



Digital Phenotyping and Mental Health: Evaluating the Reliability of Passive Smartphone Behavioral Features Using a Public Beiwe Dataset

Devesh Punetha¹

¹Bellarmino College Preparatory; San Jose, California, USA

Abstract

Background/Objective: Digital phenotyping is the use of data collected from personal digital devices to study patterns in human behavior over time. It has become useful in mental-health research because smartphones can collect information in everyday settings. The objective of this research is to examine how reliable several commonly used smartphone-based behavioral features are.

Methods: I used a publicly available dataset from the Beiwe digital phenotyping platform. The data included six participants and 49 participant-days from March 2022. I added Gaussian noise to the GPS coordinates at eight levels, with standard deviations of 0-800 meters in each direction.

Results: Added GPS noise greatly increased total distance and mean speed. Time at home fell to 39% of baseline at 100 meters of noise and 1% at 800 meters. Radius of gyration remained within 1-7% of baseline through 200 meters. Changes in time at home and radius of gyration matched the values calculated from the amount of noise. Location entropy and number of significant places were less predictable because both depend on clustering. For 33 participant-days, GPS-based and self-reported time at home had a weak correlation in hours, which increased when expressed as proportions. However, day-to-day agreement within individuals was negligible.

Conclusions: Behavioral features can respond very differently to the same sensor error. This study does not test psychiatric diagnoses. Instead, it checks how reliable smartphone measurements are before they are used in mental-health research.

Keywords: digital phenotyping, mental health, smartphone sensing, Beiwe, GPS, mobility, behavioral features, privacy, measurement reliability, machine learning

1. Introduction

Mental-health assessment often depends on information collected during appointments and on what a person remembers about their behavior between those appointments. This can make gradual behavioral changes difficult to observe. A person may spend more time at home, become less active, or change their daily routine. These gradual changes may be difficult to remember accurately when reported later.

This problem may be particularly important during adolescence. Many mental-health conditions first appear during childhood, adolescence, or early adulthood ((M. Solmi, J. Radua, M. Olivola, E. Croce, L. Soardo, G. Salazar de Pablo, J. I. Shin, J. B. Kirkbride, P. Jones, J. H. Kim, J. Y. Kim, A. F. Carvalho, M. V. Seeman, C. U. Correll, P. Fusar-Poli. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Molecular Psychiatry*. Vol. 27, pg. 281–295, 2022, <https://doi.org/10.1038/s41380-021-01161-7>)). High school is also a period when students experience major changes in academic responsibilities, social environments, sleep schedules, and independence. This makes adolescence an important area for future research on methods that can measure behavioral patterns over longer periods of time.

Digital phenotyping offers one possible way to collect this type of longitudinal information. Digital phenotyping uses data from personal devices, such as smartphones and wearables, to study behavioral and physiological patterns ((J. Torous, M. V. Kiang, J. Lorme, J.-P. Onnela. New tools for new research in psychiatry: a scalable and customizable platform to empower data-driven smartphone research. *JMIR Mental Health*. Vol. 3, pg. e16, 2016, <https://doi.org/10.2196/mental.5165>))((A. Oudin, R. Maatoug, A. Bourla, F. Ferreri, O. Bonnot, B. Millet, F. Schoeller, S. Mouchabac, V. Adrien. Digital phenotyping: data-driven psychiatry to redefine mental health. *Journal of Medical Internet Research*. Vol. 25, pg. e44502, 2023, <https://doi.org/10.2196/44502>)). Smartphones can collect much of this data passively. This can include movement, location patterns, and screen use. Participants do not need to repeatedly enter information manually.

Previous research has examined whether smartphone-derived features are associated with mental-health outcomes. For example, researchers have studied whether changes in mobility are associated with depressive symptoms ((S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, D. C. Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of Medical Internet Research*. Vol. 17, pg. e175, 2015, <https://doi.org/10.2196/jmir.4273>)). Digital phenotyping has also been explored in people with schizophrenia and major depressive disorder ((I. Barnett, J. Torous, P. Staples, L. Sandoval, M. Keshavan, J.-P. Onnela. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. *Neuropsychopharmacology*. Vol. 43, pg. 1660–1666, 2018, <https://doi.org/10.1038/s41386-018-0030-z>))((F. Matcham, C. Barattieri di San Pietro, V. Bulgari, G. de Girolamo, R. Dobson, H. Eriksson, A. A. Folarin, J. M. Haro, M. Kerz, F. Lamers, Q. Li, N. V. Manyakov, D. C. Mohr, I. Myin-Germeys, V. Narayan, B. W. J. H. Penninx, Y. Ranjan, Z. Rashid, A. Rintala, S. Siddi, S. Simblett, T. Wykes, M. Hotopf, RADAR-CNS consortium. Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. *BMC Psychiatry*. Vol. 19, pg. 72, 2019, <https://doi.org/10.1186/s12888-019-2049-z>)). Its use is not limited to mental health. Similar methods have been used to measure mobility and quality of life in patients with spine disease ((D. J. Cote, I. Barnett, J.-P. Onnela, T. R. Smith. Digital phenotyping in patients with spine disease: a novel approach to quantifying mobility and quality of life. *World Neurosurgery*. Vol.

126, pg. e241–e249, 2019, <https://doi.org/10.1016/j.wneu.2019.01.297>). AI and machine-learning methods could eventually help researchers analyze several of these signals together. This may help to identify patterns that are difficult to detect manually ((I. R. Galatzer-Levy, J.-P. Onnela. Machine learning and the digital measurement of psychological health. *Annual Review of Clinical Psychology*. Vol. 19, pg. 133–154, 2023, <https://doi.org/10.1146/annurev-clinpsy-080921-073212>)).

However, there is an important issue to consider before using these features for prediction. A machine-learning model depends on the quality of the data used to train it. If a behavioral feature changes dramatically because of missing data, GPS error, privacy-related location distortion, or the mathematical way the feature is calculated, the model may learn patterns created partly by measurement error rather than by actual human behavior.

For example, researchers can use GPS data to measure different parts of a person's daily movement. These measures include estimated distance traveled, time spent at home, the number of significant locations visited, and how spread out a person's movement was during the day. These features may all appear to describe mobility, but they are constructed differently. It is therefore possible that the same GPS error could have very different effects on each one.

The purpose of the present study is to examine this measurement problem using a publicly available dataset from the Beiwe digital phenotyping platform. This study does not attempt to diagnose any mental-health conditions. Instead, it asks a more basic question: How reliable are common smartphone-based behavioral features when location data become noisy or incomplete? Understanding their reliability is important before these features are used in future mental-health research, especially in studies involving adolescents.

1.1 Research Hypothesis

Before running the GPS-noise experiment, I grouped the behavioral features based on how they are calculated. I expected to see three main patterns.

First, point-to-point measures, such as total distance traveled and mean speed, should become larger when independent noise is added. Each GPS point is moved slightly in a random direction. When distances between consecutive points are added together, this artificial movement accumulates.

Second, boundary-based measures, such as time at home, should become smaller. A participant may actually remain at home, but added GPS noise can move readings outside the radius being used to define the home area.

Third, spread-based measures, such as radius of gyration, should be more stable. Independent noise slightly increases radius of gyration, but the change is small when a person's real movement is much larger than the noise.

This prediction is testable because it can fail. If a feature moves in the opposite direction from its predicted group, or if it behaves in a way that cannot be described by any of the three groups, the proposed classification does not explain that feature. This prediction was tested by comparing the actual changes in time at home and radius of gyration with the expected changes from the added noise (Section 2.7).

1.2 Digital Phenotyping and Mental-Health Research

Digital phenotyping makes it possible to study behavior in real-world environments using devices that people already carry ((J. Torous, M. V. Kiang, J. Lorme, J.-P. Onnela. New tools for new research in psychiatry: a scalable and customizable platform to empower data-driven smartphone research. *JMIR Mental Health*. Vol. 3, pg. e16, 2016, <https://doi.org/10.2196/mental.5165>)). Data collected through a smartphone can generally be divided into active and passive data. Active data require participants to do something, such as complete a questionnaire. Passive data can be collected by device sensors with little or no additional action from the participant. Examples include GPS, accelerometer information, power state, and screen-use events.

This type of data has several potential advantages for mental-health research. Instead of using only one clinic visit or survey, researchers can study how a person's behavior changes over several days or weeks. Machine-learning methods can then potentially analyze many signals at the same time ((I. R. Galatzer-Levy, J.-P. Onnela. Machine learning and the digital measurement of psychological health. *Annual Review of Clinical Psychology*. Vol. 19, pg. 133–154, 2023, <https://doi.org/10.1146/annurev-clinpsy-080921-073212>)).

At the same time, behavioral data should not be confused with psychiatric measurements. Spending more time at home, traveling less, or changing how someone uses their phone does not by itself mean that the person is depressed or anxious. Studying these relationships requires validated symptom measures, well-designed clinical studies, and much larger groups of participants.

1.3 The Beiwe Platform

Beiwe is an open-source platform developed by the Onnela Lab at the Harvard T.H. Chan School of Public Health for high-throughput digital phenotyping research ((J. Torous, M. V. Kiang, J. Lorme, J.-P. Onnela. New tools for new research in psychiatry: a scalable and customizable platform to empower data-driven smartphone research. *JMIR Mental Health*. Vol. 3, pg. e16, 2016, <https://doi.org/10.2196/mental.5165>))((J.-P. Onnela, C. Dixon, K. Griffin, T. Jaenicke, L. Minowada, S. Esterkin, A. Siu, J. Zagorsky, E. Jones. Beiwe: a data collection platform for high-throughput digital phenotyping. *Journal of Open Source Software*. Vol. 6, pg. 3417, 2021, <https://doi.org/10.21105/joss.03417>)). It can collect multiple smartphone data streams and allows researchers to configure which sensors are used and how frequently information is recorded ((Beiwe Service Center, Harvard T.H. Chan School of Public Health. Technology. <https://hsph.harvard.edu/research-resource/beiwe/technology/>, accessed September 5, 2026.)).

An important distinction is that Beiwe is a data-collection platform, not a single dataset. The data analyzed in this study come from a separate sample dataset released publicly for educational use ((P. Emedom-Nnamdi, K. W. Carlson, Z. Clement, M. Karas, M. Straczekiewicz, J.-P. Onnela. Sample Beiwe dataset. Zenodo. 2022, <https://doi.org/10.5281/zenodo.6471045>)). The sample dataset contains passive smartphone information as well as some survey information. Because the data are public and the location information was already privacy protected before release, the dataset provides an opportunity to study how common digital-phenotyping features behave without collecting new participant information.

1.4 Behavioral Features

Researchers have proposed many ways of converting raw GPS coordinates into behavioral features. Total distance adds the estimated distance traveled between consecutive GPS readings. Mean speed uses the distance between readings together with elapsed time. Time at home measures how much recorded time falls within a defined radius around an estimated home location. Radius of gyration describes the overall spatial spread of a participant's recorded locations around their center of movement. Location entropy describes how evenly a person's recorded time is distributed among different locations. Number of significant places attempts to estimate how many meaningful location clusters appear in the participant's movements.

Although several of these features are intended to describe mobility, their mathematical definitions are different. Wang and colleagues, for example, published definitions for behavioral features as part of the HOPES digital phenotyping platform ((X. Wang, N. Vouk, C. Heaukulani, T. Buddhika, W. Martanto, J. Lee, R. J. T. Morris. HOPES: an integrative digital phenotyping platform for data collection, monitoring, and machine learning. *Journal of Medical Internet Research*. Vol. 23, pg. e23984, 2021, <https://doi.org/10.2196/23984>)). Having explicit definitions is important because two studies using the same feature name may calculate it differently.

1.5 Smartphone and Wearable Data in Mental-Health Research

Researchers have studied smartphone features like these to see how they relate to mental health. Using data from 48 college students tracked for 10 weeks in the StudentLife study, Saeb and colleagues found that regular daily movement and more balanced time spent across frequently visited places were linked to lower depression scores ((S. Saeb, E. G. Lattie, S. M. Schueller, K. P. Kording, D. C. Mohr. The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ*. Vol. 4, pg. e2537, 2016, <https://doi.org/10.7717/peerj.2537>)). Barnett et al. found that in 15 schizophrenia patients whose data were analyzed, unusual changes in mobility and social behavior detected via the Beiwe app were 71% more frequent in the two weeks before relapse than during other periods ((I. Barnett, J. Torous, P. Staples, L. Sandoval, M. Keshavan, J.-P. Onnela. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. *Neuropsychopharmacology*. Vol. 43, pg. 1660–1666, 2018, <https://doi.org/10.1038/s41386-018-0030-z>)). These features are also being used in clinical tools. The HOPES platform helps clinicians track patients' mobility using time away from home and radius of gyration ((X. Wang, N. Vouk, C. Heaukulani, T. Buddhika, W. Martanto, J. Lee, R. J. T. Morris. HOPES: an integrative digital phenotyping platform for data collection, monitoring, and machine learning. *Journal of Medical Internet Research*. Vol. 23, pg. e23984, 2021, <https://doi.org/10.2196/23984>)).

1.6 Location Privacy and Measurement Error

GPS information presents an important privacy problem. A person's repeated location history can reveal sensitive places such as a home, school, workplace, or frequently visited locations. Research has shown that human mobility patterns can be surprisingly identifying even when

obvious personal identifiers have been removed ((Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel. Unique in the crowd: the privacy bounds of human mobility. Scientific Reports. Vol. 3, pg. 1376, 2013, <https://doi.org/10.1038/srep01376>)). Because of this, location datasets may be transformed before they are publicly released.

Two possible privacy approaches have very different consequences for behavioral analysis (Figure 1). One approach is to move every coordinate belonging to one person by the same fixed amount. This hides the person's true geographic location but preserves distances between their own observations. Another approach is to add independent noise to every individual GPS reading. This makes the person's actual location harder to identify. However, it can also make the GPS data appear to show movement even when the person stayed in the same place.

Schematic — illustrative only, not data. Two candidate GPS de-identification mechanisms and their effect on derived features

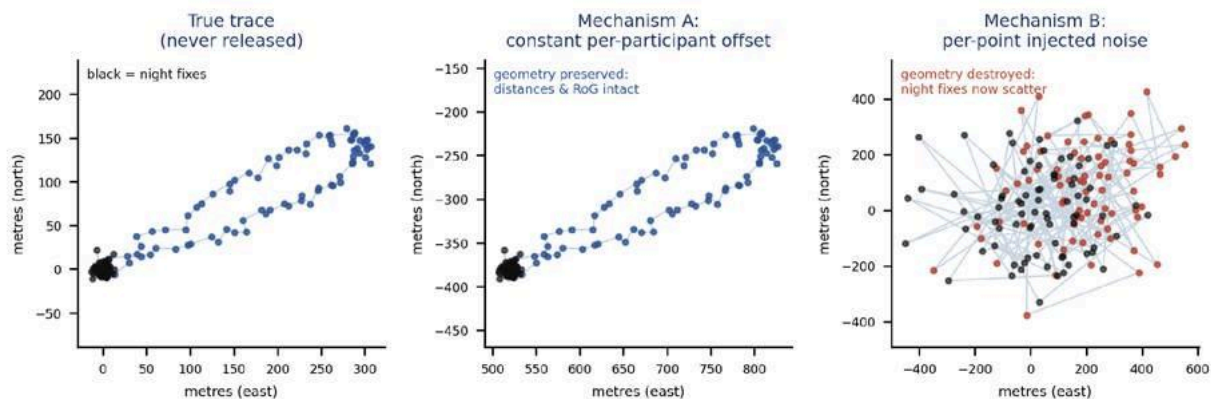


Figure 1. Schematic illustration of two possible GPS de-identification mechanisms. A fixed participant-level shift preserves the internal geometry of a trace, while independent point-level noise creates artificial movement. This figure is schematic and does not show participant data.

The released GPS coordinates data do not show exactly how the locations were changed for privacy. So, I first examined the internal GPS movement before carrying out the main noise experiment.

2. Methods

2.1 Data Source

This project uses the publicly available Beiwe sample dataset from Zenodo for a secondary analysis ((P. Emedom-Nnamdi, K. W. Carlson, Z. Clement, M. Karas, M. Straczewicz, J.-P. Onnela. Sample Beiwe dataset. Zenodo. 2022, <https://doi.org/10.5281/zenodo.6471045>)). I did not recruit participants or collect new data. The analyzed files contained observations from six participants covering March 20 through March 28, 2022. Because not every participant had usable information on every day, the final analysis contained 49 participant-days.

The dataset includes GPS data, other smartphone data such as power-state information, and available survey responses. The sample was originally released for educational use following an

internal platform test. It was not designed as a mental-health cohort or an adolescent study. Therefore, results from this analysis should not be generalized to teenagers or people with psychiatric conditions.

2.2 Ethical Considerations

No new participant data were collected for this study. Participants were not contacted, and the analysis used only data that had already been publicly released in de-identified form with location privacy protections. I made no attempt to re-identify anyone, undo the blurring, or infer a real home address beyond the relative clustering required to compute time at home. No absolute latitude or longitude coordinates appear anywhere in this paper. Every participant trace is plotted only as east-west and north-south displacement from that participant's own estimated home centroid.

2.3 Analysis Pipeline

The analysis pipeline was written in Python 3 using pandas, NumPy, SciPy, scikit-learn, and Matplotlib. GPS observations were organized chronologically by participant. Timestamps were converted into a consistent local time before observations were grouped by day. The pipeline was designed so that behavioral features could be recalculated from the raw data after different levels of simulated noise were added (Figure 2).

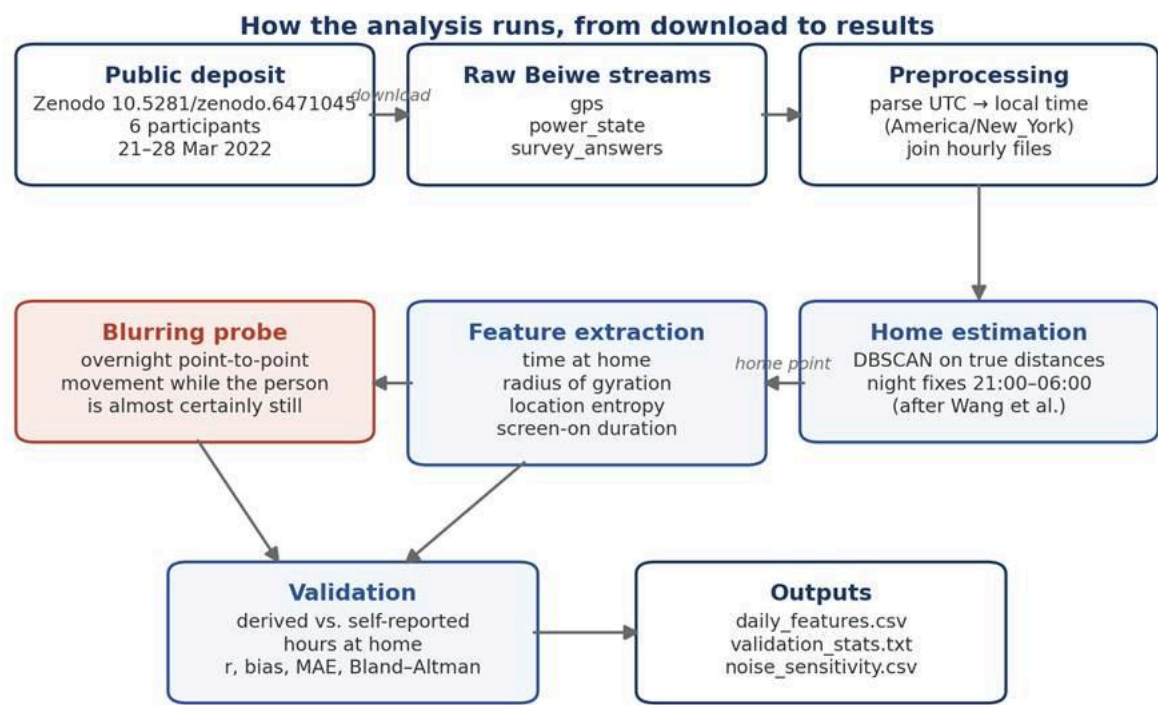


Figure 2. Analysis pipeline used for the secondary analysis of the public Beiwe sample dataset. The prerequisite blurring check and the GPS-noise sweep are the two analyses that directly test the study hypothesis.

2.4 Home Estimation and Feature Extraction

Home was estimated for each participant from overnight GPS coordinates using DBSCAN, a density-based clustering algorithm that groups nearby observations and labels isolated points as noise. Clustering used the haversine metric rather than ordinary Euclidean distance because latitude and longitude lie on a sphere. The DBSCAN neighborhood radius was 100 m, expressed in radians as $100 / 6,371,000$ using an Earth radius of 6,371,000 m, with `min_samples = 5`. The `ball_tree` algorithm was used for neighbor searches; constructing a brute-force pairwise distance matrix over dense traces exhausted memory. Following Wang et al. ((X. Wang, N. Vouk, C. Heaukulani, T. Buddhika, W. Martanto, J. Lee, R. J. T. Morris. HOPES: an integrative digital phenotyping platform for data collection, monitoring, and machine learning. *Journal of Medical Internet Research*. Vol. 23, pg. e23984, 2021, <https://doi.org/10.2196/23984>)), the home-detection window was 21:00 to 06:00 local time. The centroid of the largest qualifying overnight cluster was used as the estimated home position, and time at home was the recorded time falling within 100 m of that centroid.

The six main GPS-derived features were total distance, mean speed, time at home, radius of gyration, location entropy, and number of significant places. Significant places were found with the same DBSCAN settings on a participant's full trace. Location entropy was the Shannon entropy, in nats, of the share of readings assigned to each cluster; points labeled as DBSCAN noise were excluded. When a participant had more than 8,000 GPS readings, I used a systematic stride, taking every Nth row until the trace was below the cap. I used a deterministic stride instead of a random sample so repeated runs preserve the same observations and keep coverage spread across the study period.

Mean speed was calculated as total accepted distance divided by total accepted elapsed time, not as the arithmetic mean of instantaneous segment speeds. This choice corrected a problem in an early version of the analysis: one jittered GPS reading separated by a short time interval can create an extremely large instantaneous speed, and averaging those segment speeds can produce a daily mean that is physically inconsistent with the total distance traveled. Segments implying speeds above 200 km/h were therefore treated as sensor error and excluded. This definition makes the daily speed interpretable as distance per unit of observed travel time while preventing single short-interval spikes from dominating the result.

2.5 Analysis Settings

To make the analysis reproducible, the main thresholds and clustering settings were defined explicitly in the analysis script. The settings used for the results in this paper are shown in Table 1.

Setting	Value
Home radius	100 m
Home-detection window	21:00 to 06:00 local time
Stationary window (blurring check)	01:00 to 05:00 local time
Longest gap counted as continuous	600 s

Speed ceiling	200 km/h
Speed-limit test	noise sweep repeated without the 200 km/h ceiling; same random noise
Additional tests	reviewer_tests.py, same repository as the analysis code (speed-limit and coverage-by-hour tests)
DBSCAN epsilon	100 m, haversine metric
DBSCAN min_samples	5
Added-noise levels	0, 10, 25, 50, 100, 200, 400, 800 m
Draws per noise level	5 (1 at zero noise)
Study time zone	America/New_York
Feature-extraction trace cap	8,000 readings; systematic stride
Noise-sweep trace cap	6,000 readings; systematic stride; location entropy and significant places only
Noise seed components	participant identifier + noise level + draw number

Table 1. Configurable settings used to produce the reported analysis.

2.6 Prerequisite Test for Existing GPS Distortion

The public dataset states that geographic information was privacy protected, but the released data do not make it possible to determine the exact transformation with certainty. Before adding more noise, I therefore tested whether the original traces appeared to contain large independent point-to-point errors.

I used GPS data collected between 01:00 and 05:00, when participants were most likely asleep and not moving. I chose these hours because any apparent movement during this time is more likely to come from GPS error than from real travel. For consecutive observations within this period, I calculated apparent displacement between GPS readings. If noise had already been added to each GPS point, I would expect to see movement in the data even when the person was probably staying in the same place.

2.7 Simulated GPS-Noise Experiment

After this prerequisite check, I added independent Gaussian noise to each GPS reading. I tested eight noise levels: 0, 10, 25, 50, 100, 200, 400, and 800 meters. A noise level of X meters means each GPS reading was shifted randomly in both the north-south and east-west directions, with a standard deviation of X meters in each direction. I converted the meter offsets into changes in latitude and longitude. The east-west change was adjusted based on latitude. I ran five separate trials at each nonzero noise level. The zero-noise level was run only once because repeating it would give the same result. After each trial, I recalculated all six features from the beginning. This included estimating the home location again using the noisy GPS data. For each noise level, I divided each noisy value by the same participant's zero-noise value. I report the median of the 30 ratios (6 participants $\times$ 5 trials) and their interquartile range (IQR).

Entropy and significant places were calculated once per participant. For the repeated tests, I reduced GPS traces with more than 6,000 readings by selecting points at regular intervals.

For each trial, I seeded the NumPy random generator using the participant ID, noise level, and trial number. This made the noise generation consistent within the same run. Because Python randomizes string hashing across sessions unless PYTHONHASHSEED is fixed, the exact random coordinates are reproducible in distribution rather than necessarily bit-for-bit across machines. The 200 km/h speed ceiling was applied after noise was added. This matters because larger noise can push more segments above the ceiling, causing those segments and their distance to be discarded. The noise sweep therefore measures how the complete analysis pipeline behaves under increasing coordinate error, not an isolated mathematical feature with all filtering removed.

For time at home, the probability that a noisy home reading stays within 100 meters is $1 - \exp(-100^2/(2\sigma^2))$, based on the Rayleigh distribution.

2.8 GPS Coverage

Daily behavior can only be measured when usable smartphone data are available. For this reason, I calculated GPS coverage for each participant-day. Coverage was expressed in hours of usable GPS observation and was also compared with the full 24 hours in a day.

2.9 Agreement With Self-Report

The dataset contained self-reported hours at home for some participant-days. I joined self-report and computed values on calendar date so each computed day matched at most one reported value. I report Pearson correlation, mean bias, mean absolute error, and Bland-Altman limits of agreement only as descriptive statistics. I computed these twice. The first comparison uses raw hours from the GPS data with the self-reported hours. The second places them on a common scale by expressing computed time at home as a fraction of that day's usable coverage and self-report as a fraction of 24 hours, which tests whether any disagreement on the raw scale is explained by the mismatch in observation windows described below. The two quantities are structurally mismatched: computed time at home can never exceed that day's usable GPS coverage, while self-report refers to the full 24-hour day. The 33 paired observations also come from only six individuals, so observations are clustered within participants rather than independent. Standard Pearson correlation and ordinary Bland-Altman limits assume independence. I calculated a repeated-measures correlation ((J. Z. Bakdash, L. R. Marusich. Repeated measures correlation. *Frontiers in Psychology*. Vol. 8, pg. 456, 2017, <https://doi.org/10.3389/fpsyg.2017.00456>)) to compare day-to-day changes within each person. I also calculated a separate correlation using the average values of the six participants.

2.10 Reproducibility

The pipeline is a single documented Python script with an inspect mode that reports the input folder layout and column names before analysis. Every analysis choice used here is a named constant, including the 100 m home radius, 21:00-06:00 home window, 01:00-05:00 stationary window, 600 s maximum continuous gap, 200 km/h speed ceiling, DBSCAN epsilon and

min_samples, 8,000-reading extraction cap, 6,000-reading noise-sweep cap, noise levels, and number of draws. Table 1 records the primary settings used to produce the reported results. The analysis script is `beiwe_features-3.py` and figure generation is in `make_figures.py`, both stored in a fork of the Onnela Lab Forest library. Forest itself was not modified or called by this analysis. Code is available at https://github.com/deveshp-web/forest/blob/develop/beiwe_features-3.py.

2.11 Data and Code Availability

The dataset analyzed in this study is publicly available through Zenodo: Emedom-Nnamdi P, Carlson KW, Clement Z, Karas M, Straczekiewicz M, Onnela JP. Sample Beiwe Dataset. Zenodo. 2022. DOI: 10.5281/zenodo.6471045.

The analysis code is available at https://github.com/deveshp-web/forest/blob/develop/beiwe_features-3.py and the figure-generation code is available at https://github.com/deveshp-web/forest/blob/develop/make_figures.py.

3. Results

3.1 Released Traces Showed Little Evidence of Large Independent Point-Level Distortion

Across all six participants, the median apparent displacement between consecutive GPS observations during the overnight stationary-proxy period was 0.1 meters. The 95th percentile was 8.2 meters (Figure 3).

These relatively small values did not show evidence of the large artificial movement that would be expected if substantial independent noise had already been added separately to every GPS observation (Figure 4). The results are consistent with a privacy transformation that preserves much of the internal geometry of an individual's trace, such as a fixed participant-level shift. However, the exact privacy method cannot be determined from the released dataset alone.

One caveat is that the median of 0.1 meters is smaller than the drift often expected between independent GPS readings from a stationary device. Some readings may therefore be repeated or cached fixes rather than fresh measurements. The 95th percentile of 8.2 meters is more informative for ruling out large point-level perturbation.

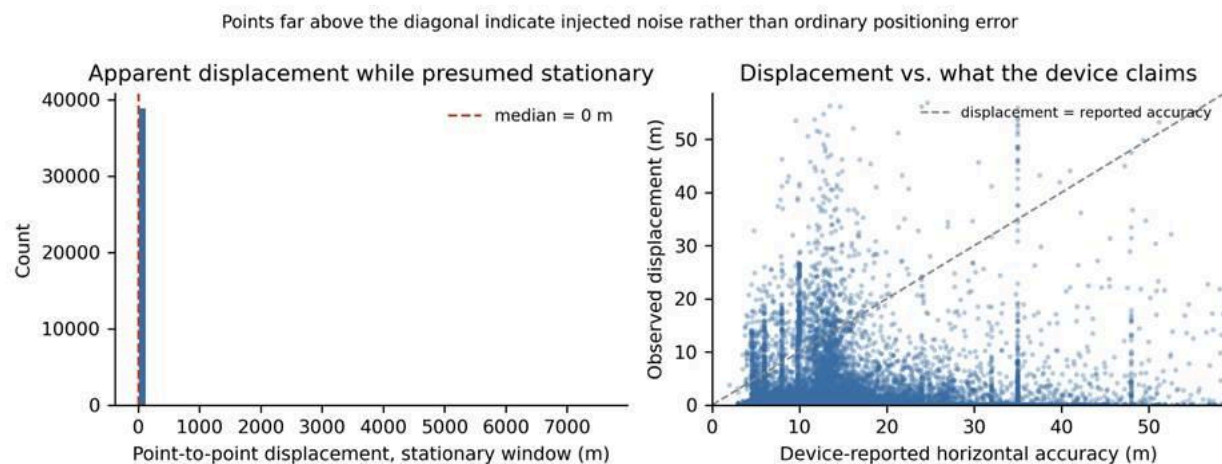


Figure 3. Apparent point-to-point displacement during the stationary-proxy window (left) and displacement compared with device-reported horizontal accuracy (right). The results show no evidence of tens-to-hundreds-of-meters independent perturbation already present in the released traces.

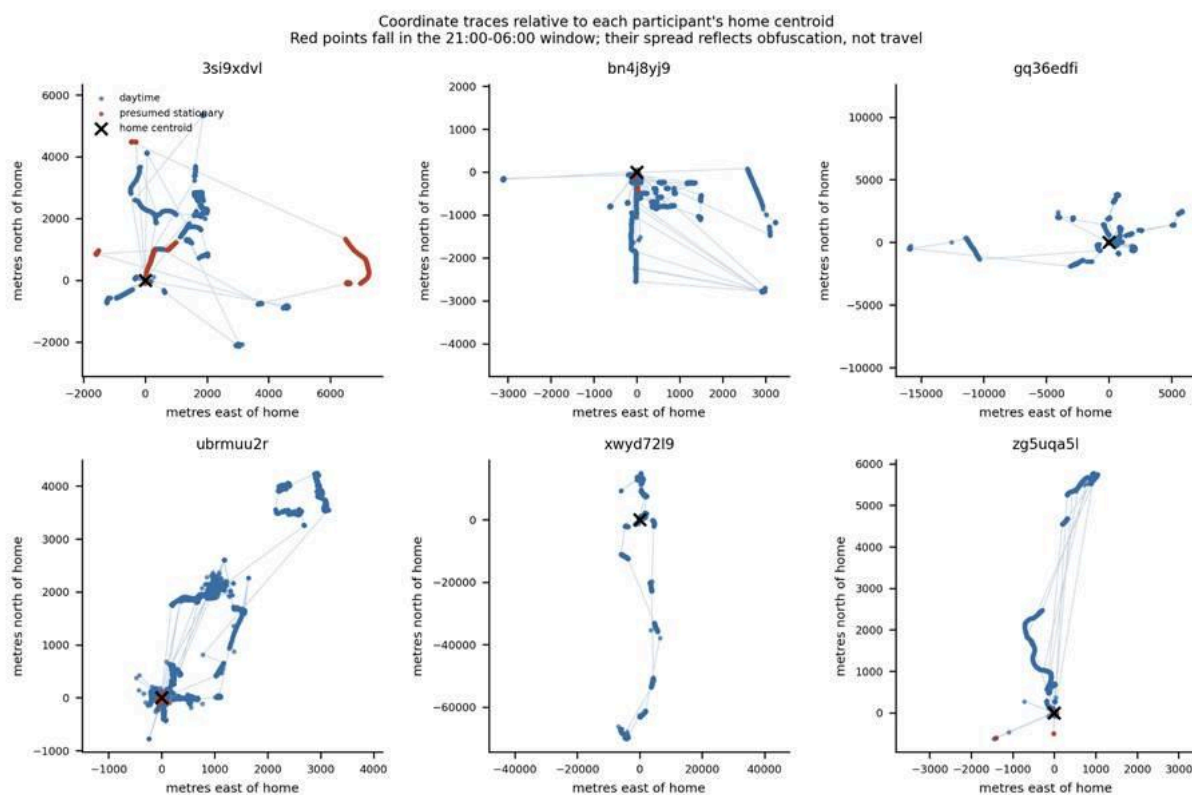


Figure 4.

GPS locations shown relative to each participant's estimated home location, measured in meters. Overnight GPS points are shown separately. No actual latitude or longitude values are displayed. This keeps the movement patterns used in the analysis without showing the actual GPS locations.

3.2 Point-to-Point Features Were Highly Sensitive to Noise

Total distance and mean speed behaved in the predicted direction (Figure 5A). At only 10 meters of added noise, both total distance and mean speed increased to 8.4 times their baseline values (IQR 7.4–13.4 for distance and 7.4–13.0 for speed). However, the increases were not steady. At 25 meters of noise, total distance reached 11.7 times baseline, while mean speed reached 13.5 times baseline.

Mean speed later declined to 10.5 times baseline at 100 meters before increasing again and reaching 21 times baseline at 800 meters. Total distance fell closer to 7.2 times baseline at 100 and 200 meters and ended at 8.0 times baseline at 800 meters. Although the exact values did not rise steadily with increasing noise, both measures remained strongly inflated throughout the experiment.

One possible explanation for part of this irregular pattern is the 200 km/h speed filter. Large GPS errors can make it appear that a person moved an unrealistic distance in a short amount of time. If the calculated speed was above 200 km/h, that segment was removed from the calculation. This may have reduced some of the increase at higher noise levels. As shown in Table 2, without the speed limit, distance and speed rose steadily with noise, while the limit removed most segments at higher noise levels, causing the uneven pattern. The rerun with the limit matched the original results, including 8.4 times baseline at 10 meters.

Noise (m)	Distance, with limit	Distance, no limit	Speed, with limit	Speed, no limit	Segments removed (%)
10	8.4 (7.4–13.4)	6.3 (5.3–10.8)	8.4 (7.4–13.0)	6.4 (5.3–11.0)	5
25	11.7 (9.6–17.7)	14.9 (12.3–26.5)	13.5 (11.6–18.8)	15.2 (12.3–26.9)	32
50	9.1 (5.7–12.0)	29.4 (24.0–52.6)	12.6 (11.8–16.2)	29.8 (24.0–53.4)	71
100	7.2 (4.6–11.0)	58.3 (47.8–104)	10.6 (9.8–16.3)	58.4 (47.9–106)	89
200	7.2 (6.1–8.0)	117 (95.4–209)	12.8 (11.3–17.6)	116 (95.6–213)	95
400	7.6 (5.3–8.0)	234 (190–418)	17.3 (12.1–18.1)	234 (191–425)	97
800	8.1 (4.0–12.5)	469 (380–834)	21.5 (15.4–23.7)	475 (381–849)	99

Table 2. Total distance and mean speed with and without the 200 km/h speed limit. Values are median ratios to each version’s own zero-noise value (IQR), based on 6 participants and 5 trials. With the speed limit, median daily distance at zero noise was 6.2 km (IQR 3.4–11.6) and mean speed was 1.8 km/h (IQR 1.2–4.1).

3.3 Time at Home Declined as Noise Increased

Time at home followed the predicted behavior for a boundary-based feature. At 50 meters of added noise, calculated time at home was 86% of the baseline value (IQR 86–87%). At 100 meters, it had fallen to 39% (IQR 38–40%). At 200 meters, it was 12% (IQR 11–12%). At 800 meters, calculated time at home was only about 1% of its baseline value (IQR 0.6–1.2%).

As more noise was added, more GPS points were pushed outside the 100-meter radius used to define home. This shows that time at home can change greatly because of GPS error, even if the person's actual behavior does not change. As shown in Table 3, the observed time at home closely matched the expected values at each noise level, suggesting that the drop was caused by GPS noise rather than a change in behavior.

3.4 Radius of Gyration Was Comparatively Stable

Radius of gyration was much less affected by GPS noise than total distance, mean speed, or time at home. With up to 200 meters of added noise, it stayed within 1–7% of its baseline value. The effect became larger only at very high noise levels. At 800 meters of noise, radius of gyration reached 1.52 times its baseline value (IQR 1.25–1.80). Overall, this matched my prediction that measures of overall movement spread would be less affected by GPS errors than the other types of features. This is expected because noise adds about $2\sigma^2$ to the squared radius of gyration, so it always increases slightly. As shown in Table 3, the expected radius of gyration closely matched the observed values, with the largest difference for any participant no greater than 0.02 at every noise level.

Noise (m)	Time at home, observed (%)	Time at home, expected (%)	Radius of gyration, observed	Radius of gyration, expected	Largest participant difference
10	100 (100–100)	100	1.00 (1.00–1.00)	1.00	0.00
25	100 (100–100)	100	1.00 (1.00–1.01)	1.00	0.00
50	86 (86–87)	86	1.01 (1.01–1.01)	1.01	0.00
100	39 (38–40)	39	1.02 (1.01–1.02)	1.02	0.00
200	12 (11–12)	12	1.07 (1.04–1.08)	1.06	0.00
400	3.3 (3.0–3.6)	3.1	1.18 (1.09–1.27)	1.18	0.01
800	0.9 (0.6–1.2)	0.8	1.52 (1.25–1.80)	1.50	0.02

Table 3. Observed and expected time at home and radius of gyration. Observed values are median ratios to the zero-noise value (IQR), based on 6 participants and 5 trials. Expected time at home was calculated using the Rayleigh distribution:

$1 - \exp(-100^2/(2\sigma^2))$. Expected radius of gyration was calculated for each participant using $\sqrt{Rg^2 + 2\sigma^2}$, then summarized by the median. The largest participant difference is the biggest gap between expected and observed radius of gyration for any participant.

3.5 Clustering-Based Features Did Not Follow the Predicted Groups

Location entropy and number of significant places did not behave as expected (Figure 5B). Both features decreased through moderate levels of noise and then became less predictable at the highest noise levels. At 200 meters of noise, location entropy fell to 0.48 times baseline (IQR 0.19–0.70), and significant places fell to 0.59 times baseline (IQR 0.50–0.74). At 800 meters, significant places rose to 2.2 times baseline (IQR 1.6–3.4). Neither could be clearly classified as a point-to-point, boundary, or spread feature.

A likely reason is that both depend on DBSCAN clustering. A small coordinate change may have little effect on one point but may cause another point to leave one cluster, enter another cluster, or become classified as noise. These changes can alter the entire cluster structure rather than simply changing a numerical distance. For this reason, clustering-based behavioral features may need to be tested directly under realistic measurement errors instead of assuming their behavior from their final mathematical formula.

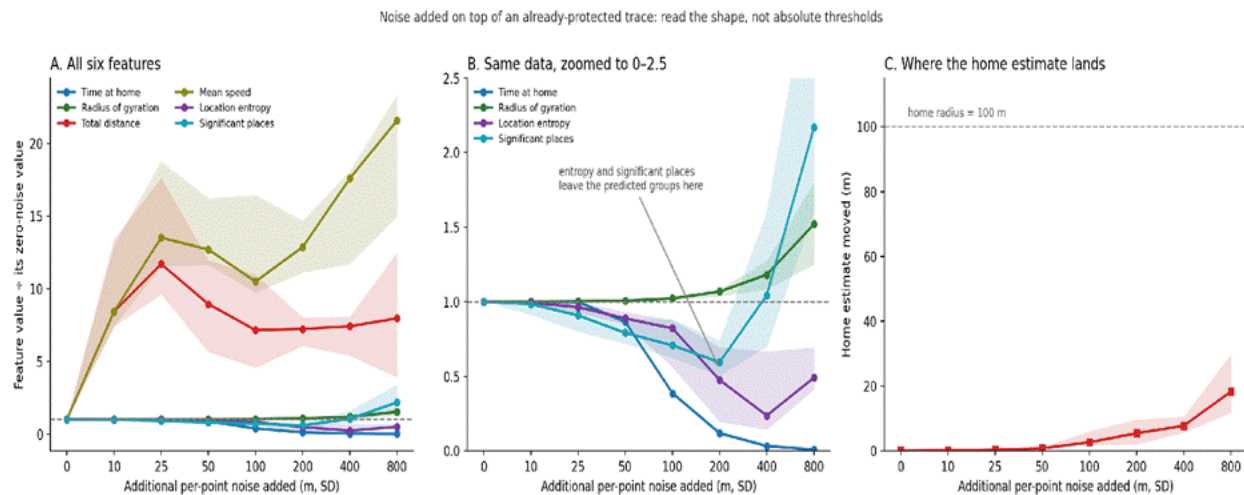


Figure 5. Effect of added GPS noise on behavioral features compared with zero-added-noise baseline. A: all six features. B: the same data with the vertical axis restricted to 0-2.5, which is the only view in which the two rejected features are legible; location entropy and number of significant places decline through moderate noise and then move unpredictably. C: displacement of the estimated home location against the 100-meter home radius. Bands show the interquartile range across participants and random draws.

3.6 Estimated Home Location Remained Stable

Through 200 meters of added noise, the estimated home location moved by a median of only 5.4 meters (IQR 2.1–9.6 meters). Even at 800 meters of noise, the median displacement was 18.28 meters (IQR 11.8–29.7 meters), which remained inside the 100-meter home radius (Figure 5C).

Therefore, the collapse in calculated time at home was not mainly caused by the algorithm completely losing track of where home was. Instead, the problem came from the boundary rule: individual observations were scattered outside the home radius and were no longer counted as time at home.

3.7 GPS Coverage Was Limited

Usable GPS coverage had a median of 3.3 hours per participant-day. The middle half of participant-days had roughly 2.1 to 4.6 hours of usable observations. Compared with a 24-hour day, the median coverage represented only about 14% of the day.

Coverage also varied considerably between participant-days (Figure 6). At least one participant-day contained essentially no usable GPS information, while even the best-covered day represented only part of the complete day. This finding places an important limitation on every daily feature reported in the dataset. For example, a measured value of two hours at home should be interpreted as two hours during the available observations, not necessarily two hours out of the complete 24-hour day.

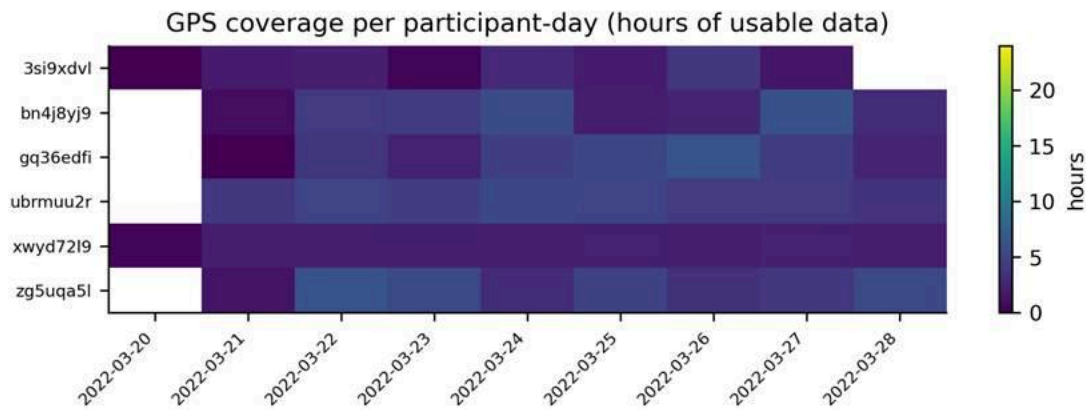


Figure 6. GPS coverage per participant-day, shown as hours of usable data. White cells represent participant-days with no usable GPS data.

3.8 Behavioral Feature Values

Across the 48 participant-days, median recorded time at home was 2.1 hours per day (IQR 1.4–3.8 hours), with values ranging from nearly zero to 6.2 hours. Median radius of gyration was 833 meters (IQR 193–1,082 meters).

Some participant-days showed much larger values (Figure 7). One participant had a radius of gyration of 29.4 kilometers on one day and 4.4 kilometers the following day. Because GPS coverage for that observation was not unusually poor, this may represent genuine long-distance movement, although the dataset does not provide enough context to determine the purpose of the trip.

Location entropy ranged from 0.13 to 1.82 nats (median 0.75 nats). The number of significant places identified over the observed week ranged from 2 to 31 (median 13.5). Screen-on time had a daily median of 2.2 hours (IQR 1.0–3.1 hours) and ranged from 0 to 4.5 hours. However, a value close to zero may represent either genuinely low phone use or missing records. The dataset does not always allow these two explanations to be separated.

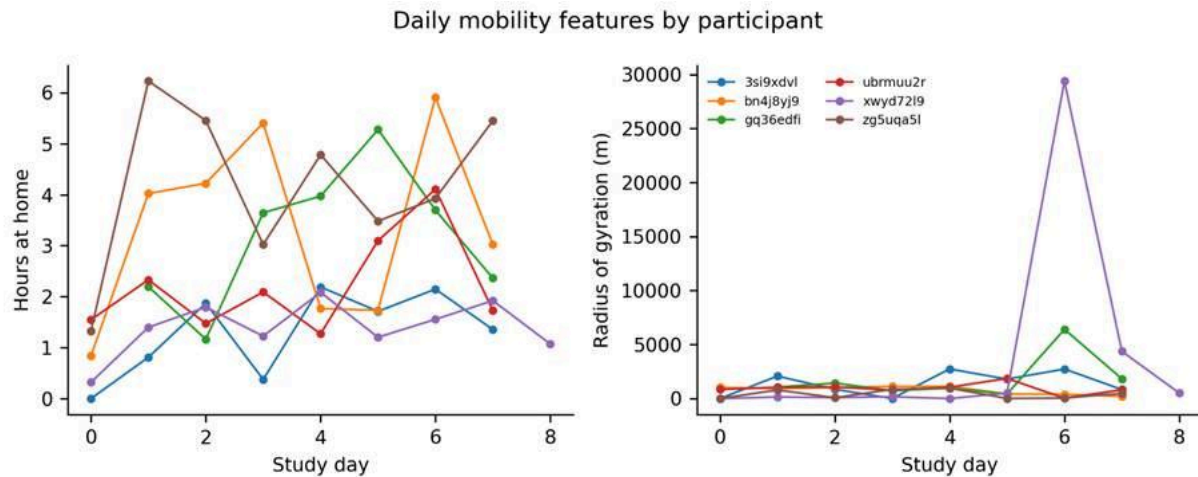


Figure 7. Daily time at home (left) and radius of gyration (right) by participant across the observed study days. These values are descriptive and should be interpreted in the context of limited GPS coverage.

3.9 Agreement Improved When Both Measures Used the Same Scale, but Only Between People

Thirty-three participant-days contained both a computed and a self-reported time-at-home value. On the raw hour scale the two quantities are not directly comparable, and the statistics reflect this mismatch. Pearson r was 0.262, mean bias was -15.59 h, mean absolute error was 15.59 h, and the exploratory 95% limits of agreement were -23.07 to -8.10 h. Every one of the 33 differences was negative, which is why the bias and mean absolute error are identical.

Figure 8A shows why the raw-hour comparison is dominated by the study design. Computed time at home is a total over recorded time and cannot exceed that day's usable GPS coverage, which reached at most 6.3 hours. Self-reported time at home describes the whole 24-hour day, and participants reported between 10 and 24 hours. The two ranges therefore do not overlap, so every computed value necessarily falls below every reported value and a negative bias was guaranteed by the mismatch in observation windows before the individual measurements were compared. Because the result is dominated by this difference in observation windows, the raw-hour comparison cannot meaningfully assess the accuracy of the feature.

Expressing both quantities on the same scale removes this structural mismatch. When computed time at home is taken as a fraction of that day's usable coverage, and self-report as a fraction of 24 hours, Pearson r rises to 0.471, mean bias falls to -0.015, and mean absolute error is 0.170. Figure 8B shows the points distributed around the line of perfect agreement rather than being constrained to the region below it. This improvement should not be interpreted as a new finding about behavior. Instead, it is consistent with the interpretation that much of the raw-scale disagreement arose from the mismatch in observation windows. Agreement was strong between people's average values ($r = 0.92$, $p = 0.010$), but not from day to day within

the same person ($r = 0.03$, $p = 0.89$; raw hours: $r = -0.18$, $p = 0.36$). This suggests that the measures agreed between people, but not in tracking daily changes.

Both analyses remain descriptive rather than inferential. The deposit also does not document a recall window for the survey question, so the span each answer covers is unknown. GPS coverage was spread across the day, with each hour accounting for 3.1–6.1% of covered time (4.2% if evenly spread) and 21:00–06:00 accounting for 40% (37.5% if even). This suggests little bias overall, but the covered hours may not represent the full day for individual participants. These statistics show that intermittent sensing and full-day self-report cannot be compared meaningfully on the raw hour scale. They do not establish that either measurement is correct.

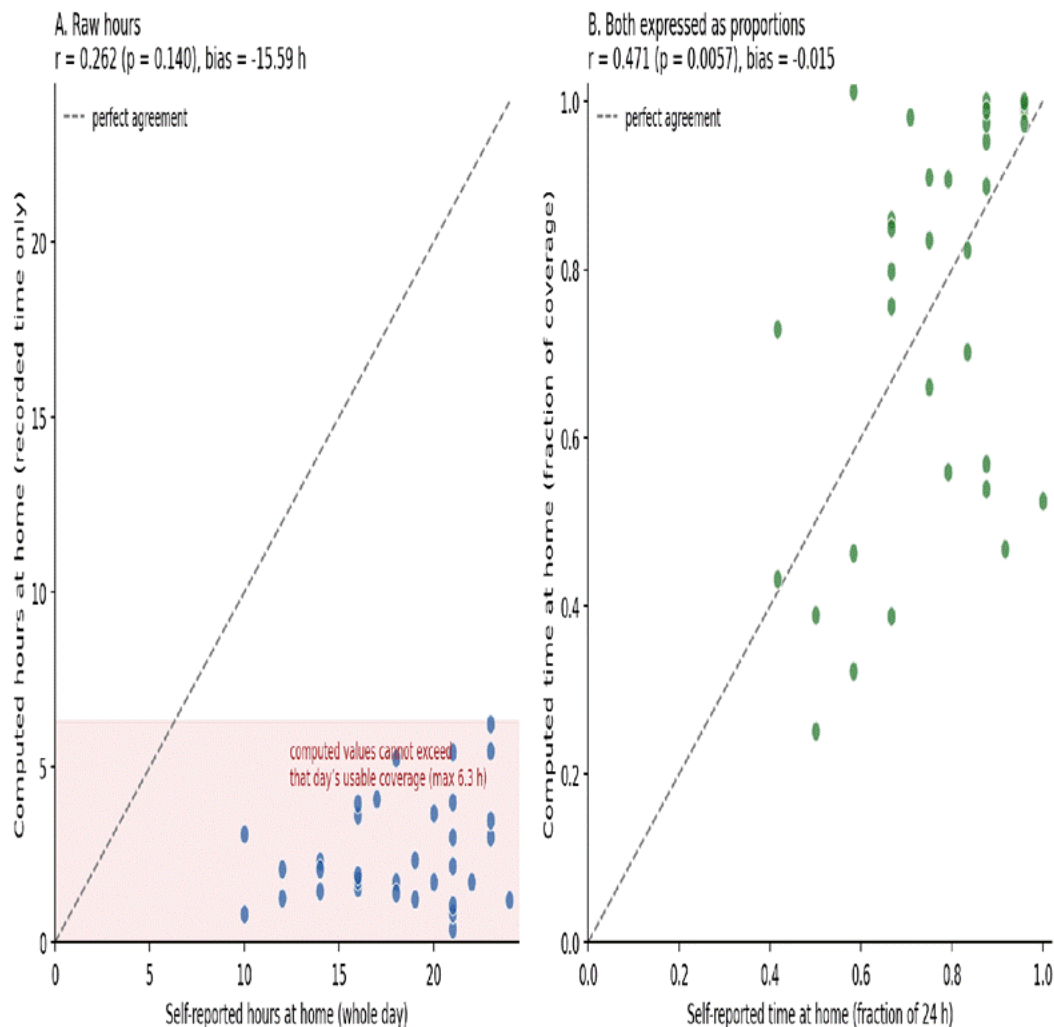


Figure 8. Computed versus self-reported time at home for the 33 participant-days with both values available. A: raw hours. The shaded band marks the maximum usable coverage observed (6.3 h); no computed value can lie above it, while every self-report lies between 10

and 24 hours, so the ranges never overlap. B: both quantities as proportions, computed against that day's coverage and self-report against 24 hours. The dashed line shows perfect agreement.

4. Discussion

4.1 Main Findings

The main finding of this study is that behavioral features calculated from the same GPS data can react very differently to the same amount of location noise. The original hypothesis predicted three types of behavior based on feature construction. The prediction worked well for four features.

Total distance and mean speed increased greatly as noise was added. Time at home decreased because more GPS points were pushed outside the fixed home boundary. Radius of gyration stayed more stable through moderate levels of noise because Radius of gyration stayed more stable at moderate noise levels because the added spread was small compared with participants' real movement. Both time at home and radius of gyration also changed as predicted by the noise model.

However, location entropy and number of significant places did not fit the proposed classification. Both depend on clustering, which makes them sensitive to nonlinear changes in how observations are grouped. The failure of these two features is important. In this experiment, two of the six features showed that the original three-group framework was incomplete.

4.2 Feature Choice Matters in Digital Phenotyping

Researchers often refer to several different calculations as mobility features. However, this study shows that they should not automatically be treated as interchangeable. Consider two studies measuring the same participant: one uses total distance traveled, while the other uses radius of gyration. If the GPS data contain independent location error, the first study could report a very large increase in apparent mobility while the second could remain relatively stable. Both calculations are based on the same GPS coordinates, but their conclusions could be very different because of how the features are constructed.

4.3 Implications for Mental-Health Research

Previous research has investigated mobility-related features as possible behavioral correlates of mental-health symptoms ((S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, D. C. Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of Medical Internet Research*. Vol. 17, pg. e175, 2015, <https://doi.org/10.2196/jmir.4273>)). Spending more time at home or moving through a smaller geographic area, for example, has sometimes been studied in relation to depressive symptoms. The present results do not test whether those relationships are valid. There are no depression or anxiety symptom labels in this dataset, and the participants were not an adolescent psychiatric cohort.

Instead, the results raise a methodological question that should be addressed before making clinical interpretations. If a study reports that someone's mobility decreased, researchers should be confident that the apparent change was not mainly created by differences in GPS coverage, privacy transformation, measurement noise, or feature construction. This matters in practice because the HOPES clinician dashboard uses time away from home and radius of gyration to help monitor mental health ((X. Wang, N. Vouk, C. Heaukulani, T. Buddhika, W. Martanto, J. Lee, R. J. T. Morris. HOPES: an integrative digital phenotyping platform for data collection, monitoring, and machine learning. *Journal of Medical Internet Research*. Vol. 23, pg. e23984, 2021, <https://doi.org/10.2196/23984>)). In this study, the same location noise affected these measures very differently, greatly increasing time away from home while barely changing radius of gyration.

This becomes even more important if these measurements are used as inputs to artificial-intelligence or machine-learning models. Machine learning can discover patterns in large datasets, but it cannot automatically determine whether an input feature represents true behavior or measurement error. A highly predictive model could still be misleading if it learns technical differences rather than meaningful behavioral changes. Measurement validation should therefore come before clinical prediction.

4.4 Relevance to Adolescent Mental Health

My original interest in digital phenotyping came from its possible use in mental-health research for teenagers. Teenagers have school schedules, transportation constraints, sleep patterns, and movement ranges that differ from those of commuting adults, so feature definitions validated in adults cannot simply be assumed to behave the same way in adolescents. Device-validation research in adolescents has shown the importance of testing digital measurements specifically within younger populations ((X. K. Lee, N. I. Y. N. Chee, J. L. Ong, T. B. Teo, E. van Rijn, J. C. Lo, M. W. L. Chee. Validation of a consumer sleep wearable device with actigraphy and polysomnography in adolescents across sleep opportunity manipulations. *Journal of Clinical Sleep Medicine*. Vol. 15, pg. 1337–1346, 2019, <https://doi.org/10.5664/jcsm.7932>)), although that work examined sleep sensing rather than smartphone mobility features.

The first version of this project proposed a system that would give advice to high school students about sleep and physical activity based on smartphone data and a wearable device like the Oura ring. It would also make a call to a helpline when needed. Before using such a system, the smartphone measurements need to be tested. This study shows that GPS errors can greatly reduce calculated time at home, making the system mistake a data error for a change in a student's routine.

However, the current study also shows why this idea has to be approached carefully. Adolescent data would be especially sensitive from a privacy perspective. Location history could reveal a student's home, school, activities, and other personal information. Researchers would therefore need strong privacy protections while also making sure that those protections do not make the behavioral measurements unreliable.

Future adolescent studies should also include validated mental-health assessments rather than attempting to infer a diagnosis directly from smartphone behavior. Digital phenotyping is more

appropriately viewed as a possible additional source of information rather than a replacement for clinicians, parents, counselors, or established mental-health assessments.

4.5 Limitations

This study has several important limitations. First, the dataset is very small. It includes only six participants across nine observed days, resulting in 49 participant-days. Second, the dataset was released as a sample from an internal platform test and was not designed as a research cohort. The findings therefore should not be generalized to a broader population.

Third, the dataset does not contain validated depression, anxiety, or other psychiatric symptom measurements. No relationship between smartphone features and mental-health symptoms can be tested from these data. Fourth, the participants should not be treated as an adolescent sample. The relevance to teenage mental-health research is a future application rather than a characteristic of the current participants.

Fifth, GPS coverage was limited. Median usable GPS data represented only around 14% of a full day. Daily totals therefore represent recorded portions of the day rather than complete observation of participants' behavior. Sixth, the exact location-privacy method applied before the dataset was released cannot be proven from the data alone. The overnight analysis suggests that large amounts of independent noise were not added to individual GPS points, but the exact privacy method is unknown.

Seventh, the 33 paired observations used in the self-report comparison are repeated measurements from only six people, so ordinary Pearson correlation and standard Bland-Altman limits violate the independence assumption. Eighth, speed limit of 200 km/h also greatly affects how noise changes total distance and mean speed (Section 3.2), so the results depend on whether this limit is used. The 600 s gap rule and the 8,000/6,000-reading caps are also pipeline choices whose interactions should be tested. Finally, only six behavioral features and one set of DBSCAN parameters were examined, so the clustering-based results may change under other definitions or settings.

4.6 Future Work

Several extensions would strengthen this work. First, the GPS-noise experiment could be repeated with different speed limits. This could show whether removing very high-speed segments caused some of the uneven changes in total distance and mean speed.

Second, different DBSCAN settings could be tested. This may help explain why location entropy and number of significant places did not follow the three predicted patterns.

Third, future experiments could also test the effect of missing GPS data, instead of only adding noise to the coordinates. Since GPS coverage was low in this dataset, missing data may be just as important as GPS error. Fourth, the analysis should be repeated with more participants and over longer periods of time.

Most importantly for the original mental-health motivation of this project, future studies could use adolescent cohorts that include validated measures of depression, anxiety, stress, or related outcomes. After the reliability of behavioral features has been established, machine-learning

models could test whether combinations of mobility and smartphone-use signals provide useful information beyond conventional assessments. This would allow the AI/ML component to be evaluated using actual clinical or symptom outcomes rather than assuming that a particular behavior represents a mental-health condition.

4.7 Conclusion

Digital phenotyping could be useful for studying changes in behavior over time. However, researchers first need to understand how reliable these measurements are before using them for mental-health prediction. This study tested how six common smartphone-based mobility features respond to added GPS noise.

Four of the six features followed the predicted pattern. Total distance and mean speed were highly sensitive to noise and became strongly inflated. Time at home decreased as GPS observations were pushed outside the home boundary. Radius of gyration remained comparatively stable through moderate levels of noise. Location entropy and number of significant places did not follow the proposed three-group classification, showing that clustering-based features can behave less predictably.

The study also found that participants had a median of 3.3 hours of usable GPS data per day, and that calculated time at home could not be meaningfully compared with self report on the raw hour scale because the two measures describe different observation windows. Once both were expressed as proportions of their respective observation windows, Pearson correlation increased substantially, while mean bias and mean absolute error decreased. However, the measures agreed between people, but not in tracking day-to-day changes within the same person, which is important for mental-health monitoring. Together, these findings show that feature selection, data completeness, privacy transformation, and measurement error are separate issues that can influence digital phenotyping results.

This project does not show that smartphone data can diagnose depression or anxiety. Instead, it addresses a step that should come earlier: determining whether behavioral features remain meaningful under realistic data-quality limitations. Future studies can build on this work by testing validated behavioral features in larger, clinically characterized populations, including adolescents, before incorporating these features into AI or machine-learning systems for mental health research.

5. Acknowledgments

I would like to thank my mentor, Dr. Kola Ogunlana, for his guidance throughout this research project, especially for helping me work through technical problems, improve my research approach, and better understand how digital phenotyping can be studied using computational methods.

I also acknowledge the Onnela Lab at the Harvard T.H. Chan School of Public Health and the contributors who released the public Beiwe sample dataset used in this analysis.

The author declares no competing interests.

6. References

1. M. Solmi, J. Radua, M. Olivola, E. Croce, L. Soardo, G. Salazar de Pablo, J. I. Shin, J. B. Kirkbride, P. Jones, J. H. Kim, J. Y. Kim, A. F. Carvalho, M. V. Seeman, C. U. Correll, P. Fusar-Poli. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Molecular Psychiatry*. Vol. **27**, pg. 281–295, 2022, <https://doi.org/10.1038/s41380-021-01161-7>.
2. J. Torous, M. V. Kiang, J. Lorme, J.-P. Onnela. New tools for new research in psychiatry: a scalable and customizable platform to empower data-driven smartphone research. *JMIR Mental Health*. Vol. **3**, pg. e16, 2016, <https://doi.org/10.2196/mental.5165>.
3. A. Oudin, R. Maatoug, A. Bourla, F. Ferreri, O. Bonnot, B. Millet, F. Schoeller, S. Mouchabac, V. Adrien. Digital phenotyping: data-driven psychiatry to redefine mental health. *Journal of Medical Internet Research*. Vol. **25**, pg. e44502, 2023, <https://doi.org/10.2196/44502>.
4. S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, D. C. Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of Medical Internet Research*. Vol. **17**, pg. e175, 2015, <https://doi.org/10.2196/jmir.4273>.
5. I. Barnett, J. Torous, P. Staples, L. Sandoval, M. Keshavan, J.-P. Onnela. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. *Neuropsychopharmacology*. Vol. **43**, pg. 1660–1666, 2018, <https://doi.org/10.1038/s41386-018-0030-z>.
6. F. Matcham, C. Barattieri di San Pietro, V. Bulgari, G. de Girolamo, R. Dobson, H. Eriksson, A. A. Folarin, J. M. Haro, M. Kerz, F. Lamers, Q. Li, N. V. Manyakov, D. C. Mohr, I. Myin-Germeys, V. Narayan, B. W. J. H. Penninx, Y. Ranjan, Z. Rashid, A. Rintala, S. Siddi, S. Simblett, T. Wykes, M. Hotopf, RADAR-CNS consortium. Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. *BMC Psychiatry*. Vol. **19**, pg. 72, 2019, <https://doi.org/10.1186/s12888-019-2049-z>.
7. D. J. Cote, I. Barnett, J.-P. Onnela, T. R. Smith. Digital phenotyping in patients with spine disease: a novel approach to quantifying mobility and quality of life. *World Neurosurgery*. Vol. **126**, pg. e241–e249, 2019, <https://doi.org/10.1016/j.wneu.2019.01.297>.
8. I. R. Galatzer-Levy, J.-P. Onnela. Machine learning and the digital measurement of psychological health. *Annual Review of Clinical Psychology*. Vol. **19**, pg. 133–154, 2023, <https://doi.org/10.1146/annurev-clinpsy-080921-073212>.
9. J.-P. Onnela, C. Dixon, K. Griffin, T. Jaenicke, L. Minowada, S. Esterkin, A. Siu, J. Zagorsky, E. Jones. Beiwe: a data collection platform for high-throughput digital phenotyping. *Journal of Open Source Software*. Vol. **6**, pg. 3417, 2021, <https://doi.org/10.21105/joss.03417>.
10. Beiwe Service Center, Harvard T.H. Chan School of Public Health. Technology. <https://hsph.harvard.edu/research-resource/beiwe/technology/>, accessed September 5, 2026.

11. P. Emedom-Nnamdi, K. W. Carlson, Z. Clement, M. Karas, M. Straczekiewicz, J.-P. Onnela. Sample Beiwe dataset. *Zenodo*. 2022, <https://doi.org/10.5281/zenodo.6471045>.
12. X. Wang, N. Vouk, C. Heaukulani, T. Buddhika, W. Martanto, J. Lee, R. J. T. Morris. HOPES: an integrative digital phenotyping platform for data collection, monitoring, and machine learning. *Journal of Medical Internet Research*. Vol. **23**, pg. e23984, 2021, <https://doi.org/10.2196/23984>.
13. S. Saeb, E. G. Lattie, S. M. Schueller, K. P. Kording, D. C. Mohr. The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ*. Vol. **4**, pg. e2537, 2016, <https://doi.org/10.7717/peerj.2537>.
14. Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel. Unique in the crowd: the privacy bounds of human mobility. *Scientific Reports*. Vol. **3**, pg. 1376, 2013, <https://doi.org/10.1038/srep01376>.
15. J. Z. Bakdash, L. R. Marusich. Repeated measures correlation. *Frontiers in Psychology*. Vol. **8**, pg. 456, 2017, <https://doi.org/10.3389/fpsyg.2017.00456>.
16. X. K. Lee, N. I. Y. N. Chee, J. L. Ong, T. B. Teo, E. van Rijn, J. C. Lo, M. W. L. Chee. Validation of a consumer sleep wearable device with actigraphy and polysomnography in adolescents across sleep opportunity manipulations. *Journal of Clinical Sleep Medicine*. Vol. **15**, pg. 1337–1346, 2019, <https://doi.org/10.5664/jcsm.7932>.