Mobility Data Representations for Spatiotemporal Tasks

Cristiano Landi

University of Pisa,
Largo Bruno Pontecorvo 3, Pisa, Italy
cristiano.landi@phd.unipi.it

Introduction

Mobility data (MD) are everywhere. Smartphones and connected cars, as well as tracking devices with GPS capabilities, produce enormous amounts of spatiotemporal data. Governments, businesses, and researchers use this data for identifying transportation modes (Lee et al. 2008), ascertaining user identities (Naini et al. 2016), and detecting suspicious movements, etc.

The most similar field in the literature is time series (TS), which involves streams of observations over a finite period. TS research is more extensively explored, particularly in classification tasks, where a wide variety of methods exist (Middlehurst, Schäfer, and Bagnall 2024). These methods can be summarized in 3 sets: (i) feature extraction methods based on task-specific ad-hoc mathematical formulas; (ii) transformation-based algorithms that aim to *represent* the data in a simplified yet effective data tabular-like format, which can then be fed into general-purpose machine learning models; (iii) deep learning methods, which, despite being the most recent, have already demonstrated results comparable to the other two approaches.

Comparing TS with mobility literature, we can observe that the former tends to focus more on report-style publications, emphasizing the results of the analysis rather than the methodologies employed. Furthermore, an analysis of the methodologies reveals that many articles still prioritize defining ad-hoc pipelines to address specific tasks, which hinders the reusability of the code (da Silva, Petry, and Bogorny 2019). In contrast, other methods rely heavily on opaque deep learning techniques, often requiring substantial training data. This poses a significant challenge in the sector, as there is a major issue regarding the availability of public data due to privacy constraints. Continuing our comparison with the TS literature, we observe that transformation-based algorithms are underexplored. This is problematic because transformation-based algorithms in TS serve several crucial purposes: they are used for explainability(Landi et al. 2023b; Ferrero et al. 2018), for enhancing performance (Middlehurst, Schäfer, and Bagnall 2024), and for enabling new types of data manipulation, ensuring specific properties (Landi and Guidotti 2024).

Another key challenge in MD analysis is achieving geographic transferability of models. A model trained on data from one region may perform poorly when applied to another due to differences in the road network patterns or population behavior. For example, a model optimized to classify transportation modes in Philadelphia may not work well in Rome, where road layouts and commuting habits differ significantly.

During my PhD, I'm focusing on developing fast, reusable, and effective trajectory *representations* suitable for multiple machine learning tasks, with an emphasis on geographic transferability and interpretability. The first method I worked on is the Trajectory Interval Forest (TIF) (Landi et al. 2023a) (published at ACM SIGSPATIAL23). It emerged from a survey where I collected and reimplemented the field's most commonly used mathematical formulas. Inspired by Middlehurst et al.'s work on time series classification (Middlehurst, Large, and Bagnall 2020), TIF transform the trajectory data into a tabular format by extracting features based on the collected mathematical formulas from random positions of each trajectory. Despite its simplicity, the method showed extremely good performance across multiple public benchmark datasets and served as a baseline for my future research.

TIF's main limitation is that it relies on a finite set of human-defined features. While it performs well across benchmark datasets, specific tasks may exist where no suitable predefined feature exists. To address this, I began developing Geolet (Landi et al. 2023b) (Published at IDA23), a shapelet-based transformation that has quickly become the core of my thesis. A shapelet is a sub-sequence of a time series extracted from the training data with strong discriminative power for the analysis tasks. Geolet builds on this concept with three key advancements: (i) the subsequence selection phase is tailored for spatiotemporal signals; (ii) Geolet introduces multiple distance measures to account for inconsistent sampling rates; (iii) it proposes two strategies to "normalize" subsequences, making them less dependent on specific geographic locations and improving geographic transferability. After selecting the discriminative subsequences, Geolet transforms the input trajectory dataset into a matrix representation, where each element reflects the maximal similarity between a trajectory and the discriminatory subsequences. The result of this transformation can then be used

with every tabular-based machine learning model.

Geolet has demonstrated performance comparable to SotA methods on classification tasks. Although it lags in runtime due to the non-linear complexity of the similarity function, it offers improved model interpretability, making it a valuable trade-off in applications where understanding the decision process is crucial.

To better understand why Geolet works, I'm currently investigating two hypotheses: (a) Is Geolet learning that a specific shape corresponds to a specific road in the road network, thus hindering geographic transferability? (b) Is Geolet learning some latent trajectory feature tied to the road's shape? In order to validate these hypotheses, in (Landi and Guidotti 2024), we addressed a preliminary question: Is it possible to map-match a normalized shapelet, i.e., a sub-trajectory, without knowledge of its original coordinates? The study's findings indicate that, given a shapelet, it is possible to reduce the space of all potential candidate roads by up to 90%, with a probability greater than 85% that the remaining roads include the correct solution[1]. The paper has been published in the BMDA workshop co-located with the VLDB23 conference. The paper has been invited to submit an extended journal version to Geoinformatica. The journal version of the paper investigates the (a) and (b) hypotheses by simulating trips using Eclipse SUMO for different vehicle types, specifically cars and bicycles, across three cities: Rome, New York, and Athens. These cities represent distinct road network topologies: curvilinear, grid-based, and mixed. Early results show that Geolet can adapt to scenarios where the target variable is influenced either by the city's structure (i.e., identifying the origin city of a trip[2]) or by movement dynamics (i.e., distinguishing between cars and bicycles). We further tested these models on different cities from where they were trained, demonstrating that trajectory normalization enables geographic transferability while maintaining high performance.

Other ongoing work is to extend Geolet by proposing rotation-invariant similarity measures and incorporating semantically enriched trajectories (Ferrero et al. 2018). The idea stems from the intuition that objects moving along roads with similar characteristics and shapes should maintain consistent movement dynamics, regardless of the direction of the movement. Inspired by (Musleh and Mokbel 2024), I'm also exploring trajectory discretization techniques to enhance the efficiency and throughput of the proposed methods. In collaboration with Prof. Pelekis from the Data Science Lab at the University of Piraeus, Greece, I am moving away from the classification task by working on a framework for benchmarking event-based sub-trajectory clustering through trajectory representation learning. This problem involves clustering segments of trajectories that exhibit common events, such as sudden braking or rapid acceleration, ignoring the spatial closeness the of trajectories.

[1]The experiments were conducted using real-world GPS traces from OSM and the road network of the entire Tuscany region in Italy.

[2]normalizing the trajectories such as everyone start from coordinates $< 0, 0 >$

Lastly, I contributed as a second author with my supervisors on an explainable clustering method, ParTree (published in DS 2023). I am also involved in ongoing projects related to fair clustering, explainable model mining (Guidotti et al. 2024), and benchmarking methods for irregular time series classification.

To sum up, during the first part of my PhD, I proposed TIF based on a survey of MD analysis, which served as one of the baselines for my work. I then introduced Geolet, virtually the first shapelet-based method for raw MD, and began investigating its capabilities and limitations. In the second part of my PhD, I plan to integrate the MD transformations I developed with the methods I am collaborating on, creating interpretable pipelines for MD analytics. I will also invest more time in deep learning, for example by using generative models to produce the discriminative sub-trajectories used by Geolet in the transformation.

References

da Silva, C. L.; Petry, L. M.; and Bogorny, V. 2019. A Survey and Comparison of Trajectory Classification Methods. In *BRACIS*, 788–793. IEEE.

Ferrero, C. A.; Alvares, L. O.; Zalewski, W.; and Bogorny, V. 2018. MOVELETS: exploring relevant subtrajectories for robust trajectory classification. In *SAC*, 849–856. ACM.

Guidotti, R.; Monreale, A.; Setzu, M.; and Volpi, G. 2024. Generative Model for Decision Trees. In *AAAI*, 21116–21124. AAAI Press.

Landi, C.; and Guidotti, R. 2024. A Shape-Based Map Matching Approach for Geographic Transferability of Discriminative Subtrajectories. In *EDBT/ICDT Workshops*, volume 3651 of *CEUR Workshop Proceedings*.

Landi, C.; Guidotti, R.; Nanni, M.; and Monreale, A. 2023a. The Trajectory Interval Forest Classifier for Trajectory Classification. In *SIGSPATIAL/GIS*. ACM.

Landi, C.; Spinnato, F.; Guidotti, R.; Monreale, A.; and Nanni, M. 2023b. Geolet: An Interpretable Model for Trajectory Classification. In *IDA*, volume 13876 of *Lecture Notes in Computer Science*, 236–248. Springer.

Lee, J.; Han, J.; Li, X.; and Gonzalez, H. 2008. *TraClass*: trajectory classification using hierarchical region-based and trajectory-based clustering. *Proc. VLDB Endow.*

Middlehurst, M.; Large, J.; and Bagnall, A. J. 2020. The Canonical Interval Forest (CIF) Classifier for Time Series Classification. In *IEEE BigData*, 188–195. IEEE.

Middlehurst, M.; Schäfer, P.; and Bagnall, A. J. 2024. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. *Data Min. Knowl. Discov.*, 38(4): 1958–2031.

Musleh, M.; and Mokbel, M. F. 2024. Let's Speak Trajectories: A Vision to Use NLP Models for Trajectory Analysis Tasks. *ACM Trans. Spatial Algorithms Syst.*

Naini, F. M.; Unnikrishnan, J.; Thiran, P.; and Vetterli, M. 2016. Where You Are Is Who You Are: User Identification by Matching Statistics. *IEEE Transactions on Information Forensics and Security*, 11(2): 358–372.