii) whether or not it uses external (or historical) data. We present a method that considers possible stop-and-move (SAM) characteristics of human trajectories during non-random gaps without using external data. The method is applicable to imputing locations over missing intervals that are reproducible in many real-world settings without stringent data requirements. The proposed method has two-step processes that can determine the likely movement state and impute location accordingly.

A proposed method for geographic imputation

The purpose of geographic imputation is to estimate 'unknown location at a given time during time gaps', denoted as $z(t_u)$, where $t_i<t_u<t_j$. It is formulated as follows:

$$z(t_u) = z(t_i) + \bar{\upsilon}\ \Delta t \text{ where } \Delta t = t_u - t_i \qquad (1)$$

Table 1. Treatment of missing data in GPS trajectories in environmental health studies.

GPS study	Imputed or not	Method of imputation	Imputation details	Summary of findings	Locality
Rodriguez *et al*., 2005	No	-	-	More PA is associated with high population density and street connectivity	US
Wiehe *et al*., 2008	Yes	ST	Set to last fix if gap <30 m; otherwise set to data point close to home	GPS can be used to determine where adolescents spent time away from home	US
Jones *et al*., 2009	No	-	-	PA is associated with gardens and street environment (farm, grassland) for urban (rural) children	UK
Maddison *et al*., 2010	Yes	ST	Set to last fix if gap <100 m	Majority of adolescents' MVPA bouts occurred within 1 km of school (71%) or 150 m of home environment (46%)	New Zealand
Oliver *et al*., 2010	No	-	-	It is feasible to combine accelerometer and GPS to measure transport-related PA	New Zealand
Almanza *et al*., 2012	No	-	-	Positive association between green-space and PA	US
Rainham *et al*., 2012	Yes	Overlay, ST, LI, manual	If last and first fix fall within known locations (*e.g*., home, school), set to those locations; if not, set to last fix if gap <10 sec; or use linear interpolation	Adolescents' commuting contributes a lot to MVPA especially in urban areas where automobiles are not used for commuting; urban adolescents are more physically active than suburban or rural adolescents	Canada
Oreskovic *et al*., 2015	Yes	Overlay, ST	Imputed to school if gap >2 h near school; imputed as last fix if gap ≤2 h within the matched GPS/PA data	Being at school, on the streets/sidewalks, in parks and playgrounds associated with greater odds of MVPA	US
James *et al*., 2017	No	-	-	PA has positive nonlinear relationship with walkability and greenness	US
Van Hecke *et al*., 2018	Yes	PALMS	Set to stop location nearby if gap >10 min and spatial clusters of data points are marked as 'stationary'; otherwise, set to part of trip	Time spent on public open spaces goes up when accompanied, for non-western Europeans	Belgium
Lee and Kwan, 2018	Yes	LI	Set to a linearly interpolated path between last fix and first fix	Random forest and gradient boosting (machine learning) can be used to predict PA type (walking, running, *etc*.) from GPS and PA data reliably	US
Barnett and Onnela, 2018	Yes	Probabilistic	Resampling from empirical distribution of observed	A proposed method nearby dataimproves accuracy of mobility measures	China (GeoLife)
Remmers *et al*., 2019	Yes	PALMS	Set to stop location nearby if gap >10 min and spatial clusters of data points are marked as 'stationary'; otherwise set to part of trip	Active transport contributes to PA; density and small activity space size is positively associated with PA; green space not associated with PA	Netherlands
Chaix *et al*., 2019	Yes	ST, LI, manual	Imputed to known stop locations nearby; manually imputed as the shortest path	Walking trip contributes much to PA	France
Tamura *et al*., 2019	No	-	-	Greenness is associated with PA	US
Allahbakhshi *et al*., 2020	Yes	LI	Set to a linearly interpolated path between last fix and first fix	Adding GPS features (such as speed) can improve classification of activity type(*e.g*., laying, walking, running, cycling, sitting) with random forest	Multiple countries

ST, stop-based interpolation; LI, linear interpolation; MVPA, moderate-to-vigorous physical activity; PALMS, personal activity and location measurement system.

where t_i is the time of last fix; t_j is the time of first fix, t_u is any time between t_i and t_j when unknown location is estimated; and $\bar{\upsilon}$ an unknown velocity during Δt. It states that $z(t_u)$ is determined by how much a person moves from the last known location $z(t_i)$. If a person is stationary during the gap (from t_i to t_j), then $\bar{\upsilon}$ can be set to zero and $z(t_u)$ set to the last fix location $z(t_i)$. This is equivalent to the simple ST method discussed above. If a person is constantly moving during the gap, then $\bar{\upsilon}$ can be set to the mean velocity ($\bar{\upsilon}_{ij}$) following the path between i and j. This is equivalent to the simple LI method discussed above. The presumption of a singular movement state (either stationary or moving) during a gap is not always held. To illustrate, it is unrealistic to have constant motion with very low speed during a gap (Figure 1A). Rather, it is more likely that a person alternates a SAM sequence during the gap (Figure 1B) (Hwang *et al*., 2018). In summary, $z(t_u)$ can be better modelled as a function of likely movement states which can be singular or composite.

States during a gap S_{ij} can vary depending on gap distance Δz_{ij} and gap velocity $\bar{\upsilon}_{ij}$. Following the sensitivity analysis conducted by Hwang *et al*. (2018), 1 m/sec and 100 m are recommended for MV (cut-off in gap velocity) and SR (cut-off in gap distance), respectively. The state is imputed as a stop if both gap distance and velocity are small (*i.e*. less than cut-off values) but imputed as a move if both gap distance and velocity are large (*i.e*. larger than cut-off values). Large gap distance and small gap velocity would be imputed as a SAM sequence since large distance indicates moving and small velocity indicates being stationary at some point. Small gap distance and large gap velocity would instead be imputed as stop or move (SOM), *i.e*., move if a person is making a short trip and stop if high velocity is an artefact of signal noise.

An unknown location is imputed differently depending on S_{ij} (Figure 2). $z(t_u)$ is set to the location of the last known point $z(t_i)$ if S_{ij} were imputed as stop but linearly interpolated between i and j at the velocity of $\bar{\upsilon}_{ij}$ if S_{ij} were imputed as move. If S_{ij} is imputed as SAM, $z(t_u)$ would be presumed to be stationary up until t_a, where $t_i<t_a<t_j$, and then move at an unknown velocity $\bar{\upsilon}$ between t_a and t_j, where t_a (the time of trip start) is determined by $\bar{\upsilon}$ and the distance between i and j. [In the previous work (Hwang *et al*., 2018) both stop-and-move and move-and-stop were considered when S_{ij} is inferred as SAM. Experiments indicate that both location imputation and state inference with move-and-stop perform poorly. Hence, we considered stop-and-move here.]. $\bar{\upsilon}$ depends on whether $S_{>j}$ (state of record following j) is stop or move. The state before or after the missing interval is determined by fuzzy inference over the minimum segment duration (MinSegDur) (Hwang *et al*., 2018). An experiment indicates that state is most accurately inferred when MinSegDur is set to 9 min. If $S_{>j}$ is move, then a person is assumed to move at the same speed as $\bar{\upsilon}_{>j}$ (mean velocity of record following j). If $S_{>j}$ is stop, the average speed on a route taken between i and j is set to $\bar{\upsilon}$. Once S_{ij} is imputed as SOM, $z(t_u)$ should be set to $z(t_i)$ if a gap is surrounded by stops at both ends; otherwise set to $z(t_i) + \bar{\upsilon}_{ij}\ \Delta t$. As the proposed method extends a path interpolation algorithm conditional upon the likely movement state, it will be referred to as conditional path interpolation (CPI) in short in following sections.

Experiments

To assess the performance of CPI and other imputation methods, we collected two-day continuous GPS trajectories from 59 participants during the warm season. Participants' activity spaces were in Chicago and its vicinity. During an orientation and with informed consent, they were provided with GPS loggers (QStarz BT-Q1000XT) and given instructions for using and completing travel surveys. The orientation included a pilot tracking under direct supervision of research staff, to ensure compliance with the protocol. Upon completing an orientation, participants went about their daily lives always carrying GPS loggers except for water activities during the tracking period (*i.e*. the full 48 hours agreed on for GPS tracking and data analysis). All participants agreed to record time and location using a voice recorder when they started and ended any trip (move) lasting at least 45 sec on a real-time basis to minimize memory bias. They were made aware that a move would be delimited by a stop of at least 2 min's length to reduce confusion over what counts as SAMs (Fillekes *et al*., 2019). They also completed travel surveys by retrieving recorded audios when amenable to them and recorded locations objectively at the time of starting and ending a trip using a button in the GPS logger. This was used to verify the accuracy of self-reported location and time in the travel surveys. Two research staff reconciled any inconsistency in travel surveys by visually inspecting raw GPS data in GIS and checking with participants. Any self-reported data that could not be verified with certainty were excluded. Thus, these validated data constitute ground truth. The sampling rate of a GPS logger is 5 sec. Any gaps ranging from 90 sec to 2 days (tracking period) are imputed. Minimum stop duration (MinStopDur) ranges from 2 min to 30 min in studies on stop detection (Hwang *et al*., 2018) and was set to 2 min in this study to increase the number of sampling points. MinSegDur is set to 9 min for fuzzy inference.

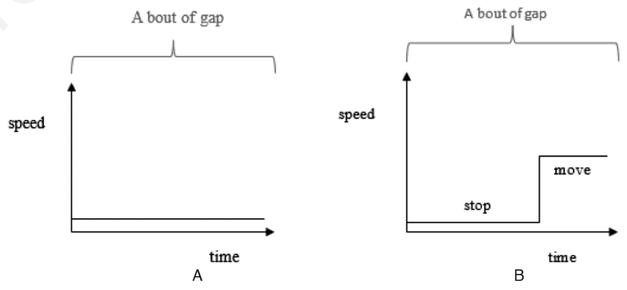
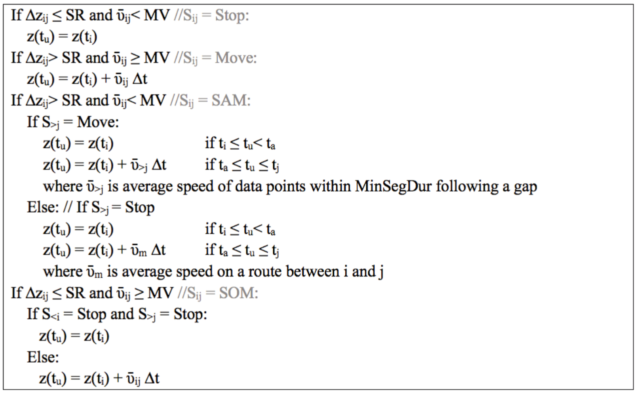


Figure 1. Path interpolation at various modes of motion. A) constant motion; B) state-varying motion.



```
If $\Delta z_{ij} \le SR$ and $\bar{\upsilon}_{ij} < MV$ //$S_{ij}$ = Stop:
  $z(t_u) = z(t_i)$
If $\Delta z_{ij} > SR$ and $\bar{\upsilon}_{ij} \ge MV$ //$S_{ij}$ = Move:
  $z(t_u) = z(t_i) + \bar{\upsilon}_{ij}\ \Delta t$
If $\Delta z_{ij} > SR$ and $\bar{\upsilon}_{ij} < MV$ //$S_{ij}$ = SAM:
  If $S_{>j}$ = Move:
    $z(t_u) = z(t_i)$                              if $t_i \le t_u < t_a$
    $z(t_u) = z(t_i) + \bar{\upsilon}_{>j}\ \Delta t$            if $t_a \le t_u \le t_j$
    where $\bar{\upsilon}_{>j}$ is average speed of data points within MinSegDur following a gap
  Else: // If $S_{>j}$ = Stop
    $z(t_u) = z(t_i)$                              if $t_i \le t_u < t_a$
    $z(t_u) = z(t_i) + \bar{\upsilon}_{m}\ \Delta t$             if $t_a \le t_u \le t_j$
    where $\bar{\upsilon}_{m}$ is average speed on a route between i and j
If $\Delta z_{ij} \le SR$ and $\bar{\upsilon}_{ij} \ge MV$ //$S_{ij}$ = SOM:
  If $S_{<i}$ = Stop and $S_{>j}$ = Stop:
    $z(t_u) = z(t_i)$
  Else:
    $z(t_u) = z(t_i) + \bar{\upsilon}_{ij}\ \Delta t$
```

Figure 2. The proposed algorithm for geographic imputation in GPS trajectories.

Some changes were made from the previous work (Hwang *et al*., 2018) in pre-processing GPS data. In order to examine the performance of imputation independent of outlier detection methods, we did not remove outliers in the current work. We retained GPS track points with 'no fix' for positioning methods (*e.g*., differential GPS, dead reckoning). While location of these data points is inaccurate, their timestamps provide a check to when a person might have gotten in or out of a built structure or tunnel that could have caused signal loss. Original coordinates with 'no fix' were replaced with coordinates of their temporally close data points with a superior positioning method.

After removing redundant (speed from the previous point <1 m/sec) or inaccurate GPS track points (horizontal dilution of position ≥ 3.5 or the number of satellites used <4), we imputed data with three different methods (*i.e*. ST, LI, CPI). GPS data is then segmented into SAM episodes using the trajectory segmentation algorithm (Hwang *et al*., 2018). A stop episode is a sub-sequence of a trajectory where a person exhibits little movement for at least MinStopDur. A move episode is a part of a trajectory that is delimited by two consecutive stops. Episodes are used to compute daily mobility measures for each participant.

To evaluate the performance of imputation methods, we first calculated location deviation (LocDev), the Euclidean distance (in meters) between estimated locations and actual locations at the time of trip start/end during time gaps. The time of trip start/end in a travel survey provides a crucial timestamp when state and location change in a GPS trajectory (*i.e*. with highest uncertainty). The performance of the imputation is thought to be the worst at these timestamps. Then, we considered to what extent the state is correctly inferred timewise at the level of episodes. Third, we accounted for how daily mobility measures, estimated for each participant during the tracking period, differed from reported mobility measures across different imputation methods. That way, we could examine both direct (*e.g*., during time gaps) and indirect (*e.g*., during the tracking period) effects of imputation methods at various levels of observations (*i.e*. that were specific to sampling points, episodes and participants).

Results

Based on 260 sampling points (verifiable locations at trip start and end over missing interval), average LocDev for CPI, LI and ST was 58 (±63 standard deviation), 79 (±339) and 135 (±499) m, respectively. The Wilcoxon signed-rank test confirms the difference between CPI and ST (z-statistics –3.583, P=0.000) but no difference between CPI and LI (z-statistics –0.008, P=0.994). In contrast to CPI and LI, LocDev measures of ST (Figure 3) had more data points exceeding 500 m. For instance, large LocDev of 5949 m by ST was remedied to 149 m when data were imputed by CPI that maps a gap to SAM instead of stop. While gaps are frequently associated with stops, ST can impute locations inaccurately when a gap is deemed to be in composite states such as SAM, with large gap distance and low gap velocity. Hence ST should be used with caution for imputing missing locations. Solely relying on ST for imputation can result in misidentification of locations of places visited.

In addition to evaluating positional accuracy at a point of time, we considered how accurately the state is classified at the level of episodes over a period, *e.g*., from time of arrival to departure. The classification accuracy is measured as the percent of correctly classified state (PCC) in sec for each of labelled episodes during gaps. The mean classification accuracy of CPI, LI and ST by participants was 76.49% (±20.82 SD), 8.52% (±16.48 SD), and 73.11% (±26.27 SD), respectively. The Wilcoxon signed-rank test confirmed that CPI improves the classification accuracy over LI (z-statistics –5.887, P=0.000) and ST (z-statistics –2.084, P=0.037).

To evaluate positional accuracy and classification accuracy jointly, we examined how the classification accuracy changes at varying values of LocDev that is assigned to each episode based on arrival time (because not all locations of trips are known). To this end, we calculated the proportion of computed episodes that were minimally matched (co-occurring with a labelled episode of the same state ≥ 2 min) and maximally matched (2 min $\leq$ co-occurring with a labelled episode of the same state $\leq$ the total duration of the labelled episode - 10 min) during gaps where the LocDev values were less than the threshold, which ranged from 25 to 900 m. Figure 4 shows that the classification accuracy of CPI (red) was consistently greater than that of LI (blue) and ST (green) at different LocDev values.

We also considered how imputation methods affect daily mobility measures calculated for the entire tracking period (beyond missing intervals). For each of the 59 participants, we computed the number of stops, time spent on stops (in sec), the number of trips, and time spent on trips (in sec) from a GPS trajectory that was imputed differently. In order to represent how computed mobility measures agree with ground truth, the mean absolute error (MAE) (*i.e*. the average of the absolute differences between reported vs. estimated mobility measures) was calculated (Table 2). The Wilcoxon signed-rank test confirmed that CPI yields less MAEs than other methods at P=0.01, except for the pair in the number of stops between CPI and LI, as well as the pair in time spent on trips between CPI and ST. LI excessively miscalculates temporal aspects of episodes (*e.g*., time spent on trips), while ST infers the number of episodes less accurately than does CPI.

All in all, CPI outperformed ST and LI in imputing location and state during gaps (Figures 3 and 4) and inferring daily mobility measures during the tracking period (Table 2). ST was more widely applicable to imputing non-random gaps than LI because more non-random gaps were associated with stops than moves. ST is, however, subject to misclassification of exposure during gaps that are better modelled as composite states than a singular state. LI negatively affects the accuracy of trajectory segmentation, resulting in inaccurate daily mobility measures. LI is therefore not recommended for inferring mobility measures especially in filling gaps with long duration given its poor performance. The experimentation results provide a proof of concept for CPI.

Discussion

CPI offers improvement over other methods (ST, LI) in imputing missing data in GPS trajectories and brings together those methods into a unified one. Statistical tests show that ST underperforms other methods in imputing locations and LI distinctively underperforms other methods in inferring states and mobility measures. Low performance of LI, especially in inferring mobility measures, is consistent with the finding of one existing study (Barnett and Onnela, 2018). To improve CPI, the presumption made about composite movement states warrants further refinement and validation. It is shown that the smaller temporal range of an imputable gap, the better imputation performs (Hwang *et al*., 2018). Depending on the margin of error allowed for exposure estimates, one can reduce the range of an imputable gap instead of filling gaps entirely. We did not evaluate positional accuracy of move in fine granularity as much as stop because participants were instructed not to record their locations while moving (*e.g.*, driving) for safety reason.

The accuracy of trajectory imputation is affected by how data quality issues are addressed prior to imputation. We anticipate that there are cascading effects of inadequately treated data quality issues on imputation, segmentation, and exposure measures (Sambasivan *et al*., 2021). The review of cases with large LocDev reveals that low performance is linked to inadequately treating imperfect data points such as outliers especially in urban canyons (*e.g*., downtown Chicago) with frequent GPS signal shortage. Further research on robust pre-processing methods is needed to develop best practices for GPS data processing and enhance spatiotemporal exposure assessment. The current study elucidates role of imputation - as part of data pre-processing - in exposure measures. To make spatiotemporal exposure assessment clinically useful, it is important to consider ways in which a person is exposed to the environment. For instance, more exposure is expected at a place where more time is spent walking than driving. Modality of exposure cannot be precisely captured at the level of raw GPS trajectories that are discretely and sparsely sampled. Instead, raw data can be turned to SAM episodes as a high-level representation of raw data. Modality of exposure can be logically inferred from these episodes with relevant attributes such as dwell time/location or mode of transportation.

A growing number of research strives to examine how health outcomes are linked to geographic contexts using GPS data as location-aware mobile devices have become commonplace. The full potential of GPS technology for improving our understanding this topic is, however, not realized yet as there is no good consensus on how to process uncertain GPS trajectory data. Imputation serves as a preliminary process for turning ‘discretely’ sampled raw GPS track points into a ‘continuous’ sequence of SAM episodes. We acknowledge that it is challenging to measure the magnitude, frequency, and duration of exposure to different environments in relation to health outcomes. Still, we can better infer how persons interact with the environment if GPS trajectories were seamlessly imputed and accurately segmented into episodes that have a bearing on the modality of exposure.

CPI can be used to fill data gaps and reduce bias in activity space measures and daily (community) mobility measures in environmental health research. As CPI does not require additional data like other methods (such as popular path or the shortest path method), CPI can be applied to processing near real-time GPS data as well as historical GPS data. Furthermore, CPI can be used for up-sampling (*i.e.* increasing the sampling rate) of low-resolution trajectory data (*e.g*., from smartphones) and compressed trajectory data (*e.g*., in the ‘cloud’). Unlike a statistical approach that accounts for random gaps with a short duration, CPI can be applied to filling non-random gaps with a long duration.

Conclusions

The proposed CPI method for geographic imputation considers both SAM characteristics of a personal trajectory during the missing interval. CPI outperforms existing imputation methods (ST, LI) in estimating location and movement states and daily mobility measures. Furthermore, ST can lead to misclassification of exposure location when gaps are associated with composite states (such as SAM), while temporal aspects of episodes (*e.g*., time spent on activity locations) will be grossly miscalculated if LI is used. Given present experimentation results, CPI is expected to reduce bias in exposure measures.

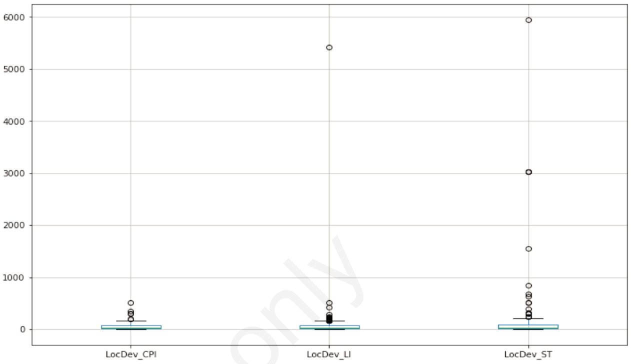


Figure 3. Boxplot of location deviation in meters by imputation methods. CPI, conditional path interpolation; LI, linear interpolation; ST, stop-based interpolation.

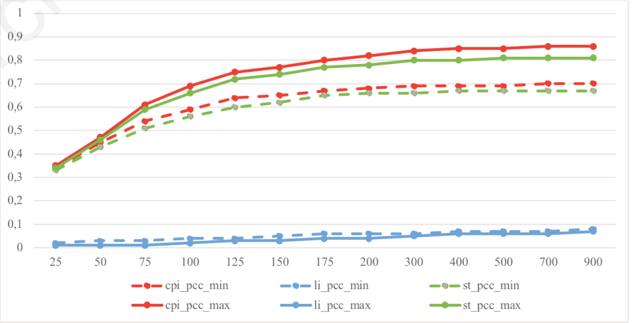


Figure 4. Classification accuracy over varying values of location deviation (LocDev). cpi_pcc_min = the proportion of minimally matched episodes by conditional path interpolation (CPI); cpi_pcc_max = the proportion of maximally matched episode by CPI; li_pcc_min = the proportion of minimally matched episodes by linear interpolation (LI); li_pcc_max = the proportion of maximally matched episode by LI; st_pcc_min = the proportion of minimally matched episodes by stop-based interpolation (ST); st_pcc_max = the proportion of maximally matched episode by ST.

Table 2. Mean absolute error of daily mobility measures by imputation methods.

	CPI	LI	ST
The number of stops	4.66	4.61	5.46
Time spent on stops (reported in hours)	5.79	20.50	5.76
The number of trips	3.49	6.05	3.85
Time spent on trips (reported in hours)	0.32	23.18	0.37

CPI, conditional path interpolation; LI, linear interpolation; ST, stop-based interpolation.

The main contribution of this study is to demonstrate how the different treatment of missing data in GPS trajectories affects the accuracy of spatiotemporal exposure assessment. Given findings of the current study, researchers should exercise caution in using ST for estimating missing locations and using LI for inferring mobility measures when working with sparsely sampled GPS trajectories. Effective imputation is the first step towards increasing the usability of GPS trajectory data for improved exposure assessment and beyond. As spatiotemporal data (including GPS trajectories) increase in volume, availability and role, more attention to the uncertainty of spatiotemporal data is warranted.

References

Allahbakhshi H, Conrow L, Naimi B, Weibel R, 2020. Using accelerometer and GPS data for real-life physical activity type detection. Sensors 20:588.

Almanza E, Jerrett M, Dunton G, Seto E, Pentz MA, 2012. A study of community design, greenness, and physical activity in children using satellite, GPS and accelerometer data. Health Place 18:46-54.

Barnett I, Onnela J-P, 2018. Inferring mobility measures from GPS traces with missing data. Biostat Oxf Engl 21:e98-e112.

Carlson JA, Jankowska MM, Meseck K, Godbole S, Natarajan L, Raab F, Demchak B, Patrick K, Kerr J, 2015. Validity of PALMS GPS scoring of active and passive travel compared to SenseCam. Med Sci Sports Exerc 47:662-7.

Chaix B, Benmarhnia T, Kestens Y, Brondeel R, Perchoux C, Gerber P, Duncan DT, 2019. Combining sensor tracking with a GPS-based mobility survey to better measure physical activity in trips: public transport generates walking. Int J Behav Nutr Phys Act 16:84.

Chaix B, Kestens Y, Perchoux C, Karusisi N, Merlo J, Labadi K, 2012. An interactive mapping tool to assess individual mobility patterns in neighborhood studies. Am J Prev Med 43:440-50.

Chaix B, Méline J, Duncan S, Merrien C, Karusisi N, Perchoux C, Lewin A, Labadi K, Kestens Y, 2013. GPS tracking in neighborhood and health studies: a step forward for environmental exposure assessment, a step backward for causal inference? Health Place 21:46-51.

Christensen A, Griffiths C, Hobbs M, Gorse C, Radley D, 2021. Accuracy of buffers and self-drawn neighbourhoods in representing adolescent GPS measured activity spaces: an exploratory study. Health Place 69:102569.

Evans CC, Hanke TA, Zielke D, Keller S, Ruroede K, 2012. Monitoring community mobility with global positioning system technology after a stroke: a case study. J Neurol Phys Ther 36.

Fillekes MP, Kim E-K, Trumpf R, Zijlstra W, Giannouli E, Weibel R, 2019. Assessing older adults' daily mobility: a comparison of GPS-derived and self-reported mobility indicators. Sensors 19:4551.

Hanke TA, Hwang S, Keller S, Zielke D, Hailey T, Nathaniel K, Evans CC, 2019. Measuring community mobility in survivors of stroke using global positioning system technology: a prospective observational study. J Neurol Phys Ther 43:175-85.

Hirsch JA, Winters M, Clarke P, McKay H, 2014. Generating GPS activity spaces that shed light upon the mobility habits of older adults: a descriptive analysis. Int J Health Geogr 13:51.

Hordacre B, Barr C, Crotty M, 2014. Use of an activity monitor and GPS device to assess community activity and participation in transtibial amputees. Sensors 14:5845-59.

Hornsby K, Egenhofer MJ, 2002. Modeling moving objects over multiple granularities. Ann Math Artif Intell 36:177-94.

Hwang S, VanDeMark C, Dhatt N, Yalla SV, Crews RT, 2018. Segmenting human trajectory data by movement states while addressing signal loss and signal noise. Int J Geogr Inf Sci 32:1391-412.

James P, Hart JE, Hipp JA, Mitchell JA, Kerr J, Hurvitz PM, Glanz K, Laden F, 2017. GPS-based exposure to greenness and walkability and accelerometry-based physical activity. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol 26:525-32.

Jankowska MM, Schipperijn J, Kerr J, 2015. A framework for using GPS data in physical activity and sedentary behavior studies. Exerc Sport Sci Rev 43:48-56.

Jansen M, Kamphuis CBM, Pierik FH, Ettema DF, Dijst MJ, 2018. Neighborhood-based PA and its environmental correlates: a GIS- and GPS based cross-sectional study in the Netherlands. BMC Public Health 18:233.

Jayaraman A, Deeny S, Eisenberg Y, Mathur G, Kuiken T, 2014. Global position sensing and step activity as outcome measures of community mobility and social interaction for an individual with a transfemoral amputation due to dysvascular disease. Phys Ther 94:401-10.

Jones AP, Coombes EG, Griffin SJ, Sluijs EM van, 2009. Environmental supportiveness for physical activity in English schoolchildren: a study using global positioning systems. Int J Behav Nutr Phys Act 6:42.

Kerr J, Duncan S, Schipperjin J, 2011. Using global positioning systems in health research: a practical approach to data collection and processing. Am J Prev Med 41:532-40.

Kerr J, Marshall S, Godbole S, Neukam S, Crist K, Wasilenko K, Golshan S, Buchner D, 2012. The relationship between outdoor activity and health in older adults using GPS. Int J Environ Res Public Health 9:4615-25.

Klinker CD, Schipperijn J, Christian H, Kerr J, Ersbøll AK, Troelsen J, 2014. Using accelerometers and global positioning system devices to assess gender and age differences in children's school, transport, leisure and home based physical activity. Int J Behav Nutr Phys Act 11:8.

Krenn PJ, Titze S, Oja P, Jones A, Ogilvie D, 2011. Use of global positioning systems to study physical activity and the environment: a systematic review. Am J Prev Med 41:508-15.

Lee K, Kwan M-P, 2018. Physical activity classification in free-living conditions using smartphone accelerometer data and exploration of predicted results. Comput Environ Urban Syst 67:124-31.

Loh M, Sarigiannis D, Gotti A, Karakitsios S, Pronk A, Kuijpers E, Annesi-Maesano I, Baiz N, Madureira J, Oliveira Fernandes E, Jerrett M, Cherrie JW, 2017. How sensors might help define the external exposome. Int J Environ Res Public Health 14:434.

Maddison R, Jiang Y, Vander Hoorn S, Exeter D, Mhurchu CN, Dorey E, 2010. Describing patterns of physical activity in adolescents using global positioning systems and accelerometry. Pediatr Exerc Sci 22:392-407.

McCrorie PR, Fenton C, Ellaway A, 2014. Combining GPS, GIS, and accelerometry to explore the physical activity and environment relationship in children and young people - a review. Int J Behav Nutr Phys Act 11:93.

McGrath LJ, Hopkins WG, Hinckson EA, 2015. Associations of objectively measured built-environment attributes with youth moderate-vigorous physical activity: a systematic review and meta-analysis. Sports Med 45:841-65.

Mennis J, Mason M, Coffman DL, Henry K, 2018. Geographic imputation of missing activity space data from ecological momentary assessment (EMA) GPS positions. Int J Environ Res Public Health 15:2740.

Meseck K, Jankowska MM, Schipperijn J, Natarajan L, Godbole S, Carlson J, Takemoto M, Crist K, Kerr J, 2016. Is missing geographic positioning system data in accelerometry studies a problem, and is imputation the solution? Geospat Health 11:403.

Miller HJ, 1991. Modelling accessibility using space-time prism concepts within geographical information systems. Int J Geogr Inf Syst 5:287-301.

Oliver M, Badland H, Mavoa S, Duncan MJ, Duncan S, 2010. Combining GPS, GIS, and accelerometry: methodological issues in the assessment of location and intensity of travel behaviors. J Phys Act Health 7:102-8.

Oreskovic NM, Perrin JM, Robinson AI, Locascio JJ, Blossom J, Chen ML, Winickoff JP, Field AE, Green C, Goodman E, 2015. Adolescents' use of the built environment for physical activity. BMC Public Health 15:251.

Pfoser D, Jensen CS, 1999. Capturing the uncertainty of moving-object representations. In: Güting RH, Papadias D, Lochovsky F (Eds.), Advances in spatial databases, lecture notes in computer science. Springer, Berlin-Heidelberg, pp. 111-131.

Quigg R, Gray A, Reeder AI, Holt A, Waters DL, 2010. Using accelerometers and GPS units to identify the proportion of daily physical activity located in parks with playgrounds in New Zealand children. Prev Med 50:235-40.

Rainham DG, Bates CJ, Blanchard CM, Dummer TJ, Kirk SF, Shearer CL, 2012. Spatial classification of youth physical activity patterns. Am J Prev Med 42:e87-96.

Remmers T, Thijs C, Ettema D, de Vries S, Slingerland M, Kremers S, 2019. Critical hours and important environments: relationships between afterschool physical activity and the physical environment using GPS, GIS and accelerometers in 10-12-year-old children. Int J Environ Res Public Health 16:3116.

Rodriguez DA, Brown AL, Troped PJ, 2005. Portable global positioning units to complement accelerometery-based physical activity monitors. Med Sci Sports Exerc 37:S572-81.

Rundle AG, Sheehan DM, Quinn JW, Bartley K, Eisenhower D, Bader MMD, Lovasi GS, Neckerman KM, 2016. Using GPS data to study neighborhood walkability and physical activity. Am J Prev Med 50:e65-72.

Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh PK, Aroyo LM, 2021. Everyone wants to do the model work, not the data work: data cascades in high-stakes AI. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1-15.

Tamura K, Wilson JS, Goldfeld K, Puett RC, Klenosky DB, Harper WA, Troped PJ, 2019. Accelerometer and GPS data to analyze built environments and physical activity. Res Q Exerc Sport 90:395-402.

Van Hecke L, Verhoeven H, Clarys P, Van Dyck D, Van de Weghe N, Baert T, Deforche B, Van Cauwenberg J, 2018. Factors related with public open space use among adolescents: a study using GPS and accelerometers. Int J Health Geogr 17:3.

Wei L-Y, Zheng Y, Peng,W-C, 2012. Constructing popular routes from uncertain trajectories. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12. ACM, New York, NY, USA, pp. 195-203.

Wheeler BW, Cooper AR, Page AS, Jago R, 2010. Greenspace and children's physical activity: a GPS/GIS analysis of the PEACH project. Prev Med 51:148-152.

Wiehe SE, Hoch SC, Liu GC, Carroll AE, Wilson JS, Fortenberry JD, 2008. Adolescent travel patterns: pilot data indicating distance from home varies by time of day and day of week. J. Adolesc Health 42:418-20.

Yoo E-H, Roberts JE, Eum Y, Shi Y, 2020. Quality of hybrid location data drawn from GPS-enabled mobile phones: Does it matter? Trans GIS 24:462-82.

Zhao P, Jonietz D, Raubal M, 2021. Applying frequent-pattern mining and time geography to impute gaps in smartphone-based human-movement data. Int J Geogr Inf Sci 0:1-29.