THE STAGA-DATASET: STOP AND TRIP ANNOTATED GPS AND ACCELEROMETER DATA OF EVERYDAY LIFE

R. P. Spang[1]*, K. Pieper[1], B. Oesterle[1], M. Brauer[1], C. Haeger[2], S. Mümken[2], P. Gellert[2], J.-N. Voigt-Antons[3,4]

[1] Quality and Usability Lab, Berlin Institute of Technology, Berlin, Germany -
(spang, kerstin.pieper)@tu-berlin.de, (max.brauer, benjamin.oesterle)@campus.tu-berlin.de
[2] Institute of Medical Sociology and Rehabilitation Science, Charité - Universitätsmedizin Berlin, Berlin, Germany -
(christine.haeger, sandra.muemken, paul.gellert)@charite.de
[3] University of Applied Sciences Hamm-Lippstadt, Germany - jan-niklas.voigt-antons@hshl.de
[4] German Research Center for Artificial Intelligence (DFKI), Berlin, Germany

Commission IV, WG IV/4

KEY WORDS: GNSS, Dataset, Annotation, Accelerometer, Benchmark, Stop Trip Classification.

ABSTRACT:

Identifying stops and trips from raw GPS traces is a fundamental preprocessing step for most mobility research applications. Thus, ensuring the excellent accuracy of such systems is of high interest to researchers designing such analysis pipelines. While there are plenty of GPS datasets available, these usually do not provide annotations and thus cannot be used for benchmarking stop/trip classifiers easily. This manuscript introduces a GPS & accelerometer dataset, including accurate stop/trip annotations. It contains 122,808 GPS samples as one continuous trajectory, spanning over 126 days. The recorded time frame includes working days, vacations, travelling, everyday life and all regular modes of transportation. During recording, a detailed mobility diary was conducted to capture each dwelling period's exact beginning and end. The position and diary data combined contain 78,900 labelled stops and 43,908 labelled trips. This serves as ground truth for stop/trip classification algorithms to test existing tools or develop new analysis methods. The introduced dataset is freely available under a CC-By Attribution 4.0 International license, the annotation tool under the BSD 3-Clause license.

1. INTRODUCTION

When we started to develop the *Stop & Go Classifier*, an algorithm to transform raw GPS samples into periods of dwelling and transit (Figure 1), we wanted to quantify its performance and the progress we made early on. Initially, parts of our team started recording GPS traces and retrospectively noting where they had been during the day. It quickly became apparent that such a memory-based approach is insufficient, especially when our goal for the *Stop & Go Classifier* was to identify stops in the sub-five minutes precision space accurately. We were researching available datasets, but all we could find were large bodies of traces without any annotations. Projects such as the GoeLife dataset (Zheng et al., 2010) or the T-Drive trajectory data sample (Yuan et al., 2010, Yuan et al., 2011) do contain plenty of GPS samples, but they do not come with annotations to serve as ground truth for stop/trip classification. Another example is the GPS Trajectories Data Set (Cruz et al., 2015) that includes separate trajectories but without the stops connecting the pathways. While such a dataset can benchmark a stop/trip classifier's accuracy on trips alone, it will not help benchmark the ability to detect stops.

When working with plain samples without annotations, manual labeling by inspecting shape and timestamps is an option to create a benchmark dataset. However, while this is certainly a start, we were looking for a precisely annotated dataset that we could employ to act as ground truth for comparing our classifier under development. We decided to create our own because the existing literature did not provide a sufficiently large dataset for this task (see Figure 2).

Through this manuscript, we contribute a comprehensive dataset consisting of GPS samples and a movement diary that provides accurate begin and end timestamps for all stops over the recording period of 126 days. This package is the STAGA dataset. It contains GPS coordinates, recording timestamps, altitude, GPS accuracy, a motion score quantifying the physical motion of the recording device, and a class label (”stop” or ”trip”). The acceleration data is provided as a separate file but covers the same time frame and contains a triple (x, y, z) of accelerometer sensor readings for each given timestamp, sampled at 1 Hz. The STAGA dataset is provided publicly and free to use. We further provide the iOS app used to create the diary data for simple stop/trip annotation while on the go. See Section 5 for further details.

1.1 Use-cases

This dataset was initially recorded to support the development and validation of the *Stop & Go Classifier* (Spang et al., 2022). The corresponding manuscript used the presented dataset to benchmark the classification performance and compare it against other libraries for stop/trip classification. While we originally wanted to create data for evaluation and comparison purposes only, we identified a gap in the literature and the corresponding available datasets: most come without any annotations or with too little information to use available data to benchmark classification capabilities. This lack motivated us to publish the test data we used during the development of our classifier. While the STAGA dataset can certainly be used for comparisons regarding stop/trip classification, the dataset as a whole opens up further possibilities for all sorts of mobility research.

* Corresponding author

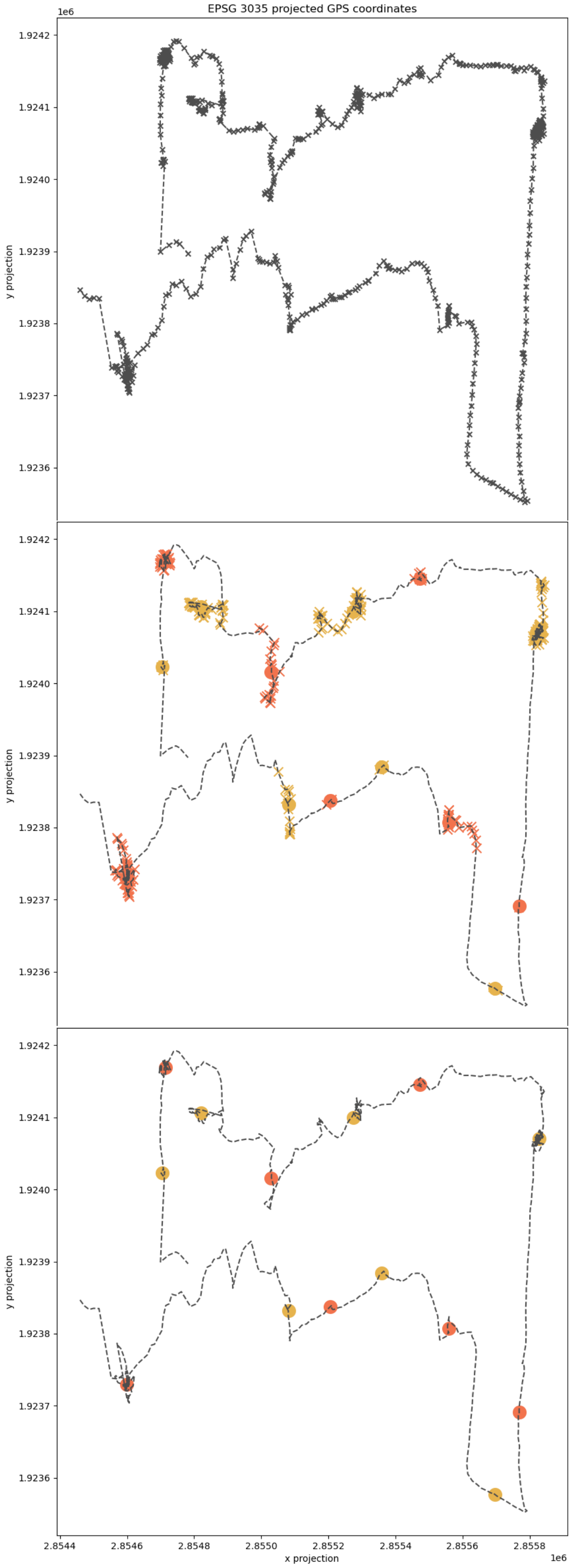


Figure 1. Example of the stop classification process. All three plots show the same trajectory, a city walk of about four hours. Top: raw samples; mid-section: all stop-classified samples; bottom: isolated stop centers.

The STAGA dataset can generally be used to study human mobility patterns. The dataset contains trajectories of everyday life and reflects the mobility patterns of middle-aged, working people with a rather active social environment. The provided accelerometer readings open up further analysis methods and techniques. With the mobility diary, stop, or trip intervals can be easily singled out to study the relation between position and motion. Lastly, the dataset can be helpful for general mobility research benefitting from a precise movement diary.

2. METHOD

2.1 Data collection

To record the GPS samples, we created a custom app based on a fork of the μlogger project[1]. This project is a good foundation as it logs GPS records on the device first and syncs them later when connected to a network. For this project, this is primarily interesting because it prolongs battery life compared to an always-online approach. We extended this codebase to capture accelerometer data and autostart logging as soon as the smartphone boots up. We were interested in the accelerometer data to research how it can facilitate the stop detection in our *Stop & Go Classifier*. The autostart feature should simplify the recording process such that the user will not have to check if the app is running and recording; this should further minimize data gaps.

The device we used for the recordings was a ZTE Blade A5 (2019). The smartphone runs Android 9 and is equipped with a 2.600 mAh battery. It was configured to record GPS samples at a minimum accuracy of 25m, so if the device could not obtain a position reading within this radius, the data point was omitted. We sampled data with a frequency of 0.1 Hz and used both network and GPS as sources for determining the position (the smartphone supports A-GPS and GLONASS). This sampling frequency was chosen primarily to balance temporal resolution and battery life. This is important to minimize user error regarding forgetting to charge the phone regularly while recording (and thus minimize data gaps as well). With a temporal resolution of 0.1 Hz, we should be easily able to capture all relevant aspects of a person's general mobility. At the same time, we obtained a battery life of approximately 1.5 days, constantly recording. However, the battery life varied in different environments.

Volunteers aged between 25 and 35 years collected the dataset. They all worked full-time at different workspaces (office building, home office and various remote offices). While the dataset primarily contains everyday life, it also holds short vacation, travel, and hiking periods. Most trips were carried out by bike. However, the dataset also contains long periods of walking, car traffic, and train rides. Over the recording period between the last half of 2021 and the first half of 2022, the data donors stayed in two European countries (primarily urban environments). After each participant finished their data acquisition, we provided a simple interface to inspect all recorded stops on a map. We asked each participant to examine each stop to ensure the most accurate annotations. In this process, all samples belonging to a stop were plotted against a map. We added two minutes of the trajectories leading to and departing from each stop in a different color for orientation. Then, the participants were given a simple tool (a web app we developed for this purpose) to change the timestamps of each stop. This process was

[1] github.com/bfabiszewski/ulogger-android

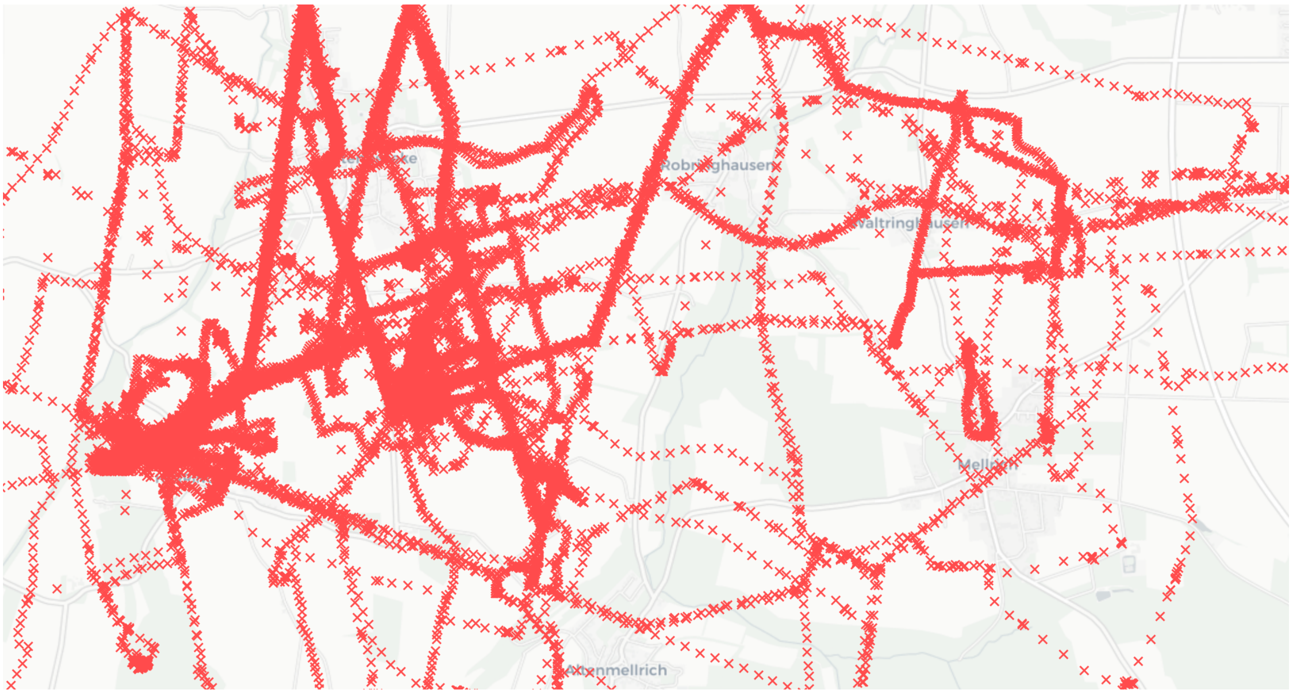


Figure 2. Excerpt of the GPS samples. This plot shows approximately only one-third of the entire dataset. Trip intervals can be identified by the continuous lines they form. Trajectories do not match streets on the background map.

efficient and helpful in correcting minor deviations during the recording of the diary. For example, one participant described a situation in which they planned to leave a building but then ran into a neighbor and chatted for several minutes in front of the entrance. However, the diary app already logged leaving the place. To compensate for situations like this, the post-hoc exploration of all the recorded stops was a helpful second step in ensuring the quality of the annotations. Figure 2 shows an excerpt of the GPS records.

2.2 Diary

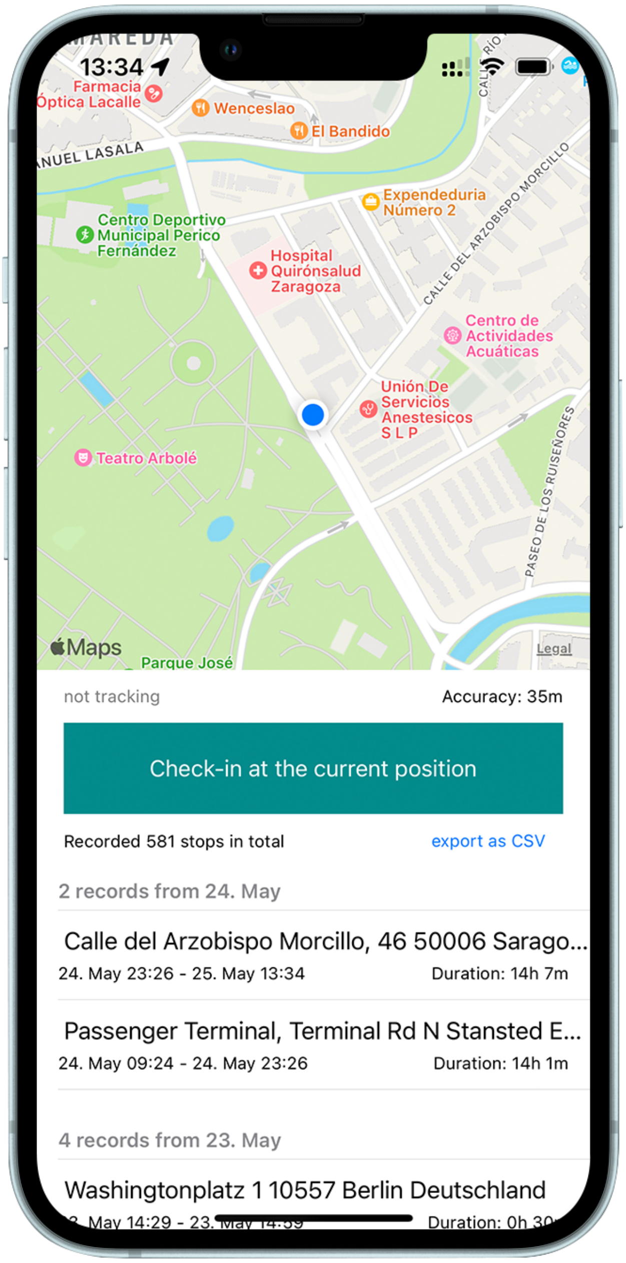
We first tried a traditional diary approach to create the dataset: four team researchers created retrospective logs at the end of each day. As this process was too error-prone, we took notes on the go, writing down addresses and times whenever they stopped. While this provided some first samples for testing, it was tedious and inaccurate since taking notes is impractical in everyday life. Furthermore, it required looking up the coordinates belonging to each noted address, which works for clearly defined urban spaces. However, it can become problematic in rural areas or parks, as addresses are no precise reference in such environments. Because of that, we developed a simple iOS app that helps annotate movements. The app contains a map to validate the current position, one button to start or end a stop, and a list of previously recorded stops. It captures the GPS position whenever a new stop is started and stores the current time as the start timestamp. The stop is completed when the button is pressed again, and the current time is stored as the stop's end timestamp. Trips are derived from the intervals between two stops (Spaccapietra et al., 2008). Records can be edited, too, for example, if a user forgot to check in but remembered shortly after. Finally, the app allows exporting the captured annotations as a CSV file. The diary was recorded using an Apple iPhone XR, see Figure 3. We did not run the diary app on the GPS tracking smartphone because we did not want to influence the accelerometer readings by using the device actively. Instead, it should be carried, ideally unconsciously, in a jacket, backpack, or the like with little to no interaction. Hence, the separate companion app to annotate the movement.

Our setup was complete with both these systems in place, the Android GPS tracker device and the iOS companion app to annotate trips. This way, we could create a GPS dataset containing precise stop/trip annotations and a reference position of the actual stop location.

2.3 Data anonymization

The Declaration of Helsinki provides important ethical principles for (medical) research involving human subjects. Amongst them, privacy and confidentiality of personal information are challenging aspects for any dataset containing individual positions: ”Every precaution must be taken to protect the privacy of research subjects and the confidentiality of their personal information.” (Association et al., 2013). Since position data, especially everyday life, contain plenty of personal information (habits, preferences, and potentially social contacts), it is of special interest to us to keep this personal information as private as possible. To do so, we omit further details describing the donors of the data. They gave full and informed consent to this publication. Moreover, we took several steps to obscure and anonymize the GPS traces contained in this dataset.

Such anonymization is a challenging task as we cannot, for example, add noise to each position (see for example (Agrawal and Srikant, 2000) who coined the term ’value distortion’ for this approach). We must keep the stop-point clusters intact, so stop/trip detection can work equally well on the resulting dataset. Hence, we developed the following approach with our data donors to keep their privacy high. We employed the following steps to make it unlikely that the multiple actual positions will be revealed.


13:34
Farmacia Óptica Lacalle
Wenceslao
El Bandido
MANUEL LASALA
Expendeduria Número 2
CALLE DEL ARZOBISPO MORCILLO
Centro Deportivo Municipal Perico Fernández
Hospital Quirónsalud Zaragoza
Centro de Actividades Acuáticas
Unión De Servicios Anestesicos S L P
Teatro Arbolé
PASEO DE LOS RUISEÑORES
Maps
Parque José
Legal
Check-in at the current position
export as CSV

not tracking Accuracy: 35m

Recorded 581 stops in total

2 records from 24. May

Calle del Arzobispo Morcillo, 46 50006 Sarago...
24. May 23:26 - 25. May 13:34 Duration: 14h 7m

Passenger Terminal, Terminal Rd N Stansted E...
24. May 09:24 - 24. May 23:26 Duration: 14h 1m

4 records from 23. May

Washingtonplatz 1 10557 Berlin Deutschland
23. May 14:29 - 23. May 14:59 Duration: 0h 30

Figure 3. Screenshot of the annotation companion app, which recorded the diary entries. On arrival at a location, the user checked in to their current position to track the start timestamp. Right before they left a dwelling spot, they logged that with a second tap to track the end of the interval.

- Timeshift: all timestamps (GPS, accelerometer and diary) have been shifted to start on January first in the year 2000 at midnight.

- Mirror & Rotation: the position data have been mirrored (flipped) and slightly rotated.

- Stop consistent spatial shifting: The previous approaches do not protect against background knowledge attacks: e.g., knowing details about a single stop in the dataset enables an attacker to roll back these measures easily. Hence, we changed the absolute position of each stop relative to each other - but kept the stops themselves intact. In this process, we clustered nearby stops together and computed an optimal position to relocate them so they would not overlap afterward. Through clustering stops, we could apply larger offsets to clusters further apart from each other and less significant offsets to close clusters. This algorithm recognizes revisiting locations and offsets all occurrences of the same place to the same new position. This allows analyzing, e.g., location importance throughout the dataset. This procedure breaks a classical aspect of the geo-indistinguishability definition that expects two geographically close locations to have close results in the obfuscated space (Andrés et al., 2013). After applying this mechanism, two formerly close places do not necessarily have spatial distance as two other equally close points. The algorithm also considers path lengths so that these are altered as minimally as possible.

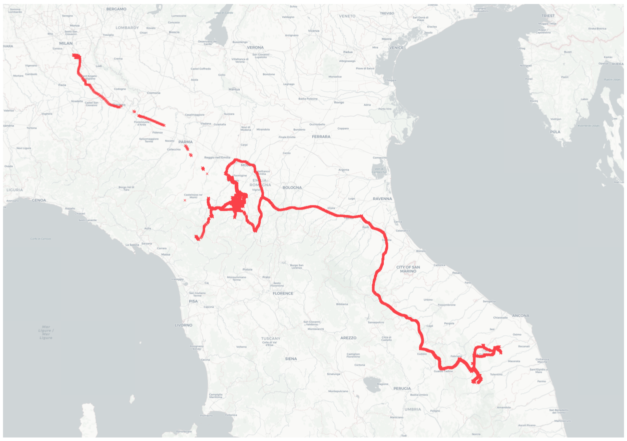


Figure 4. A plot of one of the two main areas of the STAGA GPS samples. Most data points of the dataset were projected onto Italy.

All these points help to maintain privacy against actual locations. However, the mobility diary poses several privacy risks (e.g., how often the user leaves the house or estimations about the number of social contacts, to name a few). With our mission to provide a reference dataset for stop/trip classification, the data donors and we accepted these privacy issues willingly. The resulting dataset has most samples projected onto three European countries: Portugal, Spain, and Italy (see Figure 4 and Figure 5).

3. DATASET DESCRIPTION

The dataset contains 122,808 GPS and 7,805,994 accelerometer records. The recording time spans over 126.65 days. While the recording device was set up to record GPS data at a maximum frequency of 0.1 Hz and accelerometer readings at 1 Hz, the actual sample frequency is 0.011 Hz for the GPS data and 0.714 Hz for the acceleration data. Because about two-thirds of the time was spent dwelling, mostly indoors, the average sample rate is especially reduced for stop intervals. Indoors the signal accuracy was often well above the configured threshold (25m); hence the dataset often shows gaps while dwelling.

The diary contains 692 stops and 691 trips. To provide summarization statistics of their attributes, we report details regarding the average stop and trip durations (Table 1 and Table 3), the number of samples per stop or trip (Table 2 and Table 4), and the average distance per trip (Table 5).

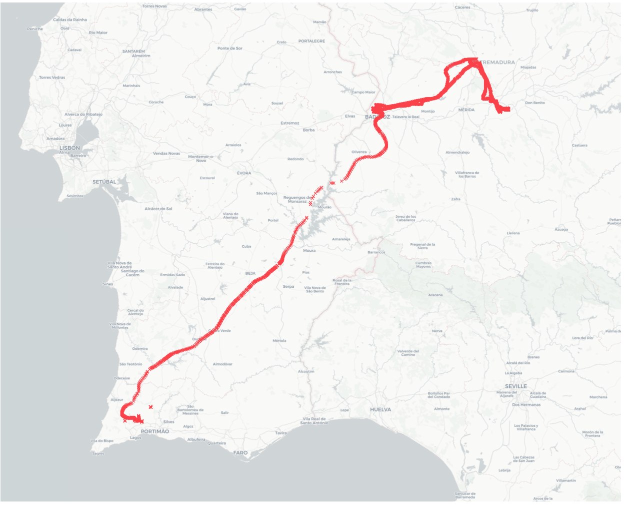


Figure 5. A plot of the other main area of the STAGA GPS samples. The dataset's second-largest chunk of interrelated samples was projected onto Portugal and Spain.

Stop duration	Value
Mean	240.7 min
Median	27.4 min
Standard deviation	432.3 min
2.5 % percentile	1.6 min
97.5 % percentile	1421.7 min

Table 1. Statistical details of all stop durations.

Samples per stop (s/s)	Value
Mean	113.9 s/s
Median	22.0 s/s
Standard deviation	321.92 s/s
2.5 % percentile	2.0 s/s
97.5 % percentile	782.3 s/s

Table 2. Statistical details of the samples per stop.

Trip duration	Value
Mean	22.8 min
Median	8.0 min
Standard deviation	121.8 min
2.5 % percentile	1.2 min
97.5 % percentile	92.8 min

Table 3. Statistical details of all trip durations.

Samples per trip (s/t)	Value
Mean	63.7 s/t
Median	39.0 s/t
Standard deviation	78.18 s/t
2.5 % percentile	4.0 s/t
97.5 % percentile	256.0 s/t

Table 4. Statistical details of the samples per trip.

Trip distance	Value
Mean	24.7 km
Median	1.3 km
Standard deviation	127.7 km
2.5 % percentile	61.7 m
97.5 % percentile	73.7 km

Table 5. Statistical details of all trip distances.

4. DISCUSSION

The presented dataset closes a gap in the literature regarding annotated stops and trips in long-time GNSS tracking. For example, shown in Figure 1, the dataset contains many stops that can be identified and compared against the annotations. Other mobility-related research questions can be investigated with the dataset as well but were not the scope of this presentation of the dataset.

The statistical analysis of the stops and trips shows a large variety of different aspects that are relevant for stop/trip classification. For example, the dataset contains a few stops without samples (likely because the achievable accuracy was above our predefined threshold, 25m). In such cases, a purely geometrical analysis approach would be sufficient to detect the stop. Other stops contain many samples but include strong outliers. Through the long timespan, the STAGA dataset includes several edge-cases that will occur in GNSS field recordings. As such, it is a good basis for benchmarking classifier performance. However, the analysis also reveals a large deviation between mean and median values, indicating strong outliers in each metric.

The annotation process using the companion app for stop labeling (Figure 3) was well received by the participants. While one participant wished to have the companion app available on their smartwatch, the process proved to be much more efficient than a retrospective labeling at, e.g., the end of the day. We believe we have recorded an accurate diary of the review process after the data acquisition phase.

The anonymization of the position data limits the possibility of studying the behavioural patterns of the data donors recording the dataset. While this is intentional, we acknowledge that a wider audience from more research areas could be addressed by providing the raw data. We tried to create a specific dataset that is as ecologically valid as possible for its primary task: providing ground truth for stop and trip classification. We compromised the width of possible applications for a long-time position tracking dataset to reach this goal.

As a result of the anonymization process, paths may have been modified. This alternation leads to changes in velocity between individual samples. This is a piece of important information, as, for example, higher than average walking speeds might be observed.

Future work on this project should investigate if resampling of the pathways were an option to restore the original velocities while maintaining anonymization regarding the actual positions. Additionally, the anonymization of large GPS datasets whilst maintaining stop-consistency should be discussed appropriately. An exciting body of literature is concerned with anonymizing GPS traces that lack best practices for maintaining, for example, stop consistency. We plan to elaborate our findings and algorithms in this field in the future.

4.1 Conclusion

The STAGA-dataset provides a large body of annotated mobility data: It was recorded over more than a third of a year, and stop intervals have been manually annotated for exact accuracy. The purpose of the dataset is to enable researchers to validate the performance of their stop/trip classification algorithms by providing ground truth labelled data. The CC-By Attribution 4.0 International license should allow researchers from all fields to use the dataset in various projects, enabling them to make data-driven decisions in developing mobility research tools.

5. DATA & CODE AVAILABILITY

The described dataset, containing GPS & acceleration records and stop/trip annotations, are publicly available at the Open Science Framework under a CC-By Attribution 4.0 International license[2].

The annotation companion app (as seen in Figure 3) we used to annotate the dataset is free software under a BSD 3-Clause license[3].

ACKNOWLEDGEMENTS

We thank Arik Grahl for his work on the adaption of the μlogger app and server. We thank Purna Dutta for her help in the quality assurance process of creating the mobility diary.

REFERENCES

Agrawal, R., Srikant, R., 2000. Privacy-preserving data mining. *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 439–450.

Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., Palamidessi, C., 2013. Geo-indistinguishability: Differential privacy for location-based systems. *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, 901–914.

Association, W. M. et al., 2013. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *Jama*, 310(20), 2191–2194.

Cruz, M. O., Macedo, H., Guimaraes, A., 2015. Grouping similar trajectories for carpooling purposes. *2015 Brazilian conference on intelligent systems (BRACIS)*, IEEE, 234–239.

Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., Vangenot, C., 2008. A conceptual view on trajectories. *Data & knowledge engineering*, 65(1), 126–146.

Spang, R. P., Pieper, K., Oesterle, B., Brauer, M., Haeger, C., Mümken, S., Gellert, P., Voigt-Antons, J.-N., 2022. Making Sense of the Noise: Integrating Multiple Analyses for Stop and Trip Classification. *Proceedings of FOSS4G, Florence, Italy*.

Yuan, J., Zheng, Y., Xie, X., Sun, G., 2011. Driving with knowledge from the physical world. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 316–324.

Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., Huang, Y., 2010. T-drive: driving directions based on taxi trajectories. *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*, 99–108.

Zheng, Y., Xie, X., Ma, W.-Y. et al., 2010. GeoLife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2), 32–39.

[2] osf.io/34sft

[3] github.com/rgreinacher/gps-diary