A Data Visualization Tool using Streamlit for the GeoLife Mobility Data and Weather Data for Beijing, China

Ekin Ugurel[1], Steffen Coenen[2], Vaibhavi Lakshmi Segu[3], Dhruvil Patel[4]

[1]THINK Lab, Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98109; e-mail: ugurel@uw.edu
[2]Sustainable Transportation Lab, Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195; e-mail: scoenen@uw.edu
[3]Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195; e-mail: vlsegu@uw.edu
[4]Sustainable Transportation Lab, Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195; e-mail: dpatel28@uw.edu

ABSTRACT

Better infrastructure has provided people with greater opportunities of choosing different modes of transportation based on their requirement or necessity. There are different modes of transportation, including walking, biking, public transportation, car, and taxi. This project focuses on the factors that lead people to choose a given mode of transportation, specifically targeting weather factors such as precipitation, temperature, wind speed, visibility, and humidity.

For analyzing the modal choice, GPS trajectory data from the GeoLife dataset for Beijing, China, and weather data were considered. The data sources were connected based on timestamps. A data visualization tool using Streamlit was used to make the exploration of the dataset accessible to anyone. The tool provides multiple visualization types on mode choices, weather conditions, and common start and end points of trips.

In addition to this tool, a multinomial logit model was used to check how each factor contributes to the mode choice. On weekends, and when the temperature is low, people tend to use their car or a taxi more often. Subways are less frequently used on weekends compared to weekdays. Also, when it is raining, the use of cars is strongly promoted, compared to other modes of transportation. At last, two cluster analyses were carried out for the mode of walking, revealing two distinct groups of walkers in each analysis.

Overall, it can be concluded that this project revealed several inferences about people's mode choice, majorly relating to weather factors, day of the week, and start and end hours.

INTRODUCTION

The past decade has brought a range of technological advancements that have made it easier to gather large sets of mobility data. These include tracks generated by GPS devices, geo-tagged posts from social media platforms, and call detail records. Such datasets have been used to quantify urban vitality [3], understand commuting patterns [2], and better predict pandemic spreading [4]. Data entries in human mobility data sets usually consist of a timestamp, latitude, and longitude, and sometimes other features like altitude and precision. A vital part of working with such datasets is inferring the mode (i.e. walking, driving, riding a bicycle) of transportation a person was using at a given point in their trajectory.

Mode inference remains a pervasive issue for many mobility researchers, who primarily work with sparse, temporally heterogeneous datasets. These attributes make mobility datasets difficult to preprocess, as the frequency of data collection is largely self-selected by the user (i.e. of a cell phone). In order to address this gap, identifying external (or "meta") factors about individuals' mobility patterns may be useful. For example, the day's weather, the time of day, or the day of the week may all act as predictor variables to infer one's mode choice.

This project integrates large-scale mobility data from individuals from Beijing, China, with weather data, providing an accessible visualization tool and inferences on travel mode choices.

DATA

For this project, we gathered data from the GeoLife dataset from 182 users from Beijing, China [5]. This dataset has information about the GPS observation, ground truth (mode choice), and trips from April 2007 to December 2011. We also gathered raw daily weather data from Visual Crossing’s Weather Data Services for the same period of time [1]. The GeoLife GPS trajectory data has three different entities which are GPS observation, Trips and Ground truth (see **Figure 1**). The GPS observations have information about the Trip ID, User ID, latitude start and end, longitude start and end, and the time stamp. The Ground truth has data on the user ID, Trip ID, start time, end time, and the mode of travel. The trips entity had information on the Trip ID, user ID, average velocity, total travel time, velocity changing rate, heading change rate, and the stop rate. The last entity is the weather entity which has data on date, precipitation, maximum, minimum, and average temperature, dewpoint, wind speed, cloud cover and visibility.

Metadata on the GPS trajectories provided information about how the data was collected, and what metrics were used to record the data. Using this metadata we were able to understand that the timestamp, start and end time were provided in GMT (Greenwich Meridian Time). With this information, we were able to convert the time from GMT to the Beijing time zone. After converting the time the 3 entities data present in the GeoLife were connected and from this a GeoLife.csv file was created. This Geolife.csv file contains 3 million rows of individual GPS observations. This data was processed in a way that each individual GPS observation was associated with one trip of a user, such that each row represents one trip. In doing so, the data was compressed from 3 million rows to approximately 5600 rows (trips).

The weather data was merged with the compressed GeoLife trips data. Here, the tables were joined with a left inner join on the dates column, with the compressed GeoLife trips data on the left and weather data on the right, assigning each trip the associated weather conditions.

From this merged dataset, some entries (rows) were removed based on the following conditions: First, transportation modes with very few appearances were removed. This included airplane, boat, run, and train. Next, some trips with outliers in the trip duration were removed. Exclusion criteria were trip duration of less than 2 minutes as well as greater than 5 hours. At last, since the focus of this analysis is on Beijing city, any trips that were recorded outside of Beijing were removed, too.

Moreover, using the start and end time of each trip, new columns were created. This included the derivation of the start and end hour, and start and end day (by day of the week). The start and end day were also classified as weekday or weekend, and start and end hours were converted as work hours (7am-5pm) or off-hours (otherwise).

For a better understanding of how these tables are connected, see the entity/relationship (E/R) diagram as shown in **Figure 1**.

- The GPS observations, Ground truth, and trips have unique or key attributes which are Trip ID and User ID.
- Along with Trip ID and User ID, GPS observations have Timestamp as a key attribute.
- GPS observations are connected to Ground truth by many to one relationship, and Trips and GPS observation are also connected by many to one relationship. Using the key attributes of these entities the 3 tables are connected.
- The key attribute for the weather entity is considered as the date attribute, and this helped us in connecting the weather table and the Ground truth entity. For these entities the relationship is many to one relationship.

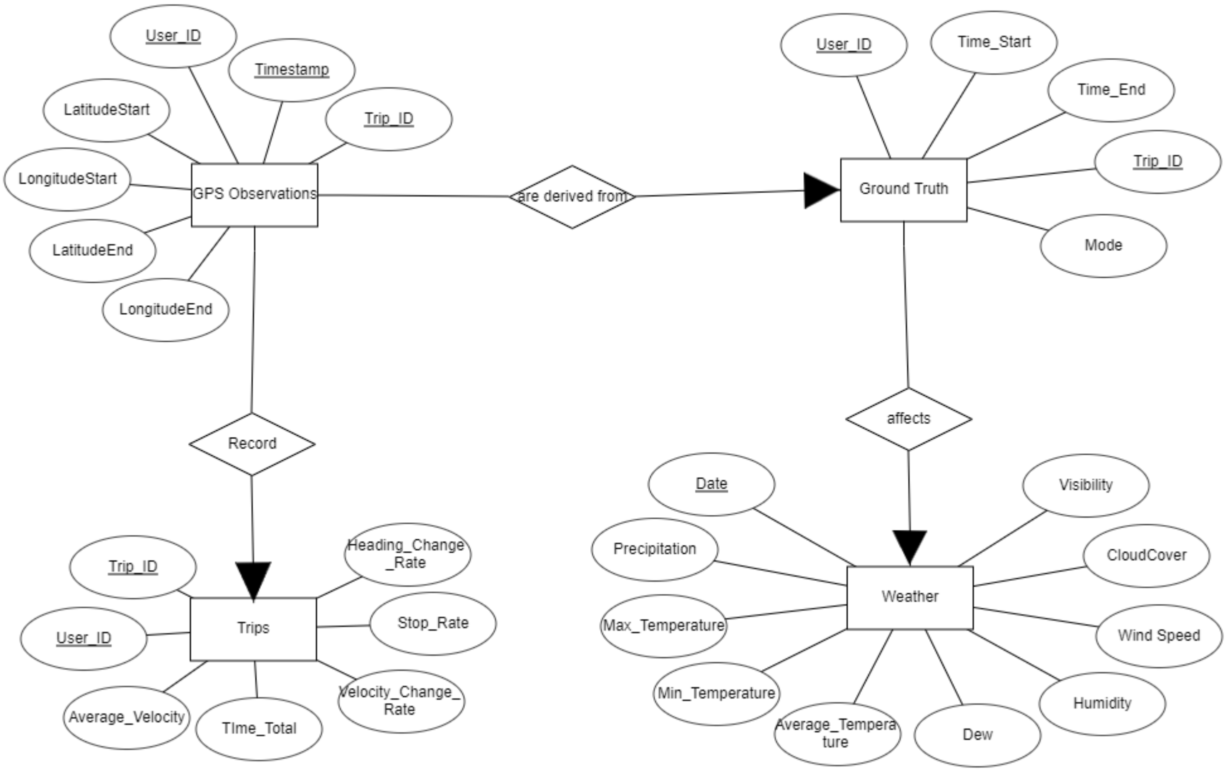


Figure 1: Entity/Relationship diagram for the 4 conceptual entities included in this analysis, along with their attributes and connecting relationships.

The corresponding relational schema for the E/R diagram is:

- GPS Observations (Trip_ID, Timestamp, User_ID, Longitude start, Longitude end, Latitude start, Latitude end, Ground Truth.Trip_ID, Ground Truth.User_ID, Trips.Trip_ID, Trips_User_ID)
- Ground Truth (Trip_ID, User_ID, Time_Start, Time_End, Mode, Weather.Date)
- Weather (Date, Precipitation, Dew, Visibility, Average Temperature, Maximum Temperature, Minimum temperature, CloudCover, Wind Speed, Humidity)
- Trips (Trip_ID, User_ID, Average_Velocity, Time_Total, Velocity_Change_Rate, Stop_Rate, Heading_Change_Rate)

A few additional variables were derived from the individual GPS observations, providing more insights into certain trip characteristics: heading change rate (the rate at which a person changes their heading direction during a trip), stop rate (the rate at which the trip is interrupted, as defined by the velocity dropping below a certain threshold), velocity change rate (the rate at which the velocity changes throughout the entire trip duration). The thresholds applied for the heading change rate, stop rate, and velocity change rate are given in **Table 1**.

Table 1: Threshold values of derived attributes.

Metric	Threshold
Heading change rate	0.33 rad/s/m
Stop rate	0.89 m/s
Velocity change rate	0.26 m/s

Names, variable names in the code, data types, and units of all attributes are listed in **Table 2**.

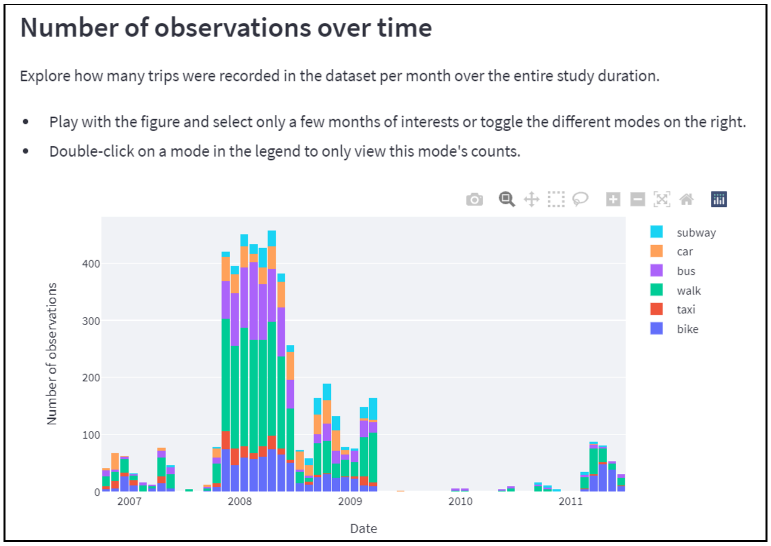
Table 2: List of all attributes, data types, and units.

Name		Variable name in code	Type	(Unit)
Temperature	T	`temp`	`num`	°C
Precipitation	P	`rain`	`num`	mm
Windspeed	V_w	`windspeed`	`num`	m/s
Humidity	H	`humidity`	`num`	%
Cloudcover	CC	`cloudcover`	`num`	oktas
Day (weekend/weekday)	d	`day`	`factor`	
Start Hour (work/off)	h_s	`start_hour`	`factor`	
End Hour (work/off)	h_e	`end_hour`	`factor`	
Average Velocity	v_{avg}	`avg_vel`	`num`	m/s
Total Distance	D_{tot}	`total_dist`	`num`	m
Total Time	t_{tot}	`total_time`	`num`	s
Heading Change Rate	hcr	`hcr`	`num`	rad/m
Velocity Change Rate	vcr	`vcr`	`num`	m/s
Stop Rate	sr	`sr`	`num`	1/s

ANALYSIS AND RESULTS

For the purpose of this project, an accessible data visualization tool[1] using Streamlit was developed on the basis of the preprocessed data, as outlined above. and can be explored by anyone. The tool's visualizations were created using Python and various packages, including matplotlib, plotly, and folium. Visualization types included bar graphs, scatter plots, line charts, histograms, maps (including heatmaps), box plots, violin plots, and data tables, among others. All plots are interactive, with the capability to hover the mouse over each visualization to retrieve more information, zoom into and select certain parts of the plot, and toggle the modes of transportation to be shown. The application contains a sidebar menu, in which the user can navigate to the different pages (*Home*, *Descriptive statistics*, *Map the data*, *Explore weather conditions*, *Explore mode choices*) within the tool. The Home page contains a short demonstration video on the use of the application.

In the section on *Descriptive statistics*, the dataset, its summary statistics, and number of trips over time stratified by mode choice (see **Figure 2**) can be explored. Under *Map the data*, start and endpoints of trips can be visualized on maps of Beijing, and corresponding heatmaps are shown for different mode choices and temperature ranges. This allows the user to gain insight into mode preferences of road users in different weather conditions. One can limit the time period from which to include data from. Using this feature, mobility patterns e.g. during the summer Olympics of 2008 in Beijing (Aug 8-24, 2008) can be revealed, such as an increased mobility around the Beijing National Stadium during this time. Weather records, showing e.g. seasonal differences in various meteorological variables, can be understood on the *Explore weather conditions* page. Lastly, the *Explore mode choices* page dives deeper into the different chosen modes of transportation and looks at how their frequency depends on different external factors, such as time of the day or weather conditions.



Number of observations over time

Explore how many trips were recorded in the dataset per month over the entire study duration.

- Play with the figure and select only a few months of interests or toggle the different modes on the right.
- Double-click on a mode in the legend to only view this mode's counts.

Figure 2: Exemplary screenshot from this project's visualization tool, showing the number of trip observations in the dataset over the study period, grouped by mode choice.

[1] It is available at https://share.streamlit.io/steffen-coe/geolife-mobility-data-app/main/app.py. Additionally, all code written to produce this tool can be found in the authors' corresponding GitHub repository at https://github.com/steffen-coe/GeoLife-Mobility-Data-App.

Further insights include that if a person is commuting or traveling during working hours, then the person is more likely to use a car than any other mode. Along with this, we were also able to understand how weather factors affect modal choice. For instance, fair and warm weather typically leads to more people choosing active modes of travel, such as walking and biking.

In addition to the data visualizations, we also performed several other statistical analyses on the data and developed a model to quantitatively understand mode choices based on different predictor variables as well as two cluster analyses.

First, variables and their influence on mode choice were assessed using correlation coefficients and heatmaps such as the one shown in **Figure 3**. The goal of these considerations was to detect endogeneity as well as multicollinearity between different predictor variables, reducing redundancy in any model approach. The LASSO method was used to identify the strongest predictor variables for mode choice, which were found to be day of the week, time of the day, temperature, and precipitation.

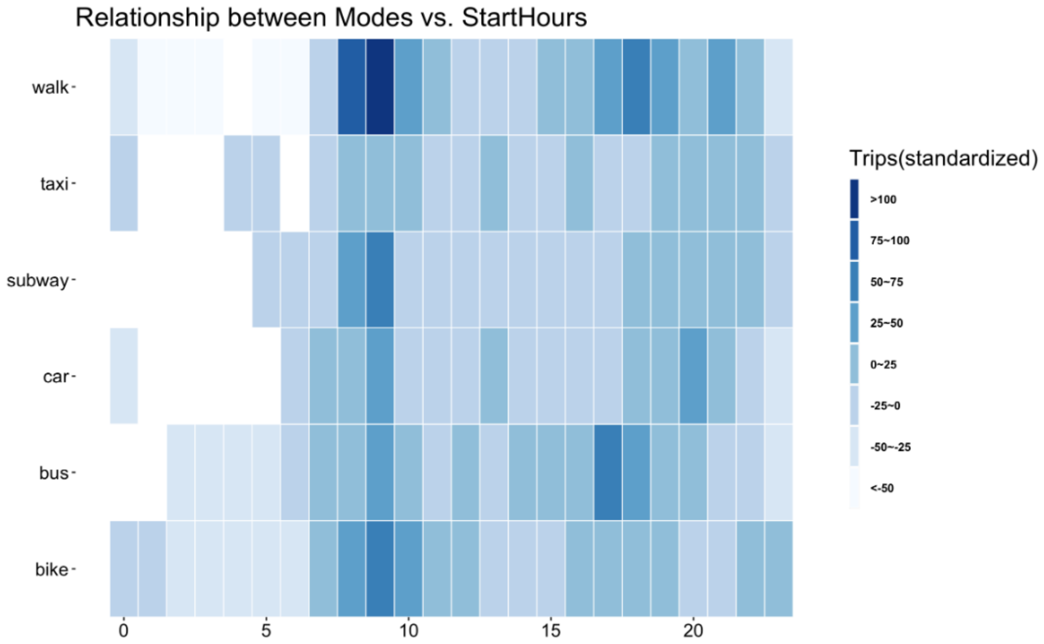
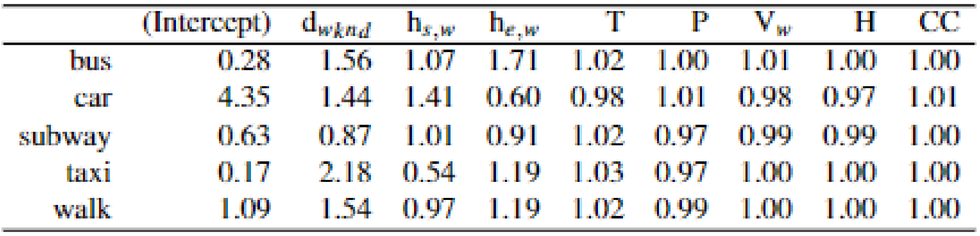


Figure 3: Heatmap of trip start hour by mode choice.

The data was then randomly split into a training set (70% of all trips) and a testing set (30%). A multinomial logit model (MLM) was run with the mode choice as the outcome variable. The reference level was set as biking. This model intentionally does not include trip metrics (e.g. average speed or stop rate) as predictor variables to understand how well mode choice can be predicted without looking at the details of the trips taken. **Table 3** shows the odds ratios for the different independent variables, that were derived from the model coefficients. The found tendencies of the variables to either increase or decrease the likelihood of each of the mode choices generally make sense and are intuitive. For instance, higher temperatures promote walking and reduce the likelihood of driving by car. Similarly, all modes are facilitated on the weekend, except for the subway, indicating that this mode is most commonly used for commuting purposes during the workdays.

Table 3: Odds ratios found for the independent variables and mode choices in the deployed multinomial logit model. d_{wknd}: day (weekend), $h_{s,w}$: start hour (work), $h_{e,w}$: end hour (work), T: temperature, P: precipitation, V_w: windspeed, H: humidity, CC: cloud coverage.

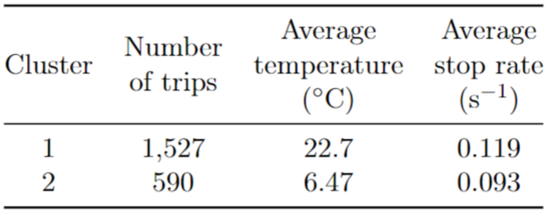


	(Intercept)	d_{wknd}	$h_{s,w}$	$h_{e,w}$	T	P	V_w	H	CC
bus	0.28	1.56	1.07	1.71	1.02	1.00	1.01	1.00	1.00
car	4.35	1.44	1.41	0.60	0.98	1.01	0.98	0.97	1.01
subway	0.63	0.87	1.01	0.91	1.02	0.97	0.99	0.99	1.00
taxi	0.17	2.18	0.54	1.19	1.03	0.97	1.00	1.00	1.00
walk	1.09	1.54	0.97	1.19	1.02	0.99	1.00	1.00	1.00

These model results were tested using the testing set, based on which a prediction accuracy of 55% could be derived. As is intuitive, this accuracy improves significantly to about 80%, when incorporating trip-specific variables, such as average speed, stop rate, or velocity change rate, into the model. A 10-fold cross validation was used to ensure that the train/test data split was not lucky, but that instead results were found to be replicable with different combinations of train/test sets.

Lastly, two cluster analyses were performed. The first cluster analysis was able to reveal two types of walkers: walkers in warm weather (higher average temperature, higher stop rate) and colder weather (lower average temperature, lower stop rate), as shown in **Table 4**. This classification is sensible, since, when walking on a warm day one might be more tempted to take stops, compared to walking on colder days when one might be more interested in arriving at one's destination quickly without too many stops.

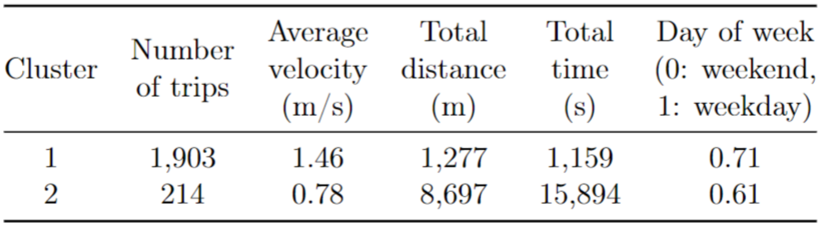
Table 4: Cluster analysis results for walkers at different temperatures.



Cluster	Number of trips	Average temperature (°C)	Average stop rate (s^{-1})
1	1,527	22.7	0.119
2	590	6.47	0.093

The second cluster analysis found walkers with different intentions: (1) walkers with a purpose and (2) strollers, see **Table 5**. Those in the latter category take their time and traverse greater distances (at slower speeds). Strollers are also more likely to be out on the weekends. Conversely, those in the first category travel at about twice the speed of strollers (on average), but only cover about 15% of the distance in 7% of the time.

Table 5: Cluster analysis results for walkers with different intentions.



Cluster	Number of trips	Average velocity (m/s)	Total distance (m)	Total time (s)	Day of week (0: weekend, 1: weekday)
1	1,903	1.46	1,277	1,159	0.71
2	214	0.78	8,697	15,894	0.61

CONCLUSION

This work dealt with geospatial mobility data from the GeoLife dataset for Beijing, China, and connected this data with relevant weather data for the city. In conclusion, the process of creating a data visualization tool using Streamlit and Python provides a plethora of opportunities for users to interactively engage with a transportation dataset on individual travel patterns and mode choices. The application is made available as part of this report. In addition to this, we yielded several significant results on people's mode choices and their dependency on weather conditions, using both a multinomial logit model and cluster analyses. We found that people generally prefer vehicles over active modes of transportation when the temperatures were low or during rainy weather. During warmer temperatures, the frequency of those choosing to bike or walk increased.

Overall, the visualizations provided by the Streamlit tool and the results of the models in this analysis enable relevant insights into mobility patterns in Beijing over several years. Future research could deploy other methods such as origin-destination matrix analysis accompanied by Google Places API, to identify trip purposes (e.g. commute and leisure trips) and hot spots of interactions.

ACKNOWLEDGEMENT

The authors would like to thank Shuyi Yin (syin@uw.edu) who taught the course CET 522 on Transportation Data Management and Visualization at the University of Washington in Winter 2022. This work is the result of the course project conducted in this class.

REFERENCES

[1] "Weather Data Services | Visual Crossing." https://www.visualcrossing.com/weather/weather-data-services (accessed Feb. 19, 2022).

[2] V. Frias-Martinez, C. Soguero, and E. Frias-Martinez, "Estimation of urban commuting patterns using cellphone network data," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, New York, NY, USA, Aug. 2012, pp. 9–16. doi: 10.1145/2346496.2346499.

[3] P. Sulis, E. Manley, C. Zhong, and M. Batty, "Using mobility data as proxy for measuring urban vitality," *Journal of Spatial Information Science*, vol. 2018, no. 16, pp. 137–162, 2018, doi: 10.5311/JOSIS.2018.16.384.

[4] Q. Hao, L. Chen, F. Xu, and Y. Li, "Understanding the Urban Pandemic Spreading of COVID-19 with Real World Mobility Data," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA, Aug. 2020, pp. 3485–3492. doi: 10.1145/3394486.3412860.

[5] "GeoLife GPS Trajectories," *Microsoft Download Center*. https://www.microsoft.com/en-us/download/details.aspx?id=52367&from=https%3A%2F%2Fresearch.microsoft.com%2Fen-us%2Fdownloads%2Fb16d359d-d164-469e-9fd4-daa38f2b2e13%2F (accessed Feb. 18, 2022).