Utrecht University

Master Thesis

Solution for data error in the Data Donation System

Improve the data quality of activity “Walking” and “Cycling” inside Google Semantic History Location Format DDPs

Author:

Jianwen Feng

First supervisor:

Dr. Erik-Jan van Kesteren

Second supervisor:

Thijs Carrière

Abstract

New advanced digital tools bring the researchers of human behavior a new way of collecting data they need by using digital trace data for their analysis. Data Donation can be an option for collection of data the human behavior researchers need without ethical risk. The application of Data Download Packages (DDPs) for Data Donation is popular in human behavior research for its secrecy, integrity, availability, and controllability features. But errors can occur in every step of data collection and wrangling. Google manages data obtained from user devices according to General Data Protection Regulation (GDPR). It creates DDPs according to its standard. One popular format among them is a Google Semantic Location History Format (GSLH) with a JSON file. The errors in the current DDPs by Google cause troubles for human behavior researchers using the Data Donation System as their expectation which means measuring the participants' geo-data accurately as their analytical foundation.

In this paper, the key question will be how to handle the data error inside Google Semantic Location History Format. We investigate the quality of logs on two activity types, namely "Walking " and "Cycling" in the JSON file of GSLH DDPs to investigate whether these activity types are classified as reasonable in the data context. Through our designed workflow pipeline, namely flagging and imputation, clear data error should be limited, the quality of data should be improved and eventually, validation of analysis results for human behavior research should be enhanced. In the end, we find that a large proportion (22.6%) of data errors about activity type in our pilot data source through our processing pipeline, the imputation can be feasible although more resources are needed for ensuring the parameters and testing the validity.

1. Introduction

Nowadays, digital trace data have been widely used to research human behavior. These data are generated while using digital platforms – such as Google, Apple, Facebook, Microsoft, Amazon, among many others – and encompass an ever-growing set of life domains – for example, interpersonal communication, politics, commerce, health, or work (Araujo et al., 2022). However, these rich digital platform data come with a challenge: the researchers face limited access to those platforms using their APIs or data-sharing initiatives to scrap data, and the ethics of using these data remains arguable.

At the same time, the growing awareness of "data rights" – including most importantly the rights for individuals to access and transfer the data that digital platforms or companies have on them – opens several opportunities for academic research (Ausloos & Veale, 2020).

Data donation as an act of an individual actively consenting to donate their personal data for research (Skatova & Goulding, 2019) may solve the above dilemma effectively. Download Data Packages (DDPs), which are machine-readable electronic files, contain the personal data requested from the data controller by the participant of Data Donation. Through voluntary donation of DDPs, all data collected by public and private entities during the course of citizens' digital life can be obtained and analyzed to answer social-scientific questions – with consent (Boeschoten et al., 2022). According to Olteanu et al. (2019), DDPs offer several potential advantages. They can measure phenomena that participants may not easily remember and include retrospective data that allow researchers to delve into the past. DDPs can also seamlessly integrate within larger surveys, facilitating necessary corrections for sample selectivity when utilizing digital trace data.

Among the different formats of DDPs, Google Semantic Location is a popular and widely used format that consists of high-level and processed information for DDPs. It was developed by Google to understand and interpret location-related information. It goes beyond traditional geolocation data by incorporating additional data sources (not merely based on GPS coordinates) and applying machine-learning techniques to infer a user's location's semantic meaning and context. By understanding the semantic meaning of location data, GSLH aims to provide more accurate, useful, and different types of information based on their reshaped geo-context which can help human behavior researchers to analyze their theme more effectively.

But errors can occur in every step of Data Donation which may lead to problematic analysis results. One concerned issue is the potential logging errors in GSLH DDPs, which means some data errors inside the recorded semantic geo-information context. It may cause inaccurate results like wrong activity type classification for some human behavior themes from the beginning stage.

In this paper, we will focus on the question of how to handle the data error inside Google Semantic Location History Format. Particularly, data errors in the Google Semantic Location History Format (GSLH) in activity "Walking" and "Cycling" inside the Data Donation System. By fixing the data errors or flagging the data errors to the data user, the quality of data inputting in human behavior analysis can be improved and eventually, the validation of analysis results can be enhanced.

In the following, we will detail our project from different aspects. First, in the remainder of section 1, we explain the Data Donation System and the structure of the GSLH file. The, we outline the data error in the Data Donation System in section 2.

1.1 Data Donation System

In this section, we try to from several aspects like the demand for Data Donation System, the advantages of GSLH, the development of the Data Donation System, and the variable of GSLH we care to give a general overview of the Data Donation System.

1.1.1 High Demand in Data Donation System

Data serves as the foundation for analyzing human behavior, and new advanced digital tools have introduced researchers to a novel approach by using digital trace data for their analyses. However, it is essential to acknowledge that these tools also come with ethical risks.

After the release of the law. Article 15 of the EU's 2018 General Data Protection Regulation (GDPR), a new way, which requires following the principles of FAIR, namely Findability, Accessibility, Interoperability, and Reusability from the site of the data owner (GO FAIR, 2022), to get the proper data for human behavior research is in high demand. The emergence of Data Donation by applying the DDPs can perfectly solve the previous problem in digital data trace from digital platforms.

Some human behavior researchers have tried to use GSLH to find people's Activity Point Locations (APL) for analyzing activity-travel behavior between pre- and post-COVID periods (Moncayo-Unda et al, 2022). The Python scripts would be applied to transport the relevant variables of the JSON file to a CSV file for making data analysis of their own theme. It may be an option for human behavior researchers to make data collection by adapting Data donation with GSLH to CSV file. But a more general data donation system with GSLH needs to be developed. It should help human behavior researchers to implement the data collection more effectively, and provide the fundamental geo-data which can serve their different research themes.

1.1.2 Difference between traditional geo-data collection method and GSLH

The GSLH will record all the geo-data information. Compared to the traditional travel history survey which highly depends on participants' memory, the GSLH record can be more reliable.

Figure 1 shows the difference between the traditional method for geo-data collection and the new method "Google Location History data."

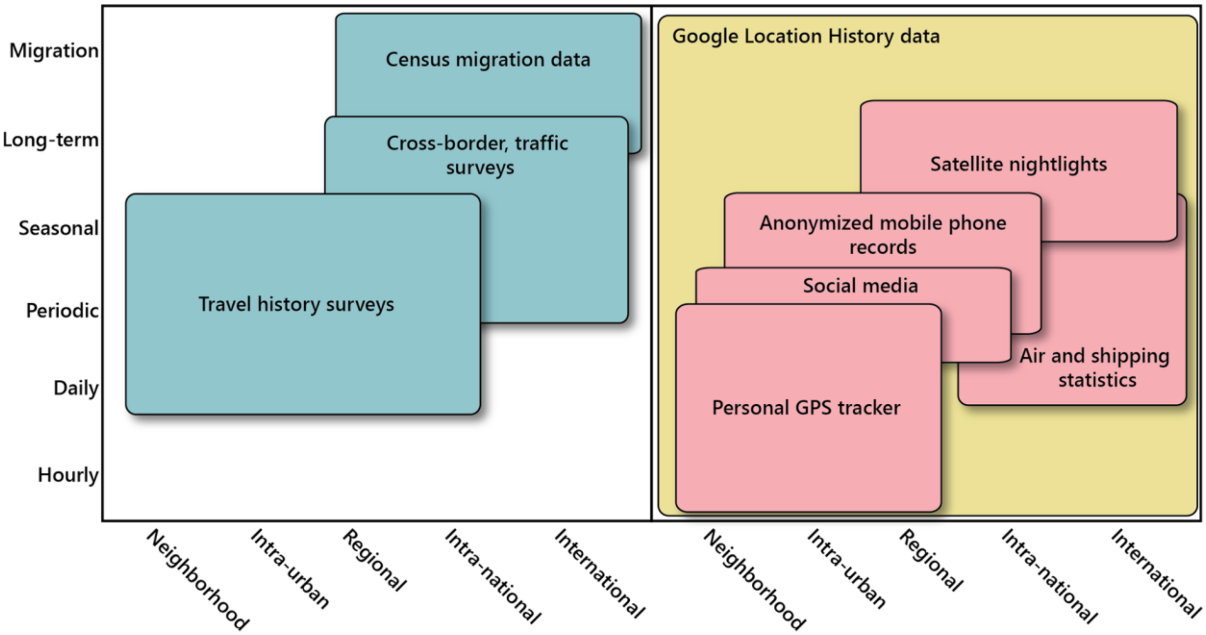


Figure 1. The information niche that Google Location History occupies. From Using Google Location History data to quantify fine-scale human mobility, by Ruktanonchai, et al, 2018. Copyright 2018 by International journal of health geographics; left includes traditional mobility data, right includes mobility data available with more recent technologies. Google Location History data (yellow) record location points similarly to GPS trackers, while spanning timescales similar to mobile phone data, and cover a breadth of time spans and spatial scales not possible in other datasets.

1.1.3 Development of Data Donation System

Some researchers have engaged in developing a general Data Donation System; for example, Web Historian by Menchen-Trevino (2016) allows individuals to visualize different parts of their Google DDPs, and has been applied by Wojcieszak et al. (2022). to investigate how internet users arrive at certain sources of news. Araujo et al. (2017) developed OSD2F, which does not allow the local processing to take place completely before uploading and donating but allows participants to inspect their data to let them decide which parts to share.

On the basis of these previous software systems for the Data Donation System, Dutch researchers developed PORT giving us more concrete insight into the ethical and practical considerations. Those allow the researcher to develop a Python script specifically for their research question. Thus, the software can easily adapt to the research question under investigation.

In our research project, we would generally base on the code of the PORT project by adapting its Python scripts to achieve our goal of controlling clear data error activity classification.

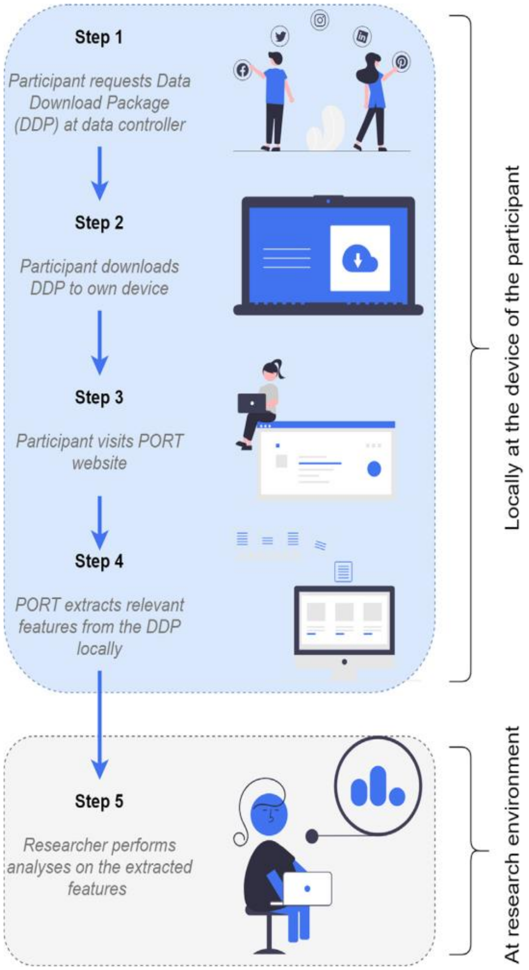

Step 1
Participant requests Data Download Package (DDP) at data controller
Step 2
Participant downloads DDP to own device
Step 3
Participant visits PORT website
Step 4
PORT extracts relevant features from the DDP locally
Step 5
Researcher performs analyses on the extracted features
Locally at the device of the participant
At research environment

Figure 2. Step-by-step illustration of the workflow that allows for a privacy-preserving analysis of data download package, locally on the device of a research participant. From Privacy-preserving local analysis of digital trace data: A proof-of-concept, by Boeschoten et al., 2022. Copyright 2022 by Patterns.

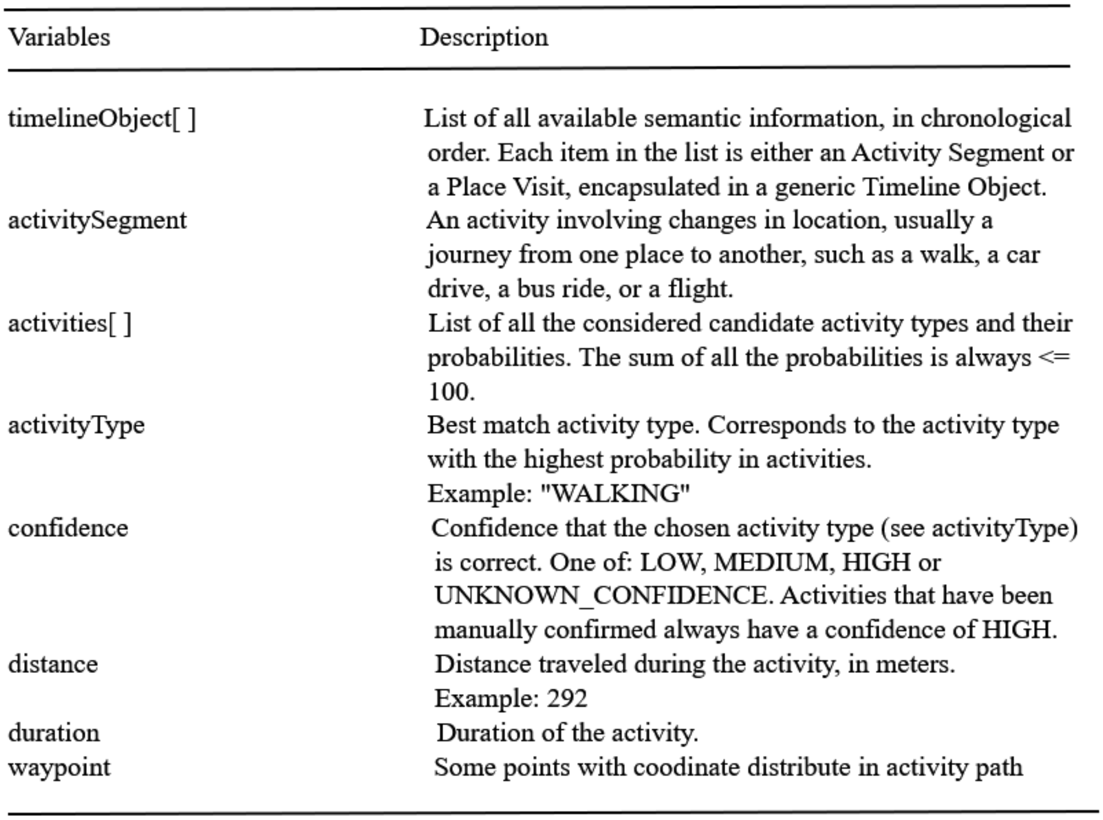
Figure 2 illustrates how PORT can be used in a data donation study. The first step of PORT is always that the participant requests the DDP in which the researcher is interested from the respective data controller. Depending on the size of the DDP, the requirement process would take a couple of seconds and up to weeks. In the second step, the participant downloads the DDP to their own device. In most cases, a confirmation email with a download link would be received, and the required data can be downloaded. In the third step, the participant should visit the researcher's project website at PORT. In the fourth step, regarded as the rather important step, the PORT system would use its extraction algorithm to display the extracted information and request consent from the participant to share this with the researchers. The PORT system basically works depending on two Python files. One will be written with the extraction algorithm, and another will mainly invoke the function from the one with the extraction algorithm to achieve the working. And this is also the step in that we will try to adapt the original Python script to improve the validation of data. In the final step, researchers can perform the analysis based on the extracted data.

1.1.4 Explanation of relevant variables in the JSON file

Knowing the structure of JSON files is also an important condition for adapting the Python script for data error detection, flagging, and fixing. The JSON file researchers get from participants would be divided by months inside the "Semantic Location History" folder inside a Zip file.

In Table 1, all the variables which are relevant to our final activity classification can be found. The description in Table 1 will briefly explain the variables and how they work in the JSON file.

Table 1
Relevant Variables In Semantic Location History



Variables	Description
timelineObject[ ]	List of all available semantic information, in chronological order. Each item in the list is either an Activity Segment or a Place Visit, encapsulated in a generic Timeline Object.
activitySegment	An activity involving changes in location, usually a journey from one place to another, such as a walk, a car drive, a bus ride, or a flight.
activities[ ]	List of all the considered candidate activity types and their probabilities. The sum of all the probabilities is always <= 100.
activityType	Best match activity type. Corresponds to the activity type with the highest probability in activities. Example: "WALKING"
confidence	Confidence that the chosen activity type (see activityType) is correct. One of: LOW, MEDIUM, HIGH or UNKNOWN_CONFIDENCE. Activities that have been manually confirmed always have a confidence of HIGH.
distance	Distance traveled during the activity, in meters. Example: 292
duration	Duration of the activity.
waypoint	Some points with coodinate distribute in activity path

Note. From Semantic Location History Format Definition, by Location History Format, 2023, (https://locationhistoryformat.com/reference/semantic/)

1.2 Data errors in the Data donation system

In this aspect, we will focus on the data errors in the Data Donation System.

We can identify the target data error and the reason why data errors may emerge by investigating the accuracy of the geo-data measurement tool, the function of the classification algorithm, and the features of “Walking” and “Cycling” from geoscience. Moreover, we can collect the key information supporting us in the methodology process.

1.2.1 The potential data errors in GSLH

As mentioned before, errors in the GSLH application are the key to our research. To enumerate the error sources associated with each step in data collection, a highly convenient framework is the total error framework (Biemer, 2016; Japec et al., 2015). Figure 3 gives us a general picture of the total error framework. Measurement error and Algorithm error on the measurement side would be our research focus, which means we will see if the activity type of “Walking” and “Cycling” are reasonable in the data context by constraining the relevant features of these activities.

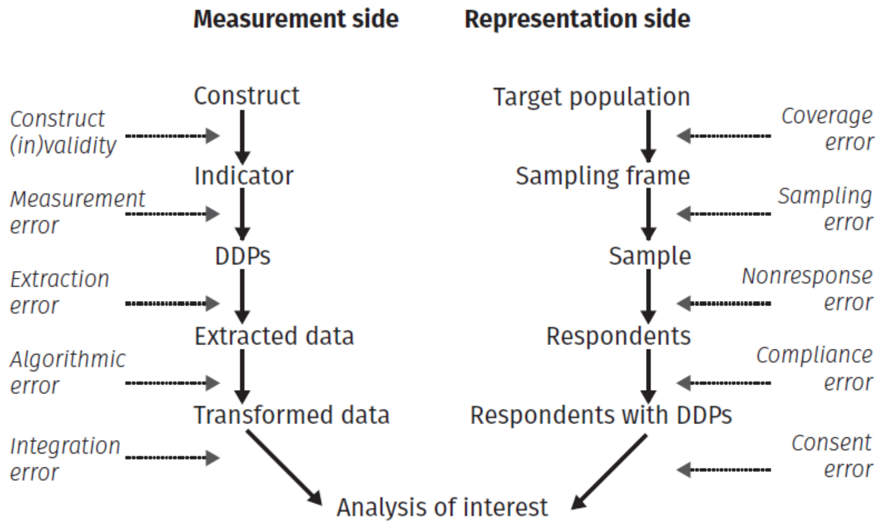


Figure 3. “Total error framework” for general social-scientific data collection. Each step in the data collection process is shown, together with the errors resulting from this step. From A framework for privacy preserving digital trace data collection through data donation, by Boeschoten et al., 2022. Copyright 2022 by Computational Communication Research.

1.2.2 Geo-data measurement tool in the field of accuracy

Compared to the GPS, which can be regarded as a professional tool for measuring geo-data variables over a long time, the performance of GSLH in geo-data is still arguable. According to Yu et al., the estimated time-weighted daily exposures to ambient particulate matter using GMLH (Google Map Location History, same as GSLH) and the GPS data logger were also similar (error less than 1.2%). But on contrast, Macarulla-Rodriguez et al. (2018) focus on the assessment of the accuracy of the locations given by Google Location History Timeline, which variables affect this accuracy, and the initial steps to develop a linear multivariate model that can potentially predict the actual error with respect to the true location considering environmental variables. Their paper shows us a rather poor performance of Google in location history with different internet environments. Although the accuracy of GMLH is not our focus, considering that it can still influence some essential variables in JSON files, we still need to take a look. For example, if the accuracy of GMLH is relatively low, the variable “distance”, “startLocation” and “endLocation” of “activitySegment” may be clearly wrong. It may lead the classification algorithm inside GMLH to get unconvincing classification results of activity type.

1.2.3 Classification Algorithm in the field of geoscience

The classification algorithm is also another important issue in our research which we have very little information about. Not like the classification algorithm in GPS which has many previous investigations. In addition, to optimize the performance, some algorithms like the Gaussian Naïve Bayers algorithm can be applied to improve the performance of classification (Bettini et al., 2020); how the classification algorithm of GSLH works remains unknown. It may increase the difficulty of finding the data error we aim for.

1.2.4 See the data error from the geographical side

As we introduced in the above two parts, the accuracy of the geo-data measurements tool and the classification algorithm itself can affect the final classification results. But considering the limitations of resources, we have few options to control them. In this context, the variables related to the classification play a rather heavy role in the detection of data errors. We shall collect the information from the side of geo-data features inside JSON files and the side of geo features about Walking and Cycling activities for the preparation of designing the methodology part.

1.2.4.1 The features of geo-data in JSON files

The key to our project is the features of geo-data variables in JSON files, particularly which set of variables ought to be chosen in the methodology part and how to constrain them reasonably according to their features.

To a large extend, the main aim of "(spatial) information networks" is data integration, which creates new information mixtures and -structures that are used for geo-communication, especially maps. Data integration needs consistent and time-appropriate geolocation, beside persistent identifiers and well-formed adaptors/keys, in order to establish valid relations to the thematic content (Döllner et al., 2019).

Furthermore, some research on spatial trajectories may also contribute to our project's goal from another side. The huge volume of spatial trajectories enables opportunities for analyzing the mobility patterns of moving objects, which can be represented by an individual trajectory containing a certain pattern or a group of trajectories sharing similar patterns. Rich knowledge can be learned from these trajectories, such as information about road networks, traffic conditions, and driver behavior, contributing to different aspects of the driving experience (Zheng & Zhou, 2011). With this information, we are more familiar with the path features in the traffic activity and it may help us to constrain the geographical data more effectively.

In our case, the variables that can influence the final activity classification result are the 2 key variables "duration", "distance" and some potential variables like "startLocation", "endLocation" and "waypoint".

The two variables "duration" and "distance" can determine the speed which is an important feature for classifying the activity type. The variables "startLocation" and "endLocation" may show us the purpose of the activity, it may give us some hidden indicators about classification. The variables "waypoint" can give us a tool to control the pathway of activity if we need to see what happens in the details of the routine.

1.2.4.2 Walking and Cycling Activity from geographical analysis

Knowing the features of activities "Walking" and "Cycling" from a geographical view can also be a useful foundation for future adaptation of Python script to handle data errors.

Millward et al. (2013) try to analyze active-transport walking behavior from 3 features namely destinations, duration, and distances. According to their paper, home is both the most common origin and destination for active-transport walks, and the most common purpose is travel-to-shop rather than travel-to-work. Most walks are to non-home locations, such as retail establishments and offices. Particularly important are restaurants and bars, grocery stores, shopping centers, banks, and other services. All major destinations show strong distance-decay effects: most walks are shorter than 600 m, and very few exceed 1200 m. In addition, Dabiri & Heaslip (2018) focused on GPS trajectories and also collected some information about different activities.

Table 2 gives us clear information about some important features of different activities in transportation.

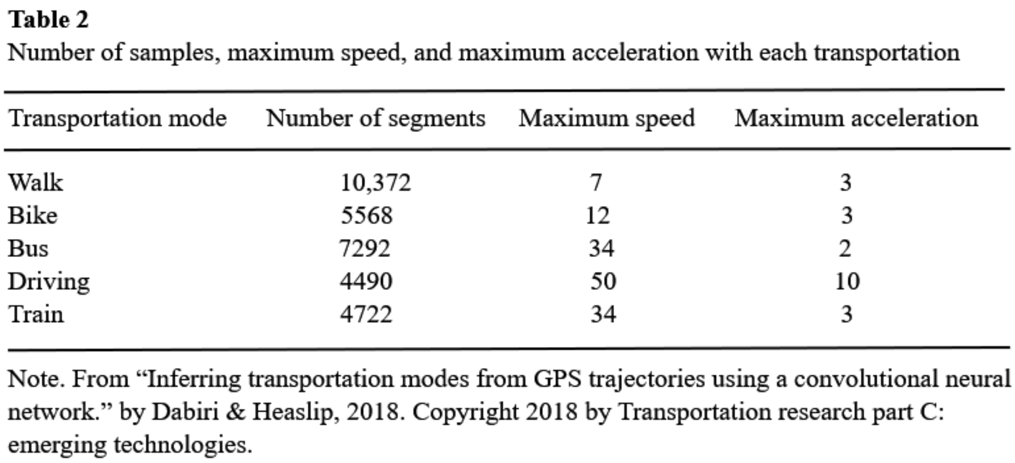


Table 2

Number of samples, maximum speed, and maximum acceleration with each transportation

Transportation mode	Number of segments	Maximum speed	Maximum acceleration
Walk	10,372	7	3
Bike	5568	12	3
Bus	7292	34	2
Driving	4490	50	10
Train	4722	34	3

Note. From "Inferring transportation modes from GPS trajectories using a convolutional neural network." by Dabiri & Heaslip, 2018. Copyright 2018 by Transportation research part C: emerging technologies.

Table 3 gives us more concrete information about active-walking and cycling.

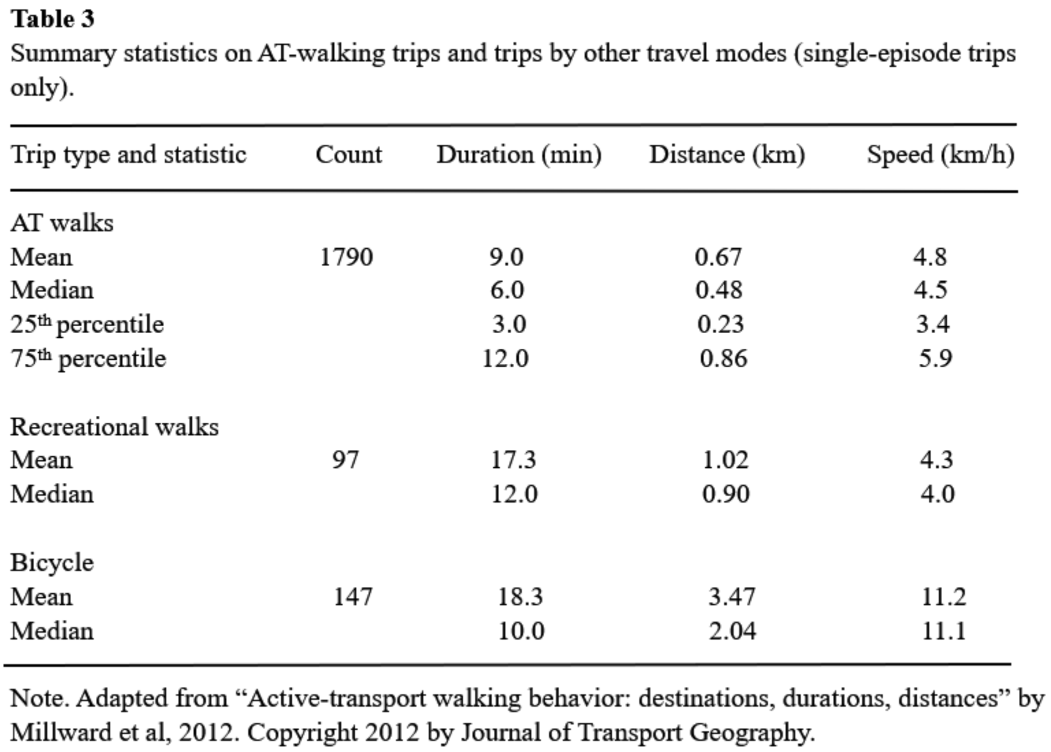


Table 3

Summary statistics on AT-walking trips and trips by other travel modes (single-episode trips only).

Trip type and statistic	Count	Duration (min)	Distance (km)	Speed (km/h)
AT walks				
Mean	1790	9.0	0.67	4.8
Median		6.0	0.48	4.5
25th percentile		3.0	0.23	3.4
75th percentile		12.0	0.86	5.9
Recreational walks				
Mean	97	17.3	1.02	4.3
Median		12.0	0.90	4.0
Bicycle				
Mean	147	18.3	3.47	11.2
Median		10.0	2.04	11.1

Note. Adapted from "Active-transport walking behavior: destinations, durations, distances" by Millward et al, 2012. Copyright 2012 by Journal of Transport Geography.

Similarly, Goel et al. (2022) try to from some simple questions like “who cycles?”, “for what purpose?”, and “how far?” to give us an overview of cycling in 17 countries across 6 continents. According to the paper, distance-based ratios, which means the ratio of cycling mode share within each distance range to the overall cycling mode share of the geography, are normally shorter than 5 km, and above 20 km is from a statistical view really rare case.

However, data on the speed of bicycles remains inconsistent. Two US studies found comparable mean speeds of 18 km/h (Dill & Gliebe, 2008) and 16 km/h (Thompson et al., 1997). Other investigations from Europe have reported mean speeds between 12 km/h and 14 km/h for conventional cyclists (Dozza & Werneke, 2014; Menghini et al., 2009).

E-bike is a new factor in cycling activity, Schleinitz et al. (2015) have shown the difference in speed of three types of bike namely the maximum 22.0 km/h for conventional bike, 31.0 km/h for Pedelec bike, and 31.9 km/h for S-Pedelec bike.

Few academic studies have investigated the max distance for recreational purposes. According to a sports website named BSX Insight, long distances for recreation are best done at 45 miles (roughly 72 km).

1.2.5 Proposed solution for error data

The proposed solution for these data errors could be flagging the data to the data user which means some relevant features will be shown and if some potential data errors are detected, there will be a statement to explain the error to the data user.

Following the flagging, data imputation means that the potential data error will be replaced by more reasonable data with certain conditions.

2. Method

This section explains the methodology used to investigate errors in GSLH data. We will first describe the test data used in this project. Then, the quality checks we impose on the data will be explained, and lastly, we introduce the structure of data processing pipelines.

2.1 Data Collection

In the intended situation, Google Semantic Location History (GSLH) data will be obtained from participants through a data donation pipeline. For the present study, however, access to several packages was limited by considerable privacy concerns. Therefore, the pipeline was developed on the basis of the author’s and member group’s GSLH data.

2.1.1 Variables for constraint conditions description

The information about the variables for constraint conditions and the types of activities there would give some additional context.

“Distance” and “Speed” play the key role in determining activity type classification. Our main concern will be the validation and reliability of variables “duration” and “distance”.

2.1.1.1 Variable Duration

The variable “duration” of GSLH in the JSON file can be regarded as rather valid with the background that few researchers raise the question in this field, we may infer the reason lies in that the measurement of duration does not need communication with the internet or satellites and every phone count the time based on its clock. However, errors can still occur, but without additional external information, we can not easily access the validity of the duration information for each activity.

2.1.1.2 Variable Distance

The variable “distance” should be paid more attention for the production of “distance” in GSLH needs more than one process. The accuracy of ordinates collected by satellites, different receipt devices, and even different distance calculations formula will affect the final results of “distance”.

The coordinates of “startLocation” and “endLocation” are given, we can base on these coordinates and the geoscience distance formula to calculate the distance between “startLocation” and “endLoaction”. The difference between this self-calculated distance and the original distance of GSLH in the JSON file can be an option to test the validation of the variable “distance”.

There are still some issues, like how the Google algorithm in GSLH calculates the distance, which can not be answered. If the Google algorithm adopts the pathway or routine rather than the coordinates to get distance, the method of testing distance difference may be problematic initially. But considering that the distance of activity “Walking” and “Cycling” normally would not be so far, the method of testing the distance difference may still be useful in our case.

2.1.1.3 Other relevant variables

Although the variables like “startLocation”, “endLocation” and “waypoint” may also contribute to limiting data error, we lack some relevant information to identify the addresses particularly the type of these places of “startLocation” and “endLocation” for only their coordinates are given in the JSON file; similarly, we may not be able to use “waypoint” because we can not put those points which also are given by coordinates in a relevant map to see if there are some wrong points in the pathway. Some methods provided by Google like Place API may be an alternative choice for identifying the type of address, but it also requires some extra sources which are beyond our accessibility.

2.2 Processing Pipeline Design

The pipeline described below was implemented in Python version and shown in Anaconda prompt, available in the supplementary materials to this paper. In general, data processing has two components: error flagging and imputation. The error flagging is mainly concerned with detecting and describing problems in the GSLH data, and in the imputation part, we create some options for solving or ameliorating the errors found in the first part.

2.2.1 Flagging the potential data error

When designing the processing pipeline for "Walking" and "Cycling," it is important to consider each activity's specific features and characteristics from a geoscience view.

2.2.1.1 Constraint standard “confidence”

The first constraint will be relevant to the variable “confidence”. We will focus on walking and cycling activities with low confidence. A low confidence level means that the classification algorithm inside Google may not be sure if this activity classification is certainly right and normally if the activity with low confidence, the probability of this activity and the second-place activity in “activities” are relatively close. Thus, the second-place activity has a chance to replace the first one.

2.2.1.2 Constraint Standards of “Distance” and “Speed” for cycling activity

Our study primarily concentrates on conventional bicycles and does not encompass the impacts of E-bikes, given that conventional bicycles still constitute the majority. Due to the inconsistent speeds of bicycles, we adopt a broad range to mitigate the risk of imposing strict constraints. Regarding distance, implementing two-level standards, one for commuting and another for recreational purposes proves viable in mitigating the influence of varying distance types.

For the activity “Cycling”, the features will be the following:

The maximum speed should be set as 25 km/h. The maximum distance that can be regarded as a standard line in many countries is 20 km for commute cycling and 75 km, a standard suggested by a sports website for recreational cycling. Although some cycling activities like "Tour de France" may be above this 75 km standard line, it is so rare that we can neglect them.

2.2.1.3 Constraint Standards of “Distance” and “Speed” for walking activity

For the activity “walking”, what makes it complicated is the different types of walking. Inside the Semantic Location History, there are different types of walking namely “Walking”, “Nordic walking” and “Hiking”.

In our study, we mainly pay attention to “Walking”. Within the context of "Walking," there exist distinct types, such as "active-transportation walking" denoting walking as a means of commuting to work, school, or public transport stations as part of the daily routine. Another type is "recreational walking," referring to walking undertaken for leisure and relaxation, but excluding activities like "Nordic walking" or "Hiking." Typically, "recreational walking" occurs within familiar surroundings, such as the vicinity of one's home or other familiar locations.

Based on Table 3 in the introduction, in the single-episode trips, the mean and median of “active-transportation walking” are 0.67 km and 0.48 km respectively, and the mean and median of “recreational walking” are 1.02 km and 0.90 km respectively. While we lack precise information on the maximum distance for walking in daily life, we can make a rough assumption based on the provided data that the maximum walking distance should not exceed 10 km. Additionally, the maximum walking speed can be set at 7 km/h, according to Table 2.

Finally, we can ensure that the constrained standards:

1. Activity with low confidence.

2. Walking activity: maximum distance 10 km, maximum speed 7 km/h

3. Cycling activity: maximum distance 20 km for commute and 75 km for recreation

maximum speed 25 km/h

2.2.2 Activity and data imputation

We set two levels framework aiming to achieve the imputation. One is about activity imputation which means the flagged activity will be replaced, another is about data imputation which means the relevant variable of the flagged activity will be replaced.

It is important to mention that the flagged distance for commute shall not be included in the activity and data imputation process because the two-level cycling distance flag is aimed at letting the data users know that the cycling distance may vary at a large scale because of the different purposes. Thus, the distance for commute cannot be regarded as our target of activity and data imputation strictly.

2.2.2.1 Activity imputation

The simplified way to make an alternative activity imputation can be to let the second-place activity in “activities” in “activitySegment” replace the original one.

One condition for activating the activity imputation to let the second-place activity replace the original one is that the constraint condition for the target activity will be much higher than the maximum speed. For example, if the speed of one walking activity is 8.5 km/h, although it is higher than the maximum speed, it is still difficult to distinguish the walking activity from the second-place activity.

Thus, we may set a rather broad standard (plus roughly 25 percent for maximum speed) for activating this activity imputation:

Distance shall be above 12.5 km, and speed shall be above 10 km/h for walking activity.

Distance for recreation shall be above 95 km, and speed shall be above 35 km/h considering the factor of E-bikes for cycling activity.

2.2.2.2 Data imputation

Data imputation serves as an alternative when activity imputation cannot be activated. When the distance or speed surpasses the maximum standard but falls short of the activity imputation threshold, we will concentrate on substituting specific data with the highest standard within the flagged activity. In our scenario, the variable subject to replacement in data imputation will be the "distance" variable, assuming the "duration" variable remains flawless.

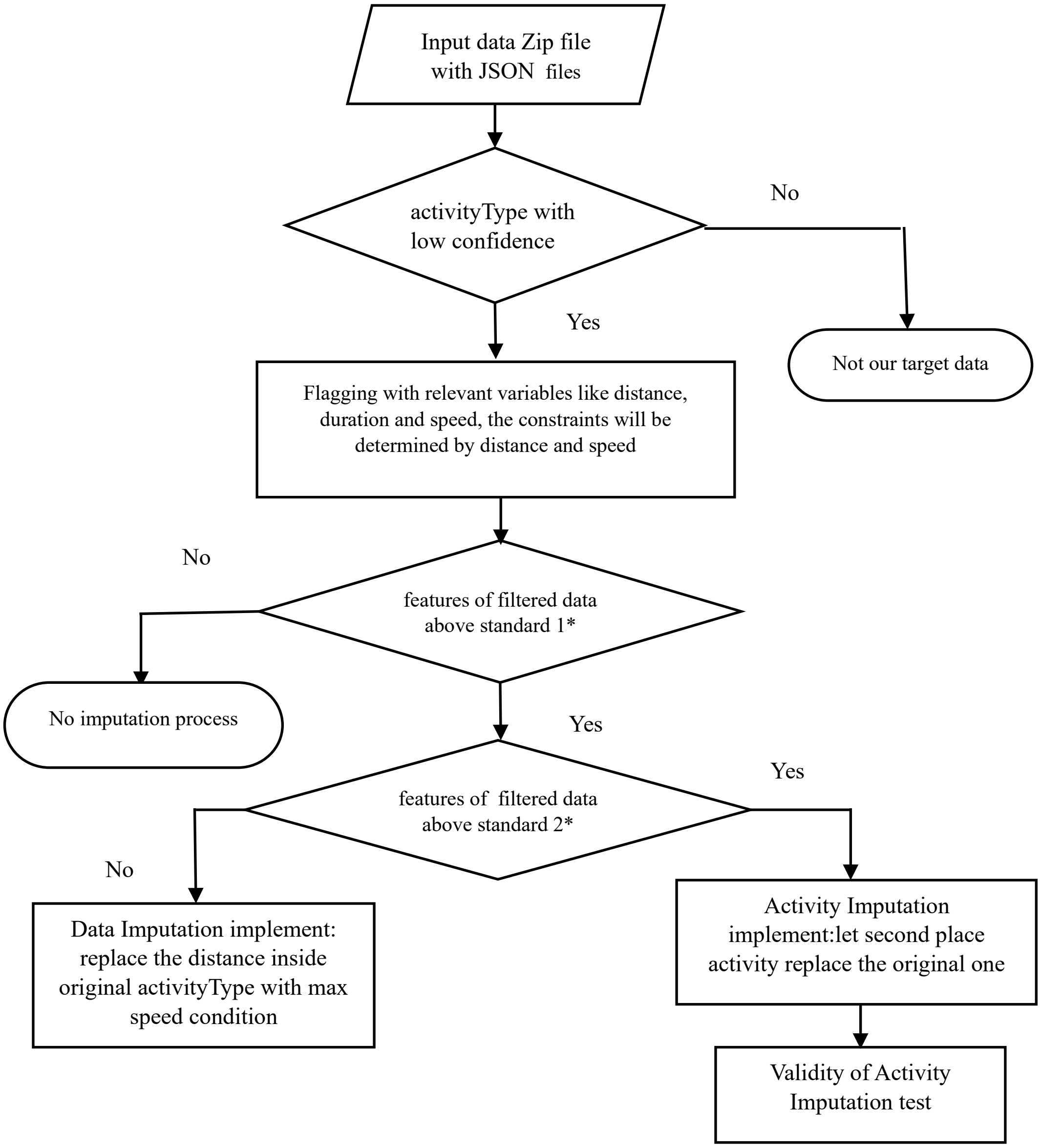
For the “Walking” activity, all the “distance” of the flagged activity should be replaced by the new distance calculated by 7 km/h multiplied the every segment duration.

For the “Cycling” activity, all the “distance” of the flagged activity should be replaced by the new distance calculated at 25 km/h multiplied the every segment duration.

Figure 4 illustrates the process of workflow for flagging and imputation.

Figure 4

Workflow for flagging and imputation of data errors


Input data Zip file with JSON files
activityType with low confidence
No
Yes
Not our target data
Flagging with relevant variables like distance, duration and speed, the constraints will be determined by distance and speed
No
features of filtered data above standard 1*
No imputation process
Yes
Yes
features of filtered data above standard 2*
No
Data Imputation implement: replace the distance inside original activityType with max speed condition
Activity Imputation implement:let second place activity replace the original one
Validity of Activity Imputation test

Note. standard 1: constraints for max distance and speed. standard 2: activation line for Activity Imputation

3. Results

In the results section, we first provide descriptive statistics on the errors detected in our pilot data. Then, how the process works by giving a certain example of a walking activity will be shown. The flagging process, activity or data imputation, and retest after imputation for cycling will generally follow the same workflow, some relevant parameters will be changed in accordance with the cycling activity constraint condition. The slight difference lies in the cycling activity being flagged with two-speed levels.

To keep the analyses comprehensible, we selected a single month (December 2018) containing various transport modes and travel behavior.

All the results will be shown in the Anaconda prompt. The version Python 3.8 will execute the Python scripts we design. The codes will be written through the IDE PyCharm.

3.1 The statistic about error

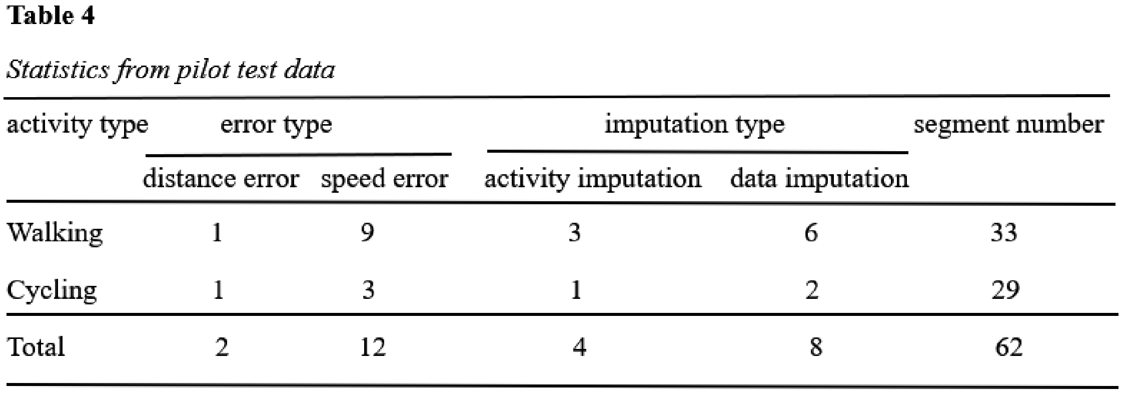


Table 4

Statistics from pilot test data

activity type	error type		imputation type		segment number
	distance error	speed error	activity imputation	data imputation	
Walking	1	9	3	6	33
Cycling	1	3	1	2	29
Total	2	12	4	8	62

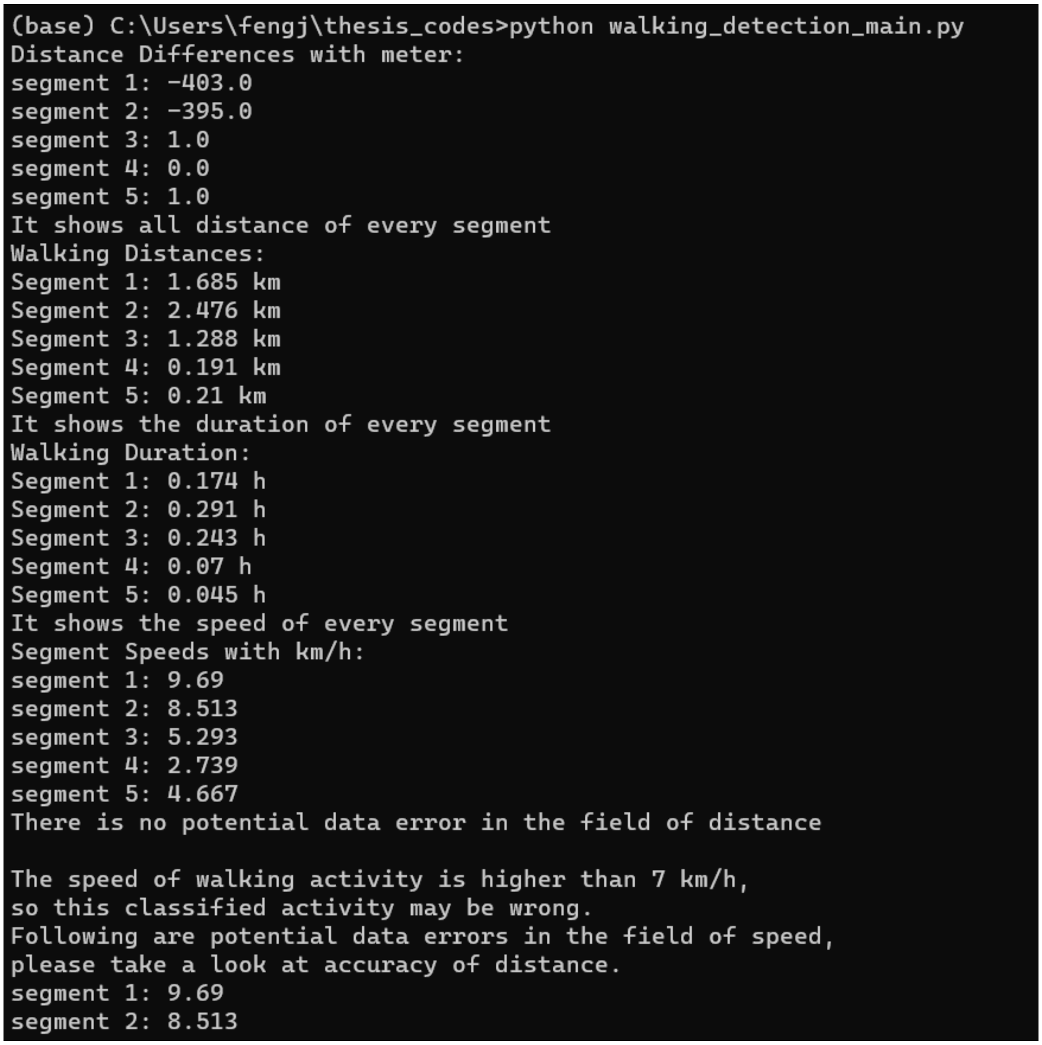
Table 4 will give the statistic from our pilot test data about error detection and imputation.

For the “Walking” activity, based on my data source, 1 error related to distance and 9 errors related to speed are found in 33 segments of 9 JSON files.

For the “Cycling” activity, based on the data source from a group member, 1 error related to distance and 3 errors related to speed are found in 29 segments of 12 JSON files.

Almost 22.6% of segments in our small-scale data source have errors. More errors in the “Walking” activity than in the “Cycling” activity and more errors can be filtered by the constraint condition of speed than the errors filtered by the constraint condition of distance. More data imputation will be implemented than activity imputation.

3.2 Results of flagging



```
(base) C:\Users\fengj\thesis_codes>python walking_detection_main.py
Distance Differences with meter:
segment 1: -403.0
segment 2: -395.0
segment 3: 1.0
segment 4: 0.0
segment 5: 1.0
It shows all distance of every segment
Walking Distances:
Segment 1: 1.685 km
Segment 2: 2.476 km
Segment 3: 1.288 km
Segment 4: 0.191 km
Segment 5: 0.21 km
It shows the duration of every segment
Walking Duration:
Segment 1: 0.174 h
Segment 2: 0.291 h
Segment 3: 0.243 h
Segment 4: 0.07 h
Segment 5: 0.045 h
It shows the speed of every segment
Segment Speeds with km/h:
segment 1: 9.69
segment 2: 8.513
segment 3: 5.293
segment 4: 2.739
segment 5: 4.667
There is no potential data error in the field of distance

The speed of walking activity is higher than 7 km/h,
so this classified activity may be wrong.
Following are potential data errors in the field of speed,
please take a look at accuracy of distance.
segment 1: 9.69
segment 2: 8.513
```

Figure 5. With running "walking_detection_main.py" Python script, the relevant features will be extracted and the note and potential error will be given. Note.

Figure 5 will show the 4 features of walking activity with low confidence.

Then, there are 2 statements about constraint features:

1. It will be about the constraint condition of maximum distance.
In our example, the statement "There is no potential data error in the field of distance" will be given because no segment has an above 10 km distance.

If there is a distance above 10 km, the statement "The distance of walking activity is more than 10 km, so this classification may be wrong. Following are potential data errors in the field of distance, Segment distance with km:" will be given.

2. It will be about the constraint condition of maximum speed.

In our example, the statement “The speed of walking activity is higher than 7 km/h, so the classified activity may be wrong. Following are potential data errors in the field of speed, please take a look at the accuracy of distance” will be given.

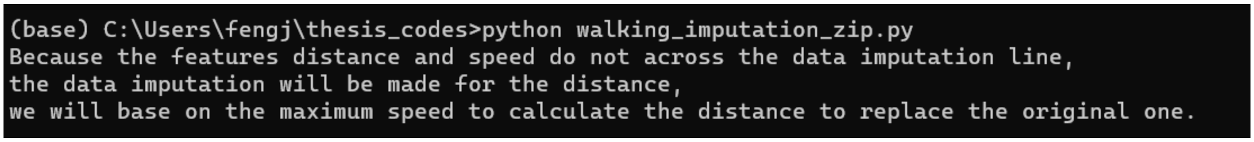
We can detect 2 segments with speed above 7 km/h through the Python script. And according to the above-shown results of “Distance Difference”, the two segments have a speed above 7 km/h which also have an unnormal “Distance Difference” namely 403 m and 395 m respectively. The warning should be given to data users that there might be a connection between these unnormal “Distance Difference” and the higher than maximum speed of the segment.

If there is no speed above 7 km/h, the statement “There is no potential data error in the field of speed” will be given.

3.3 Results of activity imputation and retest after the imputation

The replacement for the filtered walking activity in the flagging process will be implemented in this process. As mentioned in Method, we adopt one simplified way: let the second place activity in “activities” of “activitySegment” replace the walking activity if the detected features can cross the activating line we set. And the final process will be retested if the flagged walking activities are still there.

3.3.1 Activity and data imputation



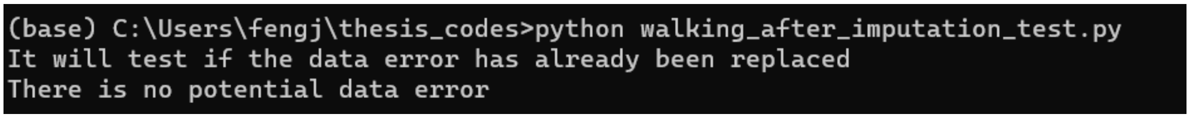
```
(base) C:\Users\fengj\thesis_codes>python walking_imputation_zip.py
Because the features distance and speed do not across the data imputation line,
the data imputation will be made for the distance,
we will base on the maximum speed to calculate the distance to replace the original one.
```

Figure 6. With running “walking_imputation_zip.py” Python script, walking activity with unnormal distance and speed can not cross the activating standard for data imputation, so the data imputation for distance will be executed.

Figure 6 shows that the walking activity imputation can not be executed because the two features distance or speed do not across the activating line. Thus, data imputation for distance rather than activity imputation will be executed.

If the features of flagging walking activities can cross the activating line, the activity imputation process will be activated. In the Anaconda prompt, the statement “The imputation has been made” and a new Zip file named “Modified Location History For Walking” will be created in the path dictionary. It contains the changed JSON file with replaced activity in “activitySegment”. If no data is flagged, the statement “There are no potential wrong data errors among these files” will be outputted in the Anaconda prompt.

3.3.2 Retest after the imputation



```
(base) C:\Users\fengj\thesis_codes>python walking_after_imputation_test.py
It will test if the data error has already been replaced
There is no potential data error
```

Figure 7. With running “walking_after_imputation_test.py” Python script. Note. It is not from my above example.

Figure 7 shows us the retest result for the “Modified Location History” zip file. If the data imputation process is activated, the statement “It will test if the data error has already been replaced. There is no potential data error.”

3.4 Conclusion for results part

For flagging, we provide the data user with more details on the relevant features as a background of the two constraint condition statements. On the condition that the duration of the activity will not be a problematic issue, the accuracy of the feature “distance” will be warned to the data user that it may take effect in the classification process of the wrong activity.

For the imputation part, the features of flagged data error can not meet the standard of activating the activity imputation process although they are above the maximum speed. Thus, a data imputation for distance in the flagged activity will be executed.

To prove the validity of the imputation, I checked the second-place activities inside the JSON file we processed. Both second-place activities for the flagged “Walking” activities are ”Vehicle” activities with a probability of 27.64 and 34.86 respectively. And based on the true routing information, it can be concluded that not activating the data imputation is reliable as no passenger vehicle was used in this trajectory.

4. Conclusion and Discussion

4.1 Conclusion

To answer the research question of how to handle the data error inside Google Semantic Location History Format, or more concretely, the problem of data errors in the GSLH, particularly in activity “Walking” and “Cycling” inside the Data Donation System, we designed processing pipeline (several Python scripts), which are generally based on the PORT project. By identifying pertinent features within the data extracted from JSON files, such as "Distance," "Duration," and "Speed," taking into account the interdependence of the preceding two attributes, as well as noting and addressing data errors, data users can gain a comprehensive understanding of the workflow and data related to the classification of "Walking" and "Cycling" activities.

Data imputation will be a supplemental way to deal with data errors after the flagging. We design Python scripts to reasonably replace those data errors detected in the phase of the flagging. The condition for activating the data imputation process will be a way to avoid the designed wrong data imputation. The data imputation process remains to be improved because we do not consider whether the second-place imputation matches the features we extracted and calculated in the flagging stage.

The system generally performs well in our pilot data source, but it is still to note that the number of errors that we flagged takes an unusual proportion, particularly, nearly 30% of the walking activity with low confidence has an error. This may be due to our private small-scale data source. However, if this phenomenon is a general situation with large data sources, we may consider updating some of the constraint parameters.

To sum up, we aimed to design a system to improve the validity of “Walking” and “Cycling” activity classification with the limitary resource in the GSLH of data donation. It can flag certain data errors to data users (in our case human behavior researchers) but may still need more work on how to replace the detected potential wrong data with reasonable results properly. We may need more data to validate our processing pipeline.

4.2 Discussion

In light of the conclusion drawn, our objective of mitigating measurement error in human behavior studies through the utilization of digital trace data has been accomplished. The ensuing discussion will enumerate certain limitations that have been identified for improvement. Furthermore, potential enhancements and future research avenues will be proposed to address these limitations.

The method design of our study presents several limitations that need to be acknowledged. These limitations can be categorized into parameters within the processing pipeline and the lack of external resources.

4.2.1 Limitations in method design

4.2.1.1 Limitations related to the parameters in method design

a. Maximum distance and speed for activities like "Walking" and "Cycling" may vary due to factors like gender, age, environment, and activity purpose.

b. The assumption of an impeccable "duration" variable may lead to potential issues, such as allowing unrealistic durations for cycling activities (e.g., over 24 hours).

c. No constraint about slow speed. More information needs to be collected to deal with a rather slow speed (e.g., 5 km/h for cycling).

d. Specific types of walking activities like "Hiking" and "Nordic walking" within Semantic Location History require more information for precise classification and understanding of their impact on maximum distance and speed. In addition, distinguishing between active-transportation walking and recreational walking requires more information to determine their influence on maximum distance and speed.

e. Semantic Location History includes only one type of cycling activity, which poses challenges in imposing restrictions on maximum distance and speed due to differences between active-transport cycling and recreational cycling. Implementing two flagging levels helps mitigate differences, but the residual risk remains. The impact of E-bikes as a special factor also needs to be addressed.

f. Bike culture, particularly in the Netherlands, where cycling is widely adopted as a means of transportation, introduces uncertainty in using global cycling data parameters. Further investigation is needed to determine if these parameters suit the Netherlands well.

4.2.1.2 Limitations related to activity and data imputation

Another limitation lies in the imputation of activities and data. While our data imputation approach is reasonable given the different standards used for flagged data, certain issues require consideration. For example, the standard for activating activity imputation may lack convincing evidence, and the compatibility between the second-place activity and the extracted features needs verification.

Furthermore, the impact of replacing only the "activityType" without considering other related variables, such as "confidence," remains unknown. We welcome feedback from human behavior researchers to enhance this aspect of our methodology.

4.2.2 Limitations of Lack of the external resources

The processing pipeline devised for managing data errors in GSLH represents a promising initial step. However, further improvements could be achieved by accessing additional external resources.

Firstly, comprehensive knowledge of Google's algorithm for classifying "activityType" within Semantic Location History is essential. A deeper understanding of the model, variables, standards, and parameters established by Google would enable us to attain precise values for the relevant features. Currently, our observations are limited to the "activityType" and its corresponding probability within the "activities" data, leaving us unaware of the underlying classification criteria set by Google.

Secondly, using relevant information from the JSON file provided by Semantic Location History is restricted. While constraints based on variables like "distance" and "duration" help establish a foundational framework for managing data errors, they may not fully capture the complexity of the data. Variables such as "startLocation" and "endLocation," which can provide insights into the purpose of an activity, as well as the "waypoint" variable, which allows for the analysis of activity routines, cannot be flagged or included in our current analysis.

Thirdly, the number of accessible JSON files in our project is limited. To validate and enhance the reliability of our Python scripts for data detection and control, it is necessary to analyze more data from various data donation participants. Furthermore, our data source might not encompass all potential problems associated with the JSON file structure.

4.3 Improvements and future works

On the basis of the points mentioned in the limitations. We provide some feasible ways to improve our method design and give some thoughts about what we do if we can access more external resources.

4.3.1 Supplemental travel history survey

More JSON files that are varied and can represent the target population are highly required to be provided to prove the validity and reliability of the Python script. Based on these data, we can use the travel history survey regarded as a normal tool in geoscience as a supplemental way to test the validity of data flagging and imputation. The participants will be asked to record their daily routine for a certain timetable. At the same time, the data collection of GSLH will also be implemented. A comparison between the data from daily routine and from GSLH will be made. Then the confusion matrix can be applied to see how the performance of accuracy.

4.3.2 Data imputation model training

Compared to the simplified way to make a data imputation which we mainly focused on before, namely letting the second place activity replace the detected potential wrong activity, training a model with external relevant variables may be more credible.

An experiment will be designed for relevant data collection. The relevant variables which may affect the classification of activities “Walking” and “Cycling” will be listed. One classification model will be chosen, the data collected by the experiment will be as training data inputted to the classification model, and the weight for each relevant variable will be trained.

Finally, the results of the classification model trained by ourselves will be compared with the results of Google activity classification in Semantic Location History. A supplemental routine questionnaire may also be used to see how the performance of accuracy in each model.

References

Araujo, T., Ausloos, J., van Atteveldt, W., Loecherbach, F., Moeller, J., Ohme, J., ... & Welbers, K. (2022). OSD2F: An open-source data donation framework. *Computational Communication Research,* 4(2), 372-387. https://doi.org/10.5117/CCR2022.2.001.ARAU

Araujo, T., Wonneberger, A., Neijens, P., & de Vreese, C. (2017). How much time do you spend online? Understanding and improving the accuracy of self-reported measures of internet use. *Communication Methods and Measures,* 11(3), 173-190. https://doi.org/10.1080/19312458.2017.1317337

Ausloos, J., & Veale, M. (2020). Researching with Data Rights. *Technology and Regulation, 2020, 136–157.* http://dx.doi.org/10.2139/ssrn.3465680

Bettini, C., Civitarese, G., & Presotto, R. (2020). Caviar: Context-driven active and incremental activity recognition. *Knowledge-Based Systems,* 196. https://doi.org/10.1016/j.knosys.2020.105816

Biemer, P. P. (2016). Errors and inference. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter,& J. Lane (Eds.), *Big data and social science: A practical guide to methods and tools* (pp. 266–297). CRC press. https://doi.org/10.18637/jss.v078.b02

Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T., & Oberski, D. L. (2022). A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research,* 4(2), 388-423. https://doi.org/10.5117/CCR2022.2.002.BOES

Boeschoten, L., Mendrik, A., van der Veen, E., Vloothuis, J., Hu, H., Voorvaart, R., & Oberski, D. L. (2022). Privacy-preserving local analysis of digital trace data: A proof-of-concept. *Patterns*, 3(3), 100444. https://doi.org/10.1016/j.patter.2022.100444

BSXINSIGHT. (2023). How Many Miles Should I Bike A Day? Complete Guide 2023. https://www.bsxinsight.com/how-many-miles-should-i-bike-a-day/

Dabiri, S., & Heaslip, K. (2018). Inferring transportation modes from GPS trajectories using a convolutional neural network. *Transportation research part C: emerging technologies,* 86, 360-371. https://doi.org/10.1016/j.trc.2017.11.021

Dill, J., Gliebe, J., (2008). Understanding and Measuring Bicycling Behavior: A Focus on Travel Time and Route Choice. https://doi.org/10.15760/trec.151

Döllner, J., Jobst, M., & Schmitz, P. (2019). Service-Oriented Mapping. *Lecture Notes in Geoinformation and Cartography*. https://doi.org/10.1007/978-3-319-72434-8

Dozza, M., Werneke, J., (2014). Introducing naturalistic cycling data: what factors influence bicyclists' safety in the real world? *Transportation research part F: traffic psychology and behaviour*, 24, 83–91. http://dx.doi.org/10.1016/j.trf.2014.04.001.

GeoPy. (n.d.). Welcome to GeoPy's documentation! https://geopy.readthedocs.io/en/stable/

GO FAIR initiative. (2022). FAIR Principles - GO FAIR. GO FAIR. https://www.go-fair.org/fair-principles/

Goel, R., Goodman, A., Aldred, R., Nakamura, R., Tatah, L., Garcia, L. M. T., ... & Woodcock, J. (2022). Cycling behaviour in 17 countries across 6 continents: levels of cycling, who cycles, for what purpose, and how far?. *Transport reviews,* 42(1), 58-81. https://doi.org/10.1080/01441647.2021.1915898

Hansen, K. B., & Nielsen, T. A. S. (2014). Exploring characteristics and motives of long distance commuter cyclists. *Transport Policy,* 35, 57-63. https://doi.org/10.1016/j.tranpol.2014.05.001

Japec, L., Kreuter, F., Berg, M., Biemer, P. P., Decker, P., Lampe, C., Lane, J., O'Neil, C., & Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly,* 79(4), 839–880. https://doi:10.1093/poq/nfv039

Kirmse, A., Udeshi, T., Bellver, P., & Shuma, J. (2011, November). Extracting patterns from location history. *In Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 397-400). https://doi.org/10.1145/2093973.2094032

Location History Format. (n.d.). Semantic Location History Format Definition. https://locationhistoryformat.com/reference/semantic/#/$defs/activityType/

Macarulla Rodriguez, A., Tiberius, C., van Bree, R., & Geradts, Z. (2018). Google timeline accuracy assessment and error prediction. *Forensic sciences research,* 3(3), 240-255. https://doi.org/10.1080/20961790.2018.1509187

Menchen-Trevino, E. (2016). Web historian: Enabling multi-method and independent research with real-world web browsing history data. *IConference 2016 Proceedings*. https://doi.org/10.9776/16611

Millward, H., Spinney, J., & Scott, D. (2013). Active-transport walking behavior: destinations, durations, distances. *Journal of Transport Geography*, 28, 101-110. https://doi.org/10.1016/j.jtrangeo.2012.11.012

Moncayo-Unda, M. G., Van Droogenbroeck, M., Saadi, I., & Cools, M. (2022). An anonymised longitudinal GPS location dataset to understand changes in activity-travel behaviour between pre-and post-COVID periods. *Data in Brief,* 45. https://doi.org/10.1016/j.dib.2022.108776

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data,* 2, 13. https://doi.org/10.3389/fdata.2019.00013

Ruktanonchai, N. W., Ruktanonchai, C. W., Floyd, J. R., & Tatem, A. J. (2018). Using Google Location History data to quantify fine-scale human mobility. *International journal of health geographics,* 17, 1-13. https://doi.org/10.1186/s12942-018-0150-z

Schleinitz, K., Petzoldt, T., Franke-Bartholdt, L., Krems, J., & Gehlert, T. (2017). The German Naturalistic Cycling Study–Comparing cycling speed of riders of different e-bikes and conventional bicycles. *Safety science,* 92, 290-297. https://doi.org/10.1016/j.ssci.2015.07.027

Skatova, A., & Goulding, J. (2019). Psychology of personal data donation. *PloS one*, 14(11). https://doi.org/10.1371/journal.pone.0224240

Wojcieszak, M., Menchen-Trevino, E., Goncalves, J. F., & Weeks, B. (2022). Avenues to news and diverse news exposure online: Comparing direct navigation, social media, news aggregators, search queries, and article hyperlinks. *The International Journal of Press/Politics,* 27(4), 860-886. https://doi.org/10.1177/19401612211009160

Yu, X., Stuart, A. L., Liu, Y., Ivey, C. E., Russell, A. G., Kan, H., ... & Yu, H. (2019). On the accuracy and potential of Google Maps location history data to characterize individual mobility for air pollution health studies. *Environmental pollution,* 252, 924-930. https://doi.org/10.1016/j.envpol.2019.05.081

Yuan, J., Zheng, Y., Xie, X., & Sun, G. (2011, August). Driving with knowledge from the physical world. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 316-324). https://doi.org/10.1145/2020408.2020462