GPS2space: An Open-source Python Library for Spatial Measure Extraction from GPS Data

Shuai Zhou[1], Yanling Li[1], Guangqing Chi[1], Junjun Yin[1], Zita Oravecz[1], Yosef Bodovski[1], Naomi P. Friedman[2], Scott I. Vrieze[3], and Sy-Miin Chow[1]

[1] The Pennsylvania State University, University Park, PA 16801, USA
sxz217@psu.edu
[2] University of Colorado Boulder, Boulder, CO
[3] University of Minnesota, Minneapolis, MN

Abstract. Global Positioning System (GPS) data have become one of the routine data streams collected by wearable devices, cell phones, and social media platforms in this digital age. Such data provide research opportunities in that they may provide contextual information to elucidate where, when, and why individuals engage in and sustain particular behavioral patterns. However, raw GPS data consisting of densely sampled time series of latitude and longitude coordinate pairs do not readily convey meaningful information concerning intra-individual dynamics and inter-individual differences; substantial data processing is required. Raw GPS data need to be integrated into a Geographic Information System (GIS) and analyzed, from which the mobility and activity patterns of individuals can be derived, a process that is unfamiliar to many behavioral scientists. In this tutorial article, we introduced GPS2space, a free and open-source Python library that we developed to facilitate the processing of GPS data, integration with GIS to derive distances from landmarks of interest, as well as extraction of two spatial features: activity space of individuals and shared space between individuals, such as members of the same family. We demonstrated functions available in the library using data from the Colorado Online Twin Study to explore seasonal and age-related changes in individuals' activity space and twin siblings' shared space, as well as gender, zygosity and baseline age-related differences in their initial levels and/or changes over time. We concluded with discussions of other potential usages, caveats, and future developments of GPS2space.

Keywords: spatial measure · twins · behavior genetics · latent growth curve model · Python.

1 Introduction

Spatial analysis is used to explain locations, attributes, and relationships of features in spatial data and has increasingly become a subject of interest in many social and behavioral science disciplines including psychology, sociology, demography, and environmental science (Chi & Zhu, 2019; Sui & Goodchild, 2011). The past three decades have witnessed the emergence and substantial growth of using spatial analysis to investigate environmental effects on behavioral changes and population dynamics. Many earlier analyses of spatial and mobility patterns were based mostly on self-reports, surveys, or administrative data (Chi & Marcouiller, 2013; Kestens et al., 2012; Vallée, Cadot, Roustit, Parizot, & Chauvin, 2011). For example, participants were usually asked to draw a map displaying their daily mobility patterns or provide locations they frequently visited in their daily routines. Recent advances in mobile technology tools (e.g., smartphones, wearable sensors) now allow researchers to collect physical location data in real-time over very short intervals (e.g., across seconds or minutes) (Kerr, Duncan, & Schipperjin, 2011; Kestens, Thierry, Shareck, Steinmetz-Wood, & Chaix, 2018; Russell, Almeida, & Maggs, 2017). Such intensive and continuous location data streams provide contextual information to elucidate the context in which (e.g., where, when, and why) individuals engage in and sustain particular behavioral and lifestyle patterns. However, the central focus of many studies in the social and behavioral sciences not only examines individuals' short-term spatial activities over hours or days, but also those that may extend over weeks, months, or even years, as well as across large populations. In such scenarios, as in the case of the Colorado Online Twin Study (CoTwins) used for demonstration in this study, the sheer quantity and density of the longitudinal Global Positioning System (GPS) data (approximately 6.65 million points from June 2016 to December 2018) make the spatial measure extraction via conventional and non-programmable spatial analysis tools highly impractical, inefficient, and irreproducible. In this article, we introduced GPS2space, a user-friendly Python package that can be used to facilitate and automate the processes of spatial data building, activity and shared space measure extraction, and fast distance query.

Myriad spatial and aspatial measures can be extracted from raw physical location data or social network data. One measure that has been found to be a useful lifestyle indicator is activity space, which has been used in studies of obesity, substance use, and mental health. Generally, these studies treat activity space as the space within which an individual engages in routine activities. This space measure may be quantified subjectively via individuals' self-reports (Buchowski, Townsend, Chen, Acra, & Sun, 1999), or objectively via location data (N. C. Lee et al., 2016). For example, using a representative sample from the Paris metropolitan area of France, Vallée et al. (2011) explored the relationship between depression and activity space as measured by individuals' daily activities. They found that depression was related to limited activity space and neighborhood characteristics such as deprivation status. Mason et al. (2010) constructed activity space from 301 Philadelphia adolescents' place-based social networks, and found that adolescents' substance use depended on their activity space, as moderated by participants' age and gender.

Another measure is shared space, which can be spatial or aspatial depending on disciplines and research questions. From a social science perspective, shared space refers to the socio-psychological or physical space within which individuals share a common identity and social belonging (Cleaveland & Kelly, 2008; Fine, 2012), or a common physical area. Studies have shown that shared space, such as coworking space shared by independent professionals, can provide social support (Gerdenitsch, Scheel, Andorfer, & Korunka, 2016). Shared space also increases neighborhood satisfaction and sense of community (Kearney, 2006).

In this study, we define an individual's activity space as the area of the minimum bounding geometry consisting of routine locations visited by the individual over a specific period of time (i.e., daily, weekly, or monthly). Accordingly, we define shared space as the overlapping areas of two individuals' activity spaces. Activity space depends on the spatial distributions of the geolocations: geolocations spanning larger areas and broader geographical regions would give rise to higher values of activity space. In contrast, geolocations that are concentrated around certain places such as home and working place would yield smaller activity space. Shared space is not necessarily linearly related to activity space because the latter is determined by the extent to which two individuals' activity spaces overlap with each other, in other words, how much they share the same area within their activity spaces.

Despite the richness of information available in location data, the mapping of raw data consisting of latitude and longitude coordinate pairs to landmarks of inferential interest requires reverse geocoding. Reverse geocoding is the process of converting machine-readable GPS coordinates into location information for geoprocessing, such as the nearest distance query, as well as specialized spatial feature extraction procedures (Yin et al., 2020). These procedures are typically implemented via specialized spatial software that may not be familiar or accessible to many social and behavioral scientists (McCormick, Lee, Cesare, Shojaie, & Spiro, 2017; Shelton, 2017; Shelton, Poorthuis, & Zook, 2015). Commercial software such as ArcGIS, TransCAD, and MapInfo (Drummond & French, 2008; Murray, Xu, Wang, & Church, 2019) are available and relatively easy to use. However, licensing restrictions may prevent broad dissemination of methodological advances and reproducibility of analytic results, and these programs are not readily available on High Performance Computing (HPC) platforms used to process data and perform large-scale analyses. ArcGIS and an open-source software, QGIS, are programmable, but their programming environments are not well developed. In contrast, R is an open-source programmable statistical language whose usage has been increasing in social and environmental sciences (Bivand, 2006). However, R poses known challenges in handling very large data sets, and often performs less satisfactorily in terms of memory management and computational speed (Patil, 2016). Taking into consideration computational speed, ease of usage, and open-source availability, we developed GPS2space in Python, a popular open-source programming language among researchers and data scientists.

The objectives of this tutorial are to introduce and demonstrate the use of GPS2space, a new, open-source Python library that we created to facilitate the construction of spatial data, simplify extraction of mobility-related measures such as activity space and shared space, and boost the nearest distance query for big data. GPS2space builds upon existing functions and includes all the necessary, tunable parameters as arguments for generating spatial measures in a straightforward and well-documented package that can be readily implemented by newer users. We used the terms library, package, and toolbox interchangeably throughout the article, as these terms all refer to reusable chunks of code but are used differently in different conventions. Likewise, we used the terms methods and functions interchangeably, in that they both refer to snippets of a library/package/toolbox that are used for specific purposes.

The remainder of the article proceeds as follows. First, we briefly introduce commonly used Python libraries for managing and analyzing GPS data and highlight the contributions of GPS2space. Then, we illustrate the utility of the GSP2space library using the CoTwins data to extract the twin siblings' activity space and shared space. These measures are used to address questions related to seasonal, age-based, gender, and zygosity effects in shaping individuals' activity space and shared space. Finally, we conclude with discussions on other potential usages, caveats, and future developments of GPS2space.

2 Contributions of GPS2space Relative to Other Commonly Used Spatial Python Packages

Like many data analysis procedures, geospatial analyses involve data reading and writing, data managing and processing, and visualization. Beyond that, geospatial analyses also deal with spatial projection and operation, Exploratory Spatial Data Analysis (ESDA), and spatial modeling. There are existing Python libraries that focus on certain specific functions useful for geospatial analysis – a brief overview is provided next.

Geospatial Data Abstraction Library (GDAL/OGR contributors, 2020) specializes in reading and writing raster and vector data, which are the two commonly used data types in GIS. It supports 168 raster data formats and 99 vector data formats at the time of writing (October 2020). Fiona (Gillies et al., 2011) and Rasterio (Gillies et al., 2013), two other popular libraries in Python, focus on reading, writing, and manipulating vector and raster data, respectively. Pyproj exclusively focuses on cartographic projections and coordinate transformations (Crickard, Toms, & Rees, 2018). Shapely specializes in spatial operations such as distance query and intersecting and overlapping analyses (Gillies et al., 2007). Python Spatial Analysis Library (PySAL) is the most commonly used library in conducting ESDA and spatial modeling (Rey, 2019; Rey & Anselin, 2007). GeoPandas, on the other hand, combines Pandas, a widely used Python data analysis library, and GIS science, providing a wide array of geospatial functions such as spatial operation, spatial projection transformation, and visualization (Jordahl, 2014). These packages are often used together to conduct a series of data managing, manipulation, visualization, and modeling tasks. For example, GeoPandas relies on Fiona to read and write spatial data and PyProj to perform spatial projection transformations. Rasterio also uses PyProj for its projection functionalities.

The packages reviewed thus far do have limitations, especially for novices who do not have strong background in programming and GIS. For example, Shapely does not provide options for coordinate system transformations, so the original units of distance and area measures are usually degrees, which may not be intuitive for non-specialist audiences. GeoPandas incorporates many useful geoprocessing methods and spatial analysis techniques and provides foundational functions for such spatial operations; however, it assumes users have GIS and programming background to perform the analyses. For example, to calculate the area of a polygon from GPS data with latitude and longitude coordinate pairs using GeoPandas, a researcher has to first build a spatial data set, project it to an appropriate coordinate reference system (CRS), and then calculate the area.

Even though we did not provide an exhaustive list of all the Python packages that can perform geospatial manipulation and analysis, we highlighted that almost all of these packages are tailored for experts with considerable spatial data handling and GIS experience, and require function customizations in multiple steps. For novices such multi-step data pre-processing and function customization processes can be challenging and error-prone. In addition, none of the above packages provides immediately available functions for constructing activity space and shared space.

In this article, we introduced GPS2space with the aim to facilitate and automate, whenever possible, the processes of spatial data building, activity and shared space measure extraction, and distance query. Specifically, GPS2space has three functionalities: (1) building unprojected spatial data from geolocations with latitude and longitude coordinate pairs using the geodf function; (2) constructing buffer- and convex hull-based activity space and shared space at different timescales using the space function; and (3) performing nearest distance query using the dist function, which incorporates cKDTree [1] and spatial indexing and R-Tree [2] algorithms to decrease execution time. GPS2space provides an easily replicable and open-source solution to building spatial data directly from latitude and longitude coordinate pairs. It also provides default parameterizations suited for many longitudinal spatial data streams that can be used to simplify and reduce the specification steps needed for extraction of activity- and shared-space-related and distance measures included in the package. GPS2space enables transparent and easily replicable ways to change these default options for experienced GIS scientists and programmers to perform custom specifications.

[1] cKDTree is a function from SciPy, a commonly used library for scientific computing in Python. cKDTree is used to rapidly look up the nearest neighbors of any point and can dramatically reduce the time needed for such processes.

[2] GeoPandas incorporated spatial indexing using the R-tree algorithm to boost the performance of spatial queries. R-tree is a tree-like data structure that groups nearby objects together along with their minimum bounding box. In this tree-like data structure, spatial queries such as finding the nearest neighbor does not have to travel through all geometries, dramatically increasing performance, especially for two data sets with different bounding boxes.

These spatial measures provide additional contextual information and expand the usages of GPS data. In sum, GPS2space provides an open-source tool to consolidate, simplify, and automate data processing and spatial measure extraction from large (e.g., intensive longitudinal) GPS data sets. In this way, replicability and reproducibility of results can be greatly enhanced – for veteran and novice researchers alike.

3 Motivating Data: The CoTwins Study

We used data from the CoTwins study to illustrate the utility of GPS2space and demonstrate how spatial activity measures can shed light on individual and dyadic activity patterns between twin siblings. Twin studies have the advantage of disentangling genetic and environmental factors for the trait of interest (Newman, Freeman, & Holzinger, 1937). Despite the increasing application of spatial thinking and spatial data in social and behavioral research, few twin studies have been designed to collect twins' location data, which often convey valuable information concerning social contexts. For instance, shared activity space and time spent with each other reflect opportunities for relationship bonding, and may thus convey the extent of emotional closeness between two individuals (Ben-Ari & Lavee, 2007). Furthermore, with twins' location data, it would be interesting to investigate how monozygotic (MZ; identical) twins and dizygotic (DZ; fraternal) twins differ in their shared activity space.

The CoTwins study comprises data on substance use among 670 twins. Twins were initially recruited at ages 14 to 17 and followed from 2015 to 2018. Throughout 2016 to 2018, the twins' geolocations were recorded and reported via their GPS enabled smartphones. iOS devices used the built-in significant-change location service to record and report geolocations whenever they detected a significant position change of 500 meters or more. Android devices recorded and reported geolocations every five minutes as long as the device was in use. Over the course of the study, the twins' spatial footprints covered locations within and outside of the United States. In this article, we only used locations in the contiguous United States, which includes the District of Columbia but excludes Alaska and Hawaii.

Figure 1 shows the spatial distribution of the twins' footprints in 2016, 2017, and 2018 across Colorado and the contiguous United States. The CoTwins study began collecting locations in June 2016 so the figure shows fewer data points in 2016. Throughout 2017 and 2018, the twins set foot in almost every state of the contiguous United States and showed a consistent pattern of footprints concentrated in Colorado and all over parts of the contiguous US, with North Dakota, Arkansas, and Alabama as the least visited states. In Colorado in 2017 and 2018 they showed consistent mobility patterns with geolocations clustered around metropolitan areas such as Denver and Colorado Springs and along major roads within the state. The border counties in Colorado such as Moffat, Rio Blanco, Yuma, Cheyenne, Kiowa, and Baca were rarely visited. The code for Figure 1 can be found in Supplementary Material.

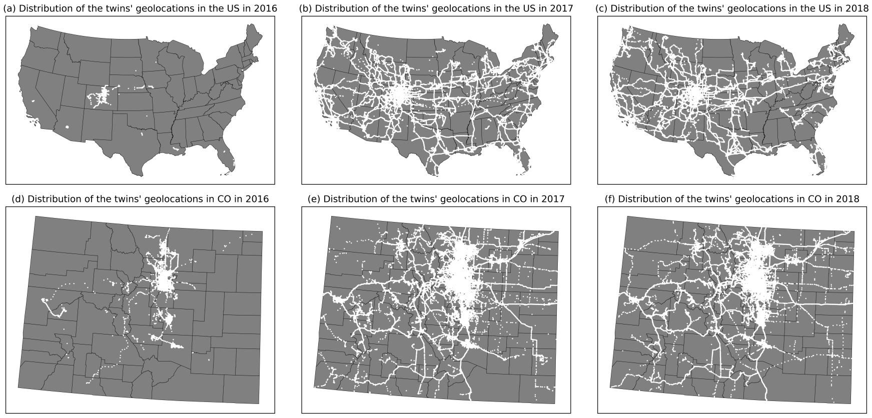

(a) Distribution of the twins' geolocations in the US in 2016
(b) Distribution of the twins' geolocations in the US in 2017
(c) Distribution of the twins' geolocations in the US in 2018
(d) Distribution of the twins' geolocations in CO in 2016
(e) Distribution of the twins' geolocations in CO in 2017
(f) Distribution of the twins' geolocations in CO in 2018

Figure 1. Distribution of geolocations in the contiguous United States and Colorado across 2016, 2017, and 2018 in the CoTwins study.

Many related works have demonstrated the spatial aspects of activity space and shared space and their impact on human behaviors such as substance use (Mason et al., 2010) and social support in a specific setting such as working space (Gerdenitsch et al., 2016); however, the temporal variations of such spatial measures and interindividual differences therein have not been thoroughly explored. Hence, we employed passive sensor (GPS) data to investigate whether meaningful seasonal, time- (e.g., weekend), and age-based variations, as well as between-individual differences in these intra-individual changes, could be meaningfully inferred from individuals' spatial measures as extracted using GPS2space. In particular, we examined (1) whether there were seasonal effects in twins' activity space/shared space; (2) whether there were weekend effects in twins' activity space/shared space; (3) inter-individual differences in initial levels of activity space/shared space, and possible associations with gender, baseline age, and twin type (MZ vs. DZ twins); and (4) age-related changes in activity space/shared space, and possible roles of gender as correlates of interindividual differences in these age-based changes.

4 Example I: Buffer- and Convex hull-based Activity Space and Shared Space

As previously defined, activity space refers to the area of individuals' routine locations over a specific time period. Practically, ellipses, convex hulls, and density kernels are often used to construct the activity space (Huang & Wong, 2016). The GPS2space library currently includes two commonly used methods for constructing activity space: the buffer method and the convex hull method. The buffer method uses a user-specified buffer distance as the radius in determining activity space, while the convex hull method lines up the outermost points to a minimum bounding geometry (J. H. Lee, Davis, Yoon, & Goulias, 2016) to represent activity space. Both buffer- and convex hull-based activity space approaches are associated with their own pros and cons. For buffer-based activity space, users have to specify a buffer distance to group and dissolve points into polygons to enable extraction of activity space. The choice of buffer distance can be arbitrary and application-specific, and it affects the sizes of activity space and shared space. However, this approach provides interpretable mobility estimates even with only one data point. In this case, activity space for that one data point is simply the area of the circle whose radius is the buffer distance. Importantly, it is less sensitive to extreme geolocations that are beyond the clusters of geolocation. Convex hull-based activity space does not require any arbitrary parameter. However, convex hull-based activity space computations require at least three non-collinear points to form an enclosed convex hull. In addition, convex hull-based activity space is sensitive to extreme geolocations, giving extreme activity space values in the presence of outliers. For example, instances where individuals travel via cars or flights from one main location to another would be outliers. The convex hull method would yield extreme activity space values in trying to construct a convex hull containing all the data points prior to, during, and after such travels, whereas the buffer-based method would use the user-specified buffer value to "group" the data points into clusters of points and compute activity and other spatial activity measures accordingly. We recommend that users consider their respective applications and contexts in detail when choosing between these two methods.

To illustrate how buffer- and convex hull-based activity space and shared space are obtained from raw GPS data with latitude and longitude coordinate pairs, we used one randomly selected twin pair, denoted herein as TwinX, and their geolocations on May 12, 2017. For buffer-based activity space, we used a buffer distance of 1000 meters based on common choices of buffer distance in other published studies (Perchoux, Chaix, Brondeel, & Kestens, 2016; Stewart et al., 2015). The process of computing activity and shared spaces can be grouped largely into 3 steps. We described each step and provided the associated code as organized by these steps.

Step 1: Conversion of raw GPS data into spatial data.

To perform spatial operations, we need to first convert raw GPS data with latitude and longitude coordinate pairs to spatial data using the *df_to_gdf* function in the GPS2space library. The *df_to_gdf* function takes three parameters: the first one is the Pandas dataframe [3] that contains GPS data with geolocation information as represented by latitude and longitude coordinate pairs; the second one is the column name of the longitude information; the third one is the column name of the latitude information. The *df_to_gdf* function returns an un-

[3] Pandas is a commonly used library for data manipulation analysis in Python. A Pandas dataframe is a 2-dimensional data structure with rows representing observations and columns representing variables. A column can have different data types in a Pandas dataframe.

projected GeoPandas dataframe [4] in the World Geodetic System 84 (WGS84). The following code imports the required libraries for the process, then reads in latitude and longitude coordinate pairs stored in two csv files comprising the two twin members' respective data, TwinXa_512.csv and TwinXb_512.csv, and finally converts the non-spatial dataframe to spatial data using the *df_to_gdf* function. One important note is that users must pass the longitude column name to x and the latitude column name to y.

```
# Import required libraries for the analyses.
import pandas as pd
import geopandas as gpd
from gps2space import geodf, space, dist

# Read TwinXa_512 and TwinXb_512 as Pandas dataframes.
df_twinXa_512 = pd.read_csv('./data/TwinXa_512.csv')
df_twinXb_512 = pd.read_csv('./data/TwinXb_512.csv')

# Convert Pandas dataframes to GeoPandas dataframes.
gdf_twinXa_512 = geodf.df_to_gdf(df_twinXa_512, x='
    longitude', y='latitude')
gdf_twinXb_512 = geodf.df_to_gdf(df_twinXb_512, x='
    longitude', y='latitude')
```

Step 2: Spatial projection and spatial measure extraction of activity space.

After successful data conversion, the next step is to project the spatial data and calculate buffer- and convex hull-based activity space using the *space.buffer_space* and *space.convex_space* functions, respectively. The *buffer_space* takes four parameters: the first is the unprojected GeoPandas dataframe; the second is a user-defined buffer distance dist, where the default value is 0; the third is dissolve, the user-specified level of timescale at which the geolocations are aggregated to form polygons, where the default value is "week"; the fourth is *proj*, the user-specified EPSG identifier [5] based on the selected spatial data for projection. The default value for *proj* is 2163 (US National Atlas Equal Area projection), a commonly used projection for the US. The *buffer_space* function returns a GeoPandas dataframe with a "buff_area" column representing the buffer-based activity space. The *proj* parameter specifies the unit for activity space, shared space, and buffer distance in the *buffer_space* function. For instance, the unit of EPSG 2163 is meter, so the unit for dist is meter; accordingly, the unit for activity space and shared space is square meter. We recommend that users choose a meter-based projection system because it provides more intuitive measurement

[4] A GeoPandas dataframe is an extension of Pandas dataframe with a "geometry" column storing geolocation information.

[5] EPSG identifiers are codes representing different spatial reference systems that can be used to project, reproject, and transform between different spatial reference systems. For example, the EPSG: 4326 is the default spatial reference system used by GPS, the EPSG: 3857 is used by Google Map and OpenStreetMap.

units than a degree-based projection system. [6] As mentioned above, the buffer distance in the *buffer_space* function is an application-specific parameter, users can refer to Browning and Lee (2017), K. Lee and Kwan (2019), Sugiyama, Kubota, Sugiyama, Cole, and Owen (2019), and Prins et al. (2014) for discussion on selecting buffer distances and their impacts on the study involved.

The *convex_space* takes three parameters: the first is the unprojected GeoPandas dataframe; the second is *group*, the level of timescale at which users want to group geolocations to form polygons, where the default value is "week"; the third is the EPSG identifier, where the default value is 2163. The *convex_space* function returns a GeoPandas dataframe with a "convx_area" column representing the convex hull-based activity space. When constructing activity space, the timescale should either be one of the variables in the dataframe, or it can be inferred and included as a variable in the dataframe from the timestamp when the geolocations are recorded. In the following example, we constructed TwinXa and TwinXb's daily activity space on May 12, 2017, and the variable "day" is inferred from the twin pairs' timestamps ranging from 5/12/2017 at 07:25 to 5/12/2017 at 20:10.

```
# Project spatial data.
gdf_twinXa_512 = gdf_twinXa_512.to_crs(‘epsg:2163’)
gdf_twinXb_512 = gdf_twinXb_512.to_crs(‘epsg:2163’)

# Buffer- and convex hull-based activity space.
buff_twinXa_512 = space.buffer_space(gdf_twinXa_512,
   dist=1000, dissolve=‘day’, proj=2163)
buff_twinXb_512 = space.buffer_space(gdf_twinXb_512,
   dist=1000, dissolve=‘day’, proj=2163)
convex_twinXa_512 = space.convex_space(gdf_twinXa_512,
   group=‘day’, proj=2163)
convex_twinXb_512 = space.convex_space(gdf_twinXb_512,
   group=‘day’, proj=2163)
```

Step 3: Extraction of shared space by overlaying activity space features.

Once we have the activity space, we can utilize the *overlay* function from GeoPandas to calculate shared space by overlaying and intersecting the activity spaces of two individuals. For instance, in the following code example, we overlaid the buffer- and convex hull-based activity space. We specified "intersection" for the how parameter to extract the intersection area between the twins' activity space. We then invoked the *area* function to obtain a column named "share_space," representing the areas of the twins' shared space. A loop to iterate over multiple activity space features to obtain shared space between one another is provided in Supplementary Material.

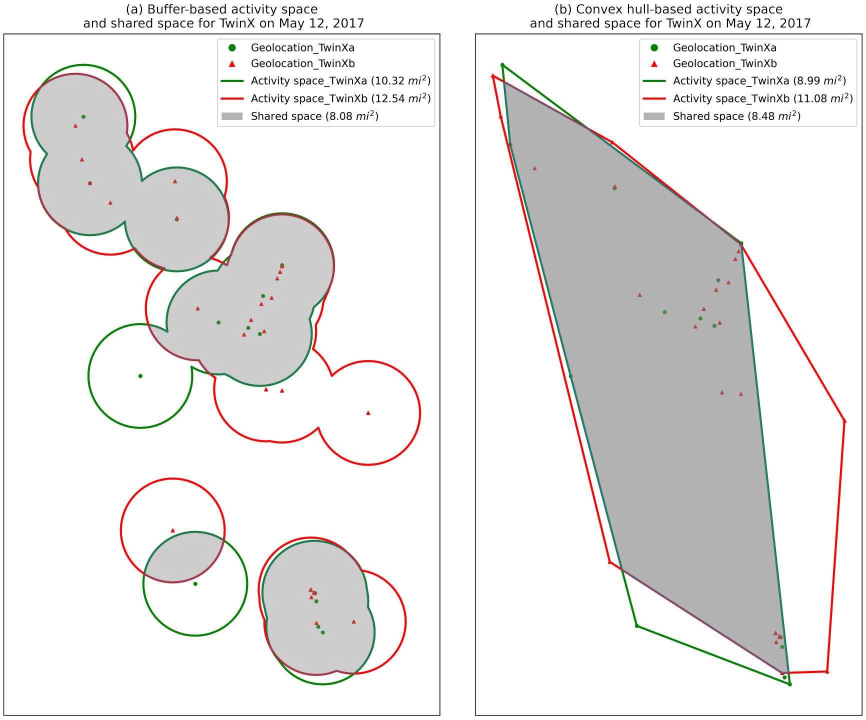
```
# Calculate shared space from activity space.
buff_share = gpd.overlay(buff_twinXa_512,
   buff_twinXb_512, how=‘intersection’)
```

[6] For the unit of different projection systems, see https://epsg.io/.

```
buff_share[‘share_space’] = buff_share[‘geometry’].area

convex_share = gpd.overlay(convex_twinXa_512,
    convex_twinXb_512, how=‘intersection’)
convex_share[‘share_space’] = convex_share[‘geometry’].
    area
```

Figure 2 shows the buffer- and convex hull-based activity space and shared space for TwinX on May 12, 2017. The buffer-based approach using 1000 meters as buffer distance gives an activity space of 10.32 and 12.54 square miles [7] for TwinXa and TwinXb, and a shared space of 8.08 square miles between them. The convex hull-based approach produces an activity space of 8.99 and 11.08 square miles for each individual of TwinX and a shared space of 8.48 square miles between them. The code for Figure 2 can be found in Supplementary Material.


(a) Buffer-based activity space and shared space for TwinX on May 12, 2017

(b) Convex hull-based activity space and shared space for TwinX on May 12, 2017

● Geolocation_TwinXa
▲ Geolocation_TwinXb
Activity space_TwinXa (10.32 mi^2)
Activity space_TwinXb (12.54 mi^2)
Shared space (8.08 mi^2)
● Geolocation_TwinXa
▲ Geolocation_TwinXb
Activity space_TwinXa (8.99 mi^2)
Activity space_TwinXb (11.08 mi^2)
Shared space (8.48 mi^2)

Figure 2. (a) Buffer-based activity space and shared space for TwinX on May 12, 2017 in Colorado. (b) Convex hull-based activity space and shared space for TwinX on May 12, 2017 in Colorado.

[7] For illustration purposes, we converted area measurement in square meters to square miles.

5 Example II. The Nearest Distance Query

The nearest distance measure is a useful indicator of accessibility of infrastructures and places that would influence behavioral and socioeconomic outcomes. For example, research has shown that the distance to the ballot drop box influences voters' turnout (McGuire, O'Brien, Baird, Corbett, & Collingwood, 2020), and access to highways affects population distribution (Chi, 2010). However, the nearest distance query can be computationally demanding and time-consuming, especially for processing data in large volumes. To boost the nearest distance query, the *dist* function in the GPS2space library incorporates two types of spatial indices to rapidly look up the nearest neighbor and calculate the distance. When the geometries of target features are points, *dist_to_point* in the dist function utilizes the cKDTree from SciPy to search for nearest neighbors; when the geometries of target features are polygons, *dist_to_poly* in the dist function utilizes the R-Tree from Geopandas to search for nearest neighbors. Both cKDTree and R-Tree algorithms create tree-like data structures from the Geopandas dataframe which enable fast nearest neighbor searching, therefore working efficiently with data sets in large volumes.

We used TwinXa's geolocations on May 12, 2017 to demonstrate the utility of the dist function and calculated the distance from each unique location to its nearest supermarket (represented as points) and park (represented as polygons) in Colorado. The supermarket and park data were obtained from OpenStreetMap (OSM). The OSM started in 2004 and its main goal is to collect and provide free access to geospatial data. The initial focus was on transportation infrastructure (streets, highways, railways, etc.), but data collection has expanded to multiple points of interest, such as buildings and community landmarks. Since most commercial data sources are expensive and have data sharing restrictions, OSM has quickly become a popular data source for geospatial-related research.

We downloaded and compiled the OSM data from Geofabrik [8], a Germany-based company specializing in processing and reorganizing free geodata created by projects like OSM. There are some concerns, however, about the quality of OSM data. For example, studies have shown that there were some disparities in data quality between urban/densely populated areas and rural/sparsely populated areas in OSM (Barron, Neis, & Zipf, 2014). In this study we compared OSM data with a high quality commercial data source called Infogroup Business Dataset, which contains more than 15 million geocoded business locations in the US. We found that the OSM data provided solid coverage when it came to major retail chains and good positional accuracy for corresponding locations. For example, comparing Infogroup and OSM data for the major Colorado supermarket chain "Safeway," 94% of the Safeway locations contained in the OSM data were also found in Infogroup. We also found similar results for two other major retail chains – "King Soopers" and "Whole Foods."

The *dist_to_point* function takes three parameters: the first one is the source GeoPandas dataframe; the second one is the target GeoPandas dataframe; and

[8] See https://www.geofabrik.de/geofabrik/geofabrik.html.

the third one is the EPSG identifier, with a default value of 2163. When *dist_to_point* function is called, the nearest neighbor search is then performed by traversing the cKDTree created on the spatial points in the target data set, which only deals with a subset of the points for the distance calculation. As shown in the following code example, we first constructed the spatial data set for the supermarket data, then we provided three parameters to the *dist_to_point* function for the nearest distance query from the TwinXa to supermarkets. The “dist” is the outcome GeoPandas dataframe with a “dist2point” column showing the distance from the source point to its nearest supermarket. All the columns from both the source and target dataframes are preserved in the outcome dataframe.

```
# Read market data into Pandas dataframes.
df_market = pd.read_csv(‘./data/market.csv’)

# Convert Pandas dataframes to GeoPandas dataframes.
gdf_market = geodf.df_to_gdf(df_market, x=‘longitude’,
    y=‘latitude’)

# The nearest distance from twinXa_512 to supermarket.
dist = dist.dist_to_point(gdf_twinXa_512, gdf_market,
    proj=2163)
```

The *dist_to_poly* function is similar to the *dist_to_point* function, except that the nearest neighbors in the *dist_to_poly* function are polygons. The *dist_to_poly* function takes four parameters: the first one is the source GeoPandas dataframe; the second one is the target GeoPandas dataframe; the third one is the EPSG identifier, with a default value of 2163; and the fourth one is a search radius in meters, with a default value of None. If the search radius is not specified, the *dist_to_poly* function employs a brute-force search to find the nearest distance, and the computation time increases significantly as the number of polygons grows. If the search radius is specified, R-tree is implemented by creating a minimum bounding box (MBR) for each target polygon. Instead of calculating the distance from the source point to every polygon in the target dataframe, the *dist_to_poly* function takes advantage of the R-tree index to only consider those polygons whose MBRs intersected with the search radius and calculate the minimum distance. If no polygon is within the search radius, then the *dist_to_poly* function returns a *NaN* value, a common way to represent missing values in Python. The *dist_to_poly* function works efficiently in calculating the nearest distance by specifying a search radius, but at the expense of missing values for points with no neighbors within the search radius. We recommend choosing an appropriate search radius based on how it can affect specific research designs.

As shown in the following code example, we read the park data in the form of shapefiles as GeoPandas dataframe, then we provided the parameters to the *dist_to_poly* function for the nearest distance query from the TwinXa to parks. The “dist_no_radius” and “dist_with_radius” are the outcome GeoPandas dataframes with a “dist2poly” column showing the distance from the source point to its nearest park.

```
# Read parks as GeoPandas dataframes.
gdf_parks = gpd.read_file ('./data/parks.shp')

# The nearest distance without search radius.
dist_no_radius = dist.dist_to_poly (gdf_twinXa_512,
    gdf_parks, proj=2163)

# The nearest distance with search radius of 5000m.
dist_with_radius = dist.dist_to_poly(gdf_twinXa_512,
    gdf_parks, proj=2163, search_radius=5000)
```

The two functions, *dist_to_point* and *dist_to_poly*, serve to provide distance measures geared respectively toward places of interest that are adequately represented as points (typically places covering smaller geographical regions such that the centroids of their enclosing polygon provide a reasonable representation, such as supermarkets, transportation terminals, and health facilities) vs. polygons (typically geographically dispersed places of interest or places that require precise definitions of boundaries, such as parks, water bodies, and administrative boundaries). Results from *dist_to_poly* and *dist_to_point* do not always agree, mainly because *dist_to_poly* and *dist_to_point* treat points within polygons differently. To illustrate the differences, we calculated the nearest distance from TwinX to the nearest park, playground, and supermarket (represented as polygons, search radius not specified) and their centroids (represented as points). Table 1 shows the results. Overall, the two functions produce similar results except for differences in minimum distance, where *dist_to_poly* may produce 0 values while *dist_to_point* rarely produces 0 values. The main reason for the differences in the minimum distance is that once *dist_to_poly* detects the point is within the polygon it assigns 0 to the nearest distance, while *dist_to_point* calculates the Euclidean distance between the two points and only returns 0 if the geolocations of the two points are identical. In sum, the distance measure between *dist_to_point* and *dist_to_poly* depends on the source data's relative position to the target polygon and the shape of the target polygon. The code for Table 1 can be found in Supplementary Material.

Table 1. Comparison between the nearest distance from TwinX to polygon boundary and polygon centroid for parks, playgrounds, and supermarkets in Colorado

Nearest distance to landmark measure	Mean	SD	Min	Median	Max
Distance to park (point)	0.57	0.39	0.01	0.46	6.60
Distance to park (polygon)	0.50	0.38	0.00	0.42	6.46
Distance to playground (point)	0.84	0.53	0.01	0.93	8.29
Distance to playground (polygon)	0.83	0.53	0.00	0.92	8.29
Distance to supermarket (point)	1.30	1.06	0.01	0.98	9.36
Distance to supermarket (polygon)	1.27	1.06	0.00	0.95	9.33

Note: The original distance measures were in meters, we converted them to miles for illustration purposes.

6 Example III. Growth Curve Analysis of Activity and Shared Spaces

6.1 Data Pre-Processing

Before extracting the daily activity space and shared space for all participants using the functions presented above, we pre-processed the GPS data following procedures implemented in the previous study (Li et al., in press). First, we excluded records with fewer than 20 valid data points within a week because these unusually low numbers of GPS points lacked sufficient variability. Then we excluded data points showing atypical travel trajectories as detected by *dbscan* (Density-Based Spatial Clustering of Applications with Noise), an R package that is commonly used to identify clusters and outlying points (Hahsler, Piekenbrock, & Doran, 2019). Then the daily activity space was calculated using a buffer distance of 1000 meters and transformed from square meters to square miles for illustrative purposes. The activity space was then log transformed to reduce skewness in the data. The log transformed activity space was referred to hereafter as LAS. For each participant, we focused on the proportion of shared space, referred to as PSS hereafter and defined as the proportion of one's daily activity space that overlapped with his/her twin sibling's daily activity space. The distributions of LAS and PSS were shown in Figure 3. The final data set consisted of 558 participants with baseline ages between 14 and 20 (mean = 17), followed between 1 to 3 years (mean = 2). 43% of the participants were males. In terms of twin types, 33% were MZ twins, 41% were DZ twins of the same sex, and 26% were DZ twins of opposite sex.

6.2 Data Analytic Plans

As mentioned before, we were interested in exploring within-individual changes of LAS and PSS and inter-individual differences in their initial levels and changes over time, including both between-individual and between-family differences. At the within-individual level, we sought to address the seasonal effect (research

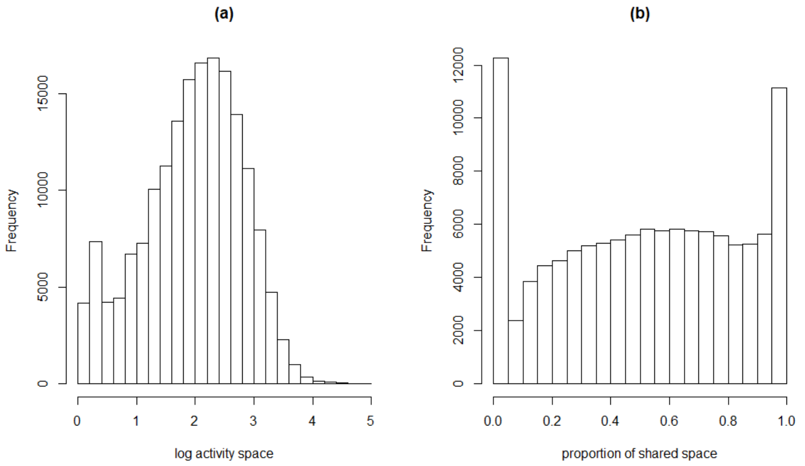


Figure 3. Distributions of (a) log activity spaces (LAS) and (b) proportions of shared space (PSS) across participants.

question 1), the weekend effect (research question 2), and age-related changes (research question 4) in LAS and PSS; at the between-individual level, we sought to explore gender differences in the initial levels and changes of LAS and PSS, as well as the effect of baseline ages on the initial levels (research questions 3-4); at the between-family level, we investigated the effect of twin zygosity (MZ vs. DZ twins) on the initial levels of PSS (research question 3). Therefore, we used three-level growth curve models (see, e.g., Enders & Tofighi, 2007; Hoffman, 2015) as implemented using the R package, *brms* (Bürkner, 2017), to study these temporal changes and levels of nesting within this data set, namely, time nested within individuals within family. In particular, we used seasonal and weekend indicators, as well as participants' ages as within-individual (or so-called level-1) predictors, gender and baseline age as between-individual (level-2) predictors, and twin zygosity as a between-family (level-3) predictor when relevant to address our questions of interest. The R code for model fitting can be found in Supplementary Material.

We first introduced the model for LAS, as shown below.

Level-1 model:

$$LAS_{itk} = \beta_{0ik}+\beta_{1ik}Age_{itk}+\beta_2 Weekend_t+\beta_3 Summer_t+\beta_4 Fall_t+\beta_5 Winter_t+e_{itk} \tag{1}$$

Level-2 model:

$$\beta_{0ik} = \gamma_{00k} + \gamma_{01k}Gender_{ik} + \gamma_{02k}Age_{i0k} + u_{0ik} \tag{2}$$
$$\beta_{1ik} = \gamma_{10k} + \gamma_{11k}Gender_{ik} + u_{1ik} \tag{3}$$

Level-3 model:

$$\gamma_{00k} = \delta_{000} + v_{0k} \tag{4}$$

$$\gamma_{01k} = \delta_{010} + v_{1k} \tag{5}$$

$$\gamma_{02k} = \delta_{020} + v_{2k} \tag{6}$$

$$\gamma_{10k} = \delta_{100} + v_{3k} \tag{7}$$

$$\gamma_{11k} = \delta_{110} + v_{4k} \tag{8}$$

with,

$$\begin{aligned}
e_{itk} &\sim N(0, \sigma^2), \\
\begin{bmatrix} u_{0ik} \\ u_{1ik} \end{bmatrix} &\sim MN(\mathbf{0}, T = \begin{bmatrix} \tau_0^2 & \\ \tau_{01} & \tau_1^2 \end{bmatrix}), \\
\begin{bmatrix} v_{0k} \\ v_{1k} \\ v_{2k} \\ v_{3k} \end{bmatrix} &\sim MN(\mathbf{0}, \boldsymbol{\Phi} = \begin{bmatrix} \varphi_0^2 & & & & \\ \varphi_{01} & \varphi_1^2 & & & \\ \varphi_{02} & \varphi_{12} & \varphi_2^2 & & \\ \varphi_{03} & \varphi_{13} & \varphi_{23} & \varphi_3^2 & \\ \varphi_{04} & \varphi_{14} & \varphi_{24} & \varphi_{34} & \varphi_4^2 \end{bmatrix})
\end{aligned}$$

The seasonal effect, weekend effect, and age-based changes in LAS were modeled in the level-1 model, where LAS_{itk} was the LAS of person i in family k on day t, and Age_{itk} was the age of person i in family k on day t, centered by subtracting the baseline age from each age instance so that 0 corresponded to the baseline age. The *Weekend*, *Summer*, *Fall*, and *Winter* variables were dummy-coded, with 1 each representing weekend, summer (June 1 to August 30), fall (September 1 to November 30), and winter (December 1 to February 28 or 29). Based on the definitions of these variables, β_{0ik} represented person i's initial LAS at baseline age on Spring weekdays; β_{1ik} was the effect of age on the LAS for person i; and β_j $(j = 2, \ldots, 5)$ represented weekend or seasonal effects, which were not set as person-specific since we focused on the overall seasonal and weekend effects in this study. Finally, the level-1 error e_{itk} followed a normal distribution with a zero mean and a variance of σ^2.

In the level-2 model, the level-1 parameters, β_{0ik} and β_{1ik}, were regressed on a *person-specific* variable, $Gender_{ik}$ (1 = male; -1 = female), to explore gender differences in the initial levels and age-based changes of LAS. In addition, β_{0ik} was regressed on the baseline age, Age_{i0k}, centered by subtracting the mean of baseline ages so that 0 corresponded to the average baseline age. Thus, the corresponding coefficient γ_{02k} represented the effect of baseline ages on the initial LAS, and γ_{00k} and γ_{10k} represented the overall initial level and growth rate of LAS across individuals, respectively, while $2\gamma_{01k}$ and $2\gamma_{11k}$ represented the corresponding gender differences, respectively. The level-2 random effects were denoted as u_{0ik} and u_{1ik}, which described person i's deviations in the values of β_{0ik} and β_{1ik} not accounted for by the predictors. Finally, the variance and covariance structure of level-2 random effects was defined in $\boldsymbol{T}$. For instance, the variance of β_{0ik}, denoted as τ_0^2, described the extent of between-individual difference in the initial LAS; the covariance between β_{0ik} and β_{1ik}, denoted as τ_{01}, described the relationship between initial levels and growth rates of LAS.

The level-3 model was built to capture between-family differences. Specifically, we would like to investigate whether twins from different families would have different initial levels and growth rates of LAS and whether the effects of gender and baseline age on the initial levels and/or growth rates of LAS would differ across families as well. Note that twin type was not included as a predictor in the level-3 model because the magnitudes of activity space were not expected to be significantly different between MZ and DZ twins (although they might be expected to differ in the degree to which they share space with their siblings, which was addressed below in the model for PSS). Among parameters in the level-3 model, δ_{010} and δ_{110} were of particular interest because they reflected the differences between males and females in terms of their average initial levels and growth rates of LAS, respectively. The level-3 random effects, v_{0k} - v_{4k}, followed a multivariate normal distribution with zero means and a covariance matrix, $\boldsymbol{\Phi}$, where the variances, denoted as φ_0^2 - φ_4^2, captured the extent of between-family differences in the overall initial LAS, the effects of gender and baseline age on the initial LAS, the overall growth rate of LAS and gender differences therein, respectively.

In terms of the model for PSS, some slight modeling adaptations were needed to capture characteristics of the PSS data. As noted, PSS was defined as the proportion of one's activity space that overlapped with his/her twin sibling's activity space, thus yielding a value ranging from 0 to 1. The model presented above, which assumed that the error term followed a normal distribution with a constant variance, might not be appropriate for the data in this scenario. However, the beta distribution is known for its flexibility in modeling proportions because its density can display different shapes as decided by the values of α and β. The beta density can be expressed as:

$$f(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1},\ 0 < y < 1, \alpha > 0, \beta > 0 \tag{9}$$

Thus, in the generalized growth curve model with PSS as the dependent variable, PSS was specified to conform to a beta distribution. Consistent with the beta regression specification proposed by Ferrari and Cribari-Neto (2004), which is similar to that of the well-known class of generalized linear models (McCullagh & Nelder, 1989), we defined $\mu = \alpha/(\alpha+\beta)$ and $\phi = \ \alpha+\beta$, then $E(y) = \mu$ and $Var(y) = \mu(1-\mu)/(1+\phi)$, where μ was the mean and ϕ was called the precision parameter. In our case, we assumed that the PSS, PSS_{itk}, followed a beta distribution with person-specific means (i.e., $E(PSS_{itk}) = \mu_{itk}$). Then we implemented a logit transformation of μ_{itk} and built a three-level growth curve model on the transformed value (i.e., η_{itk}). The level-1 model for PSS was specified as:

$$\begin{aligned}\eta_{itk} &= log(\frac{\mu_{itk}}{1-\mu_{itk}}) \\ &= \beta_{0ik} + \beta_{1ik}Age_{itk} + \beta_2 Weekend_t + \beta_3 Summer_t + \beta_4 Fall_t + \beta_5 Winter_t\end{aligned} \tag{10}$$

where $\frac{\mu_{itk}}{1-\mu_{itk}}$, denoted below as the odds of PSS, represented the average level of PSS for individual i in family k at time t relative to not sharing space with twin siblings, and $\eta_{itk} = log(\frac{\mu_{itk}}{1-\mu_{itk}})$ represented the corresponding log odds. The independent variables were as summarized in Equation 1. Note that the regression coefficients had different interpretations due to the logit transformation. For instance, β_{0ik} represented the log-odds of PSS for person i in family k at the baseline age on Spring weekdays; β_{1ik} was the age-related log-odds ratio, which means that the odds of PSS would multiply by $e^{\beta_{1ik}}$ for every 1-unit increase in Age_{itk}. Other parameters (e.g., seasonal and weekend effects) can be interpreted in a similar way.

The level-2 model for PSS was identical to the level-2 model for LAS (see Equations 2 - 3), but the regression coefficients had different interpretations for the reason stated above. For instance, the level-2 intercept, γ_{00k}, represented the overall log-odds of PSS.

In terms of the level-3 model, we hypothesized that MZ and DZ twins might have different levels of space sharing to the extent that these spatial measures reflect genetically influenced behavior/preferences. To evaluate this hypothesis, we added a predictor, twin type, to Equations 4 - 6 (i.e., the level-3 model for γ_{00k}, γ_{01k}, and γ_{02k}, which were the coefficients in the level-2 model for β_{0ik}, the log-odds of initial levels of PSS), to investigate zygosity differences in PSS and how these differences might affect the effects of gender and baseline age on PSS, as shown below.

$$\gamma_{00k} = \delta_{000} + \delta_{001}DZSS_k + \delta_{002}DZOS_k + v_{0k} \tag{11}$$

$$\gamma_{01k} = \delta_{010} + \delta_{011}DZSS_k + \delta_{012}DZOS_k + v_{1k} \tag{12}$$

$$\gamma_{02k} = \delta_{020} + \delta_{021}DZSS_k + \delta_{022}DZOS_k + v_{2k} \tag{13}$$

Specifically, we set MZ twins as the reference and added two dummy-coded, *family-specific* variables, $DZSS_k$(1 = DZ twins of the same sex) and $DZOS_k$ (1 = DZ twins of opposite sex). Thus, δ_{000}, δ_{001}, and δ_{002} represented the average log-odds of PSS for MZ twins, DZ twins of the same sex, and DZ twins of the opposite sex, respectively; δ_{010}, δ_{011}, and δ_{012} represented the corresponding gender differences in each twin type group; and δ_{020}, δ_{021}, and δ_{022} represented the effect of the baseline age on the average log-odds of PSS in each twin type group. The models for other level-2 parameters (i.e., γ_{10k}, γ_{11k}) were identical to the level-3 model for LAS (see Equations 7 - 8).

6.3 Results

With the *brms* package, the models were fitted in a Bayesian framework using Markov chain Monte Carlo (MCMC) methods. Specifically, we ran two chains, each with 5000 iterations in total and a burn-in of 2000 (discarded) iterations. On an Intel i5-8350U, 16GB RAM, Windows 10 computer, it took about 40 hours to run each model. Two diagnostic statistics were used to check the sampling quality (Gelman et al., 2013): (1) the effective sample size (ESS), which describes how many posterior draws in the MCMC procedure can be regarded as independent, and (2) $\hat{R}$, which describes the ratio of the overall variance of posterior samples across chains to the within-chain variance. The diagnostic criteria for adequate sampling and convergence were set as ESS greater than 800 and $\hat{R}$ below 1.1, respectively. Results showed that ESS was greater than 800 for most parameters, except for some random effect standard deviation parameters (e.g., $\varphi_1 - \varphi_4$), for which the average ESS was about 400, which can be deemed satisfactory. $\hat{R}$ was below 1.1 for all parameters in both models.

Table 2. Parameter estimates of the model for LAS from the CoTwins study, 2016-2018

Parameter	Estimate	SE	95% CI
Fixed effects			
Intercept, δ_{000}	1.81	0.03	[1.76, 1.86]
Gender, δ_{010}	-0.07	0.02	[-0.11, -0.01]
Baseline age, δ_{020}	0.13	0.02	[0.09, 0.16]
Age, δ_{100}	-0.01	0.01	[-0.03, 0.02]
Age*Gender, δ_{110}	0.01	0.01	[-0.01, 0.03]
Weekend, β_2	0.06	0.00	[0.05, 0.07]
Summer, β_3	0.07	0.00	[0.06, 0.07]
Fall, β_4	-0.12	0.00	[-0.13, -0.11]
Winter, β_5	-0.08	0.01	[-0.09, -0.07]
Level-2 random effects			
Intercept standard deviation, τ_0	0.25	0.01	[0.22, 0.28]
Age standard deviation, τ_1	0.19	0.01	[0.16, 0.22]
Intercept-Age correlation, $\tau_{01}/(\tau_0 * \tau_1)$	-0.31	0.08	[-0.46, -0.16]
Level-3 random effects			
Intercept standard deviation, φ_0	0.37	0.02	[0.33, 0.42]
Age standard deviation, φ_3	0.11	0.03	[0.06, 0.16]
Residual standard deviation, σ	0.72	0.00	[0.71, 0.72]

Note: SE = standard errors estimated by standard deviations of the posterior samples; CI = credible interval. N = 558 participants. The number of time points for each participant ranged from 3 to 569.

Table 2 shows the parameter estimates for LAS. In terms of the fixed effects, weekend and seasonal effects were found in the trajectory of LAS. Specifically, the participants showed greater LAS values on weekends than on weekdays ($\beta_2 = 0.06,\ \ 95\%\,CI = [0.05,\ 0.07]$), which was reasonable since most of the participants were supposed to be spending most of their time in school on weekdays, thus yielding limited activity space. Seasonally, the participants tended to display greater LAS in summer ($\beta_3 = 0.07,\ \ 95\%\,CI = [0.06,\ 0.07]$), which was likely due to summer break as well as the warmer weather. Gender differences were found in the initial levels of LAS ($\delta_{010} = -0.07,\ \ 95\%\,CI = [-0.11,\ -0.01]$), although the upper bound of the 95% credible interval was close to 0. No gender differences were found in the growth rates of LAS. Finally, older participants tended to have higher levels of LAS at baseline ($\delta_{020} = 0.13,\ \ 95\%\,CI = [0.09,\ 0.16]$), but when it comes to within-individual changes over time, participants' ages were not found to be credibly linked to their levels of LAS, as indicated by the 95% credible interval including 0.

In terms of the random effects, we found between-individual and between-family differences in both initial levels and age-based changes of LAS. These differences were indicated by the relatively high magnitude of random effect standard deviations and the credible intervals whose lower bounds were far from 0 (see, $\tau_0,\ \tau_1,\ \varphi_0$, and φ_3; random effect standard deviation parameters whose credible intervals were close to 0 were not shown in Table 2). We also found negative associations between the initial levels and growth rates at the individual level, indicating that individuals who had higher initial levels of activity space tended to experience larger decreases in activity space with age.

Table 3 shows the parameter estimates for PSS. In terms of the fixed effects, weekend and seasonal effects were found in the trajectory of PSS. Specifically, participants shared more activity space on weekdays than on weekends ($\beta_2 = -0.12,\ \ 95\%\,CI = [-0.13,\ 0.10]$). This pattern might be due to the restricted daily routines on weekdays during which twin siblings in this age range tended to spend most of their time in school and thus, showed greater PSS. Participants tended to have the largest PSS in spring, followed by winter, summer, and fall. In addition, older twins tended to share less activity space at baseline ($\delta_{020} = -0.30,\ \ 95\%\,CI = [-0.42,\ -0.18]$), and when it comes to within-individual changes over time, in contrast to the lack of age-related changes in LAS, PSS was found to decrease as twins grew older ($\delta_{100} = -0.38,\ \ 95\%\,CI = [-0.44,\ -0.31]$). Note that a small portion of twins were in the transition from high school to college, so the reduction in PSS might also reflect some of the inevitable life transitions that occur with age, such as attending colleges or working at different geographical locations. In terms of zygosity differences, both DZ twins of the same sex and opposite sex were found to share less activity space than MZ twins ($\delta_{001} = -0.29,\ \ 95\%\,CI = [-0.49,\ -0.09]$; $\delta_{002} = -0.49,\ \ 95\%\,CI = [-0.71,\ -0.27]$), indicating that there might be genetically influenced differences in PSS. Finally, no gender differences were found in the initial levels and growth rates of PSS.

Table 3. Parameter estimates of the model for PSS from the CoTwins study, 2016-2018

Parameter	Estimate	SE	95% CI
Fixed effects			
Intercept, δ_{000}	0.74	0.08	[0.58, 0.89]
Gender, δ_{010}	0.03	0.08	[-0.12, 0.19]
Baseline age, δ_{020}	-0.30	0.06	[-0.42, -0.18]
Age, δ_{100}	-0.38	0.03	[-0.44, -0.31]
Age*Gender, δ_{110}	-0.04	0.03	[-0.09, 0.01]
DZSS, δ_{001}	-0.29	0.10	[-0.49, -0.09]
DZOS,δ_{002}	-0.49	0.11	[-0.71, -0.27]
DZSS*Gender, δ_{011}	0.04	0.10	[-0.17, 0.24]
DZOS*Gender, δ_{012}	0.08	0.09	[-0.09, 0.25]
DZSS*Baseline age, δ_{021}	-0.13	0.08	[-0.28, 0.02]
DZOS*Baseline age, δ_{022}	-0.04	0.09	[-0.22, 0.13]
Weekend, β_2	-0.12	0.01	[-0.13, -0.10]
Summer, β_3	-0.31	0.01	[-0.33, -0.29]
Fall, β_4	-0.46	0.01	[-0.48, -0.44]
Winter, β_5	-0.09	0.01	[-0.11, -0.07]
Level-2 random effects			
Intercept standard deviation, τ_0	0.34	0.02	[0.30, 0.38]
Age standard deviation, τ_1	0.20	0.02	[0.17, 0.24]
Intercept-Age correlation, $\tau_{01}/(\tau_0 * \tau_1)$	-0.35	0.09	[-0.52, -0.16]
Level-3 random effects			
Intercept standard deviation, φ_0	0.58	0.11	[1.08, 1.50]
Age standard deviation, φ_3	0.40	0.03	[0.32, 0.42]
Precision parameter, ϕ	1.91	0.01	[1.89, 1.92]

Note: SE = standard errors estimated by standard deviations of the posterior samples; CI = credible interval. N = 484 participants (or 242 pairs of twins). The number of time points for each participant ranged from 3 to 569.

Results for random effects were similar to those in the LAS model. We found between-individual and between-family differences in both initial levels and age-based changes of PSS. We also found negative associations between the initial levels and growth rates at the individual level, indicating that twins who had higher initial levels of PSS tended to show more declines in PSS with age. In other words, the participants' GPS data suggested that higher physical closeness at younger ages might not persist as the twins grew older.

Finally, we conducted sensitivity analysis by re-running the analysis with the full data set (i.e., keeping the records with fewer than 20 valid data points within a week in the final data set). Results were detailed in Table S1 and Table S2 in Supplementary Material, which showed only slight differences in the magnitude of point estimates and standard errors. Both data sets yielded consistent conclusions across all parameters in terms of whether they were credibly different from zero based on their 95% credible intervals.

7 Discussion

The proliferation of real-time and longitudinal GPS data provides excellent opportunities to study human behavior (Osorio-Arjona & García-Palomares, 2019). At the same time, the GPS data also pose challenges for consolidating, automating, and analyzing data that are not only massive in their quantities but also contain spatial features that require expertise in GIS. Commercial software packages make these studies easier but may have license and reproducibility issues, and analyses with commercial software cannot be readily deployed to HPC platforms to facilitate research procedures. In this article, we reviewed and compared existing commonly used Python libraries for spatial analysis with GPS2space, our newly developed open-source Python library. GPS2space can build spatial data from GPS data with latitude and longitude coordinate pairs, construct buffer- and convex hull-based activity space and shared space, and perform the nearest distance query from user-specified locations. We demonstrated how to process spatial data and calculate buffer- and convex hull-based activity space and shared space, as well as the nearest distance, with code examples. We also discussed the pros and cons of buffer- and convex hull-based approaches and illustrated different scenarios when the two approaches could be appropriately applied. Lastly, using data from the CoTwins study, we explored intra-individual changes and between-individual differences in daily activity space and shared space with twin siblings; and gender, zygosity and baseline age-related differences in their initial levels and/or changes, using growth curve modeling techniques. We found different patterns of seasonal effects in the trajectories of LAS and PSS, less activity space shared between DZ twins compared with MZ twins, and a decrease of PSS with increasing age.

There are several limitations to the current data analysis. First, we did not allow for individual differences in the seasonal effects, so our results only provided a general description of seasonal patterns of LAS and PSS. In practice, the seasonal effects might vary across individuals and need to be considered in model specifications. Second, some other factors might affect individuals' activity space, such as time of the year (e.g., school days versus holidays) and weather (e.g., snow). Similarly, the magnitude of shared space between twin siblings depends on whether they live together or not. These factors need to be included in the models to better explain the temporal pattern of LAS and PSS as well as individual differences in these patterns. Finally, in our example, some participants were assessed for fewer than three years, while typically at least three repeated measures per individual are required in the growth curve analysis. Therefore, participants need to be followed for several more years to better investigate age-related changes at the year level. We may also assess changes of finer granularity (e.g., at the month level) based on the current data.

Although we illustrated usage of GPS2space with data from a twin study, the functions available in this package are applicable to a broad range of studies that rely on GPS data or geolocation data with latitude and longitude coordinate pairs. For example, GPS2space can be used to quantify individuals' mobility patterns using data from social media platforms. Health studies investigating the spread of contagious diseases can examine individuals' physical movements and interaction patterns with other individuals using activity space and shared space measures as derived from GPS2space. From demographic and sociological perspectives, activity space and shared space obtained using GPS2space can provide important information regarding people's sense of place, social segregation, and their impacts on a series of socioeconomic outcomes such as educational attainment and occupational status. In addition, the nearest distance measure from GPS2space can also be used to examine the effects of accessibility to food and healthcare providers. Meanwhile, researchers have shown disagreements in mobility or trajectory measures between self-reported data and GPS/Sensor data (Fillekes, Kim, et al., 2019; Fillekes, Röcke, Katana, & Weibel, 2019). GPS2space can provide information for researchers to validate and compare mobility or trajectory measures from different data sources.

Many other extensions are possible within GPS2space to circumvent some of its current limitations. For example, constructing activity space and shared space involves topological structuring, which can take other forms besides convex hull and buffer, the two methods currently available in GPS2space. Some researchers use hexagon methods to measure territorial control based on road data (Tao, Strandow, Findley, Thill, & Walsh, 2016); others also use the concave hull method to estimate crown volumes of trees from remote sensing data (Yan et al., 2019). Those approaches are useful and beneficial for certain research questions but are currently unavailable in GPS2space. To extend the GPS2space, one could include concave hull, hexagon, and network-based methods in constructing activity space and parameterize the column name variables for the spatial measures in GPS2space so that users have control of naming their desired outcomes.

With rapid developments of spatial economics, readily available spatial data sets, and the computational power of personal computer and cloud computing, spatial analyses have gained popularity in areas such as social, behavioral, and environmental studies. We provided a timely open-source solution to work with GPS data and extract spatial measures with code snippets and empirical examples using GPS2space. Overall, we have demonstrated that GPS2space can be a versatile, handy, and extendable tool for researchers to harness the spatialities of GPS data to investigate a wide array of research questions regarding spatial-temporal variations of human behavioral changes and environment-population linkages.

References

Barron, C., Neis, P., & Zipf, A. (2014). A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, *18*(6), 877–895.

doi: https://doi.org/10.1111/tgis.12073

Ben-Ari, A., & Lavee, Y. (2007). Dyadic closeness in marriage: From the inside story to a conceptual model. *Journal of Social and Personal Relationships*, *24*(5), 627–644. doi: https://doi.org/10.1177/0265407507081451

Bivand, R. (2006). Implementing spatial data analysis software tools in R. *Geographical Analysis*, *38*(1), 23–40. doi: https://doi.org/10.1111/j.0016-7363.2005.00672.x

Browning, M., & Lee, K. (2017). Within what distance does "greenness" best predict physical health? A systematic review of articles with gis buffer analyses across the lifespan. *International Journal of Environmental Research and Public Health*, *14*(7), 675. doi: https://doi.org/10.3390/ijerph14070675

Buchowski, M. S., Townsend, K. M., Chen, K. Y., Acra, S. A., & Sun, M. (1999). Energy expenditure determined by self-reported physical activity is related to body fatness. *Obesity Research*, *7*(1), 23–33. doi: https://doi.org/10.1002/j.1550-8528.1999.tb00387.x

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: https://doi.org/10.18637/jss.v080.i01

Chi, G. (2010). The impacts of highway expansion on population change: An integrated spatial approach. *Rural Sociology*, *75*(1), 58–89. doi: https://doi.org/10.1111/j.1549-0831.2009.00003.x

Chi, G., & Marcouiller, D. W. (2013). Natural amenities and their effects on migration along the urban-rural continuum. *Annals of Regional Science*, *50*(3), 861–883. doi: https://doi.org/10.1007/s00168-012-0524-2

Chi, G., & Zhu, J. (2019). *Spatial Regression Models for the Social Sciences*. Thousand Oaks: SAGE Publications.

Cleaveland, C., & Kelly, L. (2008). Shared Social Space and Strategies to Find Work: An Exploratory Study of Mexican Day Laborers in Freehold, N.J. *Social Justice*, *35*, 51–65.

Crickard, P., Toms, S., & Rees, E. v. (2018). *Mastering geospatial analysis with Python: explore GIS processing and learn to work with GeoDjango, CARTOframes and MapboxGL-Jupyter*. Birmingham: Packt Publishing Ltd.

Drummond, W. J., & French, S. P. (2008). The future of GIS in planning: Converging technologies and diverging interests. *Journal of the American Planning Association*, *74*(2), 161–174. doi: https://doi.org/10.1080/01944360801982146

Enders, C. K., & Tofighi, D. (2007). Centering Predictor Variables in Cross-Sectional Multilevel Models: A New Look at an Old Issue. *Psychological Methods*, *12*(2), 121–138. doi: https://doi.org/10.1037/1082-989X.12.2.121

Ferrari, S. L., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815. doi: https://doi.org/10.1080/0266476042000214501

Fillekes, M. P., Kim, E.-k., Trumpf, R., Zijlstra, W., Giannouli, E., & Weibel, R. (2019). Assessing older adults' daily mobility: a comparison of GPS-derived and self-reported mobility indicators. *Sensors*, *19*(20), 4551.

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 and Medicine*, *220*(October 2018), 193–202. doi: https://doi.org/10.1016/j.socscimed.2018.11.010

Fine, G. A. (2012). Group culture and the interaction order: Local sociology on the meso-level. *Annual Review of Sociology*, *38*, 159–179. doi: https://doi.org/10.1146/annurev-soc-071811-145518

GDAL/OGR contributors. (2020). *GDAL/OGR Geospatial Data Abstraction software Library.*

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian Data Analysis* (Third Edit ed.). New York: Chapman and Hall/CRC.

Gerdenitsch, C., Scheel, T. E., Andorfer, J., & Korunka, C. (2016). Coworking spaces: A source of social support for independent professionals. *Frontiers in Psychology*, *7*, 581. doi: https://doi.org/10.3389/fpsyg.2016.00581

Gillies, S., et al. (2007). *Shapely: manipulation and analysis of geometric objects.* Retrieved from `https://github.com/Toblerity/Shapely`

Gillies, S., et al. (2011). *Fiona is ogr's neat, nimble, no-nonsense api.* Retrieved from `https://github.com/Toblerity/Fiona`

Gillies, S., et al. (2013). *Rasterio: geospatial raster i/o for Python programmers.* Retrieved from `https://github.com/rasterio/rasterio`

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, *91*(1), 1–30. doi: https://doi.org/10.18637/jss.v091.i01

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change.* New York: Routledge.

Huang, Q., & Wong, D. W. (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, *30*(9), 1873–1898. doi: https://doi.org/10.1080/13658816.2016.1145225

Jordahl, K. (2014). *GeoPandas: Python tools for geographic data.*

Kearney, A. R. (2006). Residential development patterns and neighborhood satisfaction: Impacts of density and nearby nature. *Environment and Behavior*, *38*(1), 112–139. doi: https://doi.org/10.1177/0013916505277607

Kerr, J., Duncan, S., & Schipperjin, J. (2011). Using global positioning systems in health research: A practical approach to data collection and processing. *American Journal of Preventive Medicine*, *41*(5), 532–540. doi: https://doi.org/10.1016/j.amepre.2011.07.017

Kestens, Y., Lebel, A., Chaix, B., Clary, C., Daniel, M., Pampalon, R., ... p Subramanian, S. V. (2012). Association between activity space exposure to food establishments and individual risk of overweight. *PLoS ONE*, *7*(8), e41418. doi: https://doi.org/10.1371/journal.pone.0041418

Kestens, Y., Thierry, B., Shareck, M., Steinmetz-Wood, M., & Chaix, B. (2018). Integrating activity spaces in health research: Comparing the VERITAS activity space questionnaire with 7-day GPS tracking and prompted recall. *Spatial and Spatio-temporal Epidemiology*, *25*, 1–9. doi: https://doi.org/10.1016/j.sste.2017.12.003

Lee, J. H., Davis, A. W., Yoon, S. Y., & Goulias, K. G. (2016). Activity space estimation with longitudinal observations of social media data. *Transportation*, *43*, 955–977. doi: https://doi.org/10.1007/s11116-016-9719-1

Lee, K., & Kwan, M.-P. (2019). The Effects of GPS-Based Buffer Size on the Association between Travel Modes and Environmental Contexts. *ISPRS International Journal of Geo-Information*, *8*(11), 514. doi: https://doi.org/10.3390/ijgi8110514

Lee, N. C., Voss, C., Frazer, A. D., Hirsch, J. A., McKay, H. A., & Winters, M. (2016). Does activity space size influence physical activity levels of adolescents?-A GPS study of an urban environment. *Preventive Medicine Reports*, *3*, 75–78. doi: https://doi.org/10.1016/j.pmedr.2015.12.002

Li, Y., Oravecz, Z., Zhou, S., Bodovski, Y., Barnett, I. J., Chi, G., . . . Chow, S.-M. (in press). Bayesian forecasting with a regime-switching zero-inflated multilevel poisson regression model: An application to adolescent alcohol use with spatial covariates. *Psychometrika*.

Mason, M. J., Valente, T. W., Coatsworth, J. D., Mennis, J., Lawrence, F., & Zelenak, P. (2010). Place-based social network quality and correlates of substance use among urban adolescents. *Journal of Adolescence*, *33*(3), 419–427. doi: https://doi.org/10.1016/j.adolescence.2009.07.006

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E. S. (2017). Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. *Sociological Methods and Research*, *46*(3), 390–421. doi: https://doi.org/10.1177/0049124115605339

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models* (2nd ed.). Chapman and Hall.

McGuire, W., O'Brien, B. G., Baird, K., Corbett, B., & Collingwood, L. (2020). Does Distance Matter? Evaluating the Impact of Drop Boxes on Voter Turnout. *Social Science Quarterly*, *101*(5), 1789–1809. doi: https://doi.org/10.1111/ssqu.12853

Murray, A. T., Xu, J., Wang, Z., & Church, R. L. (2019). Commercial GIS location analytics: capabilities and performance. *International Journal of Geographical Information Science*, *33*(5), 1106–1130. doi: https://doi.org/10.1080/13658816.2019.1572898

Newman, H. H., Freeman, F. N., & Holzinger, K. J. (1937). *Twins: a study of heredity and environment.* Chicago: University of Chicago Press.

Osorio-Arjona, J., & García-Palomares, J. C. (2019). Social media and urban mobility: Using twitter to calculate home-work travel matrices. *Cities*, *89*, 268–280. doi: https://doi.org/10.1016/j.cities.2019.03.006

Patil, S. (2016). Big Data Analytics Using R. *International Research Journal of Engineering and Technology*, *3*(7), 78–81.

Perchoux, C., Chaix, B., Brondeel, R., & Kestens, Y. (2016). Residential buffer, perceived neighborhood, and individual activity space: New refinements in the definition of exposure areas - The RECORD Cohort Study. *Health and Place*, *40*, 116–122. doi: https://doi.org/10.1016/j.healthplace.2016.05.004

Prins, R. G., Pierik, F., Etman, A., Sterkenburg, R. P., Kamphuis, C. B., & van Lenthe, F. J. (2014). How many walking and cycling trips made by elderly are beyond commonly used buffer sizes: Results from a GPS study. *Health and Place*, *27*, 127–133. doi: https://doi.org/10.1016/j.healthplace.2014.01.012

Rey, S. J. (2019). PySAL: the first 10 years. *Spatial Economic Analysis*, *14*(3), 273–282. doi: https://doi.org/10.1080/17421772.2019.1593495

Rey, S. J., & Anselin, L. (2007). PySAL: A Python Library of Spatial Analytical Methods. *The Review of Regional Studies*, *37*(1), 7–27.

Russell, M. A., Almeida, D. M., & Maggs, J. L. (2017). Stressor-related drinking and future alcohol problems among university students. *Psychology of Addictive Behaviors*, *31*(6), 676–687. doi: https://doi.org/10.1037/adb0000303

Shelton, T. (2017). Spatialities of data: mapping social media 'beyond the geotag'. *GeoJournal*, *82*(4), 721–734. doi: https://doi.org/10.1007/s10708-016-9713-3

Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, *142*, 198–211. doi: https://doi.org/10.1016/j.landurbplan.2015.02.020

Stewart, T., Duncan, S., Chaix, B., Kestens, Y., Schipperijn, J., & Schofield, G. (2015). A Novel Assessment of Adolescent Mobility: A Pilot Study. *International Journal of Behavioral Nutrition and Physical Activity*, *12*, 18. doi: https://doi.org/10.1186/s12966-015-0176-6

Sugiyama, T., Kubota, A., Sugiyama, M., Cole, R., & Owen, N. (2019). Distances walked to and from local destinations: Age-related variations and implications for determining buffer sizes. *Journal of Transport and Health*, *15*, 100621. doi: https://doi.org/10.1016/j.jth.2019.100621

Sui, D., & Goodchild, M. (2011). The convergence of GIS and social media: Challenges for GIScience. *International Journal of Geographical Information Science*, *25*(11), 1737–1748. doi: https://doi.org/10.1080/13658816.2011.604636

Tao, R., Strandow, D., Findley, M., Thill, J. C., & Walsh, J. (2016). A hybrid approach to modeling territorial control in violent armed conflicts. *Transactions in GIS*, *20*(3), 413–425. doi: https://doi.org/10.1111/tgis.12228

Vallée, J., Cadot, E., Roustit, C., Parizot, I., & Chauvin, P. (2011). The role of daily mobility in mental health inequalities: The interactive influence of activity space and neighbourhood of residence on depression. *Social Science and Medicine*, *73*(8), 1133–1144. doi: https://doi.org/10.1016/j.socscimed.2011.08.009

Yan, Z., Liu, R., Cheng, L., Zhou, X., Ruan, X., & Xiao, Y. (2019). A Concave Hull Methodology for Calculating the Crown Volume of Individual Trees Based on Vehicle-Borne LiDAR Data. *Remote Sensing*, *11*(6), 623. doi: https://doi.org/10.3390/rs11060623

Yin, Z., Goldberg, D. W., Hammond, T. A., Zhang, C., Ma, A., & Li, X. (2020). A probabilistic framework for improving reverse geocoding output. *Transactions in GIS*, *24*(3), 656–680. doi: https://doi.org/10.1111/tgis.12623

Supplementary Material

Supplementary material including links to the source code and documentation of GPS2space, code for replicating the examples, and results from sensitivity analysis are available at `https://github.com/shuai-zhou/GPS2space_SupMaterial/blob/main/Supplementary_Material_V2.pdf`. To replicate Example I and Example II and explore data structures of the input and output data sets, please follow the Jupyter Notebook at `https://github.com/shuai-zhou/GPS2space_SupMaterial/blob/main/Example%20I%20and%20II.ipynb`.