Is missing geographic positioning system data in accelerometry studies a problem, and is imputation the solution?

Kristin Meseck,[1] Marta M. Jankowska,[1] Jasper Schipperijn,[2] Loki Natarajan,[1] Suneeta Godbole,[1] Jordan Carlson,[3] Michelle Takemoto,[1] Katie Crist,[1] Jacqueline Kerr[1]

[1]Department of Family Medicine and Public Health, University of California, La Jolla, CA, USA; [2]Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark; [3]Center for Children's Healthy Lifestyles and Nutrition, Children's Mercy Hospital-University of Missouri, Kansas City, MO, USA

Abstract

The main purpose of the present study was to assess the impact of global positioning system (GPS) signal lapse on physical activity analyses, discover any existing associations between missing GPS data and environmental and demographics attributes, and to determine whether imputation is an accurate and viable method for correcting GPS data loss. Accelerometer and GPS data of 782 participants from 8 studies were pooled to represent a range of lifestyles and interactions with the built environment. Periods of GPS signal lapse were identified and extracted. Generalised linear mixed models were run with the number of lapses and the length of lapses as outcomes. The signal lapses were imputed using a simple ruleset, and imputation was validated against person-worn camera imagery. A final generalised linear mixed model was used to identify the difference between the amount of GPS minutes pre- and post-imputation for the activity categories of sedentary, light, and moderate-to-vigorous physical activity. Over 17% of the dataset was comprised of GPS data lapses. No strong associations were found between increasing lapse length and number of lapses and the demographic and built environment variables. A significant difference was found between the pre- and post-imputation minutes for each activity category. No demographic or environmental bias was found for length or number of lapses, but imputation of GPS data may make a significant difference for inclusion of physical activity data that occurred during a lapse. Imputing GPS data lapses is a viable technique for returning spatial context to accelerometer data and improving the completeness of the dataset.

Correspondence: Kristin Meseck, Department of Family Medicine and Public Health, University of California, 9500 Gilman Drive, La Jolla, CA 92093, USA.
Tel: +1.858.246.1824 - Fax: +1.858.534.9404.
E-mail: kmeseck@ucsd.edu

Key words: GPS; GIS; Missing data; Imputation; Accelerometer.

Contributions: JK, SG, KC, data collection; KM, MJ, data analysis; SG, LN, statistical design; KM, manuscript writing; MJ, JS, MT, KC, JK, JC manuscript reviewing and revisions.

Funding: this research was funded by the National Institutes of Health, and National Cancer Institute research grant R21 CA169535 entitled *Development and Validation of Novel Prospective GPS/GIS Based Exposure Measures.*

Ethical statement: written informed consent was obtained from each participant and all institutional and governmental regulations concerning the ethical use of human volunteers were followed during all studies. All study procedures were approved by the research ethics board of the University of California, San Diego, CA, USA.

Acknowledgements: this research was supported/partially supported by the National Institutes of Health, National Cancer Institute research grant R21 CA169535 titled: Development and Validation of Novel Prospective GPS/GIS Based Exposure Measures Jacqueline Kerr, principle investigator. The results of the present study do not constitute any form of endorsement by American College of Sports Medicine.

Received for publication: 30 July 2015.
Revision received: 19 September 2015.
Accepted for publication: 25 October 2015.

©Copyright K. Meseck et al., 2016
Licensee PAGEPress, Italy

Geospatial Health 2016; 11:403
doi:10.4081/gh.2016.403

This article is distributed under the terms of the Creative Commons Attribution Noncommercial License (CC BY-NC 4.0) which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Introduction

The use of geographic positioning system (GPS) devices in physical activity (PA) and sedentary behaviour (SB) research has been steadily increasing because of its ability to determine where participants interact with built environment (Kerr *et al.,* 2011; Krenn *et al.,* 2011). These data sources have the potential to have profound impacts on public health by establishing more specific and accurate measures of environmental influences on SB and PA behaviour, and also providing better evidence for public policy change (Jankowska *et al.,* 2015). GPS can show when and how long participants are indoors or outdoors (Quigg *et al.,* 2010; Lam *et al.,* 2013), locate what routes they take for transport (Duncan and Mummery 2007; Duncan *et al.,* 2009) and identify PA and SB behaviour, such as walking, bicycling or driving in specific environments (Troped *et al.,* 2010; Oliver and Badland, 2010). However, missing GPS data due to signal lapse is a problem that may introduce significant bias into modelled relationships between environment and PA or SB. Currently no studies have assessed the bias of missing GPS data in PA and SB studies, and how that bias may influence study outcomes.

Signal lapse is inherent in GPS data, which is collected through a connection between the GPS device worn by study participants and multiple satellites in the sky to establish the geographic location of

Non commercial use only

each individual. We define a signal lapse as the interruption of continuous GPS data collection, resulting in no collection for a variable amount of time, followed by reconnection to satellites and the recommencement of GPS data collection. The reasons for lapse of the signal include physical objects such as buildings, and natural objects such as cloud coverage or dense tree canopies (Costa, 2011). Examples of signal lapse are displayed in Figure 1 [georeferenced background satellite imagery photo provided by SANDAG (2014)]. For population-level PA and SB studies, a challenge for collection of GPS data is posed by free-living study participants, who move in and out of buildings, pass through dense urban areas and engage in a variety of activities that may involve environments that block GPS signals. Signal lapse is a significant concern for PA and SB studies, as GPS data is often time-matched to accelerometer data. GPS signal lapse can result in unsupported assumptions or misclassification (*e.g.* any period of signal lapse can be due to the participant's location being indoors) and data elimination (*e.g.* the removal of a participant from analyses when too many data recordings are found to be missing), both of which may produce biased results.

Previous studies have managed signal lapses in GPS data in one of three ways. One method involves leaving the data as they were originally collected, with no alteration of the GPS data (*e.g.* Wheeler *et al.,* 2010). While this method still allows other time-matched data collected during the lapse to be kept and utilised, such as accelerometer data, such data lose spatial context. In spite of this limitation, this approach may be appropriate for studies focusing on locations of outdoor PA (*e.g.* Lachowycz *et al.,* 2012). Another technique for dealing with missing GPS data is the complete removal of study participants or days of wear time when they do not meet minimum data criteria (*e.g.* Oliver and Badland, 2010). The third technique to manage missing GPS data is imputation (Ogle and Guensler, 2002; Stopher *et al.,* 2008; Wiehe *et al.,* 2008; Troped *et al.,* 2010). Such methodologies may differ, but almost all GPS data imputation methods are based on spatial or temporal parameters of the previous points. Literature reviews on the use of GPS in PA studies have been conducted (Maddison and Ni Mhurchu, 2009; Krenn *et al.,* 2011), and found that most imputation methods utilise arbitrary decisions of time and distance to impute missing points without validation of imputation assumptions.

The goals of this study are twofold: first to assess the demographic and environmental bias of missing GPS data in PA and SB studies. Are certain people or populations more prone to having missing GPS data, and does movement in specific environmental contexts effect data quality? This is an important question for a better understanding what populations GPS and PA/SB studies are better suited for, and if certain environmental contexts are not viable for assessing environment and PA/SB relationships. Our second goal is to develop and validate a GPS imputation method as a solution to missing GPS data and test if there are significant changes of PA and SB time pre- and post-imputation in home and non-home environments.

Materials and Methods

Study sample

Data from eight studies using GPS and accelerometer devices were pooled, representing a range of participants, lifestyles and interactions with built environments. The studies were conducted in San Diego County, CA, USA between the years of 2010 and 2013, and included both baseline interventions and observational studies. Demographic data with reference to age, sex, ethnicity and employment status were collected for each participant. The total sample included 782 participants of an average age of 46 years (min 18, max 102), 12.3% Hispanic and 68.6% women.

Data collection

All data were collected using the same standardised procedures. Participants wore Qstarz (http://www.qstarz.com) GPS devices (BT-Q1000XT) and Actigraph (http://www.actigraphcorp.com) accelerometers (GT3X+). The GPS data were collected every 15 seconds, the accelerometer data 30 times a second. The GPS data were processed and joined to the accelerometer data using the Personal Activity and Location Measurement System (PALMS) (Demchak *et al.,* 2012; Carlson *et al.,* 2015). Data were aggregated and then merged at the minute level. Accelerometer data were classified into sedentary behaviour (counts per minute below 100), light activity (counts per minute between 100 and 1040) and everything above light activity (counts per minute above 1040). This relatively low upper cut-point for activity was chosen because the sample included many older (>65) adults. As this analysis is concerned with the utility of GPS data in PA and SB studies, days were only included in the analysis if they met the criteria for a valid accelerometer wear day defined as a day containing 600 or more accelerometer minutes. Non-wear was defined as 90 or more sequential minutes with an activity count of zero, allowing for 2 minute periods of activity (Heil *et al.,* 2012). In total, participants wore devices for 1-13 days (mean=5.6, standard deviation=1.89).

Data for land use in San Diego County were downloaded in 2014 from San Diego Geographic Information Source - JPA/San Diego

Non-commercial use only

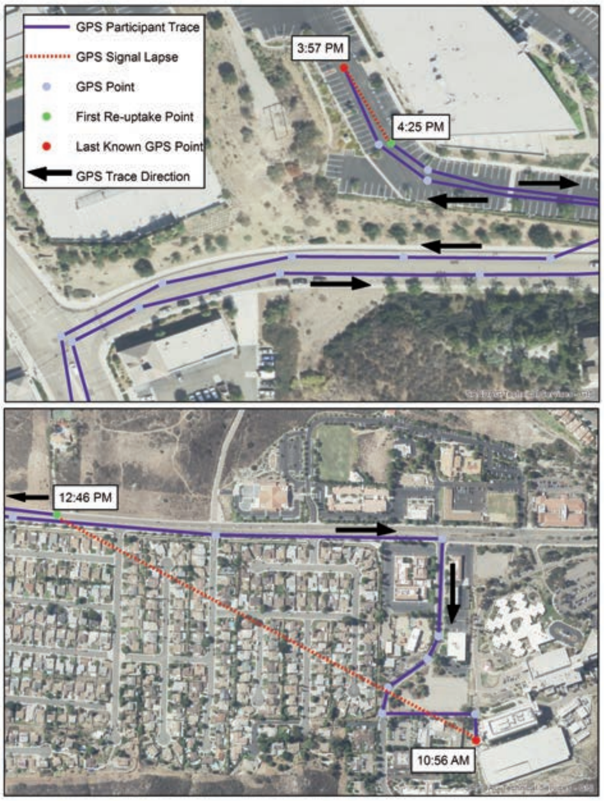

Figure 1. Examples of geographic positioning system signal lapse.

Association of Governments (SANDAG, 2014). ESRI ArcGIS, v. 10.2 software (ESRI, Redlands, CA, USA) was used to assess how areas used for residence, transportation, shopping, parks and recreation, health care, office/school, industry, hotel/resort and leisure and others, were related to signal lapse. Additionally, census data on population counts were obtained from American Community Factfinder (http://factfinder.census.gov) at the census block group level from which population density was derived. Data were matched on the minute level, where each minute or single GPS point was spatially matched with the intersecting land use and population density.

Assessing the demographic and environmental bias of missing geographic positioning system data

Missing GPS data were assessed on participant- and day-levels. To compute descriptive statistics for the periods of missing data, minutes of missing data were extracted and collapsed into lapses (defined as a period of missing GPS signal ≥1 minute), yielding multiple lapses per day, per participant. It is important to note that lapses may be caused by either environmental or technological (*e.g.* lack of battery) issues, and that these cannot be differentiated with the available data. Minutes of sedentary, light, and above light activity were assessed for each lapse, and descriptive statistics were employed to assess the amount of activity data without GPS signal.

Demographic bias was explored to identify if particular population sub-groups engaged in activities within their environments that would interfere with GPS signal more than other groups. A mixed linear effects model (days nested within person) was used to explore the relationship between the number of daily GPS lapses and the individual characteristics of participants. We tested if the number of daily GPS lapses was associated with sex, age, Hispanic ethnicity, employment status and number of daily sedentary minutes while controlling for daily GPS minutes collected. To better understand the environmental biases of GPS signal lapse, a mixed linear effect model (lapse, nested in day, nested in person) was used to test if the length of GPS lapses were associated to environmental characteristics of the lapse (*i.e.*, the land use category of the last known point before the lapse, population density of the last known point before the lapse). Due to non-normality of lapse length (many short lapses), lapse length was square root transformed and placed into the model as a continuous variable. We report the non-transformed coefficients with the transformed significance as transformed coefficients that can be difficult to interpret. This model was controlled for individual level factors found to be significant in the demographic bias model. Land use was placed into the model as a categorical variable to compare the effects of land use categories on signal lapse length. Residential was chosen as a reference category as it had the highest number of signal lapses.

Imputation algorithm, validation, and comparison

The imputation algorithm was created in the R environment (R Core Team, 2013) using functions found in the plyr (Wickham, 2011) and gmt (Magnusson, 2014) packages (imputation algorithm available upon request). Periods of missing GPS data were imputed where any lapse had at least one or more valid GPS points preceding it (to allow for points to impute from) and following it (to ensure GPS battery loss or other device malfunctioning was not imputed). The algorithm locates periods of GPS signal lapse, takes the mean centre point of the 20 GPS points that occurred before the lapse (or number of available points if less than 20) and assigned the resulting mean centre latitude and longitude to the minutes comprising the missing lapse period. All data were imputed with no limit with regard to time or distance of lapse. This decision was based on the assumption that an individual was most likely stationary during a signal lapse, and once they started moving out of the building or location, the signal would begin again. The algorithm was designed on the lapse level rather than the daily level. If a lapse would begin at 11:45 PM of one night and end at 3:05 AM the next day, then the missing data would be imputed from one day to the next. The decision to impute over several days was made because our sample consisted of many participants beyond retirement age, who might be spending all day at home for several days. The algorithm does provide distance between last known GPS point and first GPS point of data re-uptake to allow for removal of very large distances if desired.

The results of the imputation algorithm were validated against a dataset of 40 participants who wore a person-worn camera, SenseCam (Vicon Revue v1.0), in addition to GPS devices and accelerometers. SenseCam data have been employed to validate travel episodes, indoor/outdoor time, eating, and physical activity behaviours in previous studies (Ellis *et al.,* 2013; Doherty *et al.,* 2013; Kerr *et al.,* 2013). SenseCam photos were taken at least every 20 seconds. The photos were coded for a person's context (indoor, outdoor, and in-vehicle), and behaviour (walking/running, biking, sitting, standing still, and standing moving within a confined space) (Kerr *et al.,* 2013). Photo classifications were aggregated to the minute level and joined by timestamp to GPS signal lapses. If timestamps did not match, a final category of unmatched data was created. In order to assess if the imputation assumption of stationarity (the individual had not moved during the imputed lapse) were true, lapses were assessed for the amount of time spent in moving and non-moving behaviours, such as in vehicle, walking, and biking, and standing still/sedentary behaviours. These behaviours were further classified as occurring indoors or outdoors.

To better understand the utility of imputation in the context of a PA or SB study, and to explore if the method might have statistical implications for PA and SB studies, total minutes per day of non-wear, sedentary, light and above-light activity within 800 m of the home and outside of 800 m of the home were compared pre- and post-imputation using mixed linear effects models to account for days within participants.

Table 1. Mixed linear effects model results: outcome lapse length.

Variable	Coefficient	Standard error
Intercept	69.00*	127.93
Hotel/resort/leisure	-22.11	23.74
Industry	-37.22	18.50
Transportation	-4.44	7.58
Shopping	-15.17	9.26
Office/school	-27.72	9.12
Health care	-42.31	18.83
Parks and recreation	-21.01	12.65
Other land use	-18.15	18.61
Population density	-0.01**	0.0008
Female	114.44	58.60
Age	0.05	1.60
Hispanic	-26.93	93.54
Employed	-111.37	77.75
Sedentary minutes	2.03**	0.03

*P<.05; ** P<.001.

Results

Missing geographic positioning system data

The dataset was comprised of 2,007,924 missing GPS points out of 6,171,693 total GPS points, or 32.53%. About half of these missing data (17.39%) were from GPS signal lapses, where a lapse is defined by the GPS signal being re-obtained subsequent to being lost. The other half of these missing data (15.14%) occurred during the beginning of wear time without a GPS signal, or the signal was lost and never re-obtained (*i.e*., at the heads and tails of each participant's wear). After identifying and isolating GPS lapses from valid accelerometer wear days, 64.1% or 516 participants across 2033 wear days were identified as having at least one signal lapse with 15,539 total lapses identified across the entire dataset. Lapses were an average of 51.85 minutes in length (±329.60, min 1, max 9650). Participants had an average of 30 (±28.80, min 1, max 157) signal lapses over their entire wear time, with average 7.6 lapses per day (±6.40, min 1, max 39). The total missing GPS lapse time averaged 396.30 missing minutes daily (±882.41, min 1, max 1440).

Demographic and environmental bias of missing geographic positioning system data

The results of the mixed linear effects model for the number of lapses found the factor of age to be a significant factor [coeff(standard error, SE):- 0.0384 (0.012), P<0.05]. No other associations were significant; however, being female and employed had positive associations with increasing the number of lapses. Results for the mixed linear effects model for lapse length are displayed in Table 1, where *Residential* is the reference category for land use. Although this model was performed using square-root transformed values of lapse length, Table 1 reports the non-transformed coefficients with the transformed significance. None of the land use categories had any significantly different associations for increasing the length of lapses when compared to the Residential category. The Hotel/Resort/Leisure category was negatively associated with increasing lapse length [coeff(SE):-22.11 (23.74)] as were the Industry category [coeff(SE): -37.22 (18.50)], Transportation category [coeff(SE): -4.44 (7.58)], Shopping category [coeff(SE): -15.17 (9.26)], Office/School category [coeff(SE): -27.72 (9.12)], Health Care category [coeff(SE): -42.31 (18.83)], Parks and Recreation category [coeff(SE): -21.01 (12.65)], and Other Land Use category [coeff(SE): -18.15 (18.61)]. All land use categories had small associations with lapse length, with Healthcare and Industry having the largest difference, and Transportation the smallest difference from the Residential category. Population density was found to have a significantly negative association with increasing lapse length [coeff(SE): -0.01 (0.0008), P<.001]; however, the coefficient and standard error were very small. The female sex was positively associated with increasing lapse length, with a large coefficient [coeff(SE): 114.44 (58.60)], while age also had positive associations with a much smaller coefficient [coeff(SE): 0.05 (1.60)]. Hispanic ethnicity was found to have a negative association with lapse length [coeff(SE): -26.93 (93.54)]. Finally, sedentary time was found to be significantly associated with increasing signal lapse length [coeff(SE):2.03 (0.03), P<0.001], with each minute increase in sedentary time associated with a 2.03 minute increase in lapse length.

Imputation of the full sample

The algorithm imputed 100% of lapse data, or 17.39% of the entire dataset, resulting in 934,890 (15.15%) minutes of missing data remaining. On average, 396.30 (±882.41, min 1, max 10,410) minutes of data were imputed per person, per day with an average of 1561.47 (±223,304, min 1, max 11,612) points imputed over a participant's entire wear. Analysis of the accelerometer data found there to be an average of 822.81 (±198.29) of accelerometer minutes per day. Using only matched un-imputed GPS data and accelerometer data, the average minutes per day of data lowered to 560.53±343.76 minute a day. After imputation of GPS signal lapses this daily average increased to 717.71±327.91 minutes. The cut-off for the lowest quartile was 110.5 minutes of imputed data (low imputation), 570 minutes in the highest quartile (high imputation), and 570 in the middle two quartiles (medium imputation). Examples of individuals' days that fell into these quartiles are displayed in Figure 2. The figure demonstrates the variability of both when and where GPS lapses occurred throughout an individual's day.

Imputation validation

The validation of the imputation algorithm using a subset of the 40 participants and SenseCam photography found that 91.5% of the imputed minutes were classified as occurring indoors and as non-moving. Less than 1% of the data was classified as in-vehicle and less than 4% was classified as moving (Table 2). Due to conflicting time-stamps, 3.49% of the SenseCam photo data was unable to be matched with the GPS data and is therefore unclassified.

Table 3 summarises the total average daily means of activity before and after imputation, and broken down for within 800 m of the home and beyond 800 m of the home for each participant. Using a mixed linear effects model to compare pre- and post-imputation, all physical activity categories gained a significant amount of minutes after imputation was conducted. Sedentary behaviours within the home environment gained the largest amount of time [β (CI):80.37 (74.04-80.37), P<0.001] at 95% confidence. Both categories of above-light gained almost three minutes after imputation whereas light in-home gained 24.02 minutes and light outdoor gained almost 12 minutes. For all environments, 115.18 minutes of sedentary behaviour, 35.92 minutes of light activity, and 5.55 minutes of above light activity were gained after imputation.

Non-commercial use only

Discussion

The identification of periods of missing GPS data within the dataset demonstrated a high amount of GPS signal lapses on both the daily and participant levels. Although our sample was not necessarily representative of the population of San Diego county, participants did interact with a variety of different environments ranging from urban/suburban to semi-rural while engaging in free-living behaviours. Of valid accelerometer wear days, 17.39% of GPS data were missing, which means the environmental context was unknown for these usable PA and SB data. Although some participants only experienced one signal

Table 2. SenseCam photos analysed for minutes of imputed time (n=12,077).

Matched data	Moving	Non-moving
Indoors	327 (2.71%)	11,043 (91.5%)
Outdoors	152 (1.26%)	22 (0.19%)
In-vehicle	–	111 (0.92%)
Unmatched data	422 (3.49%)	

lapse, most experienced several over their entire wear with an average of 30 lapses per participant (7.6 per day). This is likely due to habitual interaction with certain environments and/or the reoccurrence of certain behaviours (*e.g*., working in offices without windows) accumulating over a participant's full device-wearing period. The variance in the length of lapses, which averaged at 51.85 minutes with a standard deviation of 329.55 minutes and a median of 2, would indicate that a variety of activities and environments may contribute to signal lapses.

Testing for demographic and environment bias of the number of lapses as well as length of lapses found little bias for the types of individuals who may engage in activities that make them more prone to GPS signal loss. This is an encouraging finding for other studies using GPS devices in a variety of adult populations. Age had a significant negative association with the number of lapses indicating that older individuals experience fewer signal lapses – a surprising finding given that many elderly individuals spend more time indoors, especially in this sample, which included elder-care communities. This may be due to less movement in and out of buildings, leading to fewer chances for dropped signals. However, age was not associated with the length of the average lapse. We also found that increased sedentary behaviour was associated with longer lapse lengths indicating that movement is an important aspect of obtaining GPS signal, and that prolonged sitting episodes will likely result in higher amounts of missing GPS signal. Participants whose behaviours do not include long bouts of sitting and do include greater movement throughout their wear day are more likely to gain and then retain a GPS signal by giving more chances for the GPS satellites to lock onto the GPS device. This has implications for those working with populations known to be more sedentary, where GPS might not be the most appropriate tool or post-collection data loss mitigation will likely be needed.

Surprisingly, land use categories were not significantly associated with the signal lapse length, where we expected to find that office and industrial land use might be locations more prone to longer signal lapses and parks or outdoor venues might be less prone. Population density was the only environmental characteristic significantly associated with the length of signal lapses. However, it was found to be inversely related with lapse length, demonstrating that lower population density is associated with longer lapses. While this result seems counterintuitive, it is likely a reflection of where the lapses are occurring, which are in residential areas with comparatively lower population density than in urban centres. A large portion of our participants were past retirement age and living in elder-care communities, and it is likely that they spent most of their time in their homes or on the campus of their care facility. Since we did find a strong association between sedentary minutes and increasing lapse length, we hypothesise that most participants are sedentary within their homes, and many lapses are occurring there. Additionally, the very small coefficient indicates that this effect is not strong.

While we did not find large demographic and environmental biases in the missing GPS data, we did find that a significant amount of PA

Non commercial use only

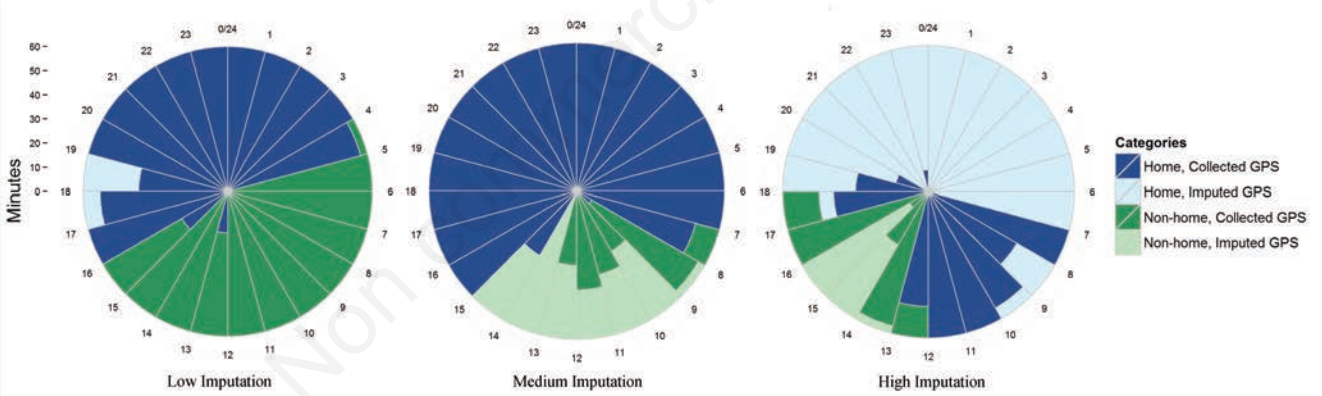

Figure 2. Total day (24 hours) geographic positioning system signal and imputation in the home (with 800 m of the home) and outside of the home (beyond 800 m) for three example participants of low-, medium-, and high-imputation days.

Table 3. Daily averages and standard deviation of minutes for valid wear days with significance of imputation significance.

		Sedentary°	Light[#]	Above-light[§]	Non-wear time
Total	Pre	361.89 (237.17)	159.10 (114.60)	38.19 (48.61)	442.81 (289.20)
	Post	477.09 (238.32)	195.02 (114.25)	43.74 (51.22)	503.02 (265.28)
	β(CI:95)	115.18 (108.55-176.79)***	35.92 (32.83-39.02)***	5.55 (4.35-6.75)***	
Within home buffer	Pre	246.80 (217.30)	104.27 (93.26)	21.26 (28.54)	412.10 (300.50)
	Post	327.12 (244.58)	128.30 (97.64)	24.12 (29.40)	467.60 (285.33)
	β(CI:95)	80.37 (74.04-80.37)***	24.02 (21.42-26.63)***	2.86 (2.07-2.86)***	
Outside of home buffer	Pre	115.10 (156.5)	54.83 (76.51)	16.93 (34.58)	30.71 (127.55)
	Post	149.96 (189.58)	66.73 (88.13)	19.62 (38.34)	35.42 (135.51)
	β(CI:95)	34.83 (30.22-39.44)***	11.90 (9.72-14.08)***	2.69 (1.76-3.61)***	

°0-99 counts per min; [#]100-1040 counts per min; [§]more than 1041 counts per min. ***P<.001.

and SB occur when the GPS signal is lost. Table 2 illustrates that missing GPS data is tied to a large proportion of light PA as well as some higher intensity PA. Before imputation, an average of 36 minutes a day of light activity and approximately 5.5 minute of above-light activity occurred during signal lapses. Added up to a weekly level that is an average of over 4 hours of light and over half an hour of above-light activity; this represents a substantial amount of PA data that lacks environmental context.

In order to regain the spatial context of this accelerometer data, we developed an imputation method based on the assumption that individuals would most likely not be moving during a GPS lapse (movement would greatly increase the chance that a satellite can pick up the signal). The results of the imputation validation using the SenseCam data subset suggest that this imputation method is accurate and a viable option for managing missing GPS data due to signal lapse. We found that 91.5% of signal lapse minutes were classified as Indoors/Non-moving. It is important to note that San Diego does not have underground transportation, nor a highly dense multi-story built environment. Imputation in cities with underground transportation would be possible by accounting for the metro grid, including entry and exit points in the imputation model (essentially imputing the metro travel path). Imputation for environments with dense high-rise structures may be a more difficult task. This diversity in urban environments does pose a challenge for the adoption of one unified imputation technique across different cities. The imputation procedure utilised in this research was an example of imputing using a general ruleset and mean centre statistic; however, several other forms of imputation have been utilised previously (Stopher *et al.,* 2008; Wiehe *et al.,* 2008; Troped *et al.,* 2010). For this example of imputation we chose not to restrict the algorithm in any way and to demonstrate what imputing the maximum amount of data would look like. Researchers can add their own limitations depending on their population and dataset that could include excluding imputation over a certain distance, or preventing imputation over multiple days. As advances continue in machine-learned algorithms to detect behaviour from accelerometry data (Ellis *et al.,* 2013), imputation could also depend on detected behaviour (sitting/standing) from the accelerometer. For any form of imputation to be fully adopted by GPS and health data analysis practitioners, decisions and justification for exact parameters and procedures used need to be reported so that the process can move towards standardisation.

Figure 2 illustrates the variety of locations and times throughout the day that GPS signal loss occurs. The figure particularly highlights that for individuals living in homes with poor signal reception, a large amount of data may be missing in the home environment. Table 3 results support this observation, and demonstrate that a large number of missing data occurs within 800 m of the home. Results from Table 2 also highlight the importance of imputation. Every category of activity, both in and outside of the home, was significantly improved through imputation. Of particular importance is the above light activity category, where 5.5 daily geo-located minutes of activity were added through imputation, evenly distributed between the home and non-home environments. These results suggest that imputation of missing GPS data may add significant PA and SB data to the model, and decrease modelling error associated with missing data.

Conclusions

Modelling of demographic and land use characteristics on presence of signal lapse and lapse length indicated that there are some demographic and behavioural characteristics associated with GPS signal lapse even after including environmental factors in the model. These biases are relatively small, but coupled with the large amount of data loss present for almost all participants. The results of this study indicate that researchers should consider the demographic and behavioural factors of their study population that may make GPS signal lapse a significant issue. Researchers should also strongly consider imputation. We found significant increases in data across all activity categories after imputation, adding up to large amounts of weekly activity. The imputation technique utilised in this study was validated and found to be highly accurate. We advocate general imputation as an effective tool for mitigating data lost through GPS signal lapse as it has the ability to return spatial context and greater utility to the data set, with the caveat that researchers must consider their research site for situations where the assumption of participant stationarity during signal loss may not apply.

References

Carlson J, 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 with SenseCam. Med Sci Sport Exer 47:662-7.

Costa E, 2011. Simulation of the effects of different urban environments on GPS performance using digital elevation models and building databases. IEEE T Int Transp Syst 12:819-29.

Demchak B, Kerr J, Raab F, Patrick K, Kruger IH, 2012. PALMS: a modern coevolution of community and computing using policy driven development. Available from: https://www.computer.org/csdl/proceedings/hicss/2012/4525/00/4525c735.pdf

Doherty A, Hodges S, King A, Smeaton AF, Berry E, Moulin CJ, Lindley S, Kelly P, Foster C, 2013. Wearable cameras in health: the state of the art and future possibilities. Am J Prev Med 44:320-3.

Duncan MJ, Badland HM, Mummery WK, 2009. Applying GPS to enhance understanding of transport-related physical activity. Sports Med Aus 12:549-56.

Duncan MJ, Mummery WK, 2007. GIS or GPS? A comparison of two methods for assessing route taken during active transport. Am J Prev Med 33:51-3.

Ellis K, Godbole S, Chen J, Marshall S, Lanckriet G, Kerr J, 2013. Physical activity recognition in free-living from body-worn sensors. Available from: eceweb.ucsd.edu/~gert/papers/par-13.pdf

Heil DP, Brage S, Rothney MP, 2012. Modeling physical activity outcomes from wearable monitors. Med Sci Sport Exer 44(Suppl.1):50-60.

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

Kerr J, Duncan S, Schipperijn 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 SJ, Godbole S, Chen J, Legge A, Doherty AR, Kelly P, Oliver M, Badland HM, Foster C, 2013. Using the SenseCam to improve classifications of sedentary behavior in free-living settings. Am J Prev Med 44:290-6.

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.

Lachowycz K, Jones AP, Page AS, Wheeler BW, Cooper AR, 2012. What can global positioning systems tell us about the contribution of dif-

Non-commercial use only

ferent types of urban greenspace to children's physical activity? Health Place 18:586-94.

Lam MS, Godbole S, Chen J, Oliver M, Badland H, Marshall SJ, Kelly P, Foster C, Doherty A, Kerr J, 2013. Measuring time spent outdoors using a wearable camera and GPS. Available from: https://ucsd-palms-project.wikispaces.com/file/view/ Lam_2013_Measuring_Time_Spent_Outdoors_Using_Wearable_Camera_GPS.pdf/545841220/Lam_2013_Measuring_Time_Spent_Outdoors_Using_Wearable_Camera_GPS.pdf

Maddison R, Ni Mhurchu C, 2009. Global positioning system: a new opportunity in physical activity measurement. Int J Behav Nutr 6:73.

Magnusson A, 2014. gmt: interface between GMT map-making software and R. Available from: http://cran.r-project.org/package=gmt

Ogle J, Guensler R, 2002. Accuracy of global positioning system for determining driver performance parameters. Transport Res Rec 1818:12-24.

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.

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.

R Core Team, 2013. R: a language and environment for statistical computing. Available from: http://www.r-project.org/

SANDAG, 2014. SanGIS/SANDAG data warehouse. San Diego Geographic Information Source - JPA/San Diego Association of Governments. Available from: http://www.sandag.org/ index.asp?subclassid=100&fuseaction=home.subclasshome

Stopher P, Fitzgerald C, Zhang J, 2008. Search for a global positioning system device to measure person travel. Transportation Res C-Emer 16:350-69.

Troped PJ, Wilson JS, Matthews CE, Cromley EK, Melly SJ, 2010. The built environment and location-based physical activity. Am J Prev Med 38:429-38.

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 51:148-52.

Wickham H, 2011. The split-apply-combine strategy for data analysis. J Stat Soft 40:1-29. Available from: https://www.jstatsoft.org/index.php/jss/article/view/v040i01/v40i01.pdf

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 Adolescent Health 42:418-20.

Non commercial use only