*Tutorial*

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Advances in Methods and
Practices in Psychological Science
April-June 2022, Vol. 5, No. 2,
pp. 1–11
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/25152459221082680
www.psychologicalscience.org/AMPPS
SAGE

# Analyzing GPS Data for Psychological Research: A Tutorial

**Sandrine R. Müller[1], Joseph B. Bayer[2,3], Morgan Quinn Ross[2], Jerry Mount[4], Clemens Stachl[5], Gabriella M. Harari[6], Yung-Ju Chang[7], and Huyen T. K. Le[8]**

[1]Data Science Institute, Columbia University, New York City, New York; [2]School of Communication, The Ohio State University, Columbus, Ohio; [3]Translational Data Analytics Institute, The Ohio State University, Columbus, Ohio; [4]IIHR - Engineering and Hydroscience, University of Iowa, Iowa City, Iowa; [5]Institute of Behavioral Science and Technology, University of St. Gallen, St. Gallen, Switzerland; [6]Department of Communication, Stanford University, Stanford, California; [7]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan; and [8]Department of Geography, The Ohio State University, Columbus, Ohio

**Abstract**

The ubiquity of location-data-enabled devices provides novel avenues for psychology researchers to incorporate spatial analytics into their studies. Spatial analytics use global positioning system (GPS) data to assess and understand mobility behavior (e.g., locations visited, movement patterns). In this tutorial, we provide a practical guide to analyzing GPS data in R and introduce researchers to key procedures and resources for conducting spatial analytics. We show readers how to clean GPS data, compute mobility features (e.g., time spent at home, number of unique places visited), and visualize locations and movement patterns. In addition, we discuss the challenges of ensuring participant privacy and interpreting the psychological implications of mobility behaviors. The tutorial is accompanied by an R Markdown script and a simulated GPS data set made available on the OSF.

**Keywords**

GPS, spatial analytics, mobile analytics, sensing, tracking, digital traces, location, open data, open materials

Received 6/29/21; Revision accepted 1/20/22

Human movement patterns have received attention across an array of psychological research areas (Ross et al., in press), including clinical psychology (e.g., depression; Cornet & Holden, 2018) and personality psychology (e.g., Big Five traits; Alessandretti et al., 2018; Stachl et al., 2020). The newfound attention to movement patterns in psychology follows the rise in access to location data collected throughout individuals' daily lives and increasing interest in the importance of situations and environments for understanding human behavior (Rauthmann, 2021). Such location information often comes in the form of cell-phone signal and global positioning system (GPS) data. GPS receivers are considered highly reliable and contained in virtually all smartphones (Crato, 2010; van Diggelen & Enge, 2015). These devices convert signals from satellites into time-stamped longitude and latitude coordinates, which allows researchers to derive mobility features, such as the number of places visited by an individual (Fillekes et al., 2018). Yet despite these developments, geospatial data, methods, and analytics are rarely incorporated into psychological studies. This may be because of perceived technical challenges in converting GPS data into variables (e.g., mobility features) that are useful to psychologists and other social scientists, which we hope to remedy through this introductory tutorial.

**Corresponding Authors:**
Sandrine R. Müller, Google, 111 8th Avenue, New York, NY 10011
Email: sandrinemuller@google.com

Joseph B. Bayer, School of Communication, The Ohio State University, 3016 Derby Hall, 154 N. Oval Mall, Columbus, OH 43210
Email: bayer.66@osu.edu

Once the challenge of turning GPS data into meaningful features has been overcome, many statistical approaches can be applied that are familiar to psychological researchers (e.g., correlation, regression, and factor analysis). For example, researchers can test relationships between mobility features and other health measures collected about participants in parallel (e.g., heart rate) or separately (e.g., life satisfaction). Beyond using mobility features as individual-level variables, the increasing availability of GPS data and spatial-analysis[1] tools offers opportunities to link spatial approaches with other psychological methods. For instance, researchers can study mobility from a situational perspective by comparing cognitive tests across key locations (e.g., home vs. work). Indeed, the ability to align objective spatial features with other approaches for understanding behavior and cognition underscores the growing potential for cross-disciplinary, convergent research on human mobility. For reviews and examples of the types of psychological studies that are well suited for using GPS data, we point readers to Harari et al. (2017), Hinds et al. (2022), and McInerney et al. (2013). Although the ways in which GPS data should be integrated with other measures will depend on the theoretical goals and study designs at hand, this tutorial describes the fundamentals of GPS data that underpin a variety of applications. Overall, we aim to illustrate how researchers can add location-based measures into their analytic tool kits without rewiring their overall agendas.

## Overview of Tutorial

This tutorial will show you how to preprocess GPS data, compute essential GPS-based variables, and navigate common challenges that are encountered when working with GPS data. To aid in this process, we provide an R Markdown tutorial file and a simulated data set, which are available at https://osf.io/rekuw. In this article, we walk through key steps covered in the R Markdown file and specify the associated section of the R Markdown in parentheses. Thus, we encourage interested readers to review this tutorial and R Markdown available on the OSF side by side to obtain a hands-on demonstration of conducting spatial analyses. All software needed to follow this tutorial are available for free. You will need R (https://cran.rstudio.com) installed on your computer, which we recommend using with the RStudio graphical interface (https://rstudio.com). For an introduction on how to use RStudio, see Venables et al. (2021). In addition, you will require a number of packages to get started (see R Markdown Chunk 1 "Setup").

## The Data Set

In practice, researchers might analyze an existing data set (e.g., the StudentLife data set; Wang et al., 2014), purchase data (e.g., from location data companies such as Foursquare or Cuebiq), or collect their own data through tracking devices or mobile applications. Common apps available for research as of 2022 include open-source apps, such as the AWARE framework and Funf; commercial solutions, such as Metricwire, Ksana Health, and Daynamica; or options offering both, such as Beiwe. Generally, Android devices provide developers with access to more data sources (see https://developer.android.com) than iOS (see https://www.apple.com/privacy). Although both operating systems currently enable GPS data collection, Android devices typically offer access to more data (e.g., screen and app usage). For information on the logistical considerations involved with running a study to collect GPS data, see Harari et al. (2016). Altogether, there are many factors that can shape a GPS data set that have implications for subsequent analyses and potential findings. At the end of this article, we provide a checklist of methodological information to consider and report when conducting studies using GPS data (see Fig. 3).

For the purposes of this tutorial, we created a simulated and simplified data set consisting of GPS data for four hypothetical participants (George, Jerry, Joe, and Josephine), who live and work in the urban area of Columbus, Ohio, USA. GPS data are typically a time-stamped series of latitude and longitude coordinates, and our data set is a representative example of how such data will look before data preprocessing (see Table 1). To provide a rich portrait of mobility behavior, we assumed that GPS data were collected using a near-continuous sampling strategy that sampled GPS coordinates every 10 s. However, we note that GPS data can be sampled near continuously (e.g., every second), periodically (e.g., every 10 min), or on an event-based sampling schedule (e.g., when the participant's location changes). The data-sampling approach will strongly influence the number of GPS records collected. Hence, whether data collection occurs at the temporal scale of seconds or hours, researchers need to be carefully attuned to how the recording rate may shape their analytic objectives.

In the tutorial data set, the time is represented as epoch time, or the number of seconds since midnight on January 1, 1970, which is a commonly used time format within operating systems. The level of accuracy is an index of how accurate the predicted GPS location is compared to the true location. An accuracy of 100 means that the true location has a 68% chance (i.e., ±1 *SD*) to fall within 100 m of the GPS location. In total, each participant had about 122,000 GPS records, and each participant had 14 days of data.

The four hypothetical participants were assumed to vary in their tendencies toward (a) taking trips away from home, and (b) visiting repeated (vs. novel) locations. For taking trips away from home, each participant was created with either a 10% (low tendency) or 80% (high tendency) likelihood of taking a trip when possible. We defined the

**Table 1.** First Rows of the Simulated Global Positioning System (GPS) Data Set Used in This Tutorial

| Name | User ID | Time stamp | Latitude | Longitude | Accuracy |
|---|---|---|---|---|---|
| George | 1010 | 1559451600 | 40.01512 | –83.02587 | 50 |
| George | 1010 | 1559451610 | 40.01509 | –83.02608 | 50 |
| George | 1010 | 1559451620 | 40.01531 | –83.02582 | 50 |
| George | 1010 | 1559451630 | 40.01515 | –83.02607 | 50 |
| George | 1010 | 1559451640 | 40.01520 | –83.02569 | 50 |
| George | 1010 | 1559451650 | 40.01515 | –83.02595 | 50 |

Note: Each row shows a single GPS record consisting of Unix epoch time stamp, geographical position in longitude and latitude coordinates, as well as their accuracy (in meters). An accuracy of 50 means that the true location has a 68% chance (i.e., ±1 SD) to fall within 50 m of the geographic coordinates provided.

possible time window for taking a trip as every 4-h period except for 12:00 a.m. to 4:00 a.m., when all participants were assumed to be sleeping. For visiting repeated locations, each participant was created with either a 10% (low repeats) or 80% (high repeats) likelihood of visiting a location he or she had previously visited. The permutations of these trip likelihoods created the behavioral tendencies for the four participants; the likelihoods were reflected in the user ID of each participant (see Table 2).

## Data Preprocessing

We begin with some basic data-management steps (see R Markdown Code Chunks 3–14 "Data Preprocessing"). It is important to recognize that there are many possibilities for researchers to consider when preprocessing raw GPS data. In addition to the steps outlined below, we encourage readers to consult Schuessler and Axhausen (2009) and Millard-Ball et al. (2019) for additional approaches. First, the data set needs to be downloaded (see R Markdown Code Chunk 2 "Load Data"). Some applications will store each user's GPS traces in a separate file so that each individual can be processed separately. In our example, the data for all individuals are found in the same file (see Table 1). Each row represents one recording of GPS data from one participant (Name/User ID) at a time (Time Stamp) and place (Latitude/Longitude) with a level of accuracy.

Second, researchers working with GPS data will need to identify and remove inaccurate records, which are likely to be present given the technical challenges associated with GPS data collection (e.g., the GPS signal might get obstructed by buildings, an area might not have enough satellite coverage, the phone might take time to acquire a signal after being turned on). Researchers must determine the level of accuracy (in meters) required for their research questions. This threshold can be set based on an inspection of the distribution of the accuracy values or the density of the study area; outliers could be removed, and denser areas may require greater accuracy to distinguish between locations. Here, we use 100 m for illustrative purposes. In addition, faulty records should be dropped, such as duplicate records or observations missing coordinates or time stamps (see R Markdown Code Chunks 3–5 "Exclusions").

Third, it is important to ensure the time stamps associated with the location coordinates are formatted so that they can be recognized correctly by R (see R Markdown Code Chunk 6 "Converting Time"). Here, we reformat the time stamp from epoch time to date time and ensure the correct time zone. We also create separate columns that index the day and hour for each record to ease future analytic steps involving time (see R Markdown Code Chunk 14 "Hour and Day Indices").

Fourth, we create a data-quality dataframe that shows how many "valid" hours of GPS data exist for each participant for each day, which allows us to remove people with insufficient data during analyses (see R Markdown Code Chunks 7–9 "Data Quality Dataframe"). Missing location data might occur because of a participant's phone running out of battery, getting switched off, or losing cellular service. Researchers may impute missing GPS data through interpolation or leveraging other data (Barnett & Onnela, 2020; Bohte & Maat, 2009; Carrel et al., 2015), or drop the cases in which missing data occurred. In turn, we use the data-quality data frame to filter out users or days that do not meet a certain threshold for a given research question. We set our threshold at 15 hr of GPS data per day so that participants could sleep for 9 hours during which GPS records may not be transmitted (for an application of this procedure, see Harari et al., 2019).

To do so, we first have to decide how many GPS traces are required to be considered a valid hour. This decision depends on the research question (e.g., the precision of the examined behaviors). It should also take into account that the number of GPS traces generated is highly dependent on the data-collection process (e.g., apps may have different sampling rates), the movement patterns of participants (e.g., being stationary may yield fewer records than traveling), and the technical platform (e.g., Android phones may record location data differently than iPhones). For the purposes of this tutorial, we do not need to capture fine-grained movement for our research goals; therefore, we set our cutoff to a minimum of one GPS recording per hour. As discussed above, the rate of real world GPS readings will vary significantly by data-collection method and research objectives. For instance, if researchers were interested in measuring moment-to-moment changes in

**Table 2.** Behavioral Tendencies in the Example Data Set

| Name | User ID | Trip frequency | Repeat visits |
|---|---|---|---|
| George | 1010 | Low | Low |
| Jerry | 1080 | High | Low |
| Joe | 8010 | Low | High |
| Josephine | 8080 | High | High |

physiological arousal, then a stricter threshold (e.g., one reading per 5 min) may be optimal.

Next, we generate the final data-quality data frame by counting the hours in which the number of GPS traces meets our cutoff for each day at the user level. We can then determine the number of days that have at least 15 hr of GPS data recorded (i.e., days that meet our minimum hourly cutoff) for each user in our data set. Or for analyses at the daily level, we could exclude participants who do not have a certain number of valid days. In our example, we set a minimum of 3 days for participants to be retained for subsequent analyses.

For the fifth and final preprocessing step, we remove geospatial outliers, such as GPS records outside of the study area. We assume that the research question focuses on mobility behaviors within a single city (Columbus, Ohio, USA), excluding movement outside of the surrounding area. Of course, the decision to restrict analyses to a single area may not always be applicable, but having data from a large geospatial area can complicate analyses. Once again, there are different approaches to narrow the scope of analysis and remove extraneous GPS points. For example, researchers can create an ellipse that captures a proportion (e.g., 68%) of all points for each participant, and excludes points outside of that area (Hinrichs et al., 2020).

Here, we eliminate all points that fall outside of the Columbus urbanized area (see R Markdown Code Chunks 10–13 “Study Area Boundary”). To accomplish this, we load a shapefile that includes the boundary of the study area. Shapefile is a file format that stores geometric data (Environmental Systems Research Institute, 2016), including projection information, which indicate the coordinate system. In our example, we reproject the data to the Universal Transverse Mercator coordinate system (Zone 17N [CRS 26917], which includes Columbus). Examples of projections and coordinate systems, and how to select a suitable projection are discussed in Lovelace et al. (2019):

```
urban <- urban_areas()
cbus <- urban %>%
 st_transform(26917)%>%
 filter(str_detect(NAMELSAD10, “Columbus,
   OH”))
```

Next, any GPS records falling outside of that area are removed:

```
b  <-  sapply(st_intersects(geodata,
  cbus),function(x){length(x)==0})
geodata <- geodata[!b,]
```

At this point, we can visualize the GPS points to verify that they are restricted to the study area and preview the data to ensure we are not overlooking other sources of noise. Because this is a computationally heavy task in R, we recommend doing this for a small subset of points. In Figure 1, we visualize all points belonging to User 8080 (Josephine), which confirms that all of her points fall within the city boundaries (i.e., the study area) and that travel paths are visible:

```
mapview(geodata[geodata$userID == 8080,])
+ mapview(cbus, alpha.regions = .1)
```

## Identifying Key Locations

GPS traces are granular and inherently scattered around the true location of a user (i.e., latitude and longitude coordinates are precise yet somewhat inaccurate estimates of a person’s true location). Therefore, raw location data cannot determine whether two GPS points are in the same “place.” Even a place like a person’s home will consist of myriad GPS points, and recordings of these points will differ at least slightly even if the person is still. We thus need to cluster the data points to determine key locations (see R Markdown Code Chunks 15–17 “Key locations”). Establishing such locations is the basis for computing many mobility features, such as the distance traveled or time spent at home. Without clustering, each individual GPS record would appear like a distinct location; computing the distance traveled between every GPS point would vastly overestimate the true distance traveled and vastly underestimate the time spent stationary.

In this case, we use the established density-based spatial clustering of applications with noise clustering algorithm (Ester et al., 1996). Other clustering algorithms include *k*-means clustering (Likas et al., 2003), hierarchical clustering (Johnson, 1967), and *k*-medoids clustering (Park & Jun, 2009). We define each key location as the centroid of an area consisting of at least 180 GPS points (i.e., approximately 30 min spent, assuming GPS is sampled every 10 s) within 25 m of their neighboring points. These are relatively strict parameters that are justifiable because the example (simulated) data are quite accurate and frequent. Researchers can test different values to determine the optimal threshold and ensure robustness given their data set and objectives.

First, a function for clustering is initialized:

```
db2 <- function(x) {
  geodata <- x %>% st_coordinates()
  cluster_20  =  dbscan::dbscan(geodata,
    eps = 25, minPts = 180)$cluster
 return  (data.frame(cluster_20m=clus
   ter_20))
}
```

Next, a function is applied to cluster GPS points by user:

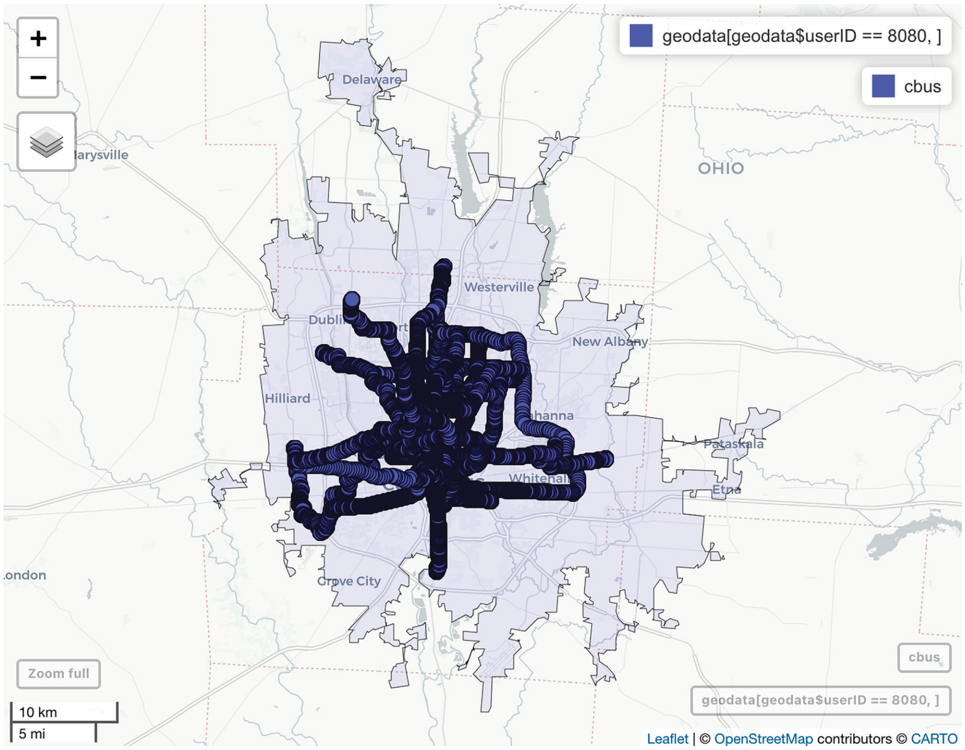


**Fig. 1.** Visualization of Josephine's global positioning system (GPS) recordings using R's *mapview* function.

```
geodata_cluster_df <- geodata %>%
  group_by(userID) %>%
  group_modify(~db2(.x)) %>%
  ungroup()
```

From this code, we now have a list that matches clusters to visited locations for each individual in the data set. With this information, we can identify and interpret key locations frequented in our participants' lives. These could include (a) a person's home, often defined as the place a person is most frequently at between 12 a.m. and 6 a.m. (see Saeb et al., 2015); (b) a person's workplace, often identified as the place a person is most frequently between typical office hours (see Hintze et al., 2017); or (c) the time a person spent in transit, which can be defined as the time spent outside of clusters (see Saeb et al., 2015).

For example, to identify the home, we can create a subset of the data with only night hours between 12:00 a.m. and 6:00 a.m.:

```
geodata_night <- with(geodata_clusters,
  geodata_clusters[geodata_clusters$hour
    .split >= 0 &
  geodata_clusters$hour.split < 6 ,])
```

We then calculate the modal location during the night for each person (i.e., the cluster most frequently inhabited during sleeping hours) to tag the participant's presumed home:

```
geodata_night <- geodata_night %>%
  group_by(userID) %>%
  mutate(clusterID, home = Mode(clusterID))
```

## Visualizing Mobility Patterns

One of the most interesting parts of collecting GPS traces is the ability to visualize naturalistic human behavior at the ground level. The mobility patterns for our four hypothetical participants are visualized in Figure 2, which displays the GPS recordings and key locations for each person. As shown, George and Joe move less overall than Jerry and Josephine. George makes more local trips, whereas Joe travels farther distances on his trips. Visualizing can help spot errors, generate ideas for follow-up analyses, and make illustrative points when presenting results.

## Computing Mobility Features

Building on the previous steps, we can compute features to quantify the overall movement patterns (see R Markdown Code Chunks 18–29 "Mobility Features"). Given the granularity and heterogeneity of our spatiotemporal data, a wide spectrum of features can be generated. For illustrative purposes, we show the code for computing two fundamental features that index individuals' mobility patterns: time spent at home and number of unique places visited (note that the R Markdown contains additional features: total distance

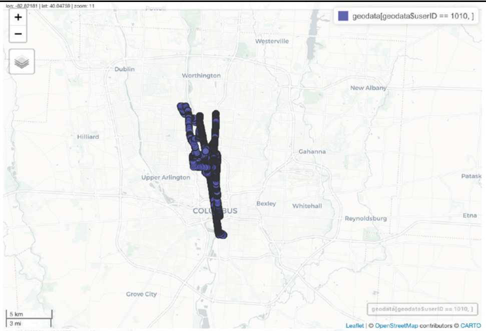
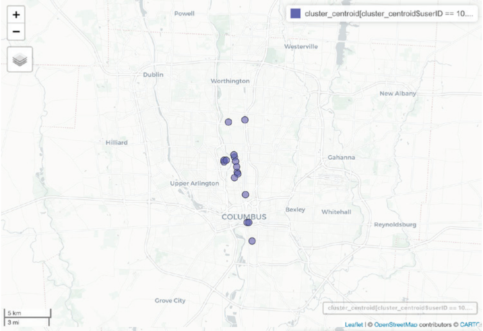
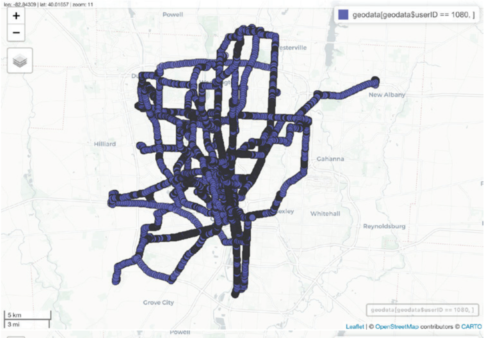
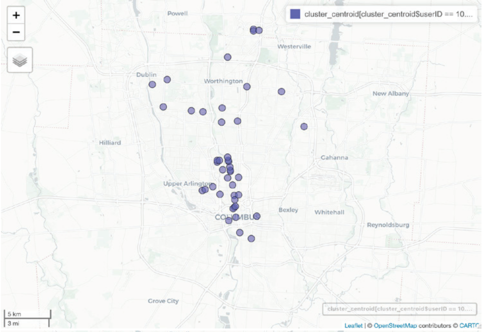
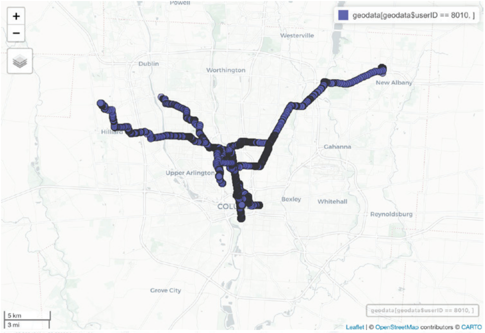
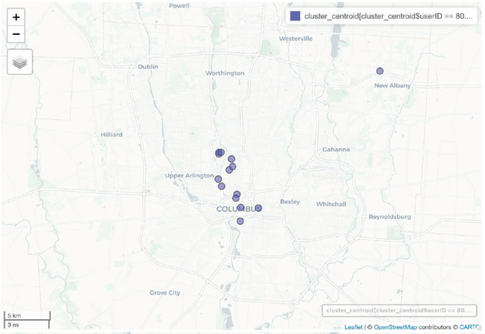
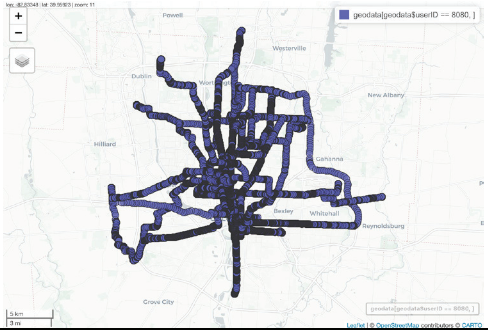
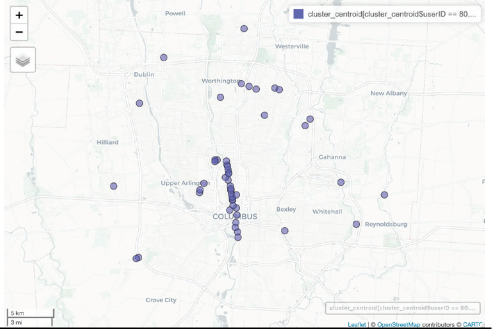
| Name | User ID | All GPS recordings | Key locations |
|---|---|---|---|
| George | 1010 |  |  |
| Jerry | 1080 |  |  |
| Joe | 8010 |  |  |
| Josephine | 8080 |  |  |

**Fig. 2.** Visualization of global positioning system (GPS) recordings and key locations.

traveled, number of places visited, time spent at key locations, time spent in transit, entropy, and routine index). Researchers can also consider a variety of other established mobility features both within and outside of psychology, such as confidence ellipses, kernel density estimates, and shortest paths networks (Müller et al., 2020; Schönfelder & Axhausen, 2003; Smith et al., 2019).

For each participant, the amount of time spent at home can be calculated by summing up the time spent in the computed home cluster (see R Markdown Code Chunk 23 "Time Spent at Home"):

```
for (i in 1:length(ids)){
  home <- geodata_clusters$home[geodata_
    clusters$userID == ids[i]][1]
  geodata_clusters$timeSpentAtHome[
    geodata_clusters$userID == ids[i]] <-
    geodata_clusters$timeSpent[geodata_
      clusters$clusterID == home[1]]
}
```

**Table 3.** Summary Descriptive Statistics for Participants in the Example Data Set

| Name | User ID | Distance traveled (in kilometers) | Unique places | Places visited | Time spent at home (in minutes) | Travel time (in minutes) |
|---|---|---|---|---|---|---|
| George | 1010 | 257<br>(4) | 16<br>(3) | 171<br>(3) | 638,330<br>(2) | 153,220<br>(3) |
| Jerry | 1080 | 1115<br>(1) | 43<br>(2) | 288<br>(2) | 179,050<br>(4) | 450,130<br>(1) |
| Joe | 8010 | 279<br>(3) | 14<br>(4) | 134<br>(4) | 675,520<br>(1) | 125,810<br>(4) |
| Josephine | 8080 | 877<br>(2) | 49<br>(1) | 496<br>(1) | 192,840<br>(3) | 362,850<br>(2) |
| *M* | | 632 | 31 | 272 | 421,435 | 273,003 |
| *SD* | | 431 | 18 | 163 | 272,402 | 158,598 |

Note: Raw mobility features are presented with ranks in parentheses. Shading reflects ranks ranging from dark blue (1) to light blue (4). Average number of global positioning system (GPS) records = 122,170 (*SD* = 0.5), and each participant has 14 days of data. Places visited differs from unique places visited because the former might include duplicate places (e.g., a participant returned home multiple times).

Next, we compute the number of unique locations, which corresponds to the total number of distinct clusters recorded for each participant (see R Markdown Code Chunk 19 “Unique Locations”):

```
ids <- unique(geodata_clusters$userID)
for (i in 1:length(ids)){
  geodata_clusters_user  <-  geodata_
    clusters[geodata_clusters$userID ==
    ids[i],]
  geodata_clusters$uniqueLocations[geoda
    ta_clusters$userID == ids[i]]
   <-  length(unique(geodata_clusters_
     user$clusterID))
}
```

Once the mobility features have been extracted, descriptive statistics can be computed to understand individual differences in mobility behavior. In Table 3, we present such descriptive statistics for each participant in our example data set, aggregated over the 14 days. Table 3 includes the two features described above (time spent at home, unique locations visited) and a subset of three other features included in the accompanying R Markdown file (distance traveled, places visited, travel time). George moved the least and spent the majority of time at home and visiting relatively few places. Similar to George, Joe spent most of his time at home and the least time traveling. Moreover, Joe typically visited the same places when leaving his home; consequently, he visited the fewest unique places. On the other hand, Jerry spent the least amount of time at home and the most time on the go. Although Josephine was simulated to be more likely to visit places repeatedly, she actually went to more unique places than Joe and visited nearly twice as many places overall. Overall, the summary statistics show how mobility features can parsimoniously quantify variability in human mobility.

## Interpreting Mobility Patterns

After computing mobility features, the next challenge is to interpret what those variables mean—and relate them to meaningful psychological factors. In this tutorial, we have concentrated on describing behavioral tendencies (e.g., Jerry’s propensity to be on the go), highlighting how location data can be used to assess individual differences in daily mobility. These behavioral tendencies can influence variables of interest to psychologists and other social scientists. Depending on the research question at hand, mobility features might cluster together in psychologically meaningful factors (Fillekes et al., 2019; Müller et al., 2020). For example, Müller et al. (2020) found that a factor consisting of distance-based features was associated with individual differences in anxiety, affect, and stress.

However, when interpreting mobility patterns and their psychological implications, it is important to account for fundamental factors that shape human movement and transportation behavior. For example, mobility patterns are likely to be influenced by whether the GPS traces were collected on a weekday or weekend. In addition, practical factors related to individuals’ career or life responsibilities can constrain daily mobility. Such factors may systematically influence data collection for certain participants (e.g., people who work from home vs. people who commute). It may be difficult to explain the intentions behind personal trips that the spatial data alone cannot reveal. Trip purposes will vary widely depending on individual differences, responsibilities, and lifestyles. For example, a trip to the local tennis courts could represent a leisure activity for one participant and a work shift for another. Another challenge involves matching timeframes in a way that is appropriate for the specific analysis (e.g., linking emotional experiences to specific locations and movements). In total, studying the psychological mechanisms of real world human mobility necessitates a careful consideration of contextual factors.

The potential uncertainty associated with these inferences highlights the value of linking GPS data to other data sources. Researchers might want to link GPS features to mobility-focused measures via surveys, daily diaries, or experience sampling, which offers the potential to cross-validate mobile data or gauge mechanisms behind recorded movement. Alternatively, researchers can draw on location-based databases (e.g., Google Maps, Foursquare, or OpenStreetMap). For instance, the Google Maps Places Library API (Google, 2021) can be used to retrieve an address or nearby points of interest (e.g., nearby coffee shops) for GPS coordinates (see R Markdown Code Chunks 30–31 “API Pull”):

```
google_places(search_string = "Coffee
  shop",
                  location=c(40.0151208778698,
                    -83.0258748810443),
                  radius = 2000)
```

Note that this approach might be affected by whether data were collected in a rural versus an urban area because it is easier to identify a person’s specific location in sparse areas. However, with future advancements in the accuracy of GPS-based methods, we may be able to overcome common challenges such as recognizing infrequent locations in dense areas and correctly inferring their semantic meaning for participants (e.g., see also Do & Gatica-Perez, 2013). Researchers can also link their GPS data to a range of other contextual information to understand other factors that may drive mobility behavior. For example, census-based socioeconomic data such as population, economic, and poverty statistics are available from the U.S. Census Bureau (2021). Another possible source of information is weather data (e.g., temperature, precipitation, wind) available from the National Centers for Environmental Information (2021) and OpenWeather (2021). By triangulating multiple data sources, researchers will be best positioned to unpack the psychological underpinnings of mobility patterns.

## Challenges to Anonymity and Privacy

Despite the potential of GPS data for psychological research, a number of ethical and practical challenges accompany the usage of location data for scientific purposes. In particular, researchers must be highly attuned to the potential for privacy violations given the specificity and sensitivity of GPS data. Thus, careful consideration should be given to how data can be anonymized and secured to minimize the risks to participant privacy.

An assortment of options for protecting participant privacy when sharing data have been discussed by the scientific community. These include removing GPS coordinates of personal locations and assigning labels instead (e.g., home and work) and providing features (e.g., distance traveled) instead of GPS coordinates. These practices reduce the risk that the data can be deanonymized while also providing the most theoretically relevant data. Indeed, many studies using location data depend on computed features rather than raw GPS records. Another approach is to store location data separately from other types of data collected because privacy risks are higher with GPS data. Participants could also be given the option to redact some or all of their data (Harari et al., 2020).

Alternatively, prior work has suggested a number of data-obfuscation strategies, including (a) adding noise, (b) obscuring time stamps, (c) removing decimals (de Montjoye et al., 2013), and/or (d) using “data guardians” to store and monitor data sets. Determining which of these strategies is appropriate will depend on sensitivity of the research questions and availability of the data set. For example, a study investigating the personality correlates of daily travel distance might choose to add noise to GPS coordinates because exact locations are not needed to examine the research question.

At the same time, it is important to balance the requirements of anonymity and privacy with the principles of open science and reproducibility. Open science has been invaluable to further the use and development of GPS-based features and will likely continue to be (see Vega et al., 2020). Thus, researchers should look for ways to advance open science and protect participant privacy in parallel. To that end, we urge researchers to seek advice from their institutional review board, and familiarize themselves with applicable local laws and regulations (e.g., the EU’s General Data Protection Regulation). As this area matures, the psychological research community should develop formal guidelines for balancing the opportunities and risks that come with GPS data. Ultimately, however, participant privacy should be the top priority.

## Summary

In this tutorial, we provided a practical overview of how to leverage GPS data for psychological research, including steps for cleaning and filtering data, identifying frequent and key locations, and extracting features to quantify individual differences in mobile behavior. In

**Information about how the GPS data were collected:**

- ○ Number of participants in the study (recruited, dropped out, excluded, and retained in the final sample used for analyses)
- ○ Geographical area covered (by design or characteristics of the data set)
- ○ Sampling frequency (how often latitude and longitude coordinates were collected)
- ○ Type of sampling strategy (event based, periodically)
- ○ Accuracy of GPS points
- ○ Duration of data collection
- ○ Operating system used to collect the data
- ○ App/data collection setup used
- ○ Number of GPS records collected (in total, per participant, per day)

**Information about how the GPS data were preprocessed and analyzed:**

- ○ Exclusions of individual records due to:
  - ○ Low accuracy records (number of records excluded, thresholds used, and rationale)
  - ○ Invalid/impossible records (number of records excluded, suspected reason for errors)
  - ○ Outlier records (number of records excluded, how they were identified)
- ○ Exclusions of geographical areas due to:
  - ○ Privacy reasons (e.g., residential areas)
  - ○ Analytical reasons (distortion of metrics such as distance covered by outliers)
  - ○ Design reasons (if an area is not of geographical interest to a study)
- ○ Exclusions of participants due to data thresholds for inclusion in the study (minimum number of records for time periods such as days or hours)
- ○ Computation (e.g., clustering algorithms)
- ○ Metrics included (e.g., total distance)
- ○ Statistical techniques employed (e.g., multilevel modeling)

**Fig. 3.** Checklist of methodological information to report for GPS research.

addition, we discussed some of the challenges and opportunities for integrating spatial variables with other methods—all while seeking to maximize participant privacy and support open-science practices. We also outlined a reporting checklist (see Fig. 3) to aid researchers in documenting GPS-based studies and to support the replicability of this burgeoning area of psychology. Beyond this high-level summary, the R Markdown file that accompanies this tutorial (along with our simulated data set of four participants) offers a more detailed and comprehensive manual for working with GPS data. Although there are many more ways to use spatial data than we have covered here, we are hopeful that our tutorial affirms how GPS methods are within grasp to a widening set of researchers in psychology and beyond.

## Transparency

*Action Editor:* Frederick L. Oswald

*Editor:* Daniel J. Simons

*Author Contributions*

S. R. Müller and J. B. Bayer generated the idea for the research project, and all authors assisted in the tutorial conceptualization. S. R. Müller, M. Q. Ross, J. Mount, and H. T. K. Le wrote the R Markdown script and conducted the analyses and simulations. S. R. Müller, J. B. Bayer, M. Q. Ross, and H. T. K. Le wrote the manuscripts, which all authors reviewed and edited.

*Declaration of Conflicting Interests*

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

*Funding*

G. M. Harari and C. Stachl thank the National Science Foundation for an award that supported their work on this research (Award SES-1758835).

*Open Practices*

Open Data: https://osf.io/2d5ep

Open Materials: https://osf.io/2d5ep

Preregistration: not applicable

All data and materials have been made publicly available via OSF and can be accessed at https://osf.io/2d5ep. This article has received the badge for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

## ORCID iDs

Sandrine R. Müller https://orcid.org/0000-0002-1226-6370

Joseph B. Bayer https://orcid.org/0000-0002-6555-4472

## Acknowledgments

We thank Weichen Wang, Ruth Appel, and Quay Au for providing feedback on earlier versions of this article and discussing technical questions.

## Note

1. Note that these established analysis approaches differ from spatial statistics, which aims to account for geographical dependence and heterogeneity when modeling.

## References

Alessandretti, L., Lehmann, S., & Baronchelli, A. (2018). Understanding the interplay between social and spatial behaviour. *EPJ Data Science*, *7*(1), 1–17. https://doi.org/10.1140/epjds/s13688-018-0164-6

Barnett, I., & Onnela, J. P. (2020). Inferring mobility measures from GPS traces with missing data. *Biostatistics*, *21*(2), e98–e112. https://doi.org/10.1093/biostatistics/kxy059

Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, *17*(3), 285–297. https://doi.org/10.1016/j.trc.2008.11.004

Carrel, A., Lau, P. S., Mishalani, R. G., Sengupta, R., & Walker, J. L. (2015). Quantifying transit travel experiences from the users' perspective with high-resolution smartphone and vehicle location data: Methodologies, validation, and example analyses. *Transportation Research Part C: Emerging Technologies*, *58*, 224–239. https://doi.org/10.1016/j.trc.2015.03.021

Cornet, V. P., & Holden, R. J. (2018). Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of Biomedical Informatics*, *77*, 120–132. https://doi.org/10.1016/j.jbi.2017.12.008

Crato, N. (2010). *Figuring it out: Entertaining encounters with everyday math*. Springer. https://doi.org/10.1007/978-3-642-04833-3

de Montjoye, Y-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, *3*, 1–5. https://doi.org/10.1038/srep01376

Do, T. M. T., & Gatica-Perez, D. (2013). The places of our lives: Visiting patterns and automatic labeling from longitudinal smartphone data. *IEEE Transactions on Mobile Computing*, *13*(3), 638–648. https://doi.org/10.1109/TMC.2013.19

Environmental Systems Research Institute. (2016). *What is a shapefile?* https://desktop.arcgis.com/en/arcmap/10.3/manage-data/shapefiles/what-is-a-shapefile.htm

Ester, M., Kriegel, H-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, & U. Fayyad (Eds.), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (Vol. 96, pp. 226–231). Association for Computing Machinery. https://doi.org/10.5555/3001460.3001507

Fillekes, M. P., Giannouli, E., Zijlstra, W., & Weibel, R. (2018). Towards a framework for assessing daily mobility using GPS data. *GI_Forum*, *6*(1), 177–186. https://doi.org/10.1553/giscience2018_01_s177

Fillekes, M. P., Röcke, C., Katana, M., & Weibel, R. (2019). Self-reported versus GPS-derived indicators of daily mobility in a sample of healthy older adults. *Social Science & Medicine*, *220*, 193–202. https://doi.org/10.1016/j.socscimed.2018.11.010

Google. (2021). *Places library*. https://developers.google.com/maps/documentation/javascript/places

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, *11*(6), 838–854. https://doi.org/10.1177/1745691616650285

Harari, G. M., Müller, S. R., Aung, M. S. H., & Rentfrow, P. J. (2017). Smartphone sensing methods for studying behavior in everyday life. *Current Opinion in Behavioral Sciences*, *18*, 83–90. https://doi.org/10.1016/j.cobeha.2017.07.018

Harari, G. M., Müller, S. R., & Gosling, S. D. (2020). Naturalistic assessment of situations using mobile sensing methods. In J. F. Rauthmann, R. A Sherman, & D. C. Funder (Eds.), *The Oxford handbook of psychological situations* (pp. 299–311). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190263348.013.14

Harari, G. M., Müller, S. R., Stachl, C., Wang, R., Wang, W., Bühner, M., Rentfrow, P. J., Campbell, A. T., & Gosling, S. D. (2019). Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. *Journal of Personality and Social Psychology*, *119*(1), 204–228. https://doi.org/10.1037/pspp0000245

Hinds, J., Brown, O., Smith, L. G. E., Piwek, L., Ellis, D., & Joinson, A. (2022). Integrating insights about human movement patterns from digital data into psychological science. *Current Directions in Psychological Science*, *31*(1), 88–95. https://doi.org/10.1177/09637214211042324

Hinrichs, T., Zanda, A., Fillekes, M. P., Bereuter, P., Portegijs, E., Rantanen, T., Schmidt-Trucksäss, A., Zeller, A. W., & Weibel, R. (2020). Map-based assessment of older adults' life space: Validity and reliability. *European Review of Aging and Physical Activity*, *17*(1), 1–9. https://doi.org/10.1186/s11556-020-00253-7

Hintze, D., Hintze, P., Findling, R. D., & Mayrhofer, R. (2017). A large-scale, long-term analysis of mobile device usage characteristics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(2), 1–21. https://doi.org/10.1145/3090078

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254. https://doi.org/10.1007/BF02289588

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global *k*-means clustering algorithm. *Pattern Recognition*, *36*(2), 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2

Lovelace, R., Nowosad, J., & Muenchow, J. (2019). *Geocomputation with R*. CRC Press. https://doi.org/10.1201/9780203730058

McInerney, J., Stein, S., Rogers, A., & Jennings, N. R. (2013). Breaking the habit: Measuring and predicting departures from routine in individual human mobility. *Pervasive and Mobile Computing*, *9*(6), 808–822. https://doi.org/10.1016/j.pmcj.2013.07.016

Millard-Ball, A., Hampshire, R. C., & Weinberger, R. R. (2019). Map-matching poor-quality GPS data in urban environments: The pgMapMatch package. *Transportation Planning and Technology*, *42*(6), 539–553. https://doi.org/10.1080/03081060.2019.1622249

Müller, S. R., Peters, H., Matz, S. C., Wang, W., & Harari, G. M. (2020). Investigating the relationships between mobility behaviours and indicators of subjective well–being using smartphone–based experience sampling and GPS tracking. *European Journal of Personality*, *34*(5), 714–732. https://doi.org/10.1002/per.2262

National Centers for Environmental Information. (2021). *Climate data online: Web services documentation*. https://www.ncdc.noaa.gov/cdo-web/webservices/v2

OpenWeather. (2021). *Weather API*. https://openweathermap.org/api

Park, H-S., & Jun, C-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, *36*(2), 3336–3341. https://doi.org/10.1016/j.eswa.2008.01.039

Rauthmann, J. F. (2021). Capturing interactions, correlations, fits, and transactions: A person-environment relations model. In J. F. Rauthmann (Ed.), *The handbook of personality dynamics and processes* (pp. 427–522). Elsevier. https://doi.org/10.1016/B978-0-12-813995-0.00018-2

Ross, M. Q., Müller, S., & Bayer, J. B. (in press). The psychology of mobile technology and daily mobility. In S. Matz (Ed.), *The psychology of technology*. American Psychological Association.

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research*, *17*(7), 1–11. https://doi.org/10.2196/jmir.4273

Schönfelder, S., & Axhausen, K. W. (2003). Activity spaces: Measures of social exclusion? *Transport Policy*, *10*(4), 273–286. https://doi.org/10.1016/j.tranpol.2003.07.002

Schuessler, N., & Axhausen, K. W. (2009). Processing raw data from global positioning systems without additional information. *Transportation Research Record*, *2105*(1), 28–36. https://doi.org/10.3141%2F2105-04

Smith, L., Foley, L., & Panter, J. (2019). Activity spaces in studies of the environment and physical activity: A review and synthesis of implications for causality. *Health & Place*, *58*, Article 102113. https://doi.org/10.1016/j.healthplace.2019.04.003

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., & Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences, USA*, *117*(30), 17680–17687. https://doi.org/10.1073/pnas.1920484117

U.S. Census Bureau. (2021). *Available APIs*. https://www.census.gov/data/developers/data-sets.html

van Diggelen, F., & Enge, P. (2015). The world's first GPS MOOC and worldwide laboratory using smartphones. In *Proceedings of the 28th International Technical Meeting of the Satellite Division of the Institute of Navigation* (pp. 361–369). https://www.ion.org/publications/abstract.cfm?articleID=13079

Vega, J., Li, M., Aguillera, K., Goel, N., Joshi, E., Durica, K. C., Kunta, A. R., & Low, C. A. (2020). *RAPIDS: Reproducible Analysis Pipeline for Data Streams Collected with mobile devices*. JMIR Preprints. https://doi.org/10.2196/preprints.23246

Venables, W. N., & Smith, D. M., & R Development Core Team. (2021). *An introduction to R*. https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., & Campbell, A. T. (2014). StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In A. J. Brush (Ed.), *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 3–14). Association for Computing Machinery. https://doi.org/10.1145/2632048.2632054