



Computational Methods for Sustainable Mobility

-

Interpretation and Prediction of Tracking Data using Graphs and Machine Learning

DISS. ETH Nr. 29680

Computational Methods for Sustainable Mobility

-

Interpretation and Prediction of Tracking Data using Graphs and Machine Learning

A dissertation submitted to attain the degree of

DOCTOR OF SCIENCE of ETH ZURICH

(Dr. sc. ETH Zürich)

presented by

Henry Martin

Master of Science, Technische Universität München
Diplôme d'ingénieur, l'École Centrale de Lille

10.07.1990

Passau, Germany

accepted on the recommendation of

Prof. Dr. Martin Raubal
Prof. Dr. Fernando Perez-Cruz
Prof. Dr. Laura Alessandretti

2023

Henry Martin

Computational Methods for Sustainable Mobility -

Interpretation and Prediction of Tracking Data using Graphs and Machine Learning

Dissertation, 2023

Supervisor: Prof. Dr. Martin Raubal

Second Supervisor: Prof. Dr. Fernando Perez-Cruz

Reviewers: Prof. Dr. Fernando Perez-Cruz and Prof. Dr. Laura Alessandretti

ETH Zürich

Professur Geoinformations-Engineering

Institut für Kartografie und Geoinformation

Departement Bau, Umwelt und Geomatik

Stefano-Franscini-Platz 5

8093 Zürich

To Paula, Jakob and Gustav

Acknowledgments

I'd like to thank the mentors that I was lucky to have during and before my doctorate. First of all, I'd like to thank Martin Raubal for giving me so much freedom to choose my own research directions and supporting me whenever necessary. I'd also like to thank David Jonietz for his guidance during the first year of my doctoral studies and for the many hands-on research lessons, Michael Kopp for the great opportunity to work at IARAI, Esra for motivating me to dare make bolder submissions, and Ludwig and Christian for keeping me motivated to keep up with machine learning. Finally, I'd like to thank Fernando Perez-Cruz for offering his close support during the complicated final part of my dissertation.

Much of this work in my dissertation would not have been possible without my colleagues from the MIE-Lab (Dominik, David J., Christian, Jannik, Nishant, Ye, Yanan, Nina, and Ayda including René and Joram as honorary members) and the geoGAZELab (Fabian, Peter, David R., Luis, Kuno, Kevin, Tiffany, Lin, Adrian, Yiwei, and Tianyi). Thank you for the intense collaborations and the many coffee and lunch breaks. A special thanks to Ruth for facilitating my work from the beginning to the end of my ETH employment. Thanks also to Henrik and Daniel from the Institute for Transport Planning and Systems, and Thomas, Merlin and Jorim from SBB, and René from the Spatial Modelling Lab for the great collaborations.

I also want to thank my friends from Passau, Munich, Lille, Austin, Kassel, and Zurich without whom the years it took me to write this dissertation would have been very boring.

Finally, I want to thank my parents and my sisters for their support and friendship, and for their frequent advice on career, kids, life, and how to bring it all together. Most importantly, I want to thank my wife, Paula, for her never-ending support, for having patience with the uncertainty related to life in academia, and for tolerating the many nights I spent on my research.

Abstract

Many of today's urgent challenges, such as greenhouse gas emissions and climate change, air quality and health, or traffic and congestion, are closely linked to the movement of people and goods. A major cause of these problems is fossil fuel based individual transport, making individual mobility behavior change a requirement for solving them. Computational methods based on data collected using Information and Communication Technologies and Location-Based Services can play key a role in supporting sustainable mobility.

The focus of this dissertation is the development and application of computational methods to support sustainable individual mobility in four different ways. Gathering empirical evidence on how Mobility as a Service affects the mobility behavior of individuals, developing a framework for more generalizable methods to preprocess tracking data, developing methods to support the modal shift of individuals toward more sustainable modes, and developing methods to support the sustainability of personal vehicles.

A core contribution of this dissertation is the formalization of a graph-based representation of individual mobility called *location graph*. This representation is compact, privacy-preserving, and can be created based on a wide range of different datasets, which simplifies the development of transferable computational methods. Based on location graphs, we developed machine learning methods for identifying user groups with similar mobility behavior and for imputing missing activity labels. These methods were applied to problems related to the management of Mobility as a Service offers, a core concept to enable modal shift for individuals.

To support the sustainability of personal vehicles this thesis includes work on traffic prediction, a critical component of an intelligent traffic management system. Furthermore, it also includes a study showing that owners of battery electric vehicles can meet most of their charging demand using power generated from their own rooftop photovoltaic system. The latter has great potential to further reduce the greenhouse gas emissions of personal vehicles.

Zusammenfassung

Der auf fossilen Brennstoffen basierende Individualverkehr ist für viele schwerwiegende Probleme unserer Zeit mitverantwortlich. Das Senken der Treibhausgasemissionen, die Verbesserung der Luftqualität in den Städten und die Reduktion von Verkehrsunfällen ist deshalb nur möglich, wenn wir unser Mobilitätsverhalten ändern.

Der Schwerpunkt dieser Dissertation liegt auf der Entwicklung und Anwendung von computergestützten Methoden zur Unterstützung nachhaltiger individueller Mobilität. Dabei konzentriert sich diese Arbeit auf vier Bereiche: das Sammeln empirischer Daten über den Einfluss von Mobility as a Service auf das individuelle Mobilitätsverhalten, die Entwicklung verallgemeinerbarer Methoden zur Aufbereitung von Tracking Daten, Methodenentwicklung für die Verlagerung von Mobilität auf nachhaltigere Verkehrsmittel und die Entwicklung von Methoden zur Unterstützung der Nachhaltigkeit von Privatfahrzeugen.

Ein zentraler Beitrag ist die Formalisierung eines graphenbasierten Modells für individuelle Mobilität, das als location graph bezeichnet wird. Location graphs sind kompakt, wahren die Privatsphäre und sind mit einer Vielzahl unterschiedlicher Datensätze kompatibel. Auf der Grundlage des location graphs und maschinellen Lernens wurden Algorithmen entwickelt, um Nutzergruppen mit ähnlichem Mobilitätsverhalten zu identifizieren und um die, in Trackingdaten oft fehlenden Informationen zu Aktivitäten, hinzuzufügen. Diese Methoden können unter anderem dazu verwendet werden, um Mobility as a Service Angebote zu verbessern.

Um die Nachhaltigkeit von Privatfahrzeugen zu unterstützen, wurden im Rahmen dieser Arbeit kurzfristige Verkehrsprognosen erstellt, die eine zentrale Komponente für intelligente Verkehrsmanagementsysteme sind. Ausserdem umfasst diese Dissertation auch eine Studie zur Kombination von Elektroautos und Photovoltaikanlagen im Privatbereich. Das Ergebnis der Studie zeigt, dass Besitzer von batteriebetriebenen Elektrofahrzeugen den Großteil ihres Ladebedarfs mit Strom aus ihrer eigenen Photovoltaikanlage decken können. Letzteres hat ein großes Potenzial zur höheren Dekarbonisierung von Privatfahrzeugen.

Publications

Articles in Peer-Reviewed Journals

Martin, H., Perez-Cruz, F., Raubal, M., (2023). “A graph-based representation for human mobility data”. *International Journal of Geographical Information Science* (under review).

Martin*, H., Wiedemann*, N., Reck, D. J., Raubal, M., (2023a). “Graph-based mobility profiling”. *Computers, Environment and Urban Systems* 100, p. 101910. DOI: 10.1016/j.compenvurbsys.2022.101910.

Martin*, H., Hong*, Y., Wiedemann*, N., Bucher, D., Raubal, M., (2023b). “Trackintel: An open-source Python library for human mobility analysis”. *Computers, Environment and Urban Systems* 101, p. 101938. DOI: 10.1016/j.compenvurbsys.2023.101938.

Wiedemann*, N., **Martin***, H., Suel, E., Hong, Y., Xin, Y., (2023). “Influence of tracking duration on the privacy of individual mobility graphs”. *Journal of Location Based Services*, pp. 1–19.

Martin, H., Buffat, R., Bucher, D., Hamper, J., Raubal, M., (2022). “Using rooftop photovoltaic generation to cover individual electric vehicle demand — A detailed case study”. *Renewable and Sustainable Energy Reviews* 157, p. 111969. DOI: 10.1016/j.rser.2021.111969.

Peer-Reviewed Full Papers at Conference

Martin, H., Bucher, D., Hong, Y., Buffat, R., Rupprecht, C., Raubal, M., (2020). “Graph-ResNets for short-term traffic forecasts in almost unknown cities”. *NeurIPS 2019 Competition and Demonstration Track*. Vol. 123. PMLR, pp. 153–163. DOI: 10.3929/ethz-b-000437682.

Peer-Reviewed Short Papers and Workshop Contributions at Conference

Martin*, H., Bucher*, D., Suel, E., Zhao, P., Perez-Cruz, F., Raubal, M., (2018).
“Graph Convolutional Neural Networks for Human Activity Purpose Imputation”.
NIPS Spatiotemporal Workshop at the 32nd Annual Conference on Neural Information Processing Systems (NIPS 2018). DOI: 10.3929/ethz-b-000310251.

* Equal contribution

Contributions Cited in This Dissertation

- Cai, H., Xin, Y., **Martin, H.**, Raubal, M., (2022). “Optimizing Electric Vehicle Charging Schedules Based on Probabilistic Forecast of Individual Mobility”. *AGILE: GIScience Series* 3, p. 3. DOI: 10.5194/agile-giss-3-3-2022.
- Hong, Y., **Martin, H.**, Raubal, M., (2022). “How do you go where? Improving next location prediction by learning travel mode information using transformers”. *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pp. 1–10. DOI: 10.1145/3557915.3560996. eprint: 2210.04095.
- Hong, Y., Xin, Y., **Martin, H.**, Bucher, D., Raubal, M., (2021). “A clustering-based framework for individual travel behaviour change detection”. *11th International Conference on Geographic Information Science (GIScience 2021)-Part II*. Vol. 208. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, p. 4.
- Jonietz, D., Bucher, D., **Martin, H.**, Raubal, M., (2018). “Identifying and interpreting clusters of persons with similar mobility behaviour change processes”. *Geospatial Technologies for All: Selected Papers of the 21st AGILE Conference on Geographic Information Science 21*. Springer International Publishing, pp. 291–307. DOI: 10.1007/978-3-319-78208-9_15.
- Kreil, D. P., Kopp, M. K., Jonietz, D., Neun, M., Gruca, A., Herruzo, P., **Martin, H.**, Soleymani, A., Hochreiter, S., (2020). “The surprising efficiency of framing geospatial time series forecasting as a video prediction task – Insights from the IARAI T4C Competition at NeurIPS 2019”. *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*. Ed. by Hugo Jair Escalante and Raia Hadsell. Vol. 123. Proceedings of Machine Learning Research. PMLR, pp. 232–241.
- Martin, H.**, Reck, D. J., Axhausen, K. W., Raubal, M., (2021a). *ETH Mobility Initiative Project MI-01-19 Empirical use and Impact analysis of MaaS: Ergebnisse*. Tech. rep. ETH Zurich. DOI: <https://doi.org/10.3929/ethz-b-000521380>.
- Martin, H.**, Reck, D. J., Raubal, M., (2021b). “Using Information and Communication Technologies to facilitate mobility behaviour change and enable Mobility as a Service”. *GI_Forum Journal for Geographic Information Science* 9.1, pp. 187–193. DOI: https://doi.org/10.1553/giscience2021_01_s187.
- Martin, H.**, Becker, H., Bucher, D., Jonietz, D., Raubal, M., Axhausen, K. W., (2019a). “Begleitstudie SBB Green Class-Abschlussbericht”. *Arbeitsberichte Verkehrs-und Raumplanung* 1439. DOI: <https://doi.org/10.3929/ethz-b-000353337>.

- Martin, H.**, Hong, Y., Bucher, D., Rupprecht, C., Buffat, R., (2019b). “Traffic4cast-Traffic Map Movie Forecasting–Team MIE-Lab”. *arXiv preprint arXiv:1910.13824*. DOI: <https://doi.org/10.48550/arXiv.1910.13824>.
- Reck, D. J., **Martin, H.**, Axhausen, K. W., (2022). “Mode choice, substitution patterns and environmental impacts of shared and personal micro-mobility”. *Transportation Research Part D: Transport and Environment* 102, p. 103134. DOI: 10.1016/j.trd.2021.103134.
- Wiedemann, N., **Martin, H.**, Raubal, M., (2022). “Unlocking social network analysis methods for studying human mobility”. *AGILE: GIScience Series* 3, p. 19. DOI: 10.5194/agile-giss-3-19-2022.

Contributions not Included in This Dissertation

- Bucher, D., **Martin, H.**, Hamper, J., Jaleh, A., Becker, H., Zhao, P., Raubal, M., (2020). “Exploring Factors that Influence Individuals’ Choice Between Internal Combustion Engine Cars and Electric Vehicles”. Vol. 1. Copernicus GmbH, pp. 1–23. DOI: 10.5194/agile-giss-1-2-2020.
- Bucher*, D., **Martin***, H., Jonietz, D., Raubal, M., Westerholt*, R., (2020). “Estimation of Moran’s I in the Context of Uncertain Mobile Sensor Measurements”. *11th International Conference on Geographic Information Science (GIScience 2021) - Part I*. Ed. by Krzysztof Janowicz and Judith A. Verstegen. Vol. 177. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2:1–2:15. DOI: 10.4230/LIPIcs.GIScience.2021.I.2.
- Deetjen, T. A., **Martin, H.**, Rhodes, J. D., Webber, M. E., (2018). “Modeling the optimal mix and location of wind and solar with transmission and carbon pricing considerations”. *Renewable energy* 120, pp. 35–50. DOI: 10.1016/j.renene.2017.12.059.
- Eichenberger, C., Neun, M., **Martin, H.**, Herruzo, P., Spanring, M., Lu, Y., Choi, S., Konyakhin, V., Lukashina, N., Shpilman, A., Wiedemann, N., Raubal, M., Wang, B., Vu, H. L., Mohajerpoor, R., Cai, C., Kim, I., Hermes, L., Melnik, A., Velioglu, R., Vieth, M., Schilling, M., Bojesomo, A., Marzouqi, H. A., Liatsis, P., Santokhi, J., Hillier, D., Yang, Y., Sarwar, J., Jordan, A., Hewage, E., Jonietz, D., Tang, F., Gruca, A., Kopp, M., Kreil, D., Hochreiter, S., (2022). “Traffic4cast at NeurIPS 2021 - Temporal and Spatial Few-Shot Transfer Learning in Gridded Geo-Spatial Processes”. *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*. Ed. by Douwe Kiela, Marco Ciccone, and Barbara Caputo. Vol. 176. Proceedings of Machine Learning Research. PMLR, pp. 97–112. DOI: <https://doi.org/10.48550/arXiv.2203.17070>.
- Franke, M., **Martin, H.**, Koch, S., Kurzhals, K., (2021). “Visual Analysis of Spatio-temporal Phenomena with 1D Projections”. *Computer Graphics Forum*. Vol. 40. 3, pp. 335–347. DOI: 10.1111/cgf.14311.
- Göbel, F., **Martin, H.**, (2018). “Unsupervised Clustering of Eye Tracking Data”. *Spatial Big Data and Machine Learning in GIScience, Workshop at GIScience 2018*, pp. 25–28.
- Hong, Y., **Martin, H.**, Xin, Y., Bucher, D., Reck, D. J., Axhausen, K. W., Raubal, M., (2023). “Conserved quantities in human mobility: From locations to trips”.

Transportation Research Part C: Emerging Technologies 146, p. 103979. DOI: 10.1016/j.trc.2022.103979.

Kopp, M., Kreil, D., Neun, M., Jonietz, D., **Martin, H.**, Herruzo, P., Gruca, A., Soleymani, A., Wu, F., Liu, Y., Xu, J., Zhang, J., Santokhi, J., Bojesomo, A., Marzouqi, H. A., Liatsis, P., Kwok, P. H., Qi, Q., Hochreiter, S., (2021). “Traffic4cast at NeurIPS 2020 - yet more on the unreasonable effectiveness of gridded geo-spatial processes”. *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*. Ed. by Hugo Jair Escalante and Katja Hofmann. Vol. 133. Proceedings of Machine Learning Research. PMLR, pp. 325–343. DOI: <https://doi.org/10.48550/arXiv.2203.17070>.

Neun, M., Eichenberger, C., Xin, Y., Fu, C., Wiedemann, N., **Martin, H.**, Tomko, M., Ambühl, L., Hermes, L., Kopp, M., (2023a). “Metropolitan Segment Traffic Speeds from Massive Floating Car Data in 10 Cities”. *IEEE Transactions on Intelligent Transportation Systems (forthcoming)*. DOI: 10.1109/tits.2023.3291737.

Neun, M., Eichenberger, C., **Martin, H.**, Spanring, M., Siripurapu, R., Springer, D., Deng, L., Wu, C., Lian, D., Zhou, M., Lumiste, M., Ilie, A., Wu, X., Lyu, C., Lu, Q.-L., Mahajan, V., Lu, Y., Li, J., Li, J., Gong, Y.-J., Grötschla, F., Mathys, J., Wei, Y., Haitao, H., Fang, H., Malm, K., Tang, F., Kopp, M., Kreil, D., Hochreiter, S., (2023b). “Traffic4cast at NeurIPS 2022–Predict Dynamics along Graph Edges from Sparse Node Data: Whole City Traffic and ETA from Stationary Vehicle Detectors”. DOI: <https://doi.org/10.48550/arXiv.2303.07758>.

Raubal, M., Bucher, D., **Martin, H.**, (2021). “Geosmartness for Personalized and Sustainable Future Urban Mobility”. *Urban Informatics*. Ed. by Wenzhong Shi, Michael F. Goodchild, Michael Batty, Mei-Po Kwan, and Anshu Zhang. Singapore: Springer Singapore, pp. 59–83. DOI: 10.1007/978-981-15-8983-6_6.

Sailer, C., **Martin, H.**, Gaia, L., Raubal, M., (2019). “Analyzing performance in Orienteering from movement trajectories and contextual information”. *15th International Conference on Location-Based Services*, p. 141.

Scartozzi, C. M., Orsi, R., Ernst, M., **Martin, H.**, (2021). “Connecting Northeast Asia: Renewable Energy and Prospects for Cooperation”. *Asian Perspective* 45.4, pp. 761–783. DOI: 10.1353/apr.2021.0040.

Zhao, P., Bucher, D., **Martin, H.**, Raubal, M., (2019). “A Clustering-Based Framework for Understanding Individuals’ Travel Mode Choice Behavior”. *22nd Conference on Geographic Information Science (AGILE 2019), Limassol, Cyprus*. Springer, Cham, pp. 77–94. DOI: 10.1007/978-3-030-14745-7_5.

* Equal contribution

Contents

Acknowledgments	v
Abstract	vii
Publications	ix
1 Introduction	1
1.1 Background	2
1.2 Research scope	5
1.3 Contributions and structure of the dissertation	5
2 Empirical data collection	9
2.1 SBB Green Class	9
2.2 Empirical use and impact analysis of MaaS	15
3 Processing and representing individual tracking data	19
3.1 Trackintel: an open-source Python library for human mobility analysis	19
3.2 A graph-based representation for human mobility data	43
3.3 Influence of tracking duration on the privacy of individual mobility graphs	71
4 Supporting modal shift (MaaS)	85
4.1 Graph-based mobility profiling	85
4.2 Graph convolutional neural networks for human activity purpose imputation from GPS-based trajectory data	111
5 Supporting sustainability of personal vehicles	119
5.1 Using rooftop photovoltaic generation to cover individual electric vehicle demand - a detailed case study	119
5.2 Graph-ResNets for short-term traffic forecasts in almost unknown cities	147
6 Conclusion	157
6.1 Summary and contributions	157
6.2 Research questions	160
6.3 Outlook	164
A Appendix	167

A.1	Supporting information for Trackintel: an open-source Python library for human mobility analysis	167
A.2	Supporting information for Influence of tracking duration on the privacy of individual mobility graphs	168
A.3	Supporting information for Graph-based mobility profiling	169
A.4	Supporting information for Using rooftop photovoltaic generation to cover individual electric vehicle demand—a detailed case study	174
	Bibliography	175
	List of Figures	205
	List of Tables	208
	Acronyms	209

Introduction

1

1.1 Background

Many of today's urgent global challenges, as defined in the United Nations Sustainable Development Goals are related to the movement of people and goods (Jaramillo et al., 2022). Action in the transport sector is required for fighting poverty and hunger (Sasidharan, 2017), to ensure health and well-being (Krzyzanowski et al., 2005; Sofiev et al., 2018), provide access to education (Hernandez, 2018), reduce gender inequality (Levin and Faith-Ell, 2019; Prati, 2018), ensure economic growth (Rodrigue, 2020) and for the creation of sustainable and livable cities (Vuchic, 1999). Most importantly, the transport sector is a major emitter of greenhouse gases and thereby contributes to climate change which by itself negatively impacts the progress of most of the SDGs (WMO, 2021).

Globally, the transport sector is responsible for about 16% of total GHG emissions (Ritchie et al., 2020) and 25% of the total final energy consumption (IEA, 2020b). In developed countries, the transport sector is usually the largest emitter of GHG with 28.5% in 2021 in the US (EPA, 2023), 28% in 2021 in the European Union (EU) (EEA, 2023) and with 30.6% in 2021 in Switzerland (BAFU, 2023). In addition to its high emissions, the transport sector is not on track for decarbonization. It is the energy sector with the highest emission growth globally (SLOCAT, 2023, 2021) and the only sector in the EU that is still increasing its emissions relative to 1990 levels (EEA, 2023). The slow decarbonization progress of the transport sector is especially problematic because the world is likely to fail to keep the increase of global average temperature below 1.5°C or even 2°C as agreed upon in the Paris Agreement (UNFCCC, 2015) as there currently is a need for additional GHG reduction measures of 17 GtCO_{2e} to limit global temperature rise to below 2°C or 25 GtCO_{2e} to keep global temperature rise to below 1.5°C (UNEP, 2022).

In order to limit global temperature rise and support the progress on the SDGs it is crucial that the transport sector and thereby our mobility becomes more sustainable. Sustainability with regards to achieving the SDGs can be defined as the “development that meets the needs of the present without compromising the ability of future generations to meet their own needs” (WCED, 1987, p.41) while balancing its economic, social and environmental dimensions (UN, 2015).

Sustainable mobility, therefore, has to continue to fulfill our society's needs for affordable transportation while reducing its environmental and social impact to a manageable level. Achieving this compromise is difficult however, research in the past decades has focused on these problems and today we know conceptually how to make our mobility sustainable. The required actions can be summarized using the three following categories: *Avoid*, *Shift*, *Improve* (Banister, 2008; Boulouchos et al., 2021; Holden et al., 2020; Jaramillo et al., 2022; SLOCAT, 2023). The Avoid-Shift-Improve (A-S-I) framework represents a set of hierarchical interventions that should be applied for the step-wise transformation of the transport sector.

Avoid

The best way to reduce emissions is to prevent them from being generated from the beginning by reducing the need to travel. There are short-term interventions to avoid travel such as incentivizing working from home and online meetings, or restrictions of car use on Sundays (Bongardt et al., 2019), and long-term planning-related measures that reduce the length of trips such as the planning of compact cities that combine mixed-use and high-density development (Burton et al., 2003).

Shift

The part of our mobility that can not be avoided should be covered by more energy-efficient modes of transport. The main source of GHG emissions in the transport sector is the usage of fossil fuels in road-based transportation. Therefore the transition of the transport sector will have to focus on a sustainable alternative for these modes. For freight transport, great potential lies in shifting goods from road to rail (Lawrence and Bullock, 2022), and for individual mobility we have to serve trips with a combination of public transport, walking, cycling, or micromobility modes instead of privately owned cars. A core concept for this shift away from the privately owned vehicle is Mobility as a Service (MaaS) which integrates shared modes with public transport to facilitate intermodal travel (Reck et al., 2020). Changing the modes that people use regularly requires individuals to change their mobility behavior significantly, which will be only acceptable for a broad population if MaaS offers are able to compete with the comfort and ease of a private car (Boulouchos et al., 2021; Raubal et al., 2021; Weiser et al., 2016). However, it is still an open question how to design MaaS offers that are appealing to many individuals and lead to a more sustainable mobility behavior at the same time (Martin et al., 2021c). Hence, more empirical data on how people use MaaS offers are required. Nowadays, these data can be collected on a large scale with comparably low costs by conducting tracking studies using GNSS devices (Zheng et al., 2008) or smartphones (Bucher et al., 2019b) while the study participants have access to a MaaS offer. However, the collected tracking data are highly sensitive and often suffer from poor data quality with location data that are often noisy or incomplete and labels for transport modes or activities that are partially implausible or missing. This all leads to a high preprocessing effort in order to correctly measure the mobility behavior.

Improve

Finally, for all mobility that can not be avoided or shifted to more sustainable modes, the energy efficiency of the employed mode should be improved through technological advances (Bongardt et al., 2019). The most important intervention in this category is the use of battery electric vehicles instead of internal combustion engine vehicles, as BEVs have significantly lower GHG emissions than ICEVs over their respective lifetime (Cox et al., 2020; Haasz et al., 2018; Ruhnau et al., 2019) which makes the large-scale roll-out of BEVs a promising path for fast decarbonization. It is important to note, that simply replacing an ICEV with a BEV leaves many challenges unresolved. Individual motorized transportation will still block large areas of public space for parking and infrastructure instead of using it for housing or recreational space and the problem of increasing traffic is left unresolved. These problems, and particularly congestion, will intensify as the world's population is expected to

continue to grow and urbanize until 2050 (UN, 2018) which will require smart management of the growing traffic volume in order to reduce its negative impact.

The combination of A-S-I measures represents a known possible pathway for the decarbonization of the transport sector (ITF, 2021) until 2050. However, their implementation is challenging as it requires an in-depth understanding of people's mobility behavior which can often only be accessed through noisy, sparse, and unlabeled tracking data. At the same time, their implementation also demands an optimal allocation and management of (spatial) resources which requires the prediction of mobility demand and the prediction of the availability of mobility assets.

A great opportunity to support their implementation lies in the ongoing digitization of our society. Today we have the possibility to collect tracking and contextual data sets on a large scale using ubiquitous Information and Communication Technologies (ICT) and leverage the advances in machine learning for their processing and analysis to gain a deep understanding of people's mobility behavior. Especially deep learning methods created disruptive breakthroughs in image (Krizhevsky et al., 2012), speech (Hinton et al., 2012), and video processing however, attempts to use deep learning with human mobility data (Chen et al., 2016b; Dabiri and Heaslip, 2018; Fan et al., 2018; Kumar and Raubal, 2021; Toch et al., 2018) did not yet lead to break-through successes that disrupted the field of human movement analysis. As their success relies on their ability to self-learn rich representations (e.g., features) from raw data (Bronstein et al., 2017; LeCun et al., 2015) a possible reason for their limited success is that deep learning methods are mostly focused on image processing while it is difficult to represent human mobility data as images.

A natural way to represent individual mobility is given by person-specific graphs based on the already visited locations (Martin et al., 2018; Pappalardo et al., 2015; Rinzivillo et al., 2014; Zheng et al., 2008). In such a graph, nodes correspond to visited locations and edges may correspond to any measure of distance between the nodes (e.g., the counts of sequential visits of two locations, spatial distance, transport time, or cost (Martin et al., 2018)). These representations offer several advantages: 1) they can be enriched with node and edge features based on the application needs, 2) they are compact and grow sub-linearly in size with increasing tracking duration, 3) they still provide rich insight into mobility behavior despite their compactness (Martin et al., 2023b; Rinzivillo et al., 2014; Wiedemann et al., 2022) and can be analyzed efficiently with graph neural networks for various applications such as activity purpose imputation (Martin et al., 2018).

The focus of this dissertation, therefore, lies in the combination of machine learning and graph-based methods to address the computational and algorithmic challenges associated with the decarbonization of mobility and transport.

1.2 Research scope

The goal of this dissertation is to develop computational methods to support sustainable mobility with a focus on the *Shift* and *Improve* measures of the A-S-I framework with a focus on individual mobility. It thereby focuses on the following two main research questions:

- How can the modal shift of individuals be supported using computational methods?
- How can the sustainability of personal vehicles be supported using computational methods?

The focus on computational methods for the support of sustainable mobility brought two additional research questions into the focus of this doctoral thesis.

- Tracking datasets of individuals are highly diverse in terms of sample size, tracking source, and available label and context data (Chen et al., 2016a). In any case, they pose a high privacy risk (Keßler and McKenzie, 2018; Song et al., 2010a) to the tracked individuals.

How can individual mobility from different data sources be represented in a compact, and privacy-friendly way such that it allows the development of computational methods?

- Preprocessing and analysis methods for tracking data have a strong influence on the results of mobility studies. However, these are currently not standardized and highly specialized on specific datasets or studies making them difficult to reuse and results difficult to generalize.

How can the reproducibility and generalizability of preprocessing and analysis methods for human mobility be improved?

1.3 Contributions and structure of the dissertation

The contribution of this dissertation is the development of computational methods with a focus on machine learning and graph methods that support the shift of individuals towards more sustainable modes using MaaS and improve the sustainability of personal vehicles.

The individual contributions of this dissertation are structured in four main chapters as shown in Figure 1.1. The following provides a brief summary of the individual chapters.

Empirical data collection

Empirical data on the impact of MaaS offers on mobility behavior and the usage of BEV are still scarce or inaccessible to researchers. As part of this dissertation, we have collected empirical tracking datasets in three case studies. The SBB Green Class E-Car (GC1) with 139 participants tracked over one year, SBB Green Class E-Bike (GC2) with 50 participants tracked over 1 year, and the yumuv study with 871 participants tracked over three months. All participants had access to a study-specific MaaS offer, a personal BEV in the GC1 study,

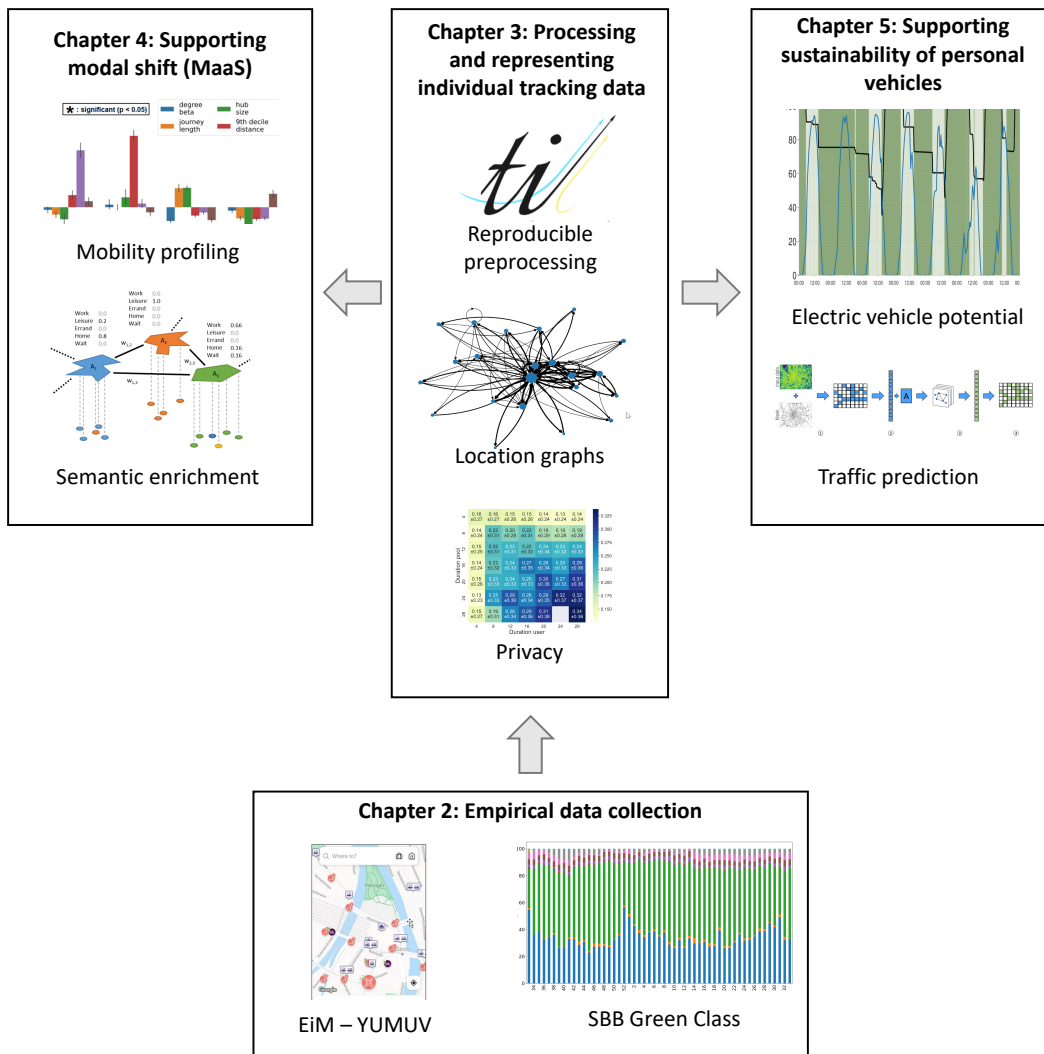


Fig. 1.1.: Organization of the dissertation.

and an e-bike in the GC2 study. These case studies are summarized in Chapter 2 of the dissertation.

Reproducible preprocessing and compact mobility representation

The collection of the different GPS-based tracking datasets in this dissertation highlighted the complexity and impact of preprocessing for further analysis. At the same time, it revealed the lack of a coherent preprocessing framework for the analysis of individual mobility data. Therefore, Trackintel¹, an open-source Python library for mobility data preprocessing and analysis, was developed as a part of this dissertation and is further described in Section 3.1

Finally, after collecting and processing the data, a core part of this dissertation is to introduce location graphs as a graph-based representation of individual human mobility based on visited locations in Section 3.2. The location graph of an individual is constructed

¹<https://github.com/mie-lab/trackintel>

using activity locations as nodes and direct transitions between activity locations (trips) as weighted edges. Section 3.2 discusses the properties, advantages, and disadvantages of this representation. The location graph also serves as the basis for the methods developed in Chapter 4. Since privacy is particularly important in the context of personalized geodata, the privacy properties of this representation are further analyzed in Section 3.3.

Supporting modal shift

The collected datasets presented in Chapter 2 and the graph representation presented in Section 3.3 are used to develop methods for the analysis and management of MaaS offers to support the shift of individuals towards more sustainable modes of transport. The work presented in Section 4.1 uses location graphs to identify groups of individuals with similar mobility behavior that appear across several datasets.

Section 4.2 presents a graph neural network based approach to predict the activity label of visited locations. The network is trained on location graphs of individuals and is used to predict the node activity labels of unknown persons. This approach can be used to enrich unlabeled tracking data to better understand the mobility behavior of individuals which may be used to personalize transportation offers.

Supporting sustainability of personal vehicles

Section 5.1 investigates to what extent a BEV owner could operate his or her vehicle solely relying on a photovoltaic (PV) system on their own roof. We used the detailed BEV data available in the GC1 dataset (consumption per trip, state of charge, charging behavior) and combined them with a fine-grained digital surface model of Switzerland to extract the detailed roof geometry and the corresponding rooftop PV generation capacity of each of the BEV owner's houses. The results demonstrate that the recorded mobility energy demand could be covered entirely by the less GHG-intensive rooftop PV generation for almost the entire study period of 10 months. However, the results strongly depend on the specific emissions of the power used for charging and on the employed charging strategy which emphasizes the need to optimize charging for renewable energy consumption in order to reduce the GHG emissions of BEVs.

Section 5.1 shows that BEVs can be used to reduce the GHG emissions of an individual by covering their consumption using local renewable energy generation. However, the growing traffic volume still has to be managed in a smart way to reduce its negative impacts which will require short-term traffic forecasts. Section 5.2 presents a graph neural network based approach to predict traffic in cities that were not part of the training set and were not previously seen by the network. For the tested traffic forecasting experiment, the generalization properties of graph neural networks were significantly better than those of previously used convolutional neural networks which facilitates the transfer of a learned network to predict the traffic status in unknown cities.

Finally, Chapter 6 summarizes the contributions of this work with respect to the research questions and offers potential avenues for future research.

MaaS is considered a key concept to support the transition away from the individual private car and to increase the sustainability of (urban) mobility. However, while MaaS has been conceptualized for some time, there is a lack of empirical data on the actual impact of MaaS offers on an individual's mobility behavior or the required computational models to support more sustainable travel behavior in the context of MaaS.

Over the course of this dissertation, two large empirical datasets on MaaS usage were collected and analyzed. This chapter summarizes the data collection, processing, and results of the SBB Green Class studies in Section 2.1 and the Empirical use and Impact analysis of MaaS (EIM) project in Section 2.2.

2.1 SBB Green Class

This Section is a translated summary of the following non peer-reviewed technical report:

Henry Martin, Henrik Becker, Dominik Bucher, David Jonietz, Martin Raubal, and Kay W Axhausen (2019b). "Begleitstudie SBB Green Class-Abschlussbericht". *Arbeitsberichte Verkehrs-und Raumplanung* 1439. DOI: <https://doi.org/10.3929/ethz-b-000353337>.

2.1.1 Background and summary

In 2016 and 2017, the Swiss Federal Railways (SBB)¹ conducted two large, one-year MaaS pilots, in each of which customers were provided with a comprehensive mobility package in return for paying a fee. In 2016, the Green Class e-car (GC1) pilot started, which included a 1st class general transport pass valid in Switzerland, a *BMW i3*² electric vehicle, a parking space near a train station of choice, and credit for using *Mobility*³ car sharing and *PubliBike*⁴ bike sharing. In 2017, the second pilot, Green Class e-bike (GC2), was launched, where customers could choose between a 1st class and 2nd class general transport pass and received an *ST2* e-bike from *Stromer*⁵ and credit for using *Mobility* car sharing. The special feature of these offers was that they represented a flat rate for mobility, with almost all costs covered by the subscription fee.

¹<https://www.sbb.ch/en/home.html>

²<https://www.bmw.com/en/index.html>

³<https://www.mobility.ch/en>

⁴<https://www.publibike.ch/en/home>

⁵<https://www.stromerbike.com/en>

As part of the pilot tests, SBB collected detailed information on the socio-demographic background, daily mobility behavior, and the use of the various service components. These were evaluated in a scientific study by the Institute of Cartography and Geoinformation (IKG) and at the Institute for Transport Planning and Systems (IVT) of ETH Zurich.

2.1.2 Data collection

Survey

Access to the *Green Class* pilots was limited. Interested parties had to apply for participation by completing an online questionnaire. This questionnaire collected, among other things, information on the socio-demographic background of the applicants as well as on vehicle and subscription ownership. This information can be used for a comparison with the Mobility and Transport Microcensus (MTMC) (BFS and ARE, 2017). For the evaluations, only the actual Green Class participants are considered, since only for them is information on daily transport behavior also available.

Travel diaries

The core of the evaluations is the travel diaries, which were recorded in both pilot studies using the *DailyTracks* app. For the GC1 pilot study, the movement behavior of 139 users was recorded between November 2016 and the end of January 2018. For the GC2 pilot study, the movement behavior of 50 users was recorded between August 2017 and August 2018. The app recorded the user's movements using GPS and segmented them into triplegs and staypoints. For each tripleg, users were suggested the most sensible means of transport, and for each staypoint, an activity label. These could then be checked, changed, and confirmed by the user. Here are some key figures on the available data from the GC1 pilot study (data on SBB GC2 in brackets): In total, over 227 million (74 million) positions were recorded using GPS. These were aggregated to 242 012 (62 470) trips with 465 195 (128 640) triplegs and 326 926 (87 884) activity locations. Paths summarize all movements between 2 activity locations and can consist of several triplegs. In total, about 5.7 million (2.15 million) kilometers were traveled within Switzerland.

BMW i3 data

In addition to the *DailyTracks* travel diaries, the electric vehicle of the GC1 participants was also available to support the analysis. The data contains information on the battery charge level, the outside temperature, the mileage as well as time stamps and coordinates (point coordinates).

2.1.3 Data processing

The tracking data that were collected during the Green Class studies required several preprocessing steps before they could be used to analyze the mobility behavior of participants. Figure 2.1 shows a schematic representation of the different preparation steps.

The preparation of the travel diaries for GC1 and GC2 was analogous except for the steps concerning the electric car. In the GC2 study no data concerning the operation of the electric bicycles were recorded. For the preparation, the data were first imported into a PostgreSQL database with PostGIS extension and separated into triplegs and staypoints (steps 1-4). Then the data were validated with the available records of the BMW i3, mainly to identify and validate the electric car mode label of triplegs (step 5). Afterward, the activity labels of staypoints were imputed using a random forest model (step 6). The DailyTracks app had difficulties in correctly detecting transfers when traveling using public transport, therefore all public transit triplegs were compared to timetable and stop data and segmented when necessary (step 7). Furthermore, access and egress walks for public transport trips were often not detected by the app and were segmented post-tracking (step 8). Then all triplegs were matched to the road and path network (step 9). In the following, plausibility tests were carried out to identify and correct incorrectly validated or incorrectly recorded data. Anomalies that could not be corrected automatically were then flagged.

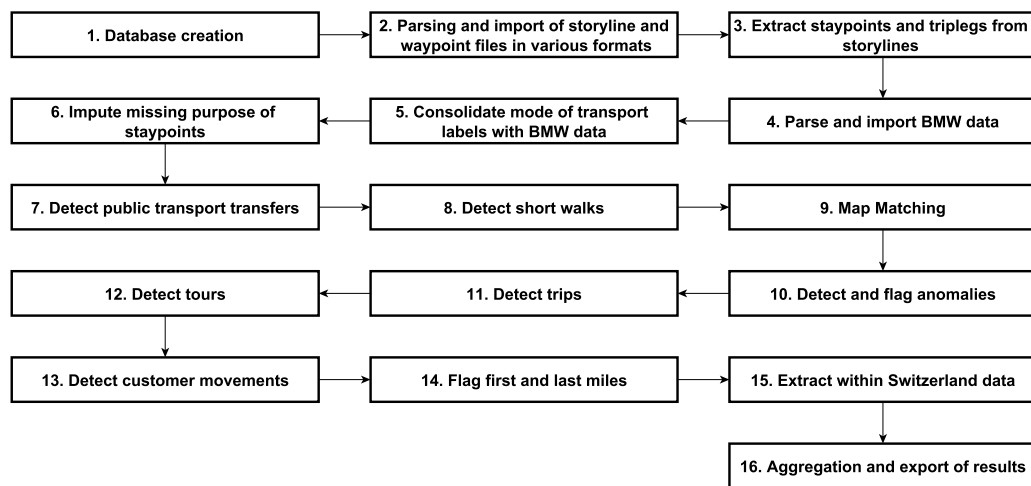


Fig. 2.1.: Data preprocessing workflow deployed for the SBB Green Class pilot studies.

2.1.4 Results

The Green Class pilot customers are predominantly male, middle-aged, and well-off.

When comparing Green Class clients with the MTMC, substantial differences emerge. Green Class customers are predominantly men (86% GC1; 82% GC2) in their mid-30s to mid-50s. 91% of GC1 customers work full-time, compared to 78% of GC2 customers and 53% in the MTMC. This is another reason why Green Class customers have an above-average household income and wealth. The latter can also be seen by the fact that only 11% of GC1 customers live for rent (MTMC: 49%; GC2: no information). 70% of Green Class customers live in detached houses (MTMC: 39%), the household size of Green Class customers is not significantly different from the MTMC.

Green Class customers travel more than the Swiss average.

Customers of both Green Class offers make an average of 4.55 trips per day. This is almost 12 % more than their respective control group sampled from the MTMC. GC1 customers cover 97 km (GC2 82 km) a day. Only the GC1 customers thus travel more than the control group. GC1 customers travel longer distances for all trip purposes, but especially for trips to work and errands. In return, they spend significantly less time at home, but more on the road, at work or for leisure activities. With GC2 customers, an exact allocation is more difficult, as a larger part of the trip purposes is undefined.

Green Class customers travel multimodal.

Green Class customers show an above-average share of multimodal travel. GC1 customers combined different modes of transport for 15% of their trips (GC2: 18%). In the control group, the share of multimodal car trips is significantly lower at a total of 5 %.

For most Green Class customers, the offer is not financially worthwhile.

For only 12 % of GC1 and 18 % of GC2 customers the package offer is cheaper than a (fictitious) individual billing of their journeys. For the majority of customers, however, it is significantly more expensive. The difference is so large that it can hardly be explained solely by the advantages of parking near the station or individual car-sharing journeys. This indicates that customers are willing to pay a rather high surcharge for a flat-rate solution such as Green Class.

The average CO₂ emissions of the GC1 participants decreased significantly with the start of the project.

After the initial tracking period without intervention (~8 weeks), the CO₂ emissions of almost all users of the GC1 bundle decreases strongly, which can be attributed to a decrease in conventional car trips. Figure 2.2 shows the development of the distance-based modal split of all GC1 users over the project period. With the start of the intervention (access to the MaaS offer), conventional car trips are replaced by trips with the BEV which has a stable share until calendar week 50 when the project officially ended. The shares of train trips remained stable over the project duration except during holiday periods.

Without the replacement of conventional cars, CO₂ emission reduction is unstable.

The GC2 pilot study participants do not show a clear trend for their average CO₂ emissions over the course of the project. In particular, there is no clear difference between the time emissions before and after the start of the project. On the one hand, the emissions of the GC2 participants were already on a low level before the project, on the other hand, conventional car trips were not replaced by the e-bike in the same manner as they were replaced by the e-car for GC1 participants.

Figure 2.3 shows the development of the distance-based modal split of all GC2 users over the project duration. At the start of the project (approx. from week 38), an increase in the share of electric bikes in the modal split can be seen which decreases again after a few

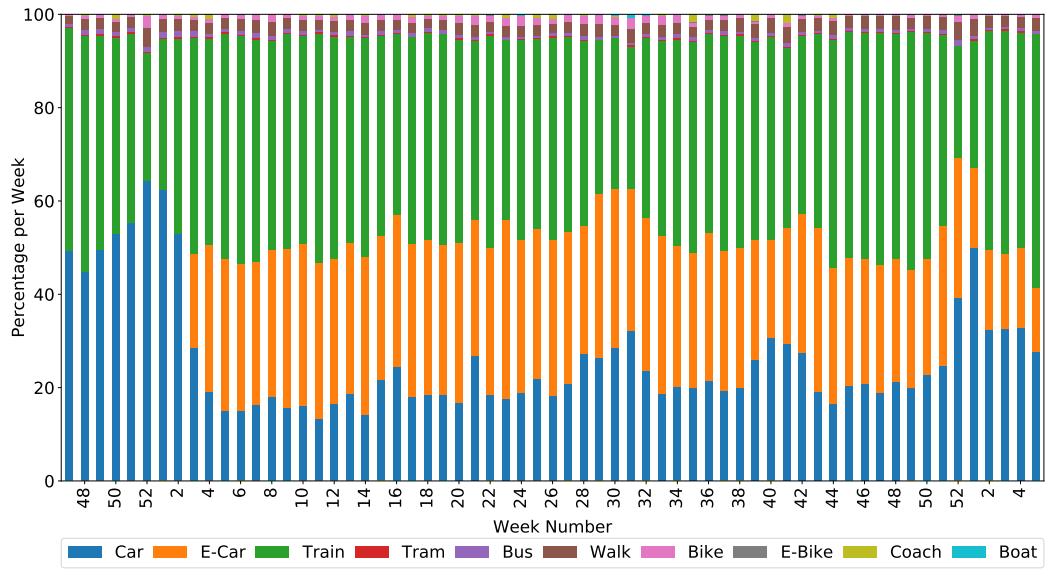


Fig. 2.2.: Distance-based modal split of GC1 participants over the course of the project.

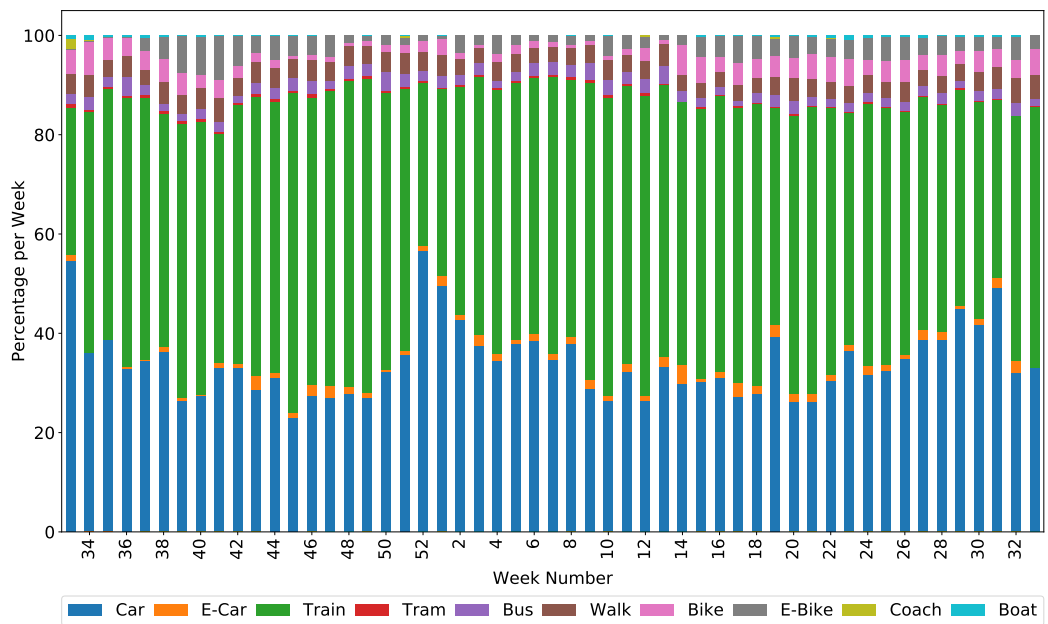


Fig. 2.3.: Distance-based modal split of GC2 participants over the course of the project.

weeks. When interpreting the shares of bicycles and electric bicycles, it should be noted that, in contrast to the GC1 pilot study, no movement data of the electric bicycles were available for validation.

2.2 Empirical use and impact analysis of MaaS

This Section is a translated summary of the following non peer-reviewed technical report:

Henry Martin, Daniel Jan Reck, Kay W Axhausen, and Martin Raubal (2021a). *ETH Mobility Initiative Project MI-01-19 Empirical use and Impact analysis of MaaS: Ergebnisse*. Tech. rep. ETH Zurich. DOI: <https://doi.org/10.3929/ethz-b-000521380>.

2.2.1 Background and summary

In 2020 the SBB together with the public transport providers in three major Swiss cities, Basel (Basler Verkehrs-Betriebe), Bern (Bernmobil), and Zurich (Verkehrsbetriebe Zürich), launched a new MaaS app called yumuv. The yumuv app integrated the local public transport with two shared e-scooter providers (Tier, Voi) and a shared e-bike provider (BOND) and allowed multi-modal trip planning and booking without leaving the app. Additionally, the app offered a subscription, also called mobility bundle, with a combined price for all mobility options (e.g., the bundle *yumuv easy* allowed for traveling 60 minutes per month with any provider without additional fees).

In parallel to the roll-out of the yumuv app, the Institute of Cartography and Geoinformation (IKG) and the Institute for Transport Planning and Systems (IVT) of ETH Zurich conducted a user study in Zurich. The study was part of the Empirical use and Impact analysis of MaaS (EIM)⁶ project to explore the influence of MaaS bundles on mobility behavior and to explore the potential of graph-based representations of an individual's mobility for mobility analytics.

The participants of the user study were separated into two groups, the treatment group (TG) and a control group (CG). Both groups participated in a survey at the beginning and the end of the study to collect socio-demographic information and used a tracking app to record their movement for three months. After one month of tracking, the TG was provided to the yumuv app and a mobility bundle. An overview of the study design is shown in Figure 2.4.

2.2.2 Data collection

Tracking data

The mobility of the participants was recorded for the duration of the study via a tracking app. The app automatically segmented the movement data and collected approximately 371 000 (TG: 112 000) triplets and approximately 248 000 (TG: 65 000) staypoints for the CG. Triplets are labeled with one of the following means of transport by the app: {tram, walk, train, motorbike, bus, boat, car, e-car, e-bicycle, bicycle, coach, ski, kick scooter}. Staypoints are automatically assigned one of the following purposes: {eat, sport, wait, work, leisure, errand, unknown, study, home}. The study participants were obliged to check the data and correct them if necessary.

⁶<https://csfm.ethz.ch/en/research/projects/eim.html>

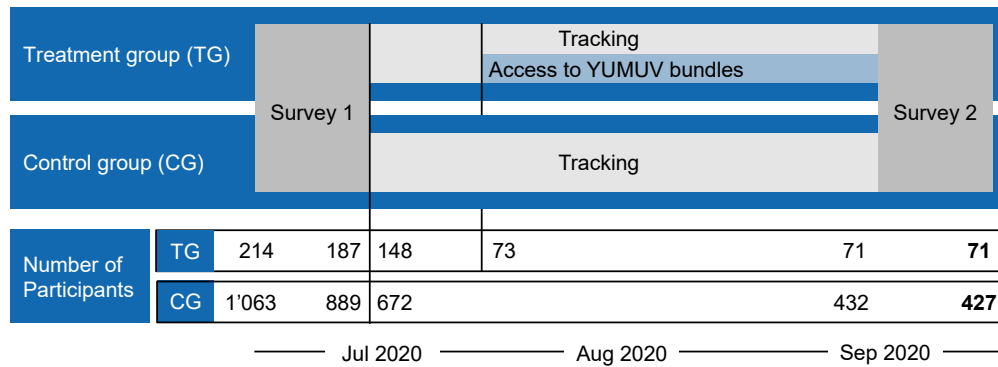


Fig. 2.4.: Overview of study design of the EIM project and the number of participants of the treatment group (TG) and the control group (CG) that successfully completed each phase of the study.

Survey data

Two online surveys were conducted at the beginning and end of the project with a total of 171 questions. The questions were divided into the following categories:

- Person-specific socio-demographic questions (e.g. date of birth, gender, educational background, employment)
- Household-specific socio-demographic questions (e.g. number of adult persons and children in the household, monthly household income, mobility tool ownership).
- Person-specific mobility questions (e.g. ownership of public transport season tickets, awareness of shared mobility services, access to shared mobility services).

Non-chosen alternatives

For the mode-choice analysis of participants, attributes about the non-chosen mode alternatives are required. Therefore, the travel distances for public transport, car, bicycle, and walking were calculated for each trip using MaTSim (Horni et al., 2016).

Booking data

For participants of the TG, the booking data for micromobility triplegs was recorded via the yumuv app. The relevant booking data included the price and time stamp at the start and end of the booking.

Context data

Additionally, the following context data were used in the analysis:

- Weather data (temperature, wind, and precipitation values) from the ERA5T dataset with an hourly temporal resolution and a 30x30 km spatial resolution.
- Elevation profile information from a digital elevation model of Switzerland with a resolution of 25 meters (Swisstopo DHM25 - Basis Model).

- Availability and positions of all relevant shared vehicles (e-scooters: VOI and Tier; e-bikes: BOND) with a temporal resolution of 5 minutes and full spatial resolution.

2.2.3 Data processing

The data processing for the EIM project was organized within an interdisciplinary team between all project partners and implemented directly in production on the infrastructure of SBB. Data were stored in a PostgreSQL database with PostGIS extension and processed using Python using the Trackintel library (Martin et al., 2023c) for movement data processing. Figure 2.5 provides an overview of the preprocessing steps applied in this project. In the first step, the movement data were summarized into trips, which are the central unit of analysis in transport planning (step 1). Then stoppoints were summarized into locations using clustering (step 2). Locations represent activity locations that are of specific importance to individuals. Next, the tracking data of the participants of the CG were matched with the booking data from the yumuv app (step 3). The booking data of micromobility modes allowed to correct mode labels, which was necessary because scooters were often misclassified as walking or biking (step 4). The map-matching algorithm and the MATSim simulation support different sets of modes of transport, the mode chains available for trips, therefore have to be aggregated for and matched to {bicycle, bus, car, tram} for map matching (step 6) and to {walk, bicycle, car, public transport} for the calculation of non-chosen alternatives using MatSim. Finally, all available context data was integrated on the individual trip level (step 7) to create the final dataset to be used for further analysis (step 8).

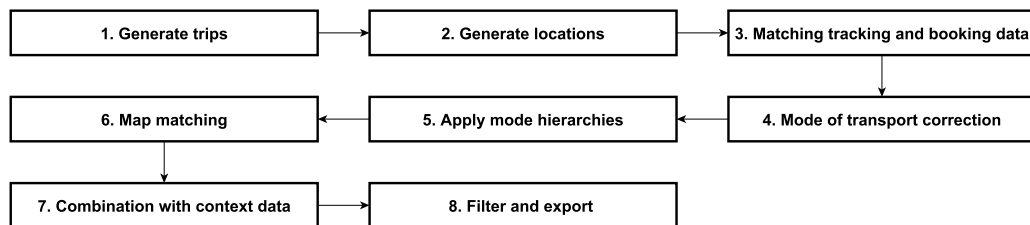


Fig. 2.5.: Data preprocessing workflow deployed for the EIM project.

2.2.4 Results

A typical user of the yumuv app is a young, male, student without a car but with a public transport pass.

As part of the study, the socio-demographic structure of the CG was compared to the TG. Participants (TG: n=71, CG: n=427) of the TG are primarily male (TG: 76%; CG: 48%), young (average age TG: 32; CG: 36), currently pursuing an education (TG: 23%; CG: 9%), do not live in a household with a car (TG: 68%; CG: 44%) but have a general transit pass (GA) (TG: 55%; CG: 14%).

Location graphs are suitable to model and analyze individual mobility behavior to analyze MaaS impact.

In the project, the location graph was formalized as a graph-based representation of individual mobility that is compact, privacy-friendly, and easy to combine with tracking data from different sources. The location graph was then used to describe the mobility behavior of the participants by mapping graph-based features to the individual mobility behavior. Furthermore, we developed a pipeline based on the graph representation that can be used to identify mobility behavior-based user groups, the impact of the MaaS offer on group membership and the attractiveness of a MaaS offer for different user groups.

The mobility bundle jointly increased the usage of micromobility and public transport.

Based on the results of a mode choice model fitted on the collected and prepared dataset, we analyzed the impact of the bundle *yumuv Abo easy*. The results of the model showed that the bundle particularly increases the use of e-scooters and to a lesser extent the use of public transport. The main substitutes are journeys with one's own bicycle and one's own e-bike. There was no significant impact on the participants' private car use.

Processing and representing individual tracking data

3.1 Trackintel: an open-source Python library for human mobility analysis

The following section is a reprint of the publication: Henry Martin*, Ye Hong*, Nina Wiedemann*, Dominik Bucher, and Martin Raubal (2023). “Trackintel: An open-source Python library for human mobility analysis”. *Computers, Environment and Urban Systems* 101, p. 101938. DOI: 10.1016/j.compenvurbsys.2023.101938.

Conceptualization: HM, YH, NW, DB; Methodology: HM, YH, NW, DB; Software: HM, YH, NW, DB; Validation: HM, YH, NW, DB; Writing – original draft: HM, YH, NW, MR; Writing – review & editing: HM, YH, NW, DB, MR; Supervision, funding acquisition: MR

The content is reproduced ‘as is’, however, formatting changes and corrections of spelling have been applied.

Abstract

Over the past decade, scientific studies have used the growing availability of large tracking datasets to enhance our understanding of human mobility behavior. However, so far data processing pipelines for the varying data collection methods are not standardized and consequently limit the reproducibility, comparability, and transferability of methods and results in quantitative human mobility analysis. This paper presents Trackintel, an open-source Python library for human mobility analysis. Trackintel is built on a standard data model for human mobility used in transport planning that is compatible with different types of tracking data. We introduce the main functionalities of the library that covers the full life-cycle of human mobility analysis, including processing steps according to the conceptual data model, read and write interfaces, as well as analysis functions (e.g., data quality assessment, travel mode prediction, and location labeling). We showcase the effectiveness of the Trackintel library through a case study with four different tracking datasets. Trackintel can serve as an essential tool to standardize mobility data analysis and increase the transparency and comparability of novel research on human mobility.

* Equal contribution

3.1.1 Introduction

Human mobility studies using large-scale human digital traces have boomed over the last decade. On the collective level, researchers revealed that human movement can be universally described using statistical distributions, i.e., the power-law distribution of consecutive displacements (Brockmann et al., 2006; Rhee et al., 2011), stationary time between displacements (Rhee et al., 2011; Song et al., 2010b), and characteristic distance traveled by individuals (i.e., the radius of gyration) (González et al., 2008; Pappalardo et al., 2015). Moreover, it has been shown that individuals exhibit markedly regular location visitation patterns (Schneider et al., 2013) with high theoretical predictability (Song et al., 2010a). People spend most of their time in a few locations (González et al., 2008; Song et al., 2010b) and maintain a stable number of important locations over time (Alessandretti et al., 2018).

To a large extent, this progress can be attributed to the widespread availability of large mobility datasets stemming from information and communications technology (ICT) and location-based services (LBS) that are now integrated into many aspects of our daily life (Huang et al., 2018a; Keßler and McKenzie, 2018). Aside from the progress on the analysis of human movement itself, the increased availability of tracking data has led to the rapid growth of studies that use human mobility data to study phenomena related to human mobility, such as understanding of residential income segregation (Moro et al., 2021), quantifying urbanization levels and city livability (Bassolas et al., 2019), classifying functional areas of a city (Yuan and Raubal, 2012), urban sensing (Ahas et al., 2015), developing infrastructure for sustainable mobility (Xu et al., 2018) and responding to epidemic spreading (Chang et al., 2021). However, the *raw* digital traces are often not the targeted unit of analysis; for example, a location where people perform an activity can not directly be derived from GPS track points or mobile phone tower data. Studies thus employ various steps to preprocess data into the desired format. These steps and their outcome are often different across studies (Chen et al., 2016a) due to the variety of the datasets and the different understanding of the definitions, which has led to a vast collection of dataset-specific preprocessing and analysis methods. For example, the study by Feng et al. (2018), which proposes the DeepMove model that is now widely accepted as a deep learning baseline model for next location prediction (Luca et al., 2021), generally regards each raw position record as a *location* and does not perform preprocessing. However, focusing on the same problem, Urner et al. (2018) extract *staypoints* (i.e., all the points where a user stayed for at least a certain duration) from GPS track points and further aggregate them into locations using the *k*-means algorithm. Solomon et al. (2021) apply a similar processing concept but introduce the mean shift algorithm to detect staypoints, which are then merged into locations according to a distance threshold. These examples show how not using a standard movement model definition and a common preprocessing standard limit the reproducibility and comparability of the methods and analysis results.

To address these problems, we present Trackintel, an open-source python library for the processing and analysis of movement data. Trackintel is based on an established model for human mobility taken from transport planning, which defines hierarchical levels of movement centred around the concept of *activities*. Trackintel standardizes the definition and

implementation of the data processing steps derived from this data model. Our work thereby makes the assumptions, parameters and filtering steps explicit and provides transparent preprocessing steps, whose implementations are known to have substantial effects on the analysis results. Trackintel further provides analysis, visualization and support functions to enrich the raw tracking data with human-mobility-specific information. Due to the versatility of the data model, Trackintel standardizes preprocessing for many types of tracking data. It thereby greatly simplifies the benchmarking of novel analysis methods, increases their reproducibility, and facilitates quantitative research based on tracking data.

The remainder of the paper is structured as follows. Section 3.1.2 provides an overview of existing libraries for analysing and preprocessing movement data. Section 3.1.3 first introduces the hierarchical model for human mobility analysis and describes its implementation in Trackintel. This section then proceeds to present the most important functionalities of Trackintel to process movement data. In section 3.1.4, we showcase the capabilities of Trackintel to simplify the analysis and comparison of several different tracking datasets. Finally, we summarize and conclude this work in section 3.1.5.

3.1.2 Related work: libraries for movement data

Due to the long history of research in transportation, human migration, and animal behavioral research, a large variety of libraries for (human) movement data processing exists. Joo et al. (2020) survey an impressive number of 58 packages for movement analysis in R. Based on this work and the overview provided by Graser (2020), we selected the libraries that aim at supporting movement analysis in Python, R and C++. In Table 3.1, these selected libraries are compared in terms of their user-friendliness (documentation and robustness), their focus and their provided functionality for human movement data analysis. To compare packages by the quality of their documentation, we evaluate them on a scale from 0-6 based on criteria used for peer-review of packages by pyOpenSci¹ and ROpenSci². See appendix A.1.1 for the list of criteria.

Many of the surveyed R libraries have a strong focus on animal behavioral analysis (Joo et al., 2020) (not all included in Table 3.1). The packages that can (also) be applied to human mobility analysis have a focus on basic statistical analysis of trajectories, such as measuring the spatial extent of animal motion (e.g., `adehabitatLT` (Calenge, 2006)), or the duration and distance of movement trajectories (e.g., `TrackR` (Frick and Kosmidis, 2017)). Currently, no coherent framework is available in R that provides the functionalities specific to human movement analysis, e.g. trip detection and transport mode labeling. Furthermore, there are several libraries available in C++, such as `Tracktable` (Wilson, 2014), `MEOS`³, and `MoveTK`⁴ that promise efficient and fast tools for trajectory data processing, although they may be less accessible for the research audience in human mobility and transportation.

¹<https://www.pyopensci.org/contributing-guide/intro.html>

²<https://ropensci.org/>

³<https://github.com/adonmo/meos>

⁴<https://github.com/movetk/movetk>

Tab. 3.1.: Comparison of movement data libraries. Packages are predominantly available open source in R and Python and they are compared with regard to their focus, documentation and functionality. While other movement analysis libraries already provide well-maintained and documented code with rich functionality for trajectory analysis, only Trackintel provides robust and flexible methods to aggregate trajectories into locations, trips and tours. (✓/✗/x : available / partially available / not available)

Package name	Focus	Programming language Python (P) R (R)	Documentation score	Test coverage (* / **: not reported but low / high)	individual (I) / collective (C)	human (H), animal (A) and/or object(O)	Staypoint detection	Aggregation to location	Aggregation to trips	Aggregation to tours	Tracking quality assessment	Transport mode labelling	Home and work labelling	Visualization	Trajectory statistics (- / + / ++ : none / basic / rich)
Trackintel	Human mobility analysis	P	6	98%	I	H	✓	✓	✓	✓	✓	✓	✓	✓	+
Scikit-mobility (Pappalardo et al., 2022)	Human mobility analysis	P	5	**	I	H	✓	✓	x	x	✗	x	✗	✓	++
Movingpandas (Graser, 2019)	Movement data analysis	P	6	96%	I	H/A/O	✓	x	✗	x	x	x	x	✓	++
PyMove	Querying and visualizing trajectories	P	5	85%	I	H/A/O	✓	x	x	x	x	x	x	✓	+
MovinPy	Mobility data analysis	P	3	0	I	H	x	x	x	x	✗	x	x	x	-
HuMobi (Smolak et al., 2021)	Human mobility prediction	P	3	0	I	H	✓	✓	x	x	✗	x	x	x	+
PTRAIL (Haidri et al., 2021)	Parallelization and feature extraction	P	4	**	I	H/A/O	x	x	x	x	x	x	x	✓	++
TransBigData (Yu and Yuan, 2022)	Transportation	P	5	90%	C	H	✓	x	x	x	✗	x	✓	✓	-
mobilityDB (Zimányi et al., 2020)	Storing and querying	SQL	6	97%	I	H/A	x	x	x	x	x	x	x	✗	+
Traja (Shenk et al., 2021)	Animal trajectories	P	6	76%	I	A	✗	x	x	x	x	x	x	✓	++
Tracktable (Wilson, 2014)	Moving object tracking	P/C++	2	**	I	O	✗	x	x	x	x	x	x	✓	++
MEOS	Spatio-temporal data analysis	C++	4	*	I	H/A/O	x	x	x	x	x	x	x	x	+
MoveTK	Computational movement analysis	C++	-	**	I/C	H/A/O	x	x	x	x	x	x	x	x	++
adehabitatLT (Calenge, 2011)	Animal habitat	R	4	**	I	A	x	x	x	x	x	x	x	x	++
moveVis	Visualization	R	6	93%	I	A	x	x	x	x	x	x	x	✓	-
stplanr (Lovelace and Ellison, 2018)	Sustainable transport planning	R	6	*	C	H	x	x	x	x	x	x	x	x	-
trajectories	Object tracking and interaction	R	5	*	I	O/H/A	x	x	x	x	x	x	x	✓	+
TrackR	Running and cycling data	R	6	52%	I	H	✓	x	x	x	x	x	x	✓	+
ArcGIS Pro	Spatial data	(P)	-	**	I/C	O/H/A	✗	✗	x	x	x	x	✗	✓	++

Furthermore, these libraries provide only highly specific functionalities and do not represent a comprehensive framework for movement data analysis.

ArcGIS Pro is a proprietary software for general spatial data processing with modules for movement data analysis such as speed and acceleration computation, trackpoint clustering and in particular trajectory visualization⁵. However, the different functionalities are scattered over different toolboxes and ArcGIS Pro does not provide a consistent framework for the analysis of movement data. Due to its proprietary nature, we could not evaluate documentation and testing as we did for the other packages, but we assume both are on a high level. We did not include QGIS⁶, a high quality open-source GIS Project, in the table, as there are no well-maintained plug-ins for movement or trajectory data analysis available. However, QGIS could be used in combination with Python libraries or the mobilityDB (Zimányi et al., 2020) library.

In Python, many open-source libraries have emerged as tools to both facilitate and standardize data processing and analysis. The geographic information science (GIScience) community in particular has benefited significantly from Python libraries, for example, the data models implemented in Shapely (Gillies, 2013) and the I/O formats for geographic data as offered in the Fiona package⁷. Most importantly, spatial data can be handled easily with the Geopandas library (Jordahl et al., 2021) that directly builds up on Pandas (The pandas development team, 2020), one of the most established Python libraries for data analysis and manipulation.

In the past years, Python has become the de-facto standard for data science and machine learning applications, which are increasingly important for the analysis of movement data (Luca et al., 2021; Toch et al., 2018). However, only a few libraries have attempted to provide preprocessing and analysis tools specifically for human mobility in a comprehensive Python package (see Table 3.1). Although many algorithms for trajectory data mining were developed in the last decade (Zheng, 2015), their open-source availability in Python is limited, and they often suffer from insufficient documentation and testing standards, such as HuMobi (Smolak et al., 2021) or MovinPy. Others are well-maintained but limited in scopes, such as Traja (Shenk et al., 2021) that targets animal movement, PTRAIL (Haidri et al., 2021) for parallel processing, and TransBigData (Yu and Yuan, 2022) which focuses on data analysis on a collective level, similar to the R library stplanr (Lovelace and Ellison, 2018).

Notable exceptions are MovingPandas (Graser, 2019) and scikit-mobility (Pappalardo et al., 2022). MovingPandas is based on Pandas and Geopandas and focuses on low-level trajectory manipulation, such as splitting, merging and visualizing trajectories. On the contrary, the scikit-mobility library targets high-level analysis functions, including computing human mobility metrics, generating synthetic trajectories and assessing privacy risks. Both libraries are actively maintained and contain various measures to ensure high code quality, but the definition of their data model implies a focus on movement trajectories (MovingPandas)

⁵<https://pro.arcgis.com/en/pro-app/2.8/tool-reference/intelligence/an-overview-of-the-movement-analysis-toolset.htm>

⁶<https://www.qgis.org/en/site/>

⁷<https://github.com/Toblerity/Fiona>

or mobility flows (scikit-mobility), which omits important concepts describing individual human mobility such as activities, trips or tours (Axhausen, 2007).

We aim to close this gap with the Trackintel framework that utilizes an established data model from the transportation literature, which incorporates different semantic aggregation levels of tracking data specific to human mobility.

3.1.3 Trackintel framework

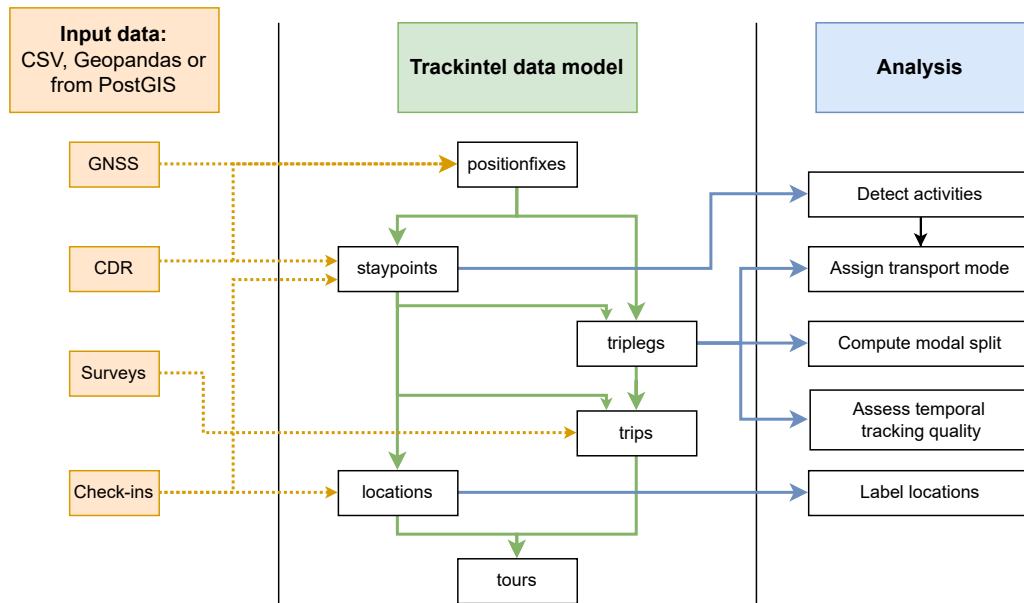


Fig. 3.1.: Overview of the Trackintel framework.

Trackintel⁸ is a library for the analysis of spatio-temporal tracking data with a focus on human mobility. The core of Trackintel is the hierarchical data model for movement data (Axhausen, 2007) that is widely adopted in GIScience (Bucher et al., 2019b), transport planning (Chen et al., 2016a) and related fields (Rout et al., 2021). We provide easy-to-use and efficient functionalities for the full life-cycle of human mobility data analysis, including import and export of tracking data of various types (e.g., GPS track points, location-based social network (LBSN) check-ins, call detail records), data model generation and preprocessing, analysis, and visualization. A conceptual overview of the different components of Trackintel can be found in Figure 3.1.

Trackintel focuses on the mobility of individual persons or objects (e.g., as opposed to crowd flows), and all functionalities are implemented as user-specific, based on unique user identifiers that link data to the respective tracked users. Trackintel is implemented in Python and is built mainly on top of Pandas (McKinney, 2010) and GeoPandas (Jordahl et al., 2021) using accessor classes, a method to extend Pandas classes⁹. This design makes Trackintel

⁸<https://github.com/mie-lab/trackintel>

⁹<https://pandas.pydata.org/docs/development/extending.html>

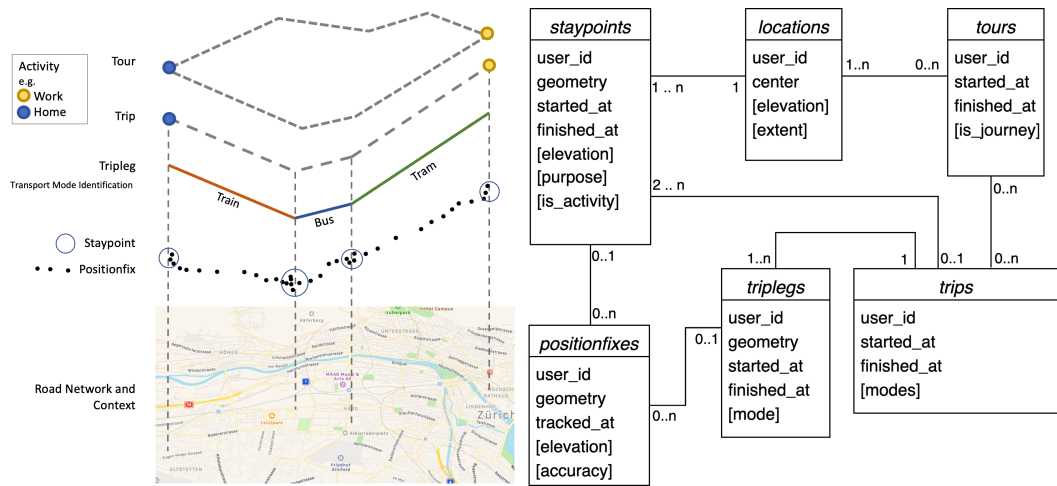


Fig. 3.2.: Semantic visualization of the Trackintel data models and their UML diagram, with mandatory and optional attributes (shown in square brackets). The relations between the different classes are shown in the connecting lines. Figure adapted from (Jonietz and Bucher, 2018)

easy to use for Python users and ensures its broad compatibility with other Python spatial analysis libraries.

The Trackintel data model

The modeling framework employed by Trackintel is based on the activity-based analysis framework in transport planning, which regards travel demand as derived from our need to perform activities at different locations. We follow the definition from (Schönfelder and Axhausen, 2016) that people’s daily mobility consists of staying at locations to perform activities and traveling between locations for the next activity (Schönfelder and Axhausen, 2016). In this definition and following the description in (Axhausen, 2007), movement is separated from activities at different semantic levels. Trackintel implements six classes to represent movement data in this hierarchical model: *positionfix*, *staypoint*, *tripleg*, *trip*, *tour*, and *location*. Figure 3.2 gives an overview of the hierarchical modeling structure and shows the classes in a UML diagram with their mandatory attributes and optional attributes in square brackets. All Trackintel classes are implemented as Pandas Dataframes or Geopandas Geodataframes. In order to be considered a valid Trackintel object, all mandatory attributes have to be present as columns with the correct names, as shown in Figure 3.2. A more detailed explanation of the required and optional attributes of the Trackintel classes is given in Table 3.2. Geometries need to be of the defined type, with the exception of the *Location* class that can have multiple geometries. Furthermore, all timestamps for the time fields required by Trackintel have to be timezone-aware¹⁰. Besides these formal requirements, classes can contain any additional information required for specific analysis. In the following, the different classes and their semantics are introduced.

¹⁰See <https://docs.python.org/3/library/datetime.html#aware-and-naive-objects> for an explanation

Tab. 3.2.: Description of the mandatory and optional columns for Trackintel data models.

Data models	Fields	Description
All	id	The unique identifier for the record
	user_id	The unique user identifier
	tracked_at	The timestamp for the point (only for positionfix)
	started_at	The starting time of the record (except for positionfix and location)
	finished_at	The ending time of the record (except for positionfix and location)
Positionfix Staypoint	geometries	Point geometry
	geometries purpose (optional)	Point geometry Purpose label for the staypoint. This could be either an activity purpose (e.g., home), or a non-activity purpose (e.g., wait).
	is_activity (optional)	Boolean flag indicating whether the staypoint is an activity
Location	center	Point geometry representing the center
	extent (optional)	Polygon geometry representing the extent
Tripleg	geometries	Line geometry
	mode (optional)	Transport mode label
Trip	origin_staypoint_id	The identifier of the starting staypoint
	destination_staypoint_id	The identifier of the destination staypoint
	primary_mode (optional)	The main transport mode label
Tour	location_id	The start and end location identifier
	journey	Boolean flag indicating whether the tour is a journey (A tour is called a journey if the start and end location is home).

Positionfix. *Positionfix* is the smallest tracking unit in the Trackintel data model, consisting of timestamped position records, for example, generated by GNSS trackers or call detailed records (CDR) data. Positionfixes are often directly transferred from raw tracking data and are thus a natural entry point to the Trackintel data model, where it can further be processed and segmented into triplegs and staypoints. No inherent semantics are included since movements and activities cannot be distinguished from Positionfix.

Staypoint. *Staypoint* represents a point in space, which is defined as an individual remaining within a defined geographical radius for a defined time. Compared to the raw *positionfix* points, staypoints can represent stationary points that carry particular semantics, such as the purpose of the stay, or they can represent an intermediate stay, such as waiting for a bus. To distinguish between these two types of staypoints, we introduce the concept of *activity*: an activity staypoint is usually the reason for a person to travel and has an important purpose with an attached activity label (e.g., home), while a non-activity staypoint only represents a trivial stationary point (e.g., waiting). The exact definition of an activity depends on the goal of the study. In Trackintel, activities are staypoints with the attribute *activity_flag* set to *True*, which can be obtained through user labels or directly inferred from data (see section 3.1.3). While activity *staypoints* are the basic unit for constructing trips, which mark the start and end of a *trip*, non-activity *staypoints* can only be part of a *trip*. Additionally, *staypoints* can be spatially aggregated to form *locations*.

Tripleg. The most basic level of movement is defined as *tripleg* (referred to as stage in Axhausen (2007)), which formally represents a continuous movement without changing

transport mode or vehicle. Therefore, triplegs contain semantics about the movement of an individual, such as the mode of transport that is stored in the attribute field *mode* if available. This information can be obtained from user labels (Hong et al., 2021; Zheng et al., 2010) or inferred using heuristics directly from the data, which is implemented as labeling functions in Trackintel (see Section 3.1.3). Triplegs can be created from positionfixes and can be aggregated to form trips.

Trip. *Trip* represents all travels between two activities and summarizes all triplegs and non-activity staypoints between two consecutive activity staypoints. Trips inherit the activity purpose from the activity label attribute of the destination staypoint. As they are often the primary quantity of interest in transport planning studies, trips, together with activities, are the core of the movement data model proposed in Axhausen (2007).

Location. Activity staypoints represent individual visits to places that are significant to the visitor. Trackintel models these significant places using the *location* class to enable the characterization of the place that is visited. While the information attached to staypoints is bound to the individual visit (e.g., the specific activity or the time of day), the semantics of locations are related to the place independent of the visit (e.g., land use or the opening hours of a shop). Locations are modeled with two different geometries, a point geometry for the center of the location and a polygon geometry to describe the extent of a location.

Tour. The mobility of individuals is centered around a few significant locations that act as the basis of their travel behavior. Individuals conduct several activities and trips if convenient but return home (or to a similar significant location) to plan their next activity. This behavior can be analyzed using the *tours* class, which is defined as “a sequence of trips starting and ending at the same location” (Axhausen, 2007, p. 4), referring to the location class defined above. A special case of a tour is the concept of *journey* that starts and ends at the home location of an individual. In Trackintel, a tour can be flagged as a journey using the *journey* attribute. A tour contains multiple trips, but one trip can also be part of several tours in case they are nested, e.g. the trip from the work location to the supermarket and back is part of a larger journey that started at home.

Data model generation

The core functionality of Trackintel is to generate all classes defined in the movement data model from the raw tracking data. In practice, this refers to the generation of the entire hierarchical movement data model from positionfix data. However, it should be noted that it is not required and often not practical to enter the framework from positionfixes - the framework can be accessed at any semantic level depending on the available data (e.g., location-based social network (LSBN) check-ins represent staypoints without the availability of positionfixes; see Figure 3.1 for examples of input levels for different tracking data types). The following section presents the implemented preprocessing steps necessary to aggregate data through the hierarchy levels. The output of all *generate* functions is a (Geo)DataFrame with the fields listed in Table 3.2.

Generate staypoints. In Trackintel, *staypoints* are generated from *positionfixes* based on the sliding window detection algorithm first reported in (Li et al., 2008), which has become a

standard algorithm for staypoint detection (Zheng, 2015). For each individual, the algorithm iterates over all positionfixes and determines groups of points that satisfy the predefined distance and time thresholds. Each output staypoint inherits the starting and ending time from the first and last positionfix that belongs to it, respectively, as well as the mean geometry coordinates of the group of positionfixes. The implemented staypoint detection algorithm extends the algorithm from (Li et al., 2008) by an option to exclude temporal gaps in the tracking data, commonly observed in many datasets due to low temporal tracking coverage. This behavior can be controlled using a parameter representing the maximum time between two consecutive positionfixes such that they are still considered to belong to the same staypoint.

Generate locations. Locations can be generated by aggregating staypoints. Existing studies proposed community detection algorithm (Aslak and Alessandretti, 2020) and spatial clustering algorithms, such as OPTICS (Yuan et al., 2013), mean shift (Solomon et al., 2021), and DBSCAN (Luo et al., 2017) to perform this processing step. Here, we implement the most commonly employed DBSCAN algorithm to aggregate staypoints that are spatially close to locations (Hariharan and Toyama, 2004; Jonietz and Bucher, 2018). DBSCAN adopts a set of neighborhood characterization parameters ϵ and the minimum number of samples (*min_samples*) to define how dense the input data has to be considered a cluster. In the context of location generation, ϵ controls the distance of which nearby staypoints will be merged into a single location, and *min_samples* determines the minimum number of staypoints to form a location (i.e., how many visits are required at the same place to consider it as significant). Generated locations are equipped with two different geometries. The *center* is a point geometry, calculated as the mean coordinates from all staypoints assigned to the cluster; the *extent* is a polygon geometry, defined as the bounding box of all belonging staypoints. Furthermore, we provide the flexibility to generate locations that are significant to a single user (Figure 3.3 right) or to all users present in a dataset (Figure 3.3 left). While user locations regard staypoints of each tracked user separately in the clustering process and prevent generating locations that are excessively large (Aslak and Alessandretti, 2020), dataset locations consider all staypoints at the same time and output locations with shared semantics across users (e.g., train stations or shopping malls). In both options, the center and the extent of the clustered staypoints are attached to the generated locations, providing geometry information that facilitates further processing and analysis tasks.

Generate triplegs. Trackintel implements an algorithm that extracts triplegs from positionfixes based on the assumption that an individual is moving if he or she is not stationary, meaning that all positionfixes that do not belong to any staypoint are assigned to a tripleg. This assignment process requires the input of positionfixes with the identifier of the already generated staypoints. Internally, the function aggregates all positionfixes between two consecutive staypoints to form a tripleg, whose line geometry is constructed by connecting the point geometries in chronological order. Similar to the generation process of staypoints, the start and end timestamp of each tripleg are inherited from the first and last positionfixes that belong to it, respectively.

Generate trips. Trackintel implements a method to generate trips based on existing staypoints and triplegs. Trips summarize all movement and all non-activity staypoints (e.g.

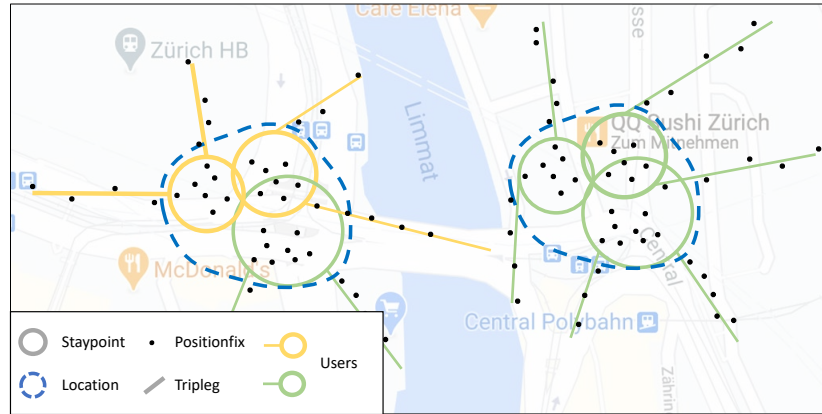


Fig. 3.3.: Semantic visualization of the relations between positionfix, staypoint and locations. Staypoints are groups of positionfixes where the users are stationary, and locations are aggregations of staypoints that the user visits multiple times. Locations can be generated across users (left) or for each user individually (right). Map data ©2022 Google.

depending on the data, this could correspond to waiting at a bus stop) between two staypoints flagged as activity. This seems trivial at first sight however, there are no easy-to-use implementations available in other libraries, and there are several special cases related to gaps in the tracking data that should be considered during the trip generation. Another important feature of the implemented trip generation is the identifier management that connects trips with their associated staypoints and triplegs.

The trip detection implemented in Trackintel can handle incomplete tracking data and supports the detection of temporal gaps. A temporal gap is defined as missing tracking signals longer than a certain time period (Zhao et al., 2021), which can be specified using the θ_{trip_gap} input parameter to the function. If a temporal gap greater than θ_{trip_gap} is detected, we assume the individual performed an unobserved activity and, therefore, the destination of the current and the origin of the next trip is unknown (NaN in the resulting table). Finally, the function provides the flexibility to specify whether the trips table should include the geometry. The geometry of a trip consists of the points for the origin and destination staypoints. If the origin is unknown, we use the first point of the first tripleg instead, or analogously the last point for the destination.

Generate tours. To the best of our knowledge, there is no standardized approach yet on how to combine trips into tours. Here, we take a rather broad definition of tours that includes nested tours as described in (Axhausen, 2007), leaving the user the choice to filter the outputs later. An example of a nested tour is shown in Figure 3.4: the tour Work-Cafe-Work is part of the longer tour Home-Work-Cafe-Work-Home. This definition implies an n-to-n relationship between trips and tours: One tour contains multiple trips, and one trip can be part of multiple tours.

Our algorithm to generate tours from trips is explained visually in Figure 3.4, and shown as pseudocode in Algorithm 2. We iterate over the trips sorted chronologically and maintain a list C of tour-starting candidates. Each trip ϕ_i is a potential candidate to start a tour. At each

Algorithm 1 Trip generation

```
1: Given
2:    $s$ : staypoints of a user
3:    $t$ : triplegs of a user
4:    $\theta_{trip\_gap}$ : Gap threshold to start new trip
5: procedure GENERATE TRIPS( $s, t, \theta_{trip\_gap}$ )
6:    $st \leftarrow$  merge  $s, t$  and sort by timestamp
7:    $n = \text{length}(st)$ 
8:    $i = 0$ 
9:    $\Phi = \{\}$  ▷ Initialize trips as empty set
10:  inTrip = False
11:  while  $i < n$  do ▷ iterate all elements of  $st$ 
12:    if (inTrip is False)  $\wedge$  ( $st[i]$  is no activity) then ▷ go to next activity
13:       $i++$ 
14:      continue with next iteration
15:    end if
16:    if (inTrip is False)  $\wedge$  ( $st[i]$  is activity)  $\wedge$  ( $i + 1 < n$ )  $\wedge$  ( $st[i + 1]$  is activity) then
17:       $i++$  ▷ Skip sequential activities w/o travel
18:      continue with next iteration
19:    end if
20:    if InTrip is False then
21:      inTrip = True ▷ Start trip
22:      initialize new trip  $\phi_{current}$ 
23:       $\phi_{current}.origin = st[i]$  ▷ set  $st[i]$  as origin of trip; add start time, activity label and
24:      geometry
25:       $i++$ 
26:      continue with next iteration
27:    end if
28:     $\delta_t = \text{startTime}(st[i]) - \text{endTime}(st[i - 1])$  ▷ compute gap duration
29:
30:    if (InTrip is True)  $\wedge$  ( $\delta_t \geq \theta_{trip\_gap}$ ) then ▷ check for gap in tracking data
31:       $\phi_{current}.destination = \text{unknown}$  ▷ set trip destination as unknown
32:       $\Phi = \Phi \cup \phi_{current}$  ▷ Add trip to collection of trips
33:      inTrip = False
34:    else if (InTrip is True)  $\wedge$  ( $st$  is no activity) then
35:      add  $st[i]$  to trip  $\phi_{current}$ 
36:    else if (InTrip is True)  $\wedge$  ( $st$  is activity) then
37:       $\phi_{current}.destination = st[i]$  ▷ set  $st[i]$  as trip destination
38:       $\Phi = \Phi \cup \phi_{current}$  ▷ Add trip to collection of trips
39:      inTrip = False
40:    end if
41:     $i++$ 
42:  end while
43: end procedure
```

Algorithm 2 Tour generation

```
1: Given
2:    $\Phi : \{\phi_0, \dots, \phi_n\}$ : trips (sort by timestamp)
3:   loc: (Optional) location function that returns the location ID given trip origin or destination
4:   start: Function that returns the geometry of the start point of a trip
5:   end: Function that returns the geometry of the end point of a trip
6:    $\theta_{max\_dist}$ : Maximum distance (in m) between the end of one trip and the start of the next trip
   on the same tour
7:    $\theta_{max\_time}$ : Maximum duration of a tour
8:    $\theta_{max\_gaps}$ : Maximum gaps that are allowed on a tour
9: procedure GENERATE TOURS( $\Phi, loc, \theta_{max\_dist}, \theta_{max\_time}, \theta_{max\_gaps}$ )
10:   $C = [\phi_0]$ : List of candidate trips to form a tour
11:   $i = 1$ 
12:  locAvail = True if loc is given, otherwise False ▷ Are locations provided or not?
13:  while  $i < n$  do
14:    ▷ Part 1: Check if the previous and the current trip are connected
15:    gapInbetween = False
16:    if locAvail then ▷ Option 1: Compare locations
17:      if  $loc(end(\phi_{i-1})) \neq loc(start(\phi_i))$  then
18:        gapInbetween = True
19:      end if
20:    else ▷ Option 2: Check distance
21:      if  $distance(start(\phi_{i-1}), end(\phi_i)) > \theta_{max\_dist}$  then
22:        gapInbetween = True
23:      end if
24:    end if
25:    if gapInbetween then
26:       $C = C.append(gap)$  ▷ Record a gap in the tour candidates
27:    end if
28:     $C.append(\phi_i)$  ▷ Add trip to tour candidate
29:    ▷ Part 2: Check if the current trip closes a tour
30:    for  $\phi_C \in C$  (iterate in reverse order) do
31:      closesTour = False ▷ Check if  $\phi_C$  is the start of a tour ending at  $\phi_i$ 
32:      if locAvail then ▷ Option 1: Compare locations
33:        if  $loc(end(\phi_i)) == loc(start(\phi_C))$  then
34:          closesTour = True
35:        end if
36:      else ▷ Option 2: Check distance
37:        if  $distance(end(\phi_i), start(\phi_C)) \leq \theta_{max\_dist}$  then
38:          closesTour = True
39:        end if
40:      end if
41:      if closesTour then
42:         $n\_gaps = \text{count gaps between } \phi_C \text{ and } \phi_i$ 
43:        if  $(n\_gaps < \theta_{max\_gaps}) \wedge (endTime(\phi_i) - startTime(\phi_C) < \theta_{max\_time})$  then ▷
44:          Tour found!
45:          Aggregate all trips from  $\phi_C$  to  $\phi_i$  into a tour
46:          end if
47:        end if
48:      end for
49:      Remove trips from  $C$  that are more than  $\theta_{max\_time}$  before  $\phi_i$ 
50:       $i++$ 
51:    end while
52: end procedure
```

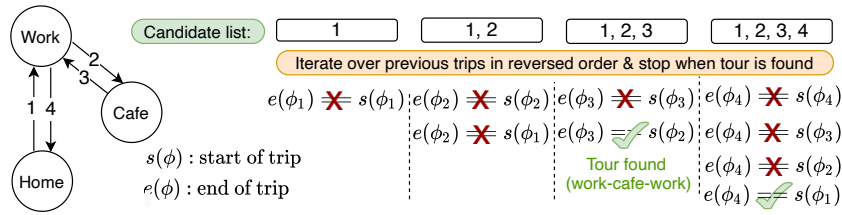


Fig. 3.4.: The algorithm of tour generation implemented in Trackintel. A list of start candidates is maintained and iteratively checked for tour-closing trips.

iteration, that is, for each trip, we first check whether there is a spatial gap between the current and the previous trip ϕ_{i-1} . Two options are implemented: If the table *staypoints* with the attribute *location_id* is provided, we compare the location identifier of the end of ϕ_{i-1} to the one of the start of ϕ_i , formally $loc(end(\phi_{i-1})) = loc(start(\phi_i))$. Alternatively, if the staypoints are not available, the predefined spatial distance threshold θ_{max_dist} controls the maximum distance between the end and start points, i.e. $distance(start(\phi_i), end(\phi_{i-1})) \leq \theta_{max_dist}$.

Additionally, our implementation offers the possibility to generate partially observed tours to accommodate tracking datasets with a low temporal tracking coverage, e.g., mobile phone data-based studies. A parameter θ_{max_gaps} determines how many spatial gaps are allowed within a single tour. Note that no gaps are allowed at the start or end of a tour, because a tour must start and end at the same location, or the start- and end-staypoints must lie within the permitted range. If the test described above yields a spatial gap between ϕ_{i-1} and ϕ_i , and $\theta_{max_gaps} = 0$, the candidate list is reset to $[\phi_i]$. Otherwise, a gap is registered.

Next, we test whether ϕ_i concludes a tour (Algorithm 2 line 15). For this purpose, we iterate over all candidates in the reversed order, such that the shortest possible tour is found first. We compare the start point of a candidate ϕ_C to the end point of ϕ_i . Again, the points are compared either by the location identifier or via the θ_{max_dist} parameter. If they are the same, the trips $\{\phi_k \mid j \leq k \leq i\}$ form a tour, subject to two further conditions: A. While iterating over candidates, the encountered gaps are counted, and the time duration is checked. The parameter θ_{max_time} is used to certify whether the tour takes place within an appropriate time period, by default 24 hours. B. When encountering more than θ_{max_gaps} in the reversed iteration, or when reaching a candidate that started more than θ_{max_time} hours ago, the loop ends, and no tour is found. Figure 3.4 shows an example where two tours are found after considering ϕ_3 and ϕ_4 respectively.

Import and export

Reading and writing data are important steps in a standard movement data analysis pipeline. To simplify this process, Trackintel provides an I/O module for accessing movement data and storing intermediate or final results in a file or database. Three methods for converting movement data with attached attribute information to Trackintel-compatible formats are provided: 1) Reading from Pandas Dataframes and Geopandas Geodataframes, 2) reading and writing from CSV file formats, and 3) reading and storing from PostgreSQL databases

with PostGIS extension. For every Trackintel data type, we provide I/O functions that internally check the validity of the input data formats. Also, Trackintel implements reading functions to convert tracking data from publicly available open-source datasets into the Trackintel data model. For example, raw tracking records from the Geolife dataset (Zheng et al., 2010) can be loaded with Trackintel into the positionfix format. In addition, we provide helper functions to attach transport mode labels, which are provided separately for some individuals in the Geolife dataset. The dataset reading functions facilitate and standardize the processing of public movement datasets using Trackintel, which also help to benchmark new methods on the same dataset.

Pre- and postprocessing

Trackintel offers several pre- and postprocessing methods. First, to smoothen the trajectory of triplegs, we employ the Douglas-Peucker algorithm (Douglas and Peucker, 1973). Furthermore, staypoints that appear consecutively at the same location can be aggregated in time. Such repetitions are a common artefact in tracking data due to noise or outliers recordings in GNSS tracking data. We propose to merge two staypoints s_1, s_2 of one individual if the following conditions hold: a) s_1 and s_2 are consecutive in time, b) s_1 and s_2 are assigned to the same location, c) there is no tripleg registered between s_1 and s_2 , and d) the time gap between the end time point of s_1 and the start of s_2 is shorter than a predefined threshold $\theta_{max_time_gap}$. The start time of s_1 and the end time of s_2 define the start and end time of the new staypoint. The aggregation of other staypoint attributes, e.g. the geometry, must be specified explicitly.

Analysis

While the main functionality of Trackintel is the implementation of the hierarchical data model, the framework also includes advanced analysis functions to label transport modes and activity purposes, as well as methods to assess the tracking quality of each individual.

Mode labeling

Applications in transport planning often require access to the travel modes of an individual (Kim et al., 2022). Since Trackintel does not assume the availability of user-provided labels, context or advanced data from the tracking device (e.g., accelerometer), we implement a simple heuristic to determine the travel mode from the tracking data. This classification is done per *tripleg* based on speed. The speed is approximated by the tripleg length (the distance of individual points in its LineString geometry) divided by its total time duration. The triplegs are labeled based on a simple division into slow mobility (<15km/h average speed), motorized mobility (<100km/h) and fast mobility (>100km/h). In future versions, a more in-depth analysis of travel patterns or map matching (Bachir et al., 2018; Huang et al., 2019; Prelipcean et al., 2017; Widhalm et al., 2012) could be incorporated into Trackintel.

Location labeling

An individual's home- and work-locations play a major role in mobility data analysis. As described in Section 3.1.3, staypoints may be associated with an activity label, but oftentimes this information is not available. We assign “home” and “work” activity labels to the staypoints with an adapted version of the OSNA algorithm proposed by Efstathiades et al. (2015). In detail, the OSNA algorithm divides weekdays into rest, work and leisure time frames. The location with the longest accumulated duration in the “rest” and “leisure” periods is labeled as home, while work is set to the most predominant location in the “work” periods. While the original algorithm derives the hours spent at a location from geo-tagged tweets, we take advantage of the *started_at* and *finished_at* attributes of a staypoint. Additionally, similar to in the *R* package proposed by Chen and Poorthuis (2021), we provide a fast method that simply assigns home and work labels to the two locations that are visited more often in the data (in this order). In both cases, the locations can optionally be pre-filtered in order to exclude locations with an insufficient number of staypoints or an insufficient length of stay.

Modal split

If mode labels for *triplefs* are available, Trackintel supports the calculation of the modal split in three different ways: Computing the modal split by count (i.e., how many triplefs with this mode exist), by duration (i.e., sum of individual's triplef duration) or by travelled distance. Furthermore, the frequency can be set according to the Pandas time series frequency syntax¹¹, and the modes can either be aggregated by user or by dataset. An example for one user is visualized in Figure 3.6 where the differences between a modal split by count (Figure 3.6(a)) and by distance (Figure 3.6(b)) stand out.

Tracking quality assessment

An important step in data analysis of tracking studies is the assessment of the tracking quality, i.e. the temporal coverage. Temporal tracking quality, here defined as the proportion of time where the user's whereabouts are recorded, is regarded as a basic measure of the temporal resolution of the dataset (Alessandretti et al., 2018). Trackintel supports the calculation of the daily, weekly or overall tracking quality of each user according to the required granularity levels, which enables individual-level temporal resolution assessment, providing support for filtering low-quality users for further analysis. Additionally, tracking quality of hours of the day and weekdays can be obtained for measuring the tracking data quality differences across time periods.

Visualization

Trackintel provides a module that supports the visualization of *positionfixes*, *staypoints* and *triplefs*. Our implementation standardizes these functions such that each data type can be displayed together with lower aggregation levels (see Figure 3.1). For example, the

¹¹https://pandas.pydata.org/pandas-docs/stable/user_guide/timeseries.html

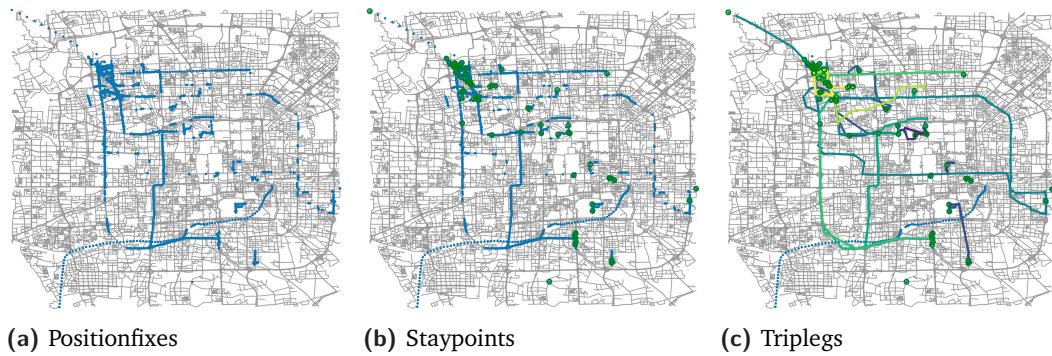


Fig. 3.5.: The Trackintel framework offers functions to plot positionfixes (a), staypoints (b), and triplegs (c) together with the road network acquired from OpenStreetMaps. This example maps the movements of one Geolife participant.

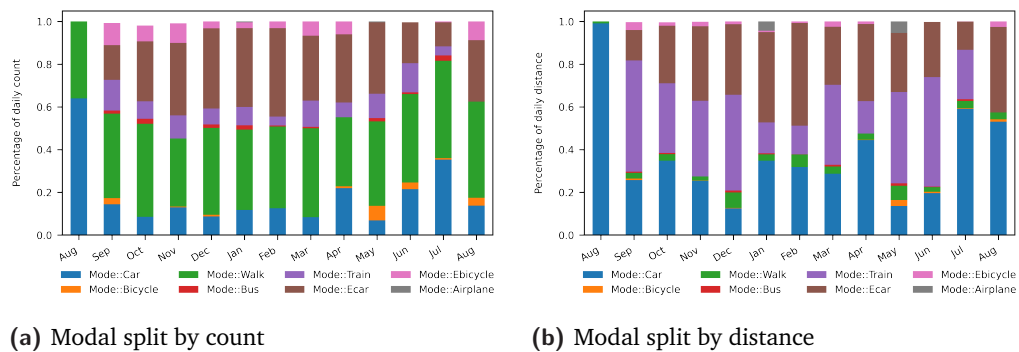


Fig. 3.6.: The visualization result of the Trackintel `plot_modal_split()` function of the triplegs recorded from one Geolife participant. Major differences can be observed between the aggregation by count (number of triplegs) (a) and distance traveled (b).

locations can be optionally shown together with *positionfixes* and *staypoints*. In that case, *staypoints* and *locations* are displayed as circles with a predefined radius. Furthermore, Trackintel integrates `osmnx` (Boeing, 2017) to optionally show the street network from Open Street Maps as background. Figure 3.5 shows example outputs of the plotting functions for *positionfixes*, *staypoints* and *triplegs* for one exemplary participant in the Geolife study.

Finally, Trackintel provides a flexible method to visualize changes in the modal split over time. The modal split by count, distance or duration, as explained in Section 3.1.3, is shown in a bar plot with one bar for each temporal bin. Different temporal resolutions (i.e., weeks and months) are handled internally. An example for one user is shown in Figure 3.6 where the modal split has been aggregated by month.

3.1.4 A case study on multiple tracking datasets

Trackintel is a framework to standardize mobility data processing and analysis. We carried out a case study on four datasets to demonstrate its capability to handle data from various tracking studies. We read all data from a PostGIS database with the I/O module, preprocess them according to the Trackintel movement data model and compare the datasets in terms of tracking quality, trip characteristics, and modal split. The code of the case study is available in the supplemental material and the public repository¹².

Tracking studies

We include the data from four tracking studies with two different tracking data types. An overview of the dataset properties is given in Table 3.3. The first study is the open-source Geolife dataset (Zheng et al., 2009) that covers the movement of employees of Microsoft Research Asia, who recorded their movement using GPS trackers. Second, we include two studies that were conducted in collaboration with the Swiss Federal Railway Systems (SBB) under the project name *SBB Green Class* (Martin et al., 2019a). In both studies, participants were given full access to all public transport in Switzerland. In addition, the participants from the first Green Class study (Green Class 1) received an electric vehicle and those from the second study (Green Class 2) an e-bike. Study participants were tracked with a GNSS-based application (app) called *Myway*¹³. The app already provides the data partially preprocessed as staypoints and triplegs. The same app was further used in our fourth dataset, the yumuv study¹⁴ which investigated the impact of a Mobility-as-a-Service app that integrates shared e-scooters, e-bikes and public transport (Martin et al., 2021b). In the yumuv study, participants were divided into control and treatment groups and were tracked for three months.

Tab. 3.3.: Overview of basic features of the considered tracking studies. Locations, staypoints, triplegs, trips and tours are given in multiples of a thousand.

	Users	Tracking period in days (std)	Input	Study type	Locations	Staypoints	Triplegs	Trips	Tours
Green Class 1	139	401 (59)	Staypoints, Triplegs	GNSS (app)	104.5	326.9	465.2	241.8	95.0
Green Class 2	50	314 (76)	Staypoints, Triplegs	GNSS (app)	35.7	87.9	128.6	61.4	22.7
Yumuv	806	87 (38)	Staypoints, Triplegs	GNSS (app)	127.3	326.3	502.3	199.7	83.0
Geolife	177	193 (443)	Positionfixes	GPS tracker	13.6	28.9	30.2	30.2	7.2

Standardized processing according to the Trackintel data model

The Trackintel framework offers a straightforward way to transform all data into the same format and aggregate the data into trips and tours with minimal code. First, the raw GPS data in the Geolife dataset are converted to staypoints and triplegs with the Trackintel

¹²https://github.com/mie-lab/trackintel/blob/master/examples/Trackintel_case_study.pdf

¹³<https://www.sbb.ch/en/timetable/mobile-apps/myway.html>

¹⁴<https://yumuv.ch/en>

`generate_staypoints()` and `generate_triplegs()` functions. Staypoints are created with a distance threshold of 100m, i.e. a user must have traveled 100 meters to generate a new staypoint, and a temporal threshold of 30 minutes, as suggested in the original paper (Li et al., 2008). Furthermore, consecutive positionfixes with a temporal gap of more than 24 hours in between cannot belong to the same staypoint.

All further preprocessing steps based on staypoints and triplegs are applied with *the same* parameters for all four datasets. This ensures the comparability of the results across datasets. More specifically, we derive the user's locations from the staypoints with the `generate_locations()` function. The method uses the DBSCAN algorithm with $\epsilon = 30$ meters and $min_samples = 1$, such that one staypoint is sufficient to form a location. Furthermore, triplegs and staypoints are aggregated to trips with the `generate_trips()` function, with input parameter $\theta_{trip_gap} = 25$ minutes. At last, tours are generated by merging trips based on a maximum distance (θ_{max_dist}) of 100m between their start and end points, and with the default parameters $\theta_{max_gaps} = 0$ and $\theta_{max_time} = 24$ hours.

Table 3.3 provides the absolute numbers of locations, staypoints, triplegs, trips and tours per dataset. These quantities decrease from triplegs to trips and tours due to the aggregation steps. Note that for Geolife our parameter choices prevent triplegs from being merged (see Table 3.3 where the number of triplegs and trips are the same); however, parameters that are more suitable for the trip generation would have decreased the quality of other parts significantly due to the low tracking quality of Geolife. In total, the considered datasets include 769,957 staypoints and 1,123,931 triplegs. These quantities depend on the number of participants in the study and the total tracking duration. While the yumuv study has the largest sample size of 806 users, the Green Class 1 study participants have the longest tracking period, with each individual tracked for more than a year on average.

Analysis and comparison of tracking datasets

We now compare the mobility behavior of the study participants of all studies on the trip level as an exemplary usage of the Trackintel *analysis* module. The insights from this analysis are summarized in Table 3.4. First, we can derive the number of daily trips per individual from the absolute numbers given above. The study participants in Green Class 1 and Green Class 2 are most active in conducting trips. The low number of trips for Geolife users may be due to the low temporal tracking coverage of the dataset. Furthermore, we compare the average trip distances and duration across datasets. Interestingly, yumuv and Geolife users take longer trips on average in terms of duration. There is also a clear effect of the bias of yumuv participants towards urban areas, where the trips cover much shorter distances. The number of trips per tour and the number of triplegs that are part of the same trip do not differ much between studies.

Another key part of tracking data analysis regards the temporal tracking quality of a dataset. Here, temporal tracking quality is defined as the temporal coverage of the tracking data (i.e., the completeness) and is computed with the Trackintel function `temporal_tracking_quality()` as explained in section 3.1.3. The results are given in the last column of Table 3.4. The three GNSS-based studies show a high coverage of more

Tab. 3.4.: Overview of the mobility statistics for the considered tracking datasets.

	Trips per day	Trips per tour	Legs per trip	Trip distance in km (std)	Trip duration (std)	Tracking quality (std)
Green Class 1	4.32	2.73	1.92	27.4 (478.7)	0.52 (0.73)	0.85 (0.17)
Green Class 2	3.80	2.66	2.09	33.7 (568.2)	0.51 (0.75)	0.75 (0.24)
Yumuv	3.13	2.11	2.51	16.9 (100.4)	0.68 (0.91)	0.77 (0.23)
Geolife	1.70	2.37	1.00	36.1 (3163.5)	0.64 (0.94)	0.4 (0.32)

than 75% on average per user, whereas Geolife data only covers about 40% of the time on average per user. Figure 3.7 shows the distribution of the tracking quality over users. In the Geolife dataset, the temporal tracking quality largely differs across individuals. In comparison, the large majority of Green Class 1 participants reached a coverage of more than 0.7. The large difference between Geolife and the other datasets can be explained by the different hardware that was used in the studies. While the Geolife individuals were equipped with dedicated GPS-only trackers that are prone to localization problems when indoors or in urban canyons, the participants in the Green Class and yumuv studies were tracked with an app on their smartphone that uses the location API of the operating system. The latter has access to all GNSS systems in addition to GPS and can fall back to other technologies such as WIFI or cell tower triangulation if no satellite is available.

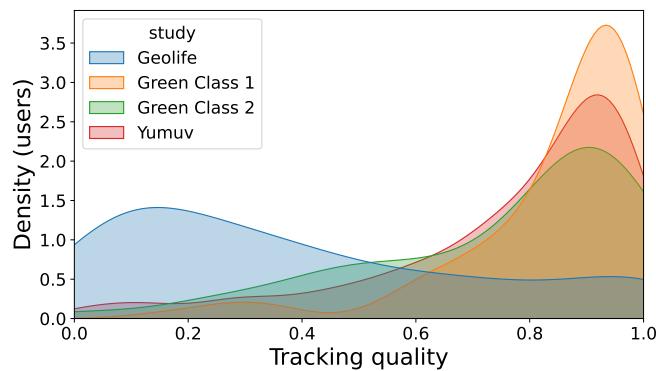


Fig. 3.7.: Distribution of the individual temporal tracking quality for the considered datasets.

We further compare the modal split of the tracking studies. The split is computed first as the number of triplegs per mode and second as the covered distance per mode. We use the Trackintel function `predict_transport_mode()` to approximate the modes for the Geolife dataset, since the original mode labels are not available for all participants and not all the time. In all other studies, high-quality mode labels are provided, and we aggregate them into the simplified categories of slow mobility (walk, bicycle, scooter), motorized mobility (tram, bus, car and motorbike) and fast mobility (airplane and train). The results are shown in Figure 3.8. The datasets differ significantly with respect to their modal split, which can be explained by the study target group, for example, Green Class participants were given full access to all public transport in Switzerland and are thus more likely to use trains (fast transport). Yumuv individuals on the other hand mostly live in urban areas and they

were using the yumuv bundle of shared e- bicycles and scooters, which explains the higher proportion of slow mobility for yumuv.

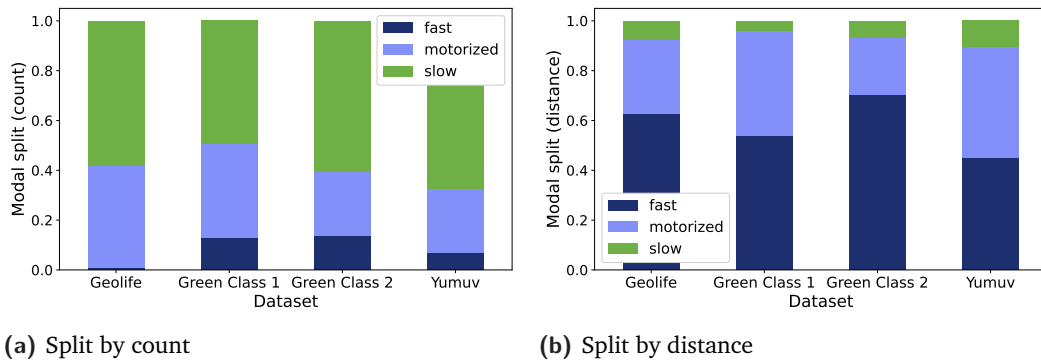


Fig. 3.8.: Comparison of modal split between datasets. The users of different studies differ considerably in terms of their usage of slow, motorized or fast transport.

Finally, we analyze the daily activity patterns of individuals. Specifically, the time periods when the individuals are at home and at work are computed. For the Green Class 1 & 2 studies, the activity label for each staypoint is provided by the participants. For the Geolife and yumuv datasets, on the other hand, we adopt the Trackintel `location_identifier()` function that implements the OSNA algorithm (Efstathiades et al., 2015) to infer the home and work locations. In Figure 3.9, the distribution of home and work staypoints over the course of a day is shown. Specifically, the average fraction of users with a staypoint labeled home (or work respectively) is shown for every minute of the day. The fraction of users at home (work) is thereby computed as the number of staypoints per day divided by the number of actively tracked users, where a user is actively tracked if there is at least one staypoint on that day. The working time between 8am and 5pm as well as the lunch breaks are clearly visible in Figure 3.9(b) for Green Class 1 & 2 and yumuv, although there are fewer work-staypoints for yumuv. While the home location is reliably identified for both yumuv and Geolife, the identification of the work location seems impaired for the Geolife dataset. As the OSNA algorithm simply selects the second-most visited location as work if the “home” and “work” labels overlap, the low tracking quality of the Geolife dataset (see Figure 3.7) could have affected the accuracy of the identification.

In summary, our study demonstrates the ease of comparing data from different sources on all levels of the movement data model and concerning various labels for the movement data. The standardized preprocessing functions implemented in Trackintel also help compare methods and explain possible discrepancies in the analysis results from the different datasets.

3.1.5 Discussion and conclusion

Quantitative analysis of human mobility currently suffers from a lack of a common model for preprocessing movement data, limiting the reproducibility and comparability of scientific studies. Existing libraries focus on data analysis, leaving seemingly easy preprocessing steps up to the user, although design choices of these steps can significantly affect the

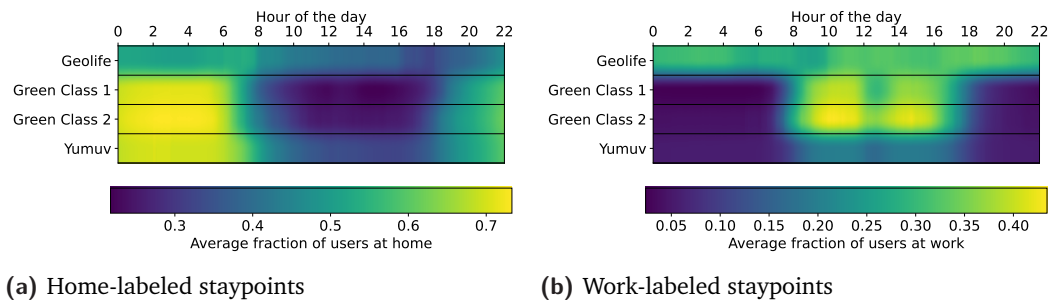


Fig. 3.9.: Distribution of activities over time.

results (Sambasivan et al., 2021). This article presented Trackintel, a new open-source tool to address these problems. Trackintel implements a widely accepted conceptual data model for movement data and provides functionalities for the full life-cycle of human mobility data analysis: import and export of tracking data collected through various methods, preprocessing, data quality assessment, semantic enrichment, quantitative analysis and mining tasks, and visualization of data and results.

A particular strength of Trackintel is that it greatly simplifies the joint analysis of several movement datasets with different properties. This was shown in a case study where four different datasets were jointly preprocessed and analyzed. We used the analysis methods implemented in Trackintel to compare the datasets with respect to their trip properties, their tracking quality, their modal split and their daily activity patterns. It was demonstrated in the supplementary material that rich insights about the characteristics of different tracking datasets could be easily obtained in Trackintel with few lines of code. Our library is thus also a response to recent calls in GIS for systematic benchmarking of new methods on several datasets (Konkol et al., 2019).

Importantly, the purpose of Trackintel is not to provide a comprehensive set of analysis functions, but rather a high-quality implementation of standard aggregation and semantics-enrichment steps that are relevant for most tracking studies. This goal is fulfilled in the current version of the library since functions for all aggregation steps in the data movement model are provided and were tested extensively on diverse datasets. Further work on the preprocessing module will focus on improvements, such as outlier filtering functions or methods to fill small gaps in the tracking data.

We plan to extend the analysis functionality of Trackintel and improve the integration with other open-source libraries. Currently, the goal of compatibility with arbitrary tracking datasets limits the capabilities of the analysis model. A good example is the transport mode prediction function provided by Trackintel, which is based on a simple heuristic. A more sophisticated and powerful method can in principle be implemented for a specific dataset, however, the applicability of this method to other datasets will be limited by the availability of specific input data or additional context data. Nevertheless, Trackintel will be continuously extended to incorporate the latest processing and analysis algorithms and to offer a wider variety of options for the preprocessing, analysis and visualization of movement data. In particular, we will work towards the integration of Trackintel with other popular

Python libraries, such as the Open Street Maps package `osmnx`. The data analysis module can be substantially improved when considering mobility-related context information, such as enriching trips with point-of-interest data for transport mode identification. Moreover, we plan to provide a basic behavioral analysis module that allows insights into users' mobility behavior, for example, user mobility profiling and detecting changes in users' mobility behavior over time.

Finally, Trackintel does not aim to cover all preprocessing and analysis needs for every movement data study. However, due to the compatibility with Pandas and Geopandas, Trackintel can easily be integrated into a larger workflow that comprises a variety of Python data and spatial analysis libraries. In particular, it is targeted at providing the same reliability as these standard libraries. This is achieved through strong compliance with Python library standards, including a high coverage of unit tests with both real and synthetic data, a code review process, and continuous integration. In this setup, new algorithms can be contributed without risking breaking existing functionality. We therefore believe that Trackintel can serve as a standard and well-trusted mobility processing tool.

Acknowledgement

This work was supported by the Swiss Data Science Center [C17-14] and the ETH Zurich Foundation [MI-01-19]; Additionally, we would like to thank Christof Leutenegger, Sven Ruf, and Nishant Kumar for their code contributions to Trackintel, David Jonietz for helping to create the idea of Trackintel, and René Buffat and Jiří Kunčar for their technical input in the early stage of this project.

3.2 A graph-based representation for human mobility data

The following section is a reprint of the publication:

Henry Martin, Fernando Perez-Cruz, and Martin Raubal (2023a). “A graph-based representation for human mobility data”. *International Journal of Geographical Information Science* (under review).

Conceptualization: HM, FPC, MR; Methodology: HM, FPC; Software: HM; Validation: HM, FPC; Writing – original draft: HM; Writing – review & editing: HM, FPC, MR

The content is reproduced ‘as is’, however, formatting changes and corrections of spelling have been applied.

Abstract

Information and communication technologies and location-based services in combination with machine learning methods will play key roles in assisting and enabling behavior change toward a more sustainable mobility behavior. To take advantage of their full potential, they require large labeled datasets. However, tracking datasets are often small, unlabeled, and can not be shared due to the high privacy concerns related to location data. In this work, we analyze location graphs that model individual mobility based on visited locations as nodes using counts of trips between locations as weighted edges. This representation is privacy-friendly because it does not require coordinates, can be created based on a wide range of different datasets which can then be combined and allows to characterize the mobility of individuals based on the graph topology. We present a guide to interpreting an individual’s mobility based on four different graph features and apply this representation to four different datasets. As validation we show that location graphs can be used to reproduce important statistical properties of individual mobility and the journey distribution of an individual with high accuracy despite being a compressed representation of individual mobility. Location graphs are suited to be used as a standardized and privacy-friendly way to store individual mobility data in the future.

3.2.1 Introduction

The transport sector is responsible for 16% of total global GHG emissions (Ritchie et al., 2020), therefore its swift decarbonization is a key requirement in order to limit global warming below 2 °C as required by the Paris Agreement. An important asset for the decarbonization of the transport sector is the digitalization of our society and the ubiquitous use of Information and Communication Technologies (ICT) (Weiser et al., 2016) and Location-Based Services (LBS) (Bucher et al., 2019b; Huang et al., 2018a). They can be used to assist individuals in changing their mobility behavior (Bucher et al., 2019b), support integrated multi-modal transport systems, and provide data required for optimized electric vehicle charging (Cai et al., 2022). However, these and other applications require computational methods for predicting and analyzing individual mobility and for the labeling of passively collected tracking data (Chen et al., 2016a). These tasks are treated in quantitative human mobility analysis (Dodge et al., 2016, 2020) and commonly rely on machine learning models in combination with (optionally enriched) tracking data (Gong et al., 2016; Kumar and Raubal, 2021; Luca et al., 2021; Toch et al., 2018; Urner et al., 2018; Wang et al., 2015). These methods learn dependencies between input and output variables from the available data and their performance is therefore closely linked to the quality and the quantity of such data. Especially the size of labeled datasets is a difficult requirement when working with human movement data. Labeled tracking data sets are often small due to the high costs of generating them via user studies and because of the high privacy concerns related to location data (Keßler and McKenzie, 2018; Montjoye et al., 2013). Additionally, the available rather small, labeled datasets are often recorded using different tracking technologies (e.g., global navigation satellite system (GNSS), Wi-Fi, call detail record (CDR) (Yuan et al., 2012)) and study designs. These problems can be mitigated by using a privacy-friendly representation of an individual's mobility data that allows merging datasets from different sources.

In this work, we formally introduce the location graph as a representation of an individual's mobility. Location graphs model individual mobility based on visited locations as nodes and by using counts of trips between locations as weighted and directed edges. Location graphs have several advantages for processing individual mobility data, they are efficient in terms of memory, they are privacy friendly, and have low requirements which means that they can be created based on a wide range of different datasets that can then be combined in the graph space.

The contributions of this work can be summarized as follows:

- Formal introduction of location graphs as a generalizable and privacy-friendly representation of individual human mobility that is based on visited locations.
- A guide on how to interpret location graphs with respect to individual mobility tested on five different tracking datasets.
- Validation of the capabilities and limits of location graphs.

The rest of the paper is organized as follows: Section 3.2.2 reviews and compares commonly used representations of individual human mobility. Section 3.2.3 formally introduces the location graph as a graph-based representation of individual human mobility. Section 3.2.4

provides an overview of the datasets used as examples in the later sections and explains how location graphs can be created from data. In Section 3.2.5 we introduce four different graph features together with a guide on how they can be used to interpret the mobility of individuals based on the topology of their location graph. Section 3.2.6 analyzes the limits of location graphs to represent individual mobility and Section 3.2.7 reviews and discusses the results. Finally, Section 3.2.8 concludes this work and presents avenues for future research.

3.2.2 Representations of individual human mobility

Individual human mobility in the sense of geography and transport planning is a broad term to describe anything related to the spatio-temporal whereabouts and movements of an individual person and its context. The modeling and analysis of human mobility is difficult when starting from such an ambiguous definition and therefore requires a compact and well-defined representation of the mobility of an individual.

In the quantitative literature, the model for the representation of individual human mobility is often not explicitly discussed but rather assumed to be the same as the data that was collected. For example, in a travel survey participants record their trips and therefore the collection of all trips of a person is assumed to be the representation of the participant's mobility. This approach is reasonable; however, it often represents a blind spot in the modeling framework.

In the following, we present the most common approaches to represent individual human mobility.

Based on activities

The State-of-the-art in transport planning is to model travel demand based on activities (Castiglione et al., 2015; Jiang et al., 2017). These activity-based models use our decisions whether, where, and when to perform an activity to model the resulting travel demand (Axhausen and Gärling, 1992).

Following this hypothesis, the optimal representation of individual mobility is based on activities. In reality, not all of our mobility such as recreational walks, can be explained by the need to perform an activity at a distant location (Mokhtarian and Salomon, 2001). However, the empirical success of activity-based modeling (Castiglione et al., 2015) shows that activity-based approaches are still a reasonable choice.

Ghosh et al. (2018) present a comprehensive graph-based framework to model the activity behavior and thereby also the mobility behavior based on temporal activity logs. The activity logs consist of records of activity labels and time stamps. This framework has a solid theoretical foundation based on activity theory (Das and Winter, 2016), however, in practice detailed information about activities is hard to obtain (Chen et al., 2016a), and the definition of activity might change significantly depending on the context and the specific application (Das and Winter, 2016).

Therefore, it is common to use the activity location instead of the *true* activity, as the former can be extracted from tracking data. The process of replacing activities by locations comes naturally in the domain of human mobility analysis, however, the relevant assumptions involved in this modeling step are often not discussed or mentioned in the literature.

Based on sequences of activities or visited locations

Individual mobility behavior is most commonly represented as the sequence of locations that were visited. This corresponds to the *location history* that was presented in (Zheng et al., 2009) which is similar to *lifeline beads* introduced in (Hornsby and Egenhofer, 2002).

The mobility profile of an individual person M_p is built by storing a location l_i in a list every time this location is visited by this person. The movement profile is then a non-unique list of locations, ordered by the time of visitation $M_p = [l_1, l_2, l_3, l_1, \dots]$. Every location in the list can then be enriched with additional information such as timestamps, context data, or semantic information. Bhattacharya and Das (2002) offer a good source for further reading on this type of model.

The location history is very expressive and highly unique so it can act like a fingerprint. It was shown that only four time-stamped locations are sufficient to uniquely identify 95 % out of 1.5 M users (Montjoye et al., 2013). Typical applications of this location history are as an input for next place prediction (Hong et al., 2022), the identification or verification of users (De Mulder et al., 2008; Lin et al., 2015), or the location management in mobile networks (Bhattacharya and Das, 2002).

The most common approach for modeling the location history is via hidden Markov models (HMM), even though it does not capture certain properties of human mobility such as scale-invariant long-distance dependencies (Kulkarni et al., 2019; Zhao et al., 2015), a problem that can be solved using deep learning approaches (Damiani et al., 2020; Hong et al., 2022). A significant disadvantage of this representation of mobility behavior is that the size of the profile grows with the tracking time as also known places are simply appended to the existing profile. A way to mitigate this problem is to split the location history into subsequences such as weeks as done in (Damiani et al., 2020).

Based on location networks/graphs

A modification of the location history that addresses the lack of a compact representation is given by approaches that construct a location graph based on the full location history. A location graph or individual mobility network is a directed weighted graph with all unique locations that have been visited by an individual as nodes that are connected with an edge if the person transitioned directly between two locations. The weight of an edge corresponds to the number of direct transitions between two places. The location graph can be enriched with context such as the spatial location of a node or aggregations of temporal data such as the average stay duration at a location.

An advantage of a location graph over the location history is that mobility data are stored in a compact way because every visited location is only stored once. This comes with the trade-off to lose some of the sequential information that is still available in the full location history. This compression also makes location graphs more privacy-friendly as it does not require the coordinates of a location but only an arbitrary identifier and it does not require information about individual visits such as start and end times. Despite being compact, (Manousakas et al., 2018) could show that the pruned topology of a location graph that is constructed by only keeping the 10 most frequently visited locations is highly unique and can act as a fingerprint of individuals. However, while it is still possible to match users to previously recorded tracking data, it is a difficult task that strongly depends on the tracking period (Wiedemann et al., 2023). These results show that the transformation of the location history into a location graph preserves a significant portion of the mobility information.

Location graphs can serve as a mobility representation that allows using the information in downstream tasks such as the classification of activity purposes as it was done in (Rinzivillo et al., 2014). Here the authors extracted network features such as various average centrality measures and showed that a random forest classifier significantly increases its performance with access to these features. (Wang et al., 2019) presented a deep learning-based approach for the automated generation of network embeddings that could be used for similar downstream tasks.

A particularly interesting analysis was performed in (Schneider et al., 2013) where the authors analyzed daily location graphs and showed that these daily graphs are not random but that only 17 out of over 1 million possible 6-node networks are required to describe over 90 % of the mobility of the daily travel patterns in a large data set. These results show that location graphs are a rich representation of human mobility, despite their simplicity.

Based on sets of visited locations

A further simplification of descriptions of human mobility is by omitting the topology of location graphs and simply storing the nodes. This corresponds to the set of unique locations available in the location history of an individual. The concept has been successfully applied to studies that analyze mobility behavior on a highly aggregated level. An example is the study of Alessandretti et al. (2018) which showed that people only visit a limited set of locations that slowly evolves over time.

Despite these very interesting results, sets of visited locations are not well suited to generate insights on an individual level due to the absence of the relation between an individual's locations.

Based on movement

Another branch of movement profiles that is not directly related to activities and therefore not further discussed in this work is based on the movement of persons instead of the visited locations. A recent example is given by (Trasarti et al., 2017) who proposes an

approach where a mobility profile is defined by the set of medoid trajectories that result from a trajectory clustering step.

Discussion of approaches

The different approaches based on activities and locations represent different trade-offs between accuracy, privacy, and generalizability. The realization of these trade-offs by choosing a representation for a study based on individual tracking data or as input for a novel method to analyze individual tracking data has a strong impact on the limitations of the respective outcome. Very importantly, the choice of representation also has a large impact on the privacy risk for an individual associated with the tracking of location data (Duckham and Kulik, 2005). However, the assumptions related to the choice of the representation are usually not discussed in detail or even not at all as data is taken as available. A notable exception is given in (Das and Winter, 2016).

In this work, we focus on the representation of individual mobility as graphs based only on visited locations. This representation has minimal requirements on the complexity of the tracking study (e.g., no labels or additional user information is needed) which allows us to easily combine tracking data from different datasets. At the same time, the minimal information requirements are privacy-friendly as the time of individual visits or the exact purpose of a visit are obfuscated through aggregation. However, location graphs still retain valuable information related to an individual's mobility that allows applications such as the personalization of services or a general analysis of mobility behavior.

3.2.3 Graph based representation of mobility

A natural way to represent individual mobility is given by person-specific graphs based on the already visited locations (Martin et al., 2018; Pappalardo et al., 2015; Rinzivillo et al., 2014; Zheng et al., 2008). In such a graph, nodes correspond to activities or visited locations, and edges may correspond to any measure of distance between the nodes (e.g., the counts of sequential visits of two locations, spatial distance, transport time or cost (Martin et al., 2018)). Using visited locations is generally advantageous, as it was shown in (Alessandretti et al., 2018) that the set of visited locations grows sublinearly over time. Furthermore, patterns of daily location visits have already been shown to offer rich insights into the mobility of individuals (Schneider et al., 2013). Both findings show that a graph representation of individual human mobility based on visited locations can capture relevant temporal patterns of human mobility.

Both are hints that personalized location graphs can be used as a compact yet information-rich model to represent individual human mobility. In such a representation, spatial information is encoded in the topology of the graph and the relative position of the nodes given the edges. Temporal information could be encoded either in the graph signal (e.g., changing values in nodes or edges over time) or by the changing topology (Holme and Saramäki, 2012; Kivelä et al., 2014). Context data relevant to human mobility are often

already represented as a graph (e.g., street networks, transportation networks, or social networks), therefore they could be directly integrated into the existing graph structure by creating a multilayer network (i.e., interconnected layers of different graphs) (Kivelä et al., 2014).

This section offers a formal introduction to a graph representation of human mobility. Given the strong relation between activities and travel, the graph is defined based on the sequence of activities of a person which are aggregated into the location graph (Figure 3.10).

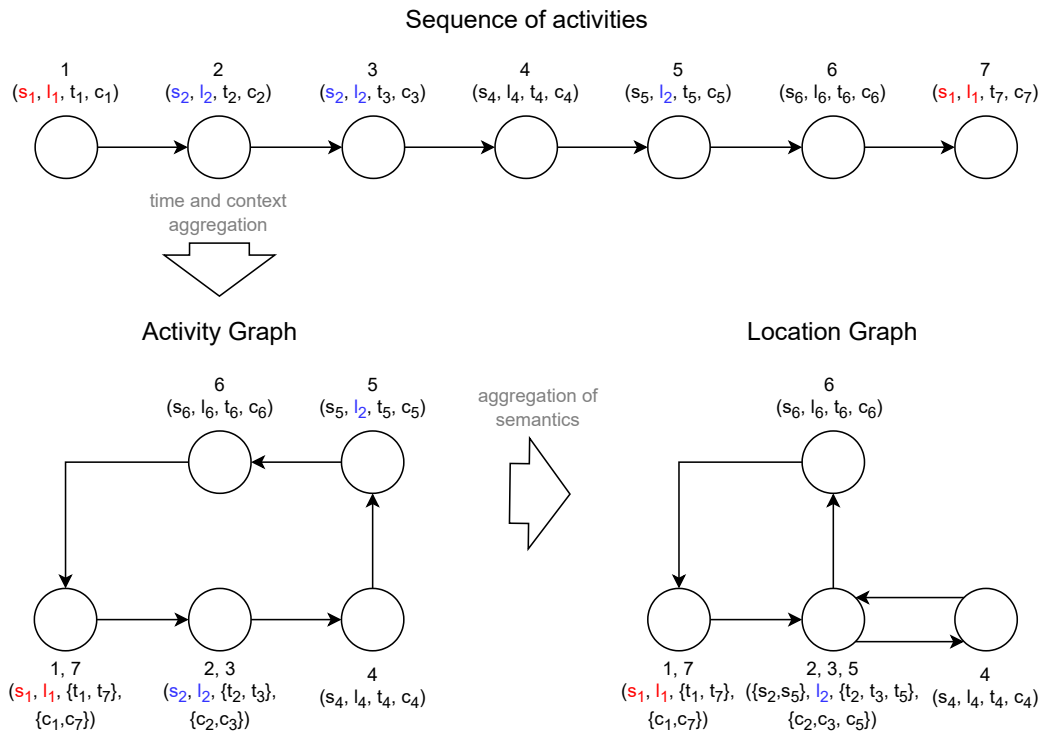


Fig. 3.10.: Location graphs are generated based on the activity sequence of a person in two aggregation steps. In the first step, activities represented by nodes are aggregated if they were performed at the same location (l_i) and share the same activity label (s_i), which leads to the activity graph. In the second step, all nodes of the activity graph are aggregated if they share the same location (l_i).

Activity graph

We follow the concept of activities used in transport planning where activities need to utilize resources and opportunities that are dispersed at different locations and possibly only accessible at certain times (Das and Winter, 2016). Furthermore, we include the context by following the argumentation of Das and Winter that the semantics of activities are context dependent. Activities in this sense have a relatively low granularity and describe what a person does throughout the day (e.g., shopping for food, going to work, being at home, etc.)

Based on these assumptions we represent the activity of an individual as a 4-tuple of a semantic activity label s , the activity location l , the activity time t , and the context of the activity c . We can then describe an arbitrary time span of an individual's behavior (in the sense of the above definition) using the sequence of activities ordered by their appearance

$$A = [a_1, a_2, \dots, a_i, \dots, a_n] \quad \text{for } a_i = (s_i, l_i, t_i, c_i). \quad (3.1)$$

In this definition, t , l , and c are conceptual terms that summarize all location, timing, and context information that is relevant to the particular application. One could argue that time and space are context information of the activity and could therefore be summarized as context. While this is generally true, we chose to model them explicitly as they are non-optional context information for the purpose of the representation of individual mobility while the relevance of other context information such as the weather, the number of friends, or the mood of a person depends on the application area of the mobility representation.

Using this definition, all activities are unique, as no two activities ever take place at the exact same time and space and with the same context. This might be true from a theoretical perspective but contradicts our experience in life where we usually consider activities that we do at the same place and with the same purpose as being the same activity. Examples are “going to work at the office” or “workout at the fitness studio”. Even though we do these things at different times and under changing context, we consider them the same activity.

We, therefore, define an aggregated activity that summarizes all activities that have the same semantic label and are taking place at the same location as

$$a'_i = (s_i, l_i),$$

the sequence of aggregated activities as

$$A' = [a'_1, a'_2, \dots, a'_i, \dots, a'_m] \quad \text{with } m \leq n,$$

and the set of aggregated activities as $\mathcal{A}' = \{A'\}$.

Every aggregated activity a'_i is now associated to a set of timing information \mathcal{T}_i and a set of context information \mathcal{C}_i that are defined as

$$\mathcal{C}_i = \{c_j \mid l_j = l_i \wedge s_j = s_i \quad \text{for } c_j, s_j, l_j \in a_j, s_i, l_i \in a'_i, \\ i \in [1, 2, \dots, m], j \in [1, 2, \dots, n]\} \quad (3.2)$$

$$\mathcal{T}_i = \{t_j \mid l_j = l_i \wedge s_j = s_i \quad \text{for } t_j, s_j, l_j \in a_j, s_i, l_i \in a'_i, \\ i \in [1, 2, \dots, m], j \in [1, 2, \dots, n]\}. \quad (3.3)$$

We further define the set of all pairs of sequential aggregated activities

$$E' = \{(a'_i, a'_j) \mid j = i + 1 \forall a'_i, a'_j \in A'\}. \quad (3.4)$$

Using the aggregated activities A' we define the *individual activity graph* of a person as the pair

$$G_{A'} = (A', E') \quad (3.5)$$

Here all aggregated activities are defined as nodes with a directed edge between all aggregated activities that were performed sequentially. Edges can be weighted by counting the number of transitions and can be undirected ignoring the edge direction, in the latter case, the weights of both directions would be summed up.

Location graph

In practice, tracking studies are often based on the passive collection of location data which means that the activity label is often unobserved. It is then impossible to create the individual activity graph, however, we can use the activity location as a proxy for the unobserved activity labels. Formally, we can define the individual location graph analogous to the individual activity graph as an aggregation of activities.

$$A'' = [a''_1, a''_2, \dots, a''_i, \dots, a''_m] \text{ with } m \leq n, \quad (3.6)$$

$$a''_i = (l_i) \quad (3.7)$$

Equations 3.6 and 3.7 show that the aggregated activity sequence A'' is the same as the sequence of visited locations L :

$$L = [l_1, l_2, \dots, l_i, \dots, l_m] \text{ with } m \leq n. \quad (3.8)$$

We further define $\mathcal{L} = \{L\}$ as the set of all visited locations, with $o = |\mathcal{L}|$ as the number of unique locations and n the number of activities of a user. For brevity and clarity, we define the individual location graph directly via the set of unique locations \mathcal{L} . Every unique location $l_i \in \mathcal{L}$ is now associated with a set of timing information \mathcal{T}_i^l , a set of context information \mathcal{C}_i^l and a set of unobserved semantics \mathcal{S}_i^l defined as

$$\mathcal{C}_i^l = \{c_j \mid l_j = l_i \text{ for } c_j, l_j \in a_j, a_j \in A, l_i \in L, \\ i \in [1, 2, \dots, o], j \in [1, 2, \dots, n]\} \quad (3.9)$$

$$\mathcal{T}_i^l = \{t_j \mid l_j = l_i \text{ for } t_j, l_j \in a_j, a_j \in A, l_i \in L, \\ i \in [1, 2, \dots, o], j \in [1, 2, \dots, n]\} \quad (3.10)$$

$$\mathcal{S}_i^l = \{s_j \mid l_j = l_i \text{ for } s_j, l_j \in a_j, a_j \in A, l_i \in L, \\ i \in [1, 2, \dots, o], j \in [1, 2, \dots, n]\} \quad (3.11)$$

We further define the set of all pairs of sequentially visited locations

$$E = \{(l_i, l_j) \mid j = i + 1 \forall l_i, l_j \in L\}. \quad (3.12)$$

We then define the *individual location graph* as the pair

$$G_L = (\mathcal{L}, E) \quad (3.13)$$

Just as the *individual activity graph*, the *individual location graph* can be weighted by counting the number of transitions and it can be undirected ignoring the edge direction, in the latter case, the weights of both directions are summed up.

Advanced modeling and possibilities for extension of the location graph

The *individual location graph* is very versatile to use in practice because of its low requirements on the data. However, often more data than the raw tracking data are available, and in this case, the *individual location graph* offers many possibilities to add additional information based on the application and the available data sources.

Nodes represent the physical locations and can act as a container that can store all information associated with the activity locations. Depending on the application, nodes can be associated with a geometry such as the position of the location or the extent of a place which might be a building footprint, a room, or a defined outdoor space such as a park. It can also be associated with arbitrary properties of this place such as the size, the density of trees in the surroundings, or the noisiness.

Edges encode the relation between the visited places and could in theory correspond to any measure of distance or proximity between the nodes such as the counts of sequential visits of two locations, the spatial distance, the travel time, travel costs, or even more abstract measures such as the cultural distance between two locations. Furthermore, location graphs could be defined as multigraphs that allow multiple types of edges in parallel.

Available temporal information can be encoded explicitly by aggregating visiting times per location as shown in Equation 3.10, in the graph signal (e.g., changing values in nodes or edges over time), or by the changing topology (Holme and Saramäki, 2012; Kivelä et al., 2014) which can be done by binning the tracking data over time as done in (Wiedemann et al., 2023).

Context data that are related to locations can be included in the graph as aggregated node features as presented in Equation 3.9. Context data that are already represented as a graph (e.g., street networks, transportation networks, or social networks), can be directly integrated into the existing graph structure by creating a multilayer network (i.e., interconnected layers of different graphs) (Kivelä et al., 2014).

3.2.4 Data and graph generation

To further explore the properties of the location graphs, we use data from four different tracking studies: the yumuv dataset (Martin et al., 2021b,c), the Green Class (GC) I and II datasets (Martin et al., 2019a), the Geolife dataset (Zheng et al., 2009) and a subset of

Study	Tracking type	Users	Number of nodes	Av. daily activities
GC1	GNSS via app	137/139	558±188	6.5±1.3
GC2	GNSS via app	49/50	460±194	6.6±1.7
Geolife	GPS tracker	59/176	159±147	2.7±1
yumuv	GNSS via app	671/813	134±113	5.1±1.6
Foursquare (1000)	LBSN checkins	1000/27252	70±94	2.4±1.2

Tab. 3.5.: Used datasets with properties. Column *users* shows participants after filtering and the total participants in the dataset. For *Number of nodes* and *Average daily activities* the average and the standard deviation over users is shown.

a Foursquare dataset¹⁵ (Yang et al., 2015, 2016). An overview of the selected datasets is shown in Table 3.5.

All data preprocessing is performed using the Trackintel python library (Martin et al., 2023c). Trackintel uses a well-known hierarchical movement data model that segments data into staypoints which represent individual visits, locations which are the places that are visited by an individual, trips which are the travel between two staypoints and tours which are a collection of all trips required to return to the same place. For the Geolife dataset staypoints are generated based on the GPS position fixes using the algorithm and parameters presented in (Li et al., 2008), for the other GNSS based datasets (yumuv and Greenclass) staypoints were already available. For all three GPS datasets, locations are extracted using DBSCAN (Ester et al., 1996) with a search radius of 30 meters and a minimum number of one point to form a cluster. Trips and tours are extracted using the algorithms available in Trackintel described in (Martin et al., 2023c). Before creating the graphs for the GNSS based datasets, we filter to include only users with at least 14 days with tracking coverage of more than 70 % of the time of the day.

For the foursquare dataset, venues are considered as locations and check-ins as staypoints. In this dataset we consider only users with high data quality to mitigate the bias introduced by the check-in behavior. Specifically, we require a user to have at least 81 check-ins (25th percentile), to have visited at least 40 different locations (25th percentile), and to have at least 24 check-ins at a labeled home location (75th percentile). There are 27'252 users with sufficiently high data quality of which we randomly sample 1'000 users that are analyzed in this study.

The location graph for each person is generated as described in Section 3.2.3. For all datasets, every location is considered a node in the location graph. For the GNSS based datasets every trip between two locations increases the weight of the edge between the two associated nodes by 1, for the Foursquare dataset two consecutive check-ins increase the weight of the edge between the corresponding locations by 1 if the check-ins are less than 12 hours apart.

¹⁵https://sites.google.com/site/yangdingqi/home/foursquare-dataset#h.p_ID_56

3.2.5 Interpretation of location graphs

Figures 3.11, 3.12, 3.13 and 3.14 show examples of location graphs. While these location graphs are hard to read at first glance, they are rich in information that can be used to analyze individual mobility. This section introduces metrics and methods to analyze mobility graphs with respect to individual mobility behavior.

In-degree distribution

The degree distribution of a graph p_k describes the probability that a randomly selected node v has degree k (Barabási, 2016). For location graphs, the weight of the in-degree represents how often an individual visited a location while the unweighted in-degree of a node represents from how many different locations the individual visited a location. Therefore, without tracking errors, their weighted in-degree equals their weighted out-degree. In the following, we analyze the unweighted in-degree distribution and explain how the weighted in-degree distribution can be analyzed accordingly.

The absolute values of the in-degrees depend strongly on the observation period. However, as the in-degree distribution of a location graph follows a power law its shape parameter β offers a way to characterize how individuals visit locations that is robust to the observation period. We follow the guidelines from (Clauset et al., 2009) and use the *Powerlaw* python package (Alstott et al., 2014) to fit a power law of the form

$$p(x) = Pr(X = x) = Cx^{-\beta} \quad (3.14)$$

to the (unweighted) in-degree distribution. Here, C is the normalization constant that includes the lower bound of the power law behavior in the data.

The shape parameter β is then a measure of how we visit places. Moving flexibly from place to place will lead to several nodes with similar in-degree and a higher β while always returning to the same node after visiting a new node will lead to a single dominant node with high in-degree and a lower β . We'll refer to this measure as the in-degree distribution β .

If we would fit on the weighted in-degree distribution, the shape parameter β would measure how we distribute our visits among places. Mostly visiting a few places leads to dominant nodes and a lower β , and a distribution among many places leads to a high β .

Figures 3.11 a) and b) show examples of two location graphs of an individual with a low in-degree distribution β (a) and an individual with a high in-degree distribution β (b). The graphs were chosen to be average (within the 25th and 75th percentile) with respect to the other features but noteworthy (below or above the 25th and 75th percentile) with respect to the in-degree distribution β and to have a similar number of nodes. The graph in Figure 3.11 (a) has multiple centers that serve as hubs while the graph in (b) has a more star-like shape with a single main node in the center. Figure 3.11 (c) shows the degrees of all nodes in a log-log plot. The graph with higher in-degree distribution β (shown in orange) has a single high-degree node while the graph with a lower in-degree distribution β (shown in blue)

has multiple nodes with a similar high degree. Figure 3.11 (d) shows the distribution of in-degree distribution β values over all considered datasets. The in-degree distribution β of most graphs lies between 2 and 2.5, therefore most of the individual location networks are scale-free ($2 \leq \beta \leq 3$ (Barabási, 2016)). Scale-free is a property that is found in many real networks such as the World Wide Web (Barabási and Albert, 1999) or e-mail communication networks (Ebel et al., 2002) and its most prominent feature is the presence of hubs in the network. For location graphs, hubs are typically places that play a prominent role in our lives such as our home or work location, and are the start and end of many of our trips. It is noteworthy that $\beta = 2$ is the theoretical minimum for graphs as here the degree of the main hub is close to the number of nodes (Barabási, 2016). Figure 3.11 (d) shows that some of the graphs are below this theoretical limit which can most likely be attributed to a low tracking coverage or errors in the tracking data processing leading to noise in the fitting process.

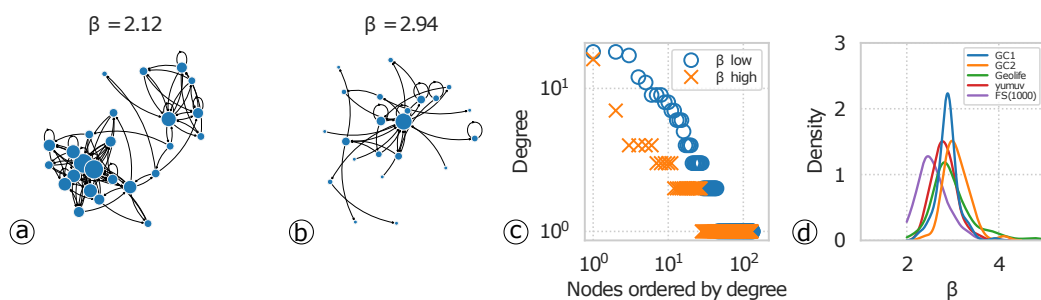


Fig. 3.11.: **a:** example of a typical location graph with a relatively low in-degree distribution β ; **b:** example of a typical location graph with a relatively high in-degree distribution β ; **c:** degree distribution of the graphs shown in a) (blue) and b) (orange); **d:** distribution of in-degree distribution β for different datasets.

Journey distribution

As described in Section 3.2.4, human mobility can be well represented using a hierarchical model for movement. Relevant units are trips, which summarize all travel between two consecutive activities, tours which summarize all consecutive trips needed to return to the start location, and journeys which are tours that start and end at the home location of a person. A journey can therefore be represented as a sequence of locations that are visited until a person returns home again. Typically, our mobility behavior is characterized by strong regularities which means that the same journeys are repeated on a monthly, weekly, or even daily basis and can therefore be observed multiple times. An interesting representation of individual human mobility is the distribution of journeys of an individual. The *journey distribution* describes the probability that a journey was observed k times over the observation period of a person. To further analyze the journey distribution, we will now define trips and tours within the notation framework introduced in Section 3.2.3.

Definition 1 (Trip). A trip t_i summarizes all movement and non-activity related stays between the origin activity $a'_{i-1} \in A'$ and the destination activity $a'_i \in A'$. A trip t_i is therefore related to an origin location $l_{i-1} \in a_{i-1}$ and a destination location $l_i \in a_i$

Definition 2 (Tour). A tour τ_k is the sequence of all sequential trips that start and end at the same activity location.

$$\begin{aligned} \tau_k = \{t_i, t_{i+1}, \dots, t_j\} \quad & \forall l_i = l_j, i \neq j \\ & l_i \neq l_k, i < k < j \\ & i < j, 1 < j < m. \end{aligned}$$

Definition 3 (Journey and sequence of journeys). A journey ϕ is a tour that starts and ends at the home location of a person. Φ is the sequence of all journeys ordered by their start time.

Definition 4 (Journey distribution). The journey distribution $p(\phi)$ describes the probability that a randomly drawn journey ϕ_i was observed k times over the observation period of a person. The journey distribution is defined as

$$p(\phi) = \frac{1}{|\Phi|} \sum_{\phi \in \Phi} I(\phi), \quad (3.15)$$

where I is the indicator function.

With access to the full tracking data of an individual, journeys and their distribution can be generated directly from the sequence of visited locations. This is not possible using only the location graph however, the journey distribution can be approximated using a random walk. Every walk that starts at the home location (or the node with the highest degree if no semantic information is available) and records all nodes until it reaches the home location again is a journey. When the approximated journey distribution stabilizes, then a sufficient number of walks were simulated.

The journey distribution can be used to characterize the mobility of an individual quantitatively by measuring the concentration of the distribution. We introduce the feature $n_{\phi 50}$ that counts how many journeys are required for 50 % of all journeys. Users with a low $n_{\phi 50}$ have a highly regular mobility behavior as they travel to the same locations in the same order most of the time, the mobility of individuals with a high $n_{\phi 50}$ is less regular and distributed more equally over different edges in the graph.

Figure 3.12 (a) and (b) show examples of location graphs of an individual with low $n_{\phi 50}$ (top) and an individual with a high $n_{\phi 50}$ (bottom). The top graph shows only a few important edges (edge weight determines line width) between the main nodes whereas the bottom graph has more important edges and more important edges that connect to the periphery of the graph. Figures (c) and (d) show a bar plot with the importance of the top 10 journeys for

the same individuals shown in (a) and (b). The individual with a low $n_{\phi 50}$ has significantly higher weight on the main journey and less on the subsequent journeys, while the journey importance is more equally distributed for the individual with high $n_{\phi 50}$. Figure 3.12 (d) shows the distribution of $n_{\phi 50}$ over all datasets. For most datasets, the peak is around $n_{\phi 50} = 10$, however for GC1 and GC2 the distribution has a second mode at around 40 which might be a hint that these datasets can be segmented into individuals with more regular and non-regular mobility behavior.

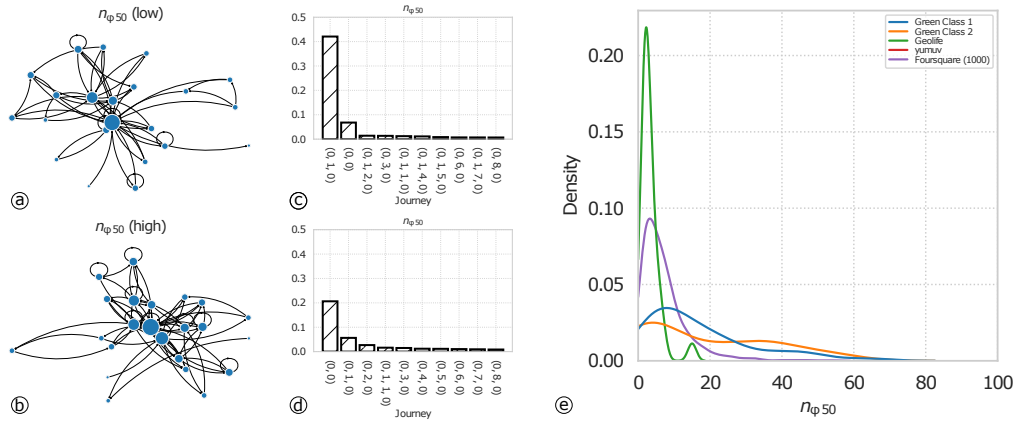


Fig. 3.12.: **a:** Location graph of an individual with low $n_{\phi 50}$; **b:** Location graph of an individual with high $n_{\phi 50}$; **c:** Histogram of frequency of top 10 journeys for the individual in a); **d:** Histogram of frequency of top 10 journeys for the individual in b); **e:** distribution of $n_{\phi 50}$ over individuals for different datasets.

Finally, the analysis based on the journey distribution is very closely related to the analysis of motifs in the mobility of individuals (Schneider et al., 2013). The main difference lies in different aggregation concepts, a motif captures all location visits within a day until a user returns home at night and does not distinguish between visits to different locations as long as they show the same pattern (e.g., $A \rightarrow B \rightarrow A$ equals $A \rightarrow C \rightarrow A$). A journey on the other hand only captures mobility until an individual returns home, therefore there could be several journeys per day and journeys distinguish between visits at different locations (e.g., $A \rightarrow B \rightarrow A$ does not equal $A \rightarrow C \rightarrow A$). The journey distribution, therefore, allows for a more detailed analysis of an individual’s mobility patterns but it may be more difficult to identify patterns that generalize to a wide range of individuals. Finally, it is important to note that the concept of motifs in mobility graphs as introduced in (Schneider et al., 2013) is related but not identical to the well-known concept of motifs used in network science introduced in (Milo et al., 2002)

Connectedness

An important property of graphs is the clustering of nodes that describes how well connected or well *meshed* a graph is. The connectedness of a graph can be measured using the average of the clustering coefficient which measures for every node how well connected its neighbors

are among each other. The clustering coefficient cc_i for node i in an unweighted and undirected graph is defined as

$$cc_i = \frac{2L_i}{k_i(k_i - 1)}$$

and the average clustering coefficient of the graph is defined as

$$\overline{cc_i} = \frac{1}{N} \sum_i^N cc_i,$$

where L_i is the number of triangles through node i .

This concept can be extended to weighted graphs by replacing the number of triangles L_i with the sum of the geometric mean of the weights of the triangles (Onnela et al., 2005), to directed graphs by counting all possible directed triangles (Fagiolo, 2007), or to both by combining both measures (Fagiolo, 2007).

All versions of the clustering coefficient can be used to analyze location graphs but they focus on different aspects of an individual's mobility. The choice between the directed and undirected coefficient relies on the definition of when two locations should be considered connected and whether the order of visitation is important ($A \rightarrow B$ vs. $B \rightarrow A$) and can be used to analyze differences in visitation patterns of whether a person visits places in both directions or rather in one direction.

The choice between the weighted or unweighted clustering coefficient is more impactful as it changes whether the focus is on the general visitation structure or on the visitation frequency (cf. the analysis of the degree distribution in Section 3.2.5). The unweighted clustering coefficient of a graph is high if the nearest neighbors of a node are connected on average while the weighted clustering coefficient is high if the edges with high weights form triangles. In the location graph, edges with high weights describe our most common trips (e.g., from home to work).

A well-connected or meshed graph is an indication of flexible mobility behavior, as we travel between places in multiple manners. The weighted clustering coefficient is therefore a measure of how flexible we are in our regular mobility behavior while the unweighted clustering coefficient is a measure of how flexible we are in our overall mobility behavior.

In the following, we analyze the unweighted and undirected clustering coefficient. Figure 3.13 shows a location graph for a user with a relatively low clustering coefficient and a user with a relatively high clustering coefficient. The location graph with a high average clustering coefficient shows a more connected center of high-degree nodes but also more connections between the periphery of the graph and several nodes in the center, showing that the person is more flexible in the order he or she visits different locations while the person with the low average clustering coefficient does rather go back and forth between the main node and other nodes. Figure 3.13 also shows the distribution of the average clustering coefficients for all graphs in different datasets.

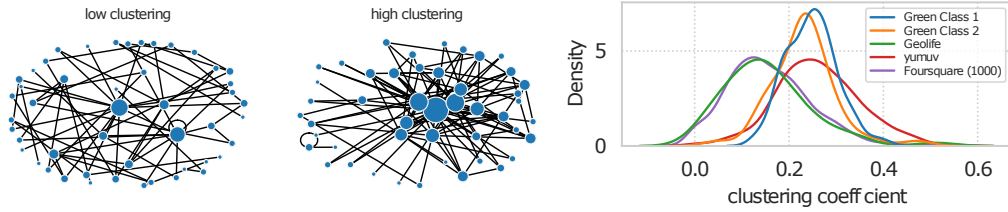


Fig. 3.13.: Examples of (unweighted and undirected) location graphs with low and high unweighted and undirected clustering coefficient and the distribution of unweighted and undirected clustering coefficients over users for different datasets.

Hubbiness

The most important property of scale-free networks is the tendency to build hubs. In the case of individual location graphs, we can explain the tendency to form hubs by the fact that some locations play a much bigger role in our lives than other locations. For most people, these locations are the home and work locations. However, some people might have changing work locations, work in a different city during the week, or have several other locations of great importance such as the house of a partner, a family member, or a close friend. An indication of these differences is already given by the distribution of the power law exponents shown in Figure 3.11, however, to analyze this dimension of mobility behavior in greater detail we propose to use the PageRank algorithm (Page et al., 1999). PageRank assigns an importance to every node based on the structure of incoming links that can be interpreted as the probability of arriving at a node after a random walk in the network. With this in mind, PageRank can be used to identify and quantify hubs (of incoming edges). In order to quantify the hubbiness of a graph we calculate how many nodes are required to cover more than 50 % of the probability which we will refer to as $PR_{0.5}$. If an individual has strong hubs and concentrates his mobility on a few hubs, this number will be low; if a person often visits different locations, his location importance will be more distributed and the number of locations required to reach 0.5 of the PageRank probability will be high.

Figure 3.14 shows the location graph of an individual with a low $PR_{0.5}$ number and a location graph of a person with a high $PR_{0.5}$ number. Figure 3.14 shows the distribution of $PR_{0.5}$

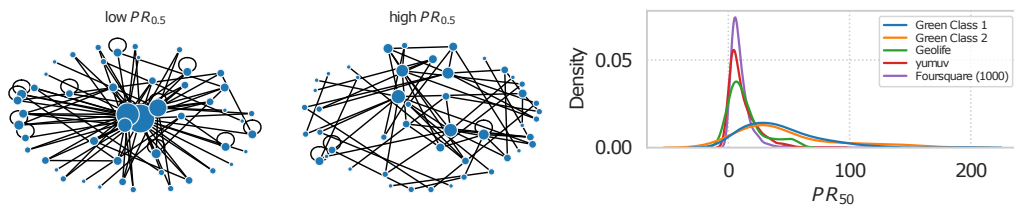


Fig. 3.14.: Examples of (unweighted and undirected) location graphs with low and high $PR_{0.5}$ and the distribution of $PR_{0.5}$ over users for different datasets. Graphs are drawn unweighted and undirected for better visibility

3.2.6 Validation

The creation of the location graph of an individual based on tracking data is not lossless and we now examine the mobility-related information retained in the model. In this section, we will first analyze how well the information stored in the location graph can be used to reproduce important statistical patterns characteristic of human mobility and then see whether a location graph is a suitable model to approximate the journey distribution of an individual.

Reproducing stylized facts of human mobility

Since the advent of large-scale tracking data sets, more generalizable statistical properties of human mobility often referred to as *laws* have been discovered. These findings can often be reproduced on a wide range of different tracking data sets and it is therefore likely that they relate to an underlying process that is specific to individual human mobility.

In the following, we will present a set of well-known statistical characteristics of (individual) human mobility and analyze to what extent these are preserved in the location graph model of an individual's mobility. We hypothesize that if these statistical laws are preserved during the transformation of tracking data into a location graph, then information relevant to describe individual human mobility is preserved in the graph. The power laws in this section are fitted using the python package *powerlaw*¹⁶ following the guidelines established in (Clauset et al., 2009) and provided in the publication associated to the *powerlaw* package (Alstott et al., 2014) using the Akaike information criterion (AIC) (Akaike, 1998) to identify the best fitting distribution.

Visitation frequency

A statistical regularity reported in (González et al., 2008) and among others confirmed in (Song et al., 2010b) is that the frequency of location visits of an individual follows Zipf's law. This means that the number of visits at a location follows a power law when the locations are ordered by their visitation frequency. This shows that people generally visit a few locations very often (e.g., home or work locations) while most locations are visited only a few times (e.g., running an exotic errand).

For a quantitative evaluation, we compare the shape parameter of the power law fit on the tracking data-based visitation frequency with the power law fit of the graph data on a per-user level. 3.6 shows that there are no significant differences for the shape parameters of both fits except for the Foursquare (1000) dataset.

Distribution of displacements

Brockmann et al. (2006) analyzed a large dataset of banknote locations over time and noticed that the geographical distance of their displacements follows a power law. As these bank notes were carried by people, they could relate these findings to the distribution of trip

¹⁶<https://github.com/jeffalstott/powerlaw>

Tab. 3.6.: Average and standard deviation of power law fitting parameters over users for the distribution of visitation frequency. The asterisk indicates a significant difference on a $p = 0.05$ level.

	β_{true}		β_{graph}	
	mean	std	mean	std
Green Class 1	1.34	0.05	1.34	0.05
Green Class 2	1.35	0.10	1.34	0.09
Geolife	1.40	0.10	1.41	0.11
yumuv	1.46	0.12	1.46	0.12
Foursquare (1000)*	1.42	0.08	1.43	0.09

distances of a population. These findings were later reproduced on several large tracking datasets (Alessandretti et al., 2017; González et al., 2008; Song et al., 2010b). To test the reproducibility of this empirical finding we collect all displacements of all individuals per dataset as ground truth. Under the assumption that the coordinates of the locations are available as node features in the context information C_i of a node i (cf. Equation 3.9) we calculate the distance matrix between all nodes and repeat every distance between two nodes a number of times that corresponds to the weight of their shared edge as displacements. It is important to note that this corresponds to the linear distance approximation while the original data uses the distance recorded in the tracking data. We then fit power laws separately for the ground truth displacements and the displacements based on location graphs.

Figure 3.15 shows the distribution of displacements and the fit of a log-normal distribution of the original trip-based data and the graph-based data for 4 datasets. It can be seen that both datasets are fitted with the same type of distribution with the same or very similar parameters for all datasets. For GC2, Geolife, and yumuv the fringes of the distribution cannot be well recovered. This is especially true for very short trips below 100 meters. A likely reason for this is that because locations are extracted from the tracking data via clustering, close-by locations might be merged and trips between them are lost. At the same time, graph-based distances do not include trips that start and end at the same location. While the graph does record self-loops in such cases, their distance will always be zero. The results show that the distribution of displacements is generally very well preserved in the graph with the exception of short trips.

Radius of gyration

The radius of gyration as a metric to characterize the travel behavior of individuals was introduced in (González et al., 2008) where the authors showed that it can be used to standardize the distribution of displacements of individuals. Since its introduction, the radius of gyration has become a standard metric to characterize the movement of individuals, e.g., to identify user groups (Pappalardo et al., 2015) or as personalized features for forecasting trip distances (Cai et al., 2022).

An individual's radius of Gyration is defined as



Fig. 3.15.: Distribution of displacements and best-fit comparison between original data and graph-based approximation for different datasets.

$$r_g = \sqrt{\frac{1}{n} \sum_j^n (l_j - l_{cm})^2}$$

where n is the total number of activities and $l_{cm} = \frac{1}{n} \sum_j^n l_j$ is the geographical center of mass of all visits to activity locations.

If the coordinates of the locations are available and stored in the nodes of the location graph as context information c_i as described in Section 3.2.3, the radius of gyration for the location graph can be calculated as

$$r'_g = \sqrt{\frac{1}{\sum_i^k d_{in,i}} \sum_i^k (d_{in,i} (l_i - l_{cm,g})^2)}$$

where k is the number of nodes, $d_{in,i}$ is the in-degree of the i^{th} node related to the i^{th} location l_i and

$$l_{cm,g} = \frac{1}{m} \sum_i^k d_{in,i} \cdot l_i$$

is the geographical center of mass of the nodes in the location graph weighted by the in-degree.

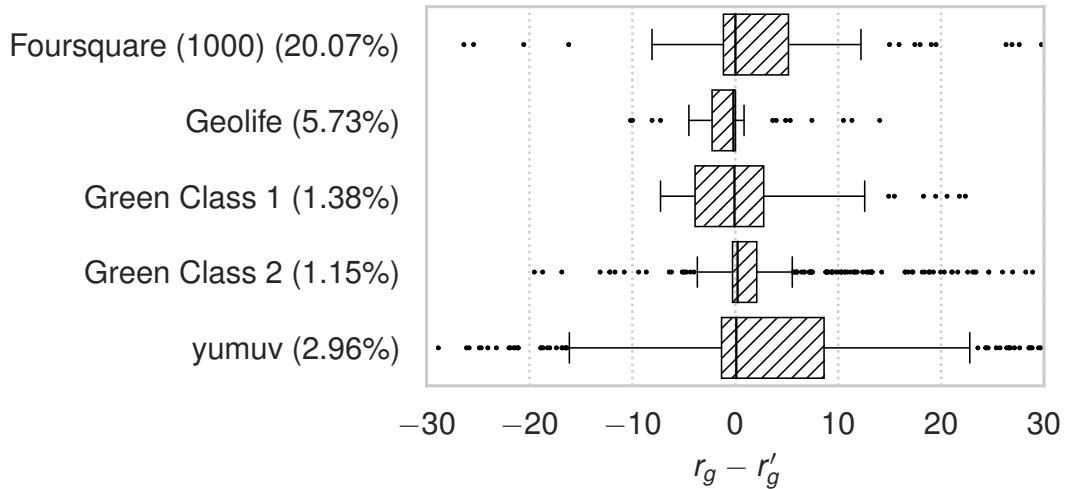


Fig. 3.16.: Box plot of the difference between the radius of gyration and the graph-based radius of gyration over users for different datasets (absolute average e_{rel} shown in parentheses).

We calculate the radius of gyration of every user based on the original tracking data and based on the location graph and evaluate their similarity using the relative error

$$e_{rel} = \frac{|r_g - r'_g|}{|r_g|}.$$

Figure 3.16 shows the difference between the radius of gyration based on location visits and based on graphs in km per user for different datasets. The relative error is given in parentheses next to the dataset labels and shows that the location graph with node coordinate preserves the radius of gyration of a person very well with $e_{rel} < 5\%$, meaning that the characteristic travel distance of an individual can be recovered from the location graph with coordinates. The Foursquare (1000) dataset is an exception where $e_{rel} = 20.07\%$. This can be attributed to uniquely visited locations that are far away and within a temporal gap of more than 12 hours (e.g., because the journey to get there took longer than that). These locations are then filtered as described in Section 3.2.4.

Journey based evaluation

Journeys are a central part of our daily mobility behavior as we are constantly returning to where we started our travels. The journey distribution was introduced in Section 3.2.5 as a feature to describe the mobility of an individual that can be derived from a location graph.

However, the location graph which is the 1st order Markov chain (MC) approximation of the sequence of visited locations, can not perfectly reconstruct the journey distribution and we are therefore interested in how well the graph-based journey distribution approximates the true journey distribution of a user and to what extent a higher-order MC might be a better fit to represent the mobility of a user.

Background

This problem can be formulated as a variational inference problem which can be described in the following way (Murphy, 2012):

Given a set of inputs $D = \{x_i\}_{i=1}^N$, we want to approximate a potentially intrac distribution $p^*(x) = p(x|D)$ by a trac parameterized distribution $q(x)$. We chose the free parameters of $q(x)$ such that the Kullback-Leibler-divergence (KL) between q and $p^*(x)$ is minimized:

$$\mathbb{KL}(p^*||q) = \sum_x p^*(x) \log \left(\frac{p^*(x)}{q(x)} \right) \quad (3.16)$$

which can be reformulated as

$$L(q) = -\mathbb{KL}(q||p^*) + \log(p(D)), \quad (3.17)$$

where $L(q)$, is the evidence lower bound (ELBO), which is a lower bound on the log-likelihood of the data (Murphy, 2012), and $\log(p(D))$ is the evidence.

For this work, we define $p^*(x)$ as the true journey distribution of a person which we want to approximate by using a simpler distribution from a family of candidate distributions.

A Markov chain that operates on a set of $S = s_1, s_2, \dots, s_n$ different states with $|S| = n$ can be represented by a transition matrix $P \in \mathbb{R}^{n \times n}$ with the transition probabilities between states $p_{ij} = p(x_j|x_i)$, $i, j \leq n$ as elements (Singer et al., 2014).

$$p_{ij} = \frac{n_{ij}}{\sum_j^n n_{ij}}$$

Given the true journey distribution of an individual $p^*(\phi)$ and the journey distribution related to the k^{th} order Markov model approximation $q(\phi|k, \theta)$ we want to choose the parameters k and θ such that the ELBO is maximized:

$$\max_{k, \theta} L(q) = \min_{k, \theta} \mathbb{KL}(q(\phi|k, \theta)||p^*(\phi)) + \log p(D) \quad (3.18)$$

As $\log p(D)$ is fixed with respect to k and θ , we have to minimize $\mathbb{KL}(q(\phi|k, \theta))$ which is done by calculating the maximum likelihood fit of the candidate model given the data. For a Markov model with fixed order k all parameters θ are represented by the transition probabilities and its maximum likelihood fit is calculated as

$$p_{ij} = \frac{n_{ij}}{\sum_j n_{ij}}, \quad (3.19)$$

where n_{ij} is the number of times that a transition from state s_i to state s_j was observed in the data.

A higher-order Markov chain will potentially fit the data better and might therefore reproduce the journey distribution leading to a lower Kullback–Leibler (KL) divergence by default. In order to evaluate the goodness of fit while considering the disadvantages of a higher order model such as increased complexity and the potential to overfit, we calculate the Akaike Information Criterion (AIC) (Akaike, 1998) which evaluates the trade-off between model complexity and gain in performance.

The AIC for MC is defined in (Tong, 1975) as

$$AIC(k) = {}_k\eta_m - 2(|S|^m - |S|^k)(|S| - 1) \quad (3.20)$$

where ${}_k\eta_m$ is the likelihood ratio, $|S|$ is the number of states, m is the order of a reference MC and k is the order of the MC that is tested. The optimal order according to this criterion is the k that minimizes the AIC (see (Singer et al., 2014) for a Markov chain-related example).

Reproduction of journey distribution

For the evaluation, we collect and count all journeys separately for each user as ground truth. We then prepare a single sequence of visited locations by concatenating the locations visited during all journeys of a user. As all journeys have the home location as the start and endpoint, we delete the endpoint during concatenation to avoid artificially increasing the home location count. We then fit a series of MCs of different orders $k \in [0, 1, \dots, 6]$ on this sequence for each user. We enforce irreducibility for each MC by isolating the largest strongly connected component of the graph associated to the transition matrix of the MC.

For each MC, we sample a sequence of 10000 visited locations by setting the first visited location of the sequence to home for $k \in [0, 1]$ or by sampling from all states that start with the home location from the stationary distribution of the MC $k \geq 2$. In both cases, we extract journeys from the sampled location sequence, count the number of their appearances and normalize the count by the total number of journeys which yields the simulated journey distribution for a user and an order k . Finally, we calculate the KL divergence as defined in Equation 3.16 with the simulated journey distribution as q and the true journey distribution as p^* for each user. For the calculation of the KL divergence we define $0 \log(0) = 0 \log(\frac{0}{0}) = 0$ (Basterrech and Woźniak, 2022) and use additive smoothing on the elements of q with $\alpha = 1e - 6$ and renormalize the vector so that the sum of the elements is one.

Figure 3.17 shows the average KL divergence per dataset as a function of the MC's order. The KL divergence declines strongly when increasing the order to one and continues to decline moderately until it goes into saturation for orders larger than two. The trajectory is similar for all datasets but those with a longer tracking period and therefore more complex location sequences have a higher KL divergence for k equaling zero and one. The evaluation of the AIC suggests an optimal order of zero for all users except for two users of the yumuv dataset. However, Figure 3.18 which plots the AIC as a function of k divided by $AIC_{k=0}$, shows that the AIC is very similar for orders ≤ 2 suggesting that the trade-off of model complexity and performance is very close for these orders. The AIC stays similar for orders ≤ 3 for the GC1 and GC2 datasets which suggests that they could accommodate more complex models as they have more tracking data available because of their longer tracking periods.

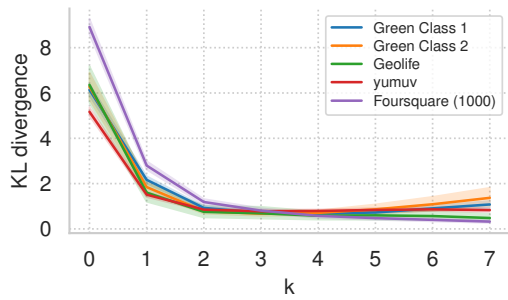


Fig. 3.17.: KL divergence for journey distributions generated from Markov models of different orders.

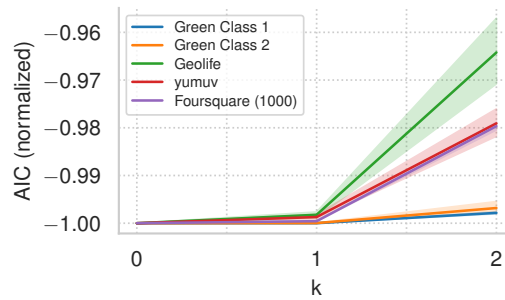
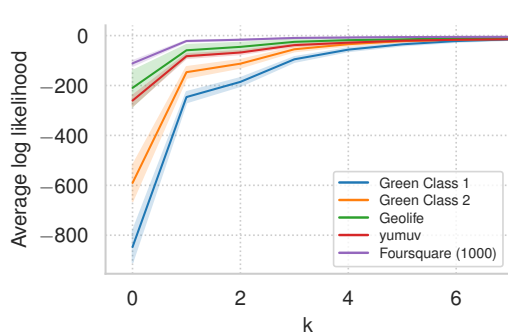
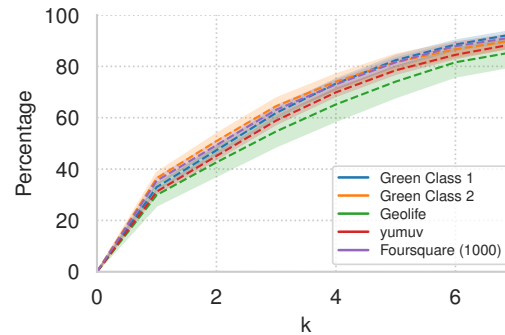


Fig. 3.18.: AIC for different k normalized by division by AIC of $k = 0$.



(a) Likelihood



(b) Percentage of transitions with probability zero

Fig. 3.19.: Log likelihood evaluated for Markov models with different order evaluated for unseen test data. Impossible transitions (as defined by the mm) return no log-likelihood.

Likelihood evaluation of hold-out journeys

To further evaluate the suitability of the location graph to represent journey information we calculate the likelihood of a hold-out set of journeys using cross-validation. This allows quantifying the degree of overfitting of a MC of a given order. We use 5-fold cross-validation without shuffling to separate all journeys of a user into a training and a testing set. Both sets of journeys are concatenated as described in paragraph 3.2.6. The training sequence is then used to fit MCs of varying order $k \in [0, 1, \dots, 7]$ using maximum likelihood estimation. The total log-likelihood is calculated as the sum of the log-likelihood of all transitions of the test sequence and the likelihood of the initial state calculated using the stationary distribution, impossible transitions (e.g, states with a transition probability of zero) are not evaluated but counted and reported later. The cross-validation results are averaged for each user.

Figure 3.19(a) shows the results of the likelihood evaluation on hold-out data for different orders k . The likelihood strongly increases when increasing the order of the MC and then continues to rise slowly for higher orders, analogously to the KL shown in Figure 3.17. Figure 3.19(b) however, reveals an overfitting issue with higher orders of the MC. Already first order MCs have a high number of impossible evaluations of around 30%, which continues to increase strongly with the increasing order of the MC. These impossible

evaluations occur if a transition in the test set was never seen in the training set and the associated transition probability is therefore zero.

3.2.7 Discussion

Dataset requirements-utility trade-off

One of the motivations that led to the in-depth analysis of location graphs to represent individual human mobility are the low technical and semantic requirements for their creation and interpretation (c.f. Section 3.2.5). The creation of a location graph does not require high-resolution tracking data or user-provided labels but only sequential records of visited locations. This allows the creation and combination of location graphs based on different types of tracking data such as CDR, GNSS, location based social network (LBSN) check-ins, geotagged tweets, or travel surveys. However, these low requirements have the drawback that the analyses that can be performed using location graphs are limited. Most notably, the location graph as presented here does not represent trajectories or routes and does therefore not allow the analysis of mobility behavior with respect to trajectories or routes.

Privacy-utility trade-off

Location graphs aggregate sequences of visited locations and by that obfuscate information that is critical from the perspective of individual geo-privacy. Notably, the times of individual visits at locations are not accessible using the location graph. Furthermore, the location graph model retains utility for the analysis of individual mobility behavior without storing coordinates. This is shown by the interpretation of the different graph features in Section 3.2.5 of which none requires location information. However, there is necessarily a trade-off between utility and privacy and it is, for example, not possible to analyze how individuals spend their days as it is done in time budget studies. Furthermore, it is important to note that while location graphs are significantly more privacy friendly than other representations they are not an anonymized representation as it was shown in (Manousakas et al., 2018) and (Wiedemann et al., 2023)

Reproduction of statistical properties

Section 3.2.6 showed that location graphs can reproduce important statistical properties of an individual's mobility data with high accuracy. This shows that the location graph preserves mobility-related information which can be made available again using the features presented in Section 3.2.5 or by other features that might be proposed in future work. Section 3.2.6 also revealed limits of location graphs as very short and very long displacements were not well preserved for some datasets. Knowledge of these limitations is important to correctly evaluate the explanatory power of applications based on location graphs.

Suitability of first order MC approximation

The decision of the optimal order of MC to model the location sequences for individual mobility with respect to journeys is difficult. On the one hand, Figure 3.17 shows a strong

decrease in the KL and Figure 3.19(a) shows respectively a strong increase in the likelihood gained by increasing the order to 1 and above, on the other hand, Figure 3.19(b) shows that even a first-order MC already yields around 30% invalid transitions during an out-of-sample validation. This behavior can be explained by the state space, created by the visited locations, which is relatively large compared to the number of samples and which is poorly covered due to the power law distribution of observation across the locations. This property of the visitation frequency of locations is discussed in paragraph 3.2.6 and leads to most locations only being visited once or twice while some few important locations are visited very often. The result is that an out-of-sample dataset may contain unseen locations as shown in Figure 3.19(b).

The unevenly sampled state space makes it difficult to draw general conclusions about the optimal order of a MC to model individual mobility based on locations. This is also reflected in the literature where the optimal order of a MC is still discussed. E.g, Lu et al. (2013) show that a MC does not improve prediction accuracy for orders greater than one while Gambs et al. (2012) claimed that second-order MC and Chen et al. (2014) reported a third order MC as the model with highest performance.

Generally, this issue could be mitigated by collecting more samples. However, the state space for individual human mobility determined by the visited locations is not stable but evolves slowly over time (Alessandretti et al., 2020). A more suitable approach for individual human mobility would be to further aggregate locations by using semantics or topologically, e.g., by using node roles (Rossi and Ahmed, 2014) if semantics are not available.

3.2.8 Conclusion

This work formally introduced, analyzed, and validated location graphs as a compact, rich, and privacy-friendly representation of individual mobility suitable for different types of tracking datasets. Location graphs are built based on the visited locations of an individual as nodes with the counts of trips between locations as weighted and directed edges.

Despite the low requirements, location graphs still describe important dimensions of individual mobility behavior. In this work, we present four different graph features (connectedness, hubbiness, journey distribution, and in-degree distribution) and showed how they relate to the mobility of an individual, showcased on five different datasets.

We further find that location graphs retain information necessary to reproduce three well-known statistical properties of individual human mobility, namely the power law distribution of location visitation frequency, the power law distribution of displacements, and the radius of gyration of an individual. Furthermore, we can show that location graphs, which are the first order MC approximation of the location history, present a good trade-off between model complexity and expressive power given the properties of individual mobility data.

Future work should focus on three different directions, namely the properties, possible extensions, and applications of location graphs. This work has related a set of graph and node properties of the location graph of a person to his mobility behavior. This work can further

be extended by analyzing the sources of the complexity of a location graph as, for example, measured by the entropy (Dehmer und Mowshowitz, 2011), the analysis of community structure in location graphs, or the analysis of further graph features or node centralities. A second research direction should focus on the analysis of extensions of the location graph. This work concentrated on location graphs with minimal requirements. In practice, additional information such as coordinates, time stamps, or geographic contexts such as land-use or point-of-interest data are often available and could provide additional insights into the mobility behavior of individuals (Zhang und Raubal, 2022). However, currently, it is unknown how they would influence the complexity-utility trade-off of location graphs. Finally, there are many applications that would benefit from the unique privacy properties of location graphs and their ability to be used on multiple different datasets. However, currently, there are only a few examples where location graphs are used for semantic enrichment of passively tracked data e.g., by assigning activity labels to nodes (Martin et al., 2018; Rinzivillo et al., 2014), or for the segmentation of individuals into groups based on their mobility behavior (Ben-Gal et al., 2019; Martin et al., 2023b). This small body of literature should be extended to further explore the potential of location graphs as a standardized and privacy-friendly way to store individual mobility data in the future.

Data availability statement

The data, codes, and instructions that support the findings of this study are available at the link¹⁷. The Geolife¹⁸ and Foursquare¹⁹ datasets are publicly available. The Green Class and yumv datasets are not publicly available due to confidentiality agreements with the participants under the European General Data Protection Regulation (GDPR).

¹⁷https://github.com/henrymartin1/graph_representation

¹⁸<https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/>

¹⁹https://sites.google.com/site/yangdingqi/home/foursquare-dataset#h.p_ID_56

3.3 Influence of tracking duration on the privacy of individual mobility graphs

The following section is a reprint of the publication:

Nina Wiedemann*, Henry Martin*, Esra Suel, Ye Hong und Yanan Xin (2023). “Influence of tracking duration on the privacy of individual mobility graphs”. *Journal of Location Based Services*, S. 1–19.

Conceptualization Conceptualization: NW, HM, ES, YH, YX; Methodology: NW, HM; Software: NW, HM; Visualization: NW; Writing - Original Draft: NW, HM, ES; Writing - Review & Editing: NW, HM, ES, YH, YX;

The content is reproduced ‘as is’, however, formatting changes and corrections of spelling have been applied.

Abstract

Location graphs, compact representations of human mobility without geocoordinates, can be used to personalize location-based services. While they are more privacy-preserving than raw tracking data, it was shown that they still hold a considerable risk for users to be re-identified solely by the graph topology. However, it is unclear how this risk depends on the tracking duration. Here, we consider a scenario where the attacker wants to match the new tracking data of a user to a pool of previously recorded mobility profiles, and we analyze the dependence of the re-identification performance on the tracking duration. Our experiment uses a one-year-long tracking dataset of 137 users divided into subsets of varying durations (1, 2, 4, 8, 16, 20, 24, and 28 weeks). We find that the top-1 re-identification accuracy is between 0.41% and 20.97% depending on the pool- and test-user tracking duration. We further show that re-identification performance is affected by both the pool duration and the test-user tracking duration, it is greater if both have the same duration, and it is not significantly affected by socio-demographics such as age or gender, but can to some extent be explained by different mobility and graph features such as the radius of gyration and the hub-size in graphs depending on the durations. Overall, the influence of tracking duration on user privacy has clear implications for data collection and storage strategies, even if only minimal information is being used. We advise data collectors to limit the tracking duration or to reset user IDs regularly when storing long-term tracking data.

* Equal contribution

3.3.1 Introduction and background

Companies are increasingly gathering and using spatio-temporal location data from personal mobile devices. User location data have substantially improved location-based services (LBS) and personalized offers (Keßler and McKenzie, 2018). However, detailed mobility traces collected from individuals may contain sensitive personal data associated with high privacy risks (Banerjee, 2019; Primault et al., 2018). A particular concern is the increasing integration of user data from different sources (Thompson and Warzel, 2019), enabling companies to build more detailed and complete user profiles (Melendez and Pasternack, 2019). Therefore, identifiability (and matching) of individuals from different datasets is a critical dimension of data privacy risk (Keßler and McKenzie, 2018).

Previous studies showed that removing basic identity information from mobility traces is insufficient in this context, as users can be re-identified using the information on frequently visited locations (De Mulder et al., 2008; Gambs et al., 2014; Golle and Partridge, 2009; Montjoye et al., 2013; Rossi et al., 2015; Zang and Bolot, 2011). One solution proposed in the literature is to obscure the geographic coordinates to guarantee ϵ -differential privacy (Andrés et al., 2013; Duckham and Kulik, 2005; Haydari et al., 2021; Wang et al., 2017) or k -anonymity (Charleux and Schofield, 2020; Gruteser and Grunwald, 2003; Shokri et al., 2010; Sweeney, 2002). For reviews of geoprivacy attacks and protection methods, we refer readers to Kounadi et al. (2018) and Fiore et al. (2020). Nevertheless, location obfuscation and related methods only provide limited privacy protection. For example, Tong et al. (2022) extend the notion of “location uniqueness” to “trajectory uniqueness” and show that full trajectories may be exploited for improving re-identification, and Tu et al. (2019) argue that k -anonymity does not protect from a semantic inference about visited locations.

Another promising possibility for privacy-preserving storage and processing of individual tracking data is given with so-called *location graphs* or *mobility networks* (Raubal et al., 2021; Rinzivillo et al., 2014). In these graphs, nodes represent visited locations, and edge weights correspond to the number of observed movements between these locations. Graph representations offer several benefits: 1) they can be enriched with node and edge features based on the application needs, 2) they are compact and grow sub-linearly in size with increasing tracking duration, 3) they still provide rich insight into mobility behavior despite their compactness (Martin et al., 2023b; Rinzivillo et al., 2014; Wiedemann et al., 2022) and can be analyzed efficiently with graph neural networks for various applications such as activity purpose imputation (Martin et al., 2018).

However, the privacy and unique identifiability properties of individual mobility graphs are not well understood. Recently, Manousakas et al. (2018) showed that the graph topology of personalized mobility graphs, even when all coordinate and time stamp information is removed from its nodes, is often uniquely identifiable. In this paper, we build upon their work and aim to understand the dependency of privacy preservation on tracking duration. Intuitively, location graphs over short periods contain less information about users and may reduce the risk of deanonymization. To investigate this possibility, we divide a tracking dataset of 137 users into distinct periods of different durations and analyze attack scenarios where a new location graph is matched to a pool of location graphs of known users. Our

experiments indeed show that matching performance depends on the tracking duration of both pool data and new data; however, there is a considerable re-identification risk even with just a few weeks of tracking duration.

3.3.2 Materials and methods

Data and preprocessing

We analyze the time dependency of topology privacy on a high-quality tracking dataset, collected through the SBB Green Class 1 tracking study (Martin et al., 2019a). The study was conducted by the Swiss Federal Railways (SBB) to evaluate the impact of a mobility-as-a-service offer on individuals' travel behavior. Study participants are predominantly male with above-average income. All study participants were tracked over a full year using an application installed on their phone that segments tracking data into stationary periods called *staypoints*, labeled with activity purpose, and movement behavior called *triplelegs*, labeled with transport modes. All preprocessing is done in Python and PostgreSQL using the Trackintel movement data processing library (Martin et al., 2023c). The staypoints are clustered into locations with the DBSCAN algorithm with the parameter $\epsilon = 30m$, and a minimum number of one point per cluster, i.e. each staypoint is assigned to a location. The Trackintel library merges consecutive staypoints and triplelegs into trips as long as they are not interrupted by an activity (staypoints with duration >25 min or labeled with a purpose other than wait and unknown) or by a temporal gap (here 25 minutes). Finally, when constructing the graph, we filter out users with low tracking coverage during the selected time period. The users are required to have a tracking coverage of at least 70% in at least one-third of the days. In our experiments, this leads to a varying number of 132-137 users depending on the time periods used.

Based on the sequence of locations and trips of a user, we construct the individual location graph (or mobility network) as described by Manousakas et al. (2018): In the graph $G(V, E)$, each location is one node, and each trip between two locations increases the weight of the directed edge by one. The edge weight $w(e)$ thus corresponds to the number of transitions during the observation period. To analyze the impact of different tracking periods, we build the graphs on subsets of the dataset that are created by binning the dataset into non-overlapping time periods of 1, 2, 4, 8, 16, 20, 24, and 28 weeks (see Figure 3.20).

Furthermore, we use the SBB Green Class 2 study (Martin et al., 2019a), which was a smaller follow-up study where 50 different participants were tracked under similar conditions for a full year directly after the Green Class 1 study. The data is processed in the same way as the Green Class 1 data, but due to the lower number of users, we will only use this dataset to validate our results in Section 3.3.3 in the Paragraphs *Ablation of approximate graph matching workflow*, *Validation of matching methodology based on related work and the Green Class 2 dataset* and *Intra-user vs inter-user variability of re-identification performance*.

Feature based graph matching

Graph matching describes the problem of either identifying if two graphs are isomorphic (exact graph matching) or identifying the best match from a set of candidate graphs (inexact graph matching) (Riesen et al., 2010). The exact solutions for both problems are computationally intractable, therefore we rely on heuristics to accomplish inexact graph matching. Related works have proposed so-called R-convolution graph kernels (Haussler, 1999) that measure the difference between two graphs in terms of counts of certain substructures, such as paths. Similarly, we compare the distributions of selected graph features to approximate the graph similarity. We represent each graph in a fixed-size vector $v(G)$ that expresses graph characteristics, e.g., the distribution of node in-degrees. Two graphs G_1 and G_2 are compared in terms of the distance between their vector representations, $d(v(G_1), v(G_2))$. As distance metrics d , we test a simple Mean Squared Error (MSE), Kullback-Leibler divergence, and Wasserstein distance.

We experiment with five vector-based graph representations $v(G)$:

- $v_{indegree}$: Distribution of (unweighted) node in-degrees, i.e., the number of connections of one location to other locations. The distribution of in-degrees over the 20 most popular locations is used.
- $v_{outdegree}$: Similar to the in-degree, the distribution of out-degrees over the 20 locations with the highest out-degree is computed.
- $v_{transition}$: The distribution of transition weights over the 20 most popular trips. Intuitively, some users commute between very few locations more frequently than other locations, whereas some users transit more evenly among locations (Pappalardo et al., 2015).
- $v_{shortest_path}$: The distribution of shortest-path lengths in the graph. All-pairs shortest paths were computed with the Floyd-Warshall algorithm (Floyd, 1962; Warshall, 1962). The ratio of shortest paths with length x for $x \leq 10$ is reported in $v_{shortest_path}$.
- $v_{centrality}$: The betweenness centrality (Freeman, 1977) of a node denotes its centrality in terms of network hops with respect to other nodes, which is bounded between 0 and 1. Since many nodes have low centrality in mobility graphs, we construct 10 bins from 0 to 1 in log space and report the number of nodes per centrality bin.

Finally, we concatenate all five graph descriptors into one combined vector v_{comb} .

Experiment design

We analyze the following privacy attack scenario: The adversary is a data broker with access to a pool of users and their tracking data. The attacker then gets access to additional tracking data of a test user, which she wants to match to the correct user in the pool to create a combined user profile. All tracking data are represented as weighted and directed individual location graphs without node or edge features such as coordinates. In the following, we define u_i^{pool} ($i \in [1..n]$) as the i -th user in a pool of n users, and u_j^{test} ($j \in [1..m]$) as a user

of the test dataset, $D_{test} = \{u_j^{test}\}$. Let G_i^{pool} and G_j^{test} further denote the corresponding location graphs.

The adversary now aims to find the best match out of the pool users for each test user u_j^{test} . This is accomplished by computing the distance of the graph descriptors presented in Section 3.3.2. The pairwise distances from a test user to all users of the pool are computed as $d(v(G_j^{test}), v(G_i^{pool}))$ and the pool users are ranked according to their distances. As a result, we obtain the rank assigned to the true match of a user in the pool. In other words, we are only interested in the rank that was assigned to the user in the pool that corresponds to the test user ($u_i^{pool} = u_j^{test}$) and the assigned rank $r_j = r(u_j^{test})$ means that this user had the r_j -highest similarity to herself compared to all other users in the pool.

To obtain statistically robust results, we evaluate the scenario on all possible tracking period combinations for the pool and the test user. Figure 3.20 gives an overview of the experimental setup and demonstrates that the tracking period combinations are not unique. For example, for our total tracking time of 56 weeks, there are 14 distinct 4-week periods and 7 distinct 8-week periods. We do not evaluate all possible combinations (here 98) but regard only combinations where the test user is matched to the closest, directly preceding tracking period in the pool. This choice of valid pool and test user pairs is exemplified by the black arrows in Figure 3.20. In Section 3.3.3, we additionally consider periods that are not *directly* successive in order to understand the effect of temporal gaps between the pool and test user.

For every valid time bin combination for a given combination of tracking periods, we match every available test user to the users from the pool and evaluate the matching success using the metrics introduced below. All code for the experiments is publicly available²⁰, however, we can not publish the tracking dataset to protect the privacy of the study participants.

Metrics for re-identification performance

To evaluate the success of the matching attack, we employ two metrics: the top-k matching performance and the mean reciprocal rank (MRR) (Voorhees, 1999). Both rely on the rank assigned to the true match of a test user in the pool as introduced above, $r(u_j^{test})$.

We then report the top-k matching performance in one set of test users D_{test} as

$$Acc(D_{test}, k) = \frac{1}{|D_{test}|} \sum_{u_j \in D_{test}} \mathbb{1}\{r(u_j^{test}) \leq k\}.$$

This considers a match as successful if the true match of the test user is among the top-k closest users in the pool.

²⁰https://github.com/mie-lab/topology_privacy

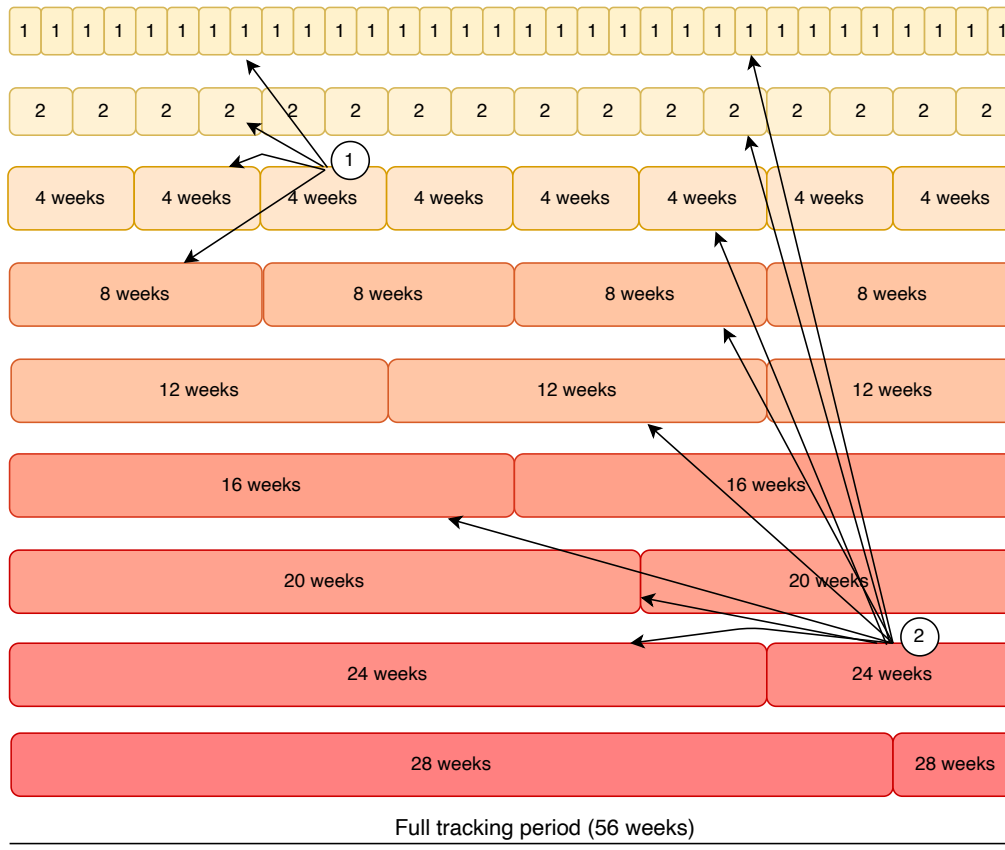


Fig. 3.20.: Experimental setup: The tracking data, comprising 56 weeks, are split into non-overlapping bins of varying duration. In the attack scenario, new tracking data from one period is matched to a pool of users at a previous time period. In example 1) the test data of four weeks length can be compared to the pool in the preceding 1, 2, 4 and 8 weeks. In the second example (marked as 2), a test user with tracking data from the second 24-week period is matched to users from all directly preceding tracking data, which includes one from each tracking duration except for 28 weeks.

Furthermore, we use the MRR as a second evaluation metric, defined as the average of the inverse of the ranks in a test dataset. It is a common metric in information retrieval and re-identification tasks (Craswell, 2009). The MRR of a test set is

$$MRR(D_{test}) = \frac{1}{m} \sum_{u_j \in D_{test}} \frac{1}{r(u_j^{test})}.$$

The MRR can be interpreted as the harmonic mean of the ranks, with the property that good matches (high rank) have a much higher influence than bad matches (low rank).

3.3.3 Results and discussion

We run the experiment described in Section 3.3.2 for all combinations of tracking periods and consecutive start times, resulting in 827 combinations. For each of these combinations, we attempt a matching for every user available in the dataset, which results in over 13 million user-to-user comparisons (Green Class 1). We find that the best matching performance is achieved with the combined graph descriptor v_{comb} and the mean squared error (MSE) as the similarity metric d . See Table 3.8 and Section 3.3.3 for more details on this choice.

In the following, we report the MRR and top-k matching accuracy for each combination of the pool- and test-user tracking duration. We report the average result and the standard deviation if several accuracy results for a tracking period combination are obtained (due to multiple time bin combinations).

Effect of tracking period on re-identification performance

Figure 3.21 shows the average matching performance and the standard deviation for all duration combinations of the pool and the test users. All metrics show a significant dependency on both the duration of the pool and the duration of the test user data. This result implies that privacy-friendly applications should be designed such that their tracking duration is as short as possible. This is especially true when new tracking data is to be collected because a privacy-concerned person does not have control over the duration of the pool in our scenario, as the pool represents data already collected by a third party.

Furthermore, even for the shortest tracking duration that was tested (i.e., one week combined with one week), the re-identification capability of our simple matching strategy is substantially better than random (see Figure 3.21). A random rank assignment would result in a top-10 accuracy of 7.6%, compared to the accuracy of 19.4% from the shortest tracking duration. Thus, the graph representation, even without any additional context or coordinate information, is not anonymous, which is in line with the conclusion reported from (Manousakas et al., 2018).

We further analyzed the importance of the pool duration, the test user duration, and the difference between their durations, using linear regression with the duration as the independent variable and the average performance as the dependent variable. The resulting coefficients are shown in Table 3.7. While both duration variables positively impact the performance, the influence of test duration is slightly stronger. For every additional week of test tracking duration, the top-10 identification accuracy increases by 1.06% on average. As the pool is not under the user's control, a potential solution to minimize the privacy risk is to require data brokers to reset user IDs after a specific tracking period. Notably, Table 3.7 also reveals a major effect from the similarity of pool and test tracking duration, corresponding to the strong performance on the diagonals in Figure 3.21. This can be explained by the higher similarity of graphs constructed from the same tracking duration, making it easier to match the correct user.

For the interpretation of the results, it is important to note that the results with small bins are statistically more robust than those with large bin combinations because more

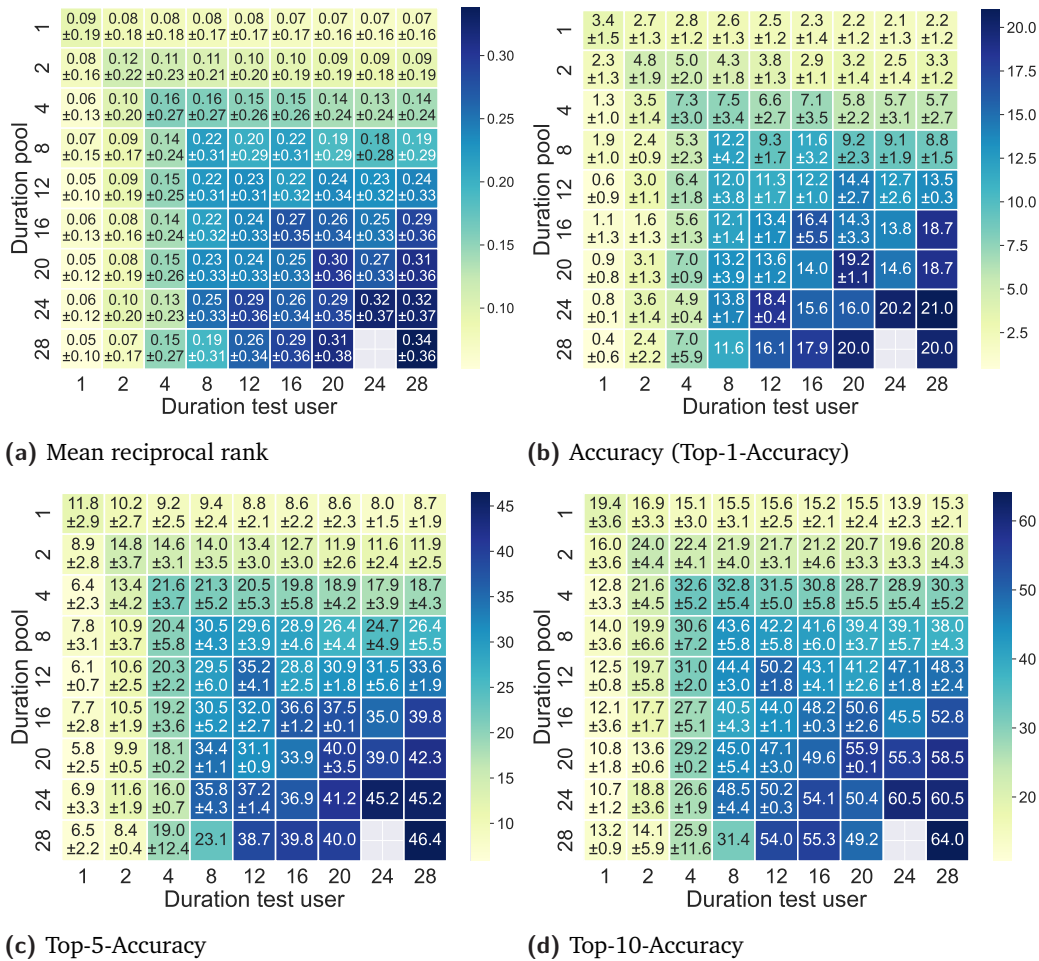


Fig. 3.21.: Dependency of matching performance on tracking duration. Top-k accuracy and MRR increase with both the tracking duration of the pool users as well as the test user.

bins are available. For several combinations of large bins, only one trial was available; therefore, no standard deviation was reported, and no distinct time bins were available for the combination of 28 weeks pool duration and 24 weeks test tracking duration.

	test duration	pool duration	absolute difference between pool and test duration	intercept	R2 score
MRR	0.01	0.01	-0.01	0.09	0.90
1-Accuracy	0.41	0.43	-0.42	2.42	0.90
5-Accuracy	0.83	0.82	-0.87	10.89	0.89
10-Accuracy	1.06	0.97	-1.10	18.76	0.87

Tab. 3.7.: Regression analysis of the effect of the pool- and test-user tracking duration on the matching performance. Both positively affect the re-identification performance (=negative impact on privacy); however, the effect of the test duration is slightly higher. The matching performance is higher if the absolute difference between the pool and test user duration is low. All results are significant (p-values $\ll 0.01$).

Ablation of approximate graph matching workflow

In Section 3.3.2, we proposed several graph descriptors to calculate the distance between graphs. Table 3.8 lists the matching performance of different graph features and distance functions. We note that the distance function does not strongly affect the matching performance. In contrast, the features result in very different re-identification abilities. The transition weight and in-degree distribution are the most useful features, whereas node centrality obtains low matching capability. Based on the results in Table 3.8, we chose the MSE of all features combined, as this performs best on average according to three out of four error metrics. While our focus is on the time-dependency of privacy preservation, future work could analyze the limits of re-identification of location graphs by using more complex matching methods such as deep graph kernels (Yanardag and Vishwanathan, 2015).

Distance metric d	$v(G)$	Recip. rank		1-Accuracy		5-Accuracy		10-Accuracy	
		Mean	Max	Mean	Max	Mean	Max	Mean	Max
KL-divergence	transition	0.10	0.19	3.92	9.60	13.19	28.46	21.73	40.80
	in_degree	0.10	0.20	3.29	10.40	12.05	24.80	20.83	41.60
	out_degree	0.09	0.17	3.07	7.32	11.74	24.80	21.06	36.29
	shortest_path	0.07	0.11	1.77	4.13	7.45	14.52	13.59	26.61
	centrality	0.04	0.06	0.79	2.02	4.14	8.06	8.52	15.32
	combined	0.16	0.35	7.96	23.33	21.41	51.20	31.25	62.40
MSE	transition	0.10	0.19	3.89	9.24	12.98	28.00	21.80	41.60
	in_degree	0.10	0.18	3.49	9.76	12.22	22.40	20.73	35.48
	out_degree	0.09	0.16	2.76	6.61	11.48	25.60	20.52	39.20
	shortest_path	0.07	0.11	1.89	4.04	7.80	15.32	14.16	29.03
	centrality	0.05	0.07	1.17	3.25	5.26	11.16	9.67	16.74
	combined	0.17	0.34	8.40	20.97	22.36	46.40	32.73	64.00
Wasserstein distance	transition	0.10	0.19	3.82	9.21	13.32	30.08	21.50	41.60
	in_degree	0.10	0.19	3.38	10.40	12.18	25.60	20.96	36.00
	out_degree	0.09	0.17	2.96	8.13	11.64	24.00	20.70	40.32
	shortest_path	0.06	0.11	1.64	4.20	6.48	16.00	11.94	24.00
	centrality	0.05	0.09	1.14	4.13	5.03	11.38	9.69	16.53
	combined	0.15	0.36	7.14	24.00	19.71	52.80	28.94	61.60
Sum all metrics	combined	0.16	0.36	8.07	22.50	21.81	52.80	31.67	62.40

Tab. 3.8.: Matching performance of different combinations of features, distance functions, and evaluation metrics. The highest matching accuracy is achieved with an R-convolution kernel that computes the MSE between all graph-features distributions combined.

Validation of matching methodology based on related work and the Green Class 2 dataset

We first validate our results by conducting the same experiment on the Green Class 2 data described above. The full pool-user-duration matrices can be found in Figure A.1 in Appendix A.2.1. The results show a similar dependency on pool and tracking duration, but, due to the lower number of users, the re-identification accuracy is generally higher (up to 82% top-10 accuracy) and the results are less stable.

We further compare our results on both datasets to the results reported by Manousakas et al. (2018). In their longitudinal study, Manousakas et al. (2018) split the tracking data user-wise into two parts at a random point in time, sampled uniformly between 30% and 70% of the whole period (around one year). The most comparable experiment from our study is the one where both the pool and the test duration are 28 weeks. Following the evaluation by Manousakas et al. (2018), we show the distribution of ranks and the “privacy loss” in Figure 3.22. Although the absolute ranks are not informative due to the different pool sizes (132 users / 27 users²¹ for our dataset versus 1500), the re-identification ability can be compared in terms of the shift of true rank. Specifically, the mean of the true rank is shifted from 62 (random) to 17.1 (informed adversary) for Green Class 1 and from 13 (random) to 7.6 (informed adversary) for Green Class 2, whereas the experiment in (Manousakas et al., 2018, p. 13) yields a shift from 750 to 140. The study by Manousakas et al. (2018) also reported a median privacy loss of 2.52 which means “the informed adversary can achieve a median deanonymization probability 3.52 times higher than an uninformed adversary” (Manousakas et al., 2018, p. 14). In our experiments, the median privacy loss is 2.85 for the Green Class 1 data and 1.31 for the Green Class 2 data. Overall, we reproduced the results successfully and extended their results with additional analysis of the impact of tracking durations.

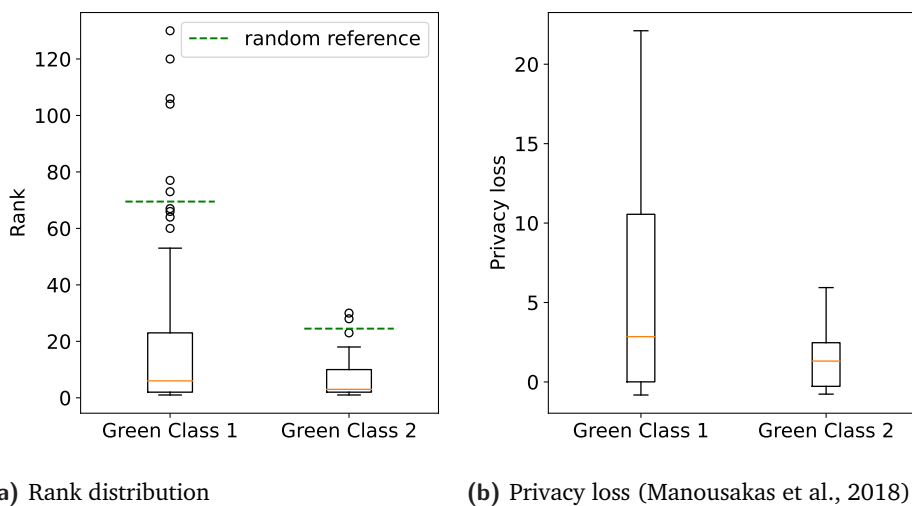


Fig. 3.22.: Evaluation of rank distribution and privacy loss as proposed by Manousakas et al. (2018).

Intra-user vs inter-user variability of re-identification performance

The main results of this study (Figure 3.21) are reported as average matching performance. We now further analyze the sources of variance of the matching performance by analyzing the variance of the rank assigned to users during the matching. In particular, we aim to answer the following question: Is the variance due to strong differences between users

²¹For long time bin durations, not all users matched the criteria set for the tracking coverage.

(e.g., easy-to-match vs. hard-to-match users), or due to a change in a user’s re-identification ability over time? To answer this question, we calculate the standard deviation between different users in the same timesteps (inter-user) and for the same user over several timesteps (intra-user).

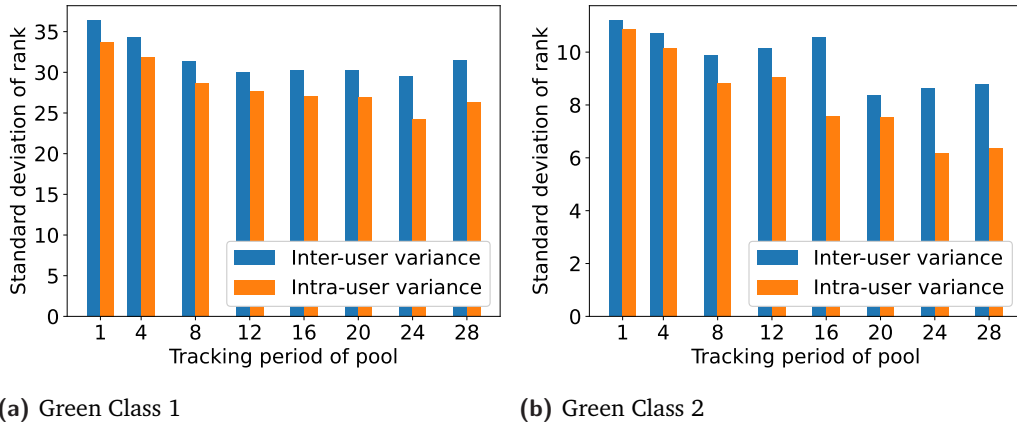


Fig. 3.23.: Inter vs intra person variability of matching performance. The variance over users is higher than the variance over time bins. Intra-user variance decreases with growing tracking duration.

Figure 3.23 shows that the inter-user standard deviation is consistently higher than the intra-user standard deviation for both datasets. This indicates the existence of user groups that are consistently hard or easy to match. Moreover, the intra-user standard deviation in general decreases as the tracking duration increases for both datasets, which can be explained by the higher stability of long-term location graphs.

Factors that explain re-identification performance

Given the high variance in the re-identification ability over users, we further analyze features that could drive the degree of recognition of a user. For this purpose, we compute features commonly used to describe individual mobility behavior, such as the radius of gyration, the typical trip distances, and the entropy of location sequences (random and real entropy) (Song et al., 2010a). Additionally, we compute graph features proposed by Martin et al. (2023b), which describe the complexity and centeredness of the location graphs. Last, we regard socio-demographics extracted from surveys in the context of the Green Class 1 and Green Class 2 studies, namely age, gender and whether the user subscribed to a public transport subscription in Switzerland (PT). Note that all features are computed as a single value for all users since there is only one value per user for sociodemographics and classical mobility features. We use the average value over both 28-week bins for the graph features to describe the user’s stable graph topology.

In Table 3.9, the coefficients of a regression analysis with the above-presented features as independent variables and the normalized rank as the dependent variable are given. The normalized rank is the true user’s rank in the matching process, normalized by the total

number of users, which allows to combine the users of Green Class 1 with the ones from Green Class 2 in this study. We further checked the correlation r between attributes to exclude potential collinearity issues, but $r < 0.6$ for all pairs. A significant positive coefficient indicates that a feature hampers the re-identification ability since it leads to a higher rank. The model is fitted separately for each tracking duration (1, 2, 4, ..., 28), whereby we only consider scenarios with the same pool- and user tracking duration, corresponding to the diagonal of the matrices in Figure 3.21, and we average all available rank predictions for each user (i.e., average over time bins).

According to the regression coefficients (Table 3.9), socio-demographics do not affect the rank significantly. A higher radius of gyration makes a user harder to identify which might be related to an increased variability of the location graph over time due to a higher level of travel activity. For long durations, a high random entropy increases the identification performance. The random entropy increases if time is spent at many different locations which increases the complexity and uniqueness of a graph and therefore makes it easier to match. The graph features, in particular the journey length, also significantly affect the rank, but in an unexpected direction: More star-shaped graphs, indicated by low journey length, low hub size, and high transition β , yield higher ranks.

Duration	Classical mobility features					Graph features				Sociodemographics			R^2
	Const.	Radius of gyration	Random entropy	Real entropy	Median distance	Journey length	Hub size	Degree β	Transition β	Age	Sex	PT	
1	32.62 (*)	1.54 (*)	-0.73	0.48	-0.04	-0.83	-0.65	0.04	0.77	-0.25	-0.39	-0.48	0.13
2	29.8 (*)	1.83 (*)	-0.62	0.19	0.06	-1.46 (*)	-0.1	-0.6	-0.46	-0.09	-0.61	-0.64	0.11
4	26.01 (*)	2.7 (*)	-0.93	-0.19	0.51	-2.5 (*)	-0.44	-1.06	-1.17	0.14	0.2	-0.98	0.15
8	20.56 (*)	2.65 (*)	-3.2 (*)	2.35	-1.44	-2.75 (*)	-1.23	-3.41 (*)	-0.01	-0.39	0.84	-1.16	0.17
12	19.24 (*)	2.52 (*)	-3.64 (*)	-0.18	-2.73 (*)	0.52	-1.92	0.35	2.28	1.63	0.6	1.09	0.14
16	20.33 (*)	2.98	-1.67	1.68	-1.05	-2.71	1.16	-3.66	2.34	0.56	0.95	0.88	0.07
20	17.75 (*)	1.92	-3.68	1.43	-0.51	-0.45	2.12	-1.8	0.42	1.69	-0.21	0.82	0.04
24	16.3 (*)	2.82	-4.18	1.63	0.34	1.72	0.93	-0.26	5.62 (*)	0.71	-0.12	3.02	0.08
28	16.53 (*)	2.58	-7.27 (*)	2.99	0.53	0.43	4.7 (*)	-1.45	6.43 (*)	0.09	0.43	3.46	0.12

Tab. 3.9.: Effect of mobility behavior and socio-demographics on re-identification accuracy, i.e., the rank of a user. A linear model is fitted and the coefficients are reported. Significant coefficients (p-value below 0.05) are marked with (*).

Influence of temporal difference between pool and user tracking period

The experiments reported so far were restricted to consecutive time periods (see Figure 3.20). Here, we further analyze the effect of a temporal gap between the tracking periods. Since the number of possible combinations becomes very high in this setting, we restrict the analysis to one pool- and user duration combination and analyze the 1.56 million combinations where pool- and user duration are four weeks and the pool was recorded before the user duration. Figure 3.24 shows the results, where the top-10 re-identification accuracy is shown by the temporal gap. As expected, the matching performance decreases as we increase the duration of the gap. However, it stabilizes already at around 16 weeks between pool and test user, and remains surprisingly high even for the longest gap of 56 weeks. This finding implies that saved location data can be exploited by an attacker for a long time.

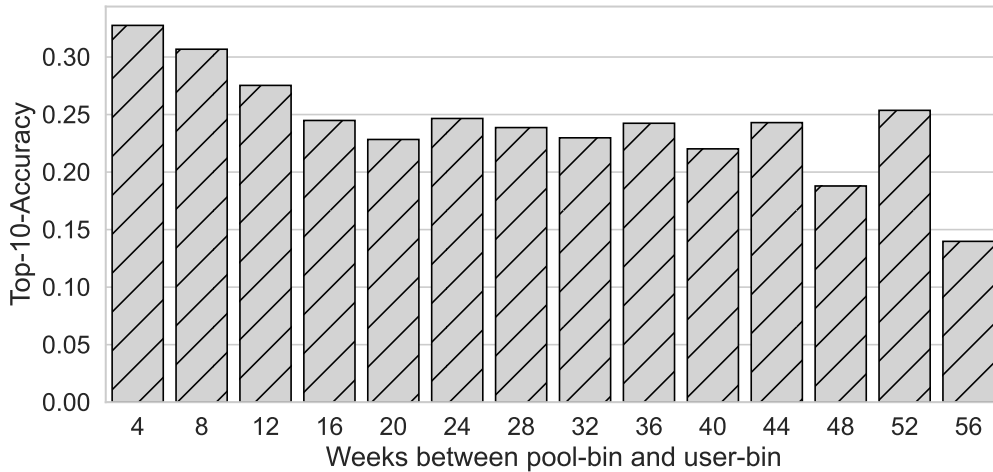


Fig. 3.24.: The re-identification accuracy decreases when there is a larger temporal gap between pool-bin and user-bin. However, the accuracy converges slowly and retains more than half of its former value even after one year.

3.3.4 Conclusion

In this work, we present a set of experiments to analyze how tracking duration influences the re-identification ability of individual location graphs. Our experiments on time-binned subsets of one-year tracking data show that the tracking duration indeed has a strong effect on the success of a privacy attack, with the re-identification accuracy at the longest tracking duration (28 weeks) being more than 3 times higher than when matching 1-week tracking data. We further show that the re-identification ability increases in roughly equal parts with increased tracking duration of the pool of candidate users on the one hand, and increased tracking duration of the test user on the other hand. Therefore, privacy-friendly applications should only require tracking data over periods that are as short as possible, and data brokers should be required to reset the user IDs of their data regularly to limit the pool duration. On top of that, long-term storage of tracking data should be impeded, since the re-identification accuracy only slowly decreases with increasing time between pool and test tracking period.

More generally, we confirm results from Manousakas et al. (2018) that location graphs without coordinates or additional context information are sufficient to re-identify users with a success rate significantly higher than random. At the longest tracking duration, the de-anonymization probability of an informed adversary is 3.85 times higher than the one of an uninformed adversary for our dataset. Our work reveals many opportunities for further work on location-graph privacy. For example, we found that certain users are consistently hard or easy to be identified. Characterization of these user groups should be explored in future work. We take a step in this direction with our analysis of the relation to different mobility-behavior features and socio-demographics, but our results hint at more complex characteristics that make a user hard to re-identify. Evidence from more diverse datasets may help to find such influence factors. The reproduction of our experiments on

new datasets is straightforward as the individual location graphs have very few requirements (e.g., no specific features or labels needed). At the same time, future work could also regard the re-identification risk of more complex location graphs, e.g., amended with temporal information.

Finally, it is important to mention that we only employed a simplistic matching strategy, and a more sophisticated matching approach, such as learning graph similarities with deep neural networks (Ma et al., 2021), could lead to even higher success rates for matching. The results should therefore be considered as a lower bound of possible matching success. The presented analysis however augments the understanding of the privacy risk of tracking data - even if it is reduced to topology - and can improve the regulation of anonymization practices.

Informed consent statement

The Green Class 1 and 2 studies were conducted by SBB. The participants provided informed consent to be tracked over the study period, and for their data to be shared for research purposes.

Acknowledgement

This work was supported by the ETH Zurich Foundation [MI-01-19] and by the Hasler Foundation through grant 1-008062;

Supporting modal shift (MaaS)

4.1 Graph-based mobility profiling

The following section is a reprint of the publication:

Henry Martin*, Nina Wiedemann*, Daniel J. Reck, and Martin Raubal Martin (2023). “Graph-based mobility profiling”. *Computers, Environment and Urban Systems* 100, p. 101910. DOI: 10.1016/j.compenvurbsys.2022.101910.

Conceptualization: HM, NW; Methodology: HM, NW; Software (preprocessing): HM; Software (experiments): NW; Data collection: HM, DR; Writing - Original Draft: HM, NW; Writing - Review & Editing: DR, MR; Supervision: MR

The content is reproduced ‘as is’, however, formatting changes and corrections of spelling have been applied.

Abstract

The decarbonization of the transport system requires a better understanding of human mobility behavior to optimally plan and evaluate sustainable transport options (such as Mobility as a Service). Current analysis frameworks often rely on specific datasets or data-specific assumptions and hence are difficult to generalize to other datasets or studies. In this work, we present a workflow to identify groups of users with similar mobility behavior that appear across several datasets. Our method does not depend on a specific clustering algorithm, is robust against the choice of hyperparameters, does not require specific labels in the dataset, and is not limited to specific types of tracking data. This allows the extraction of stable mobility profiles based on several small and inhomogeneous tracking data sets. Our method consists of the following main steps: Representing individual mobility using location-based graphs; extraction of graph-based mobility features; clustering using different hyperparameter configurations; group identification using statistical testing. The method is applied to six tracking datasets (Geolife, Green Class 1+2, yumuv and two Foursquare datasets) with a total of 1070 users that visit about 3’000’000 different locations with a total tracking duration of over 200’000 days. We can identify and interpret five mobility profiles that appear in all datasets and show how these profiles can be used to analyze longitudinal and cross-sectional tracking studies.

* Equal contribution

4.1.1 Introduction

Individual motorized transportation is a major contributor to global GHG emissions (Chapman, 2007; Creutzig et al., 2015) and linked to additional problems such as the creation of microplastics (Evangelidou et al., 2020), injuries, an increase of impervious cover for infrastructure (Gössling, 2020), and more traffic and congestion which already results in high economic costs (Reed, 2019).

Tackling these problems requires covering people's growing mobility needs using fewer resources like energy, cars, or space. This is the goal of several novel mobility concepts such as MaaS but apart from being more sustainable, these services will need to be comparable with personal cars in terms of comfort and flexibility in order to convince people to change their mobility behavior.

This challenge will require knowledge about the mobility behavior of people and the ability to predict it in the near future in order to optimally allocate shared mobility resources. With the recent success of machine learning algorithms (LeCun et al., 2015), research in computational movement analysis (Long et al., 2018) shifted towards using machine learning methods to support data interpretation (e.g., labeling (Toch et al., 2018), clustering tasks (Ben-Gal et al., 2019; Jonietz et al., 2018)) or prediction tasks (Kreil et al., 2020; Kumar and Raubal, 2021; Luca et al., 2021).

While large tracking datasets of human movement have become available in recent years, they are oftentimes unlabeled (Chen et al., 2016a), thereby preventing the use of supervised machine learning methods. Furthermore, available datasets are often different in key properties such as the duration of a tracking study, the deployed tracking technology, and its spatio-temporal resolution of trackpoints, and sample biases.

A particularly difficult problem in this situation is the identification of stable groups of users with similar mobility behavior, which are comparable across datasets. Finding such mobility types can help us to enhance our understanding of mobility behavior (Pappalardo et al., 2015), to measure regional similarities between cities (McKenzie and Romm, 2021) or neighborhoods (Calafiore et al., 2021), and to detect changes in mobility behavior over time (Hong et al., 2021; Jonietz et al., 2018). However, existing solutions are often based on dataset-specific features and can either not be applied to different datasets, or results from different datasets are not comparable.

In this work, we develop an approach to identify mobility types with minimal dataset specific assumptions, which facilitates the application to different datasets. The problems stated above are solved with a graph-based approach that uses a compact representation, does not require labeled data, and allows to easily merge different datasets. Our method is tested on six datasets to demonstrate its general validity independent of the specifics of one tracking study.

In summary, our contributions are the following:

- We propose a set of features that are based on a compact graph representation. They describe integral dimensions of individual mobility behavior and are robust to dataset properties such as tracking duration or spatio-temporal resolution of trackpoints.

- We develop a method that uses statistical testing on multiple clustering results of the same dataset and yields stable user groups.
- We apply our method to six tracking datasets and extract five mobility profiles that appear in all datasets. These profiles are robust against the parameters and initialization of the clustering algorithm.
- We demonstrate in two applications how to use graph-based mobility profiles to analyze longitudinal and cross-sectional tracking studies.

The remainder of this paper is structured as follows: In Section 4.1.2 we describe related research on human mobility profiling, clustering, and graph representations. In Section 4.1.3, our graph features and the clustering approach are explained. Next, in Section 4.1.4 the data and preprocessing steps are outlined, and the results and applications are presented and discussed in Section 4.1.5. Section 4.1.6 includes further experiments that validate our proposed methodology. Finally, we summarize our conclusions in Section 4.1.7.

4.1.2 Related work

Representing individual human mobility

In transport planning, human mobility is commonly modeled based on the hypothesis that travel demand is derived from the need to perform different activities at different locations (Jiang et al., 2017). This activity-based perspective interprets travel demand as a result of people’s decisions whether, where and when to perform activities (Axhausen and Gärling, 1992; Castiglione et al., 2015). In practice, data about human mobility are often collected passively to avoid asking users to perform time consuming labelling tasks. Therefore, additional information such as activity labels are often not available in datasets (Chen et al., 2016a).

To circumvent this problem, most approaches fall back on available tracking information such as the activity location as proxy for the *true* activity. A common way to represent an individual’s mobility behavior is based on a sequence of visited locations such as the concept of location history mentioned in (Zheng et al., 2009) or the concept of lifeline beads introduced in (Hornsby and Egenhofer, 2002). In this case the movement profile of a person is a list of locations, ordered by the time of visitation. Depending on the definition, the model can include context data for each visit such as temporal information like start time or duration, spatial information such as coordinates, or semantics such as an associated POI category. Some further variations of this model can be found in (Bhattacharya and Das, 2002).

A major downside of this representation is that it grows quickly in size because the raw data are appended to the sequence for every visit. Furthermore, this representation is privacy sensitive as it contains information such as the time and duration of each individual visit. Representing individual movement profiles using a location graph of visited locations can solve these problems as it can be stored and processed efficiently. In such a graph,

nodes correspond to visited locations (as a proxy for activities) and edges correspond to the transition count between two locations. Alessandretti et al. (2018) showed that people only visit a limited set of locations that slowly evolves over time and Schneider et al. (2013) demonstrated that our daily mobility can be described by a small set of sequential location visiting patterns (motifs). Furthermore Yan et al. (2017) created a model based on a graph of visited locations that reproduced important scaling laws of human mobility. This provides evidence that a personalized graph that is based on the visited locations can parsimoniously represent individual human mobility.

Graphs based on visited locations of individual persons have already been explored in the past. For example, (Zheng et al., 2008) transformed GPS tracking data into a graph representation to support the prediction of the transport mode of transitions between nodes. (Rinzivillo et al., 2014) transformed a large GPS tracking data set of about 150k vehicles from Tuscany into individual graph representations. They then combined structural features extracted from the graphs and classical features, such as length or duration, to show that including graph features increases the performance of trip purpose classification. Furthermore, (Martin et al., 2018) used graph representations of individuals in combination with graph neural networks to predict the distribution of activity labels at visited locations. Even though all these examples show promising results, the literature in this area is still sparse, especially with regards to unsupervised learning applications such as the identification of groups with similar mobility behavior based on graphs.

Clustering based on mobility behavior

Research on the identification of similarities based on movement data is mostly used for the discovery of previously unknown patterns and insights. Studies that do not focus on individuals often analyze movement at a city scale, such as in (Yuan and Raubal, 2012) where CDR data enriched with demographic data are used to classify different urban areas in a city in China. Their approach allows to identify areas of the city in which people move alike. Similarly, Ratti et al. (2006) analyze urban activities from mobile phone data in Italy, and Sulis and Manley (2018) use a combination of twitter data and smart card data to cluster places in London according to their travel activity patterns, which can be used to analyze the *daily rhythms* of places in a city.

Studies that focus on the movement of individuals usually present workflows that are used to mine patterns from specific situational datasets such as in (El Mahrssi et al., 2016) where the authors use public transport smart-card data to cluster users by their travel behavior with respect to time and frequency of trips. They identify 13 different passenger clusters which they further analyze to identify fine grained commute patterns. Xin and MacEachren (2020) present a methodology to extract mobility patterns to characterize different groups of football fans from twitter data. These studies are insightful; however, the methods often rely on very specific features such as the mode of transport of a trip or the content of a twitter message related to a trip destination. This makes the methods difficult to apply to different datasets or other types of tracking data where the required information might be unavailable. There are exceptions such as (Pappalardo et al., 2015) who group individuals

in *returners* and *explorers* based on their (k-) radius of gyration. These groups can be found across many datasets, but their work does not contain an approach to identify novel groups based on mobility behavior that generalize over different datasets, and further research showed that the results may depend on dataset properties such as the study duration (Wang et al., 2021). A potentially generalizable clustering approach is presented in Ben-Gal et al. (2019) who develop a lifestyle-based clustering method. They identify five patterns, namely *home*, *sweet home*, *working 9 to 5*, *traveling salesmen*, and *commuters* based on a large CDR dataset. Their activity-based approach could be applied to different datasets but they do not provide any analysis in this regard.

We fill this gap by presenting a workflow that allows to identify groups of individuals based on their mobility behavior. Our approach only requires minimal assumptions on tracking data, no labels and it permits to integrate different tracking datasets to identify their overall user groups.

4.1.3 Methods

Location based graph representation

In contrast to sequential tracking data, graph representations are compact, privacy-preserving, easy to process and still rich in information. Motivated by these properties, we choose a graph-based representation of individual human mobility. The graph is constructed based on the location sequence of a person, where a location is seen broadly as a place of interest that a person visited to perform an activity. The location sequence of a user $L_{seq} = [l_1^1, l_2^2, l_1^3 \dots, l_m^n]$ is a list of visits at locations ordered by their visitation time. L_{seq} contains n visits to m unique locations and l^i is defined as the i^{th} element of L_{seq} . Based on the location sequence we define the set of all visited locations as $L = \{l_1, l_2, \dots, l_m\}$ as the collection of all visited locations without repetition, therefore $|L| = m$.

The graph is constructed using unique locations as nodes and the number of direct transitions between pairs of nodes as weighted and directed edges. More formally, given the location sequence and the set of all visited locations of a person we define the weighted directed *individual location graph* as the pair (G_L, W) where

$$G_L = \{L, E\}, \text{ with } e = (l_i, l_j) \in E(G_L) \forall (l_i^k, l_j^{k+1}) \in L_{seq} \mid k < n \quad (4.1)$$

and the elements w of the weight matrix $W \in \mathbb{R}^{|L| \times |L|}$ are

$$w_{ij} = \sum_{k=1}^{n-1} \theta \quad \text{with } \theta = \begin{cases} 1 & \text{if } (l^k, l^{k+1}) = (l_i, l_j) \\ 0 & \text{otherwise.} \end{cases} \quad (4.2)$$

Examples of individual location graphs are given in Figure 4.1, where the transition count is proportional to the edge line width. Creating the graph representation only requires L_{seq} , it does not require specific label or context information. It can therefore be applied to

datasets that differ in properties such as the data collection methods (e.g., GNSS vs. CDR). A notable exception would be a significant bias in the sampling of visited locations as it might be present, for example, in public transport smart card data, where only visits at public transport stations are recorded and important locations such as the home location are missing systematically. This would lead to a systematically different graph structure that is incomparable to graphs based on other collection methods.

The graph representation compresses the location sequence of an individual significantly as we are mostly revisiting known locations (Alessandretti et al., 2018; González et al., 2008). However, despite the compression we know from previous work that the topology of location graphs is highly unique for each individual (Manousakas et al., 2018) and that human mobility can be well represented by substructures of such a graph (Schneider et al., 2013).

Graph-based mobility features

In order to characterize human mobility, we leverage the topology of the graph representation. Network characteristics extracted from the individual location graph can yield insights into a user's mobility behavior, despite relying on a compressed version of the raw movement data.

We propose a set of non-redundant and interpretable features that each represent separate dimensions of human mobility behavior. The features are motivated by a set of questions that address individual mobility behavior along the dimensions of the *role of base locations*, the *complexity*, the *regularity* and the *geometry* of individual mobility behavior. The numbers in square brackets link to the corresponding features from Figure 4.1 that relate to the specific question.

Role of home bases:

- Does a person have a single home base where he starts her trips from or several such bases? [1, 6]
- How home-centered is the person's behavior? Does he return home after each activity, or rather move from place to place? [1, 5]

Complexity:

- Are the activities of the person focused on few locations and trips, or distributed over many? [1, 2, 6]
- Are most trips of the user between the same locations? [2]

Regularity and geometry:

- Is the user flexible, or does he have a very regular mobility behavior? [5, 2]
- How far does a user usually travel? How far does he travel exceptionally? [3,4]

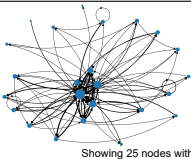
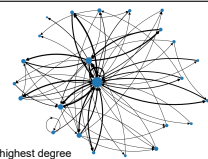
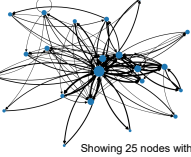
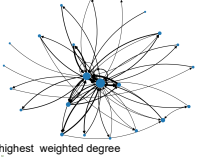
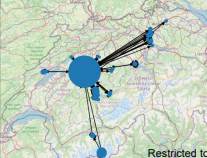
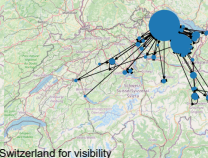
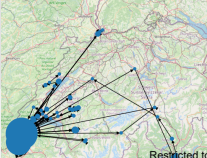
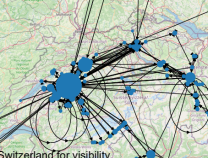
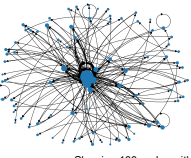
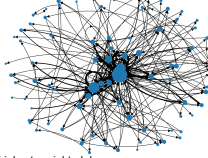
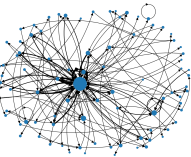
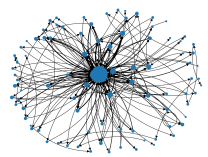
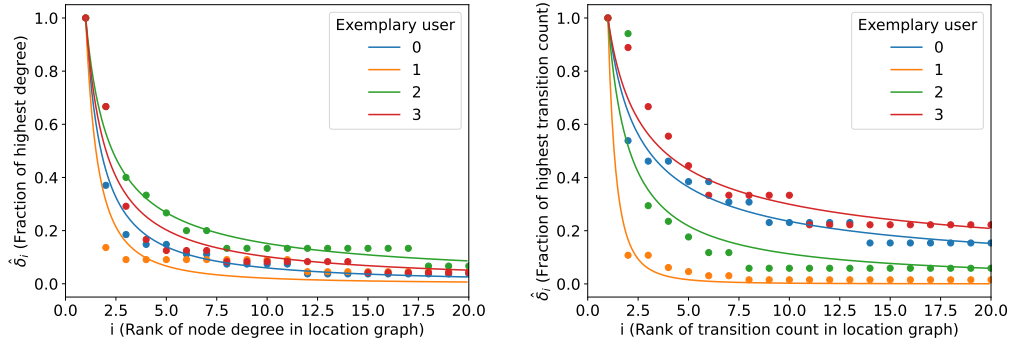
	Example low	Example high
<p>1) Degree dist. β</p> <p>Most trips starting from a single node (high) vs. many nodes have high degree (low)</p> $\hat{\delta}_i = i^{-\beta}$	 <p>Showing 25 nodes with highest degree</p>	 <p>Showing 25 nodes with highest degree</p>
<p>2) Transition count γ</p> <p>Most activities are concentrated to one edge (high) vs many edges are transitioned often</p> $\hat{t}(e)_i = i^{-\gamma}$	 <p>Showing 25 nodes with highest weighted degree</p>	 <p>Showing 25 nodes with highest weighted degree</p>
<p>3) Median trip distance</p> <p>Median linear distance between locations, weighted by transition count</p>	 <p>Restricted to Switzerland for visibility</p>	 <p>Restricted to Switzerland for visibility</p>
<p>4) Highest decile distance</p> <p>Highest decile of the trip distances, i.e. the value x where only 10% of the trips are longer than x</p>	 <p>Restricted to Switzerland for visibility</p>	 <p>Restricted to Switzerland for visibility</p>
<p>5) Journey length</p> <p>Average number of hops before returning to home location</p>	 <p>Showing 100 nodes with highest weighted degree</p>	 <p>Showing 100 nodes with highest weighted degree</p>
<p>6) Hub size</p> <p>Number of locations that account for 80% of all activity on a random walk</p>	 <p>Showing 100 nodes with highest weighted degree</p>	 <p>Showing 100 nodes with highest weighted degree</p>

Fig. 4.1.: Overview of graph features with an example graph from the Green Class 1 dataset for which the feature is rather low or rather high.



(a) Fit node degrees

(b) Fit transition counts

Fig. 4.2.: Power law fit for location graphs. The degrees are normalized by the highest degree found and ranked, and a power law is fitted.

Degree distribution β

The feature *degree distribution* β measures whether users start most of their trips from a single location or have several base locations. This will allow us to distinguish users who prefer to return home before visiting a new location from users who are more flexible, e.g., go to different places directly after being at work.

The (unweighted) out-degree of a node u in the location graph is defined as the number of locations that are visited starting from node u . If a single node has a very high out-degree compared to the other nodes, the user starts most of his trips from the same location. If multiple nodes have a high out-degree, the user has several *base locations* from where he starts his trips.

The importance of locations ranked by visitation frequency follows a power law distribution (González et al., 2008). We therefore propose to fit a power law to the distribution of ranked node out-degrees. Let $\delta_1, \delta_2, \dots, \delta_n$ be the degrees of a graph with n nodes, sorted in descending order ($\delta_1 \geq \delta_2 \geq \dots \geq \delta_n$). The values are normalized by the highest degree: $\hat{\delta}_i = \frac{\delta_i}{\delta_1}$. We then fit a simple power law following the rule $\hat{\delta}_i = i^{-\beta}$, where $0 \leq \beta \leq 5$ ($\beta > 5$ is not realistic in our data and was thus excluded to accelerate the optimization). Note that we defined $\hat{\delta}_1 = 1^{-\beta} = 1$.

A small β therefore describes graphs that have a multi-hub behavior, meaning that users start trips from several base locations. A high β characterizes user behavior that is centralized to few base locations. Figure 4.2(a) shows the power law fit for exemplary location graphs and Figure 4.1 shows example graphs of a user with high and a user with low *degree distribution* β .

Transition γ

This feature γ measures whether most trips (i.e., transitions between two locations) of a user are between the same locations or between many different locations. This information provides insights about the variety of a person's mobility behavior and visiting patterns. The transitions are stored as edge weights w_{ij} in the location graph, and their rank distribution

follows a power law. Thus, we can measure this feature with the parameter of a power law distribution fitted on the sorted transition weights, where the weights are normalized by the maximum ($\max_{i,j} w_{ij}$). Examples are shown in Figure 4.2(b). In the commuter example, we expect high *transition* γ , in contrast to low *transition* γ for a salesman because trips are very distributed.

Median and 9th decile trip distance

The *median trip distance* describes how far users usually travel in their day-to-day mobility while the *9th decile trip distance* describes the travel behavior for non-everyday trips. The distribution of trip distances follows a power law (Brockmann et al., 2006). It is therefore important to use robust measures such as the median or the 9th decile as metrics.

The median or 9th decile are computed over the distances between each pair of nodes, weighted by their transition counts w_{ij} . Note that we use the Haversine distance between node coordinates and no map-matched distances, since the latter cannot be recovered from the location graph alone and would require information on the transport mode.

Average journey length

The average *journey length* of a user measures how flexible a user is in moving from place to place. If a user oftentimes visits multiple places without returning home, his graph will become more connected and show more (highly weighted) edges. We measure this quantity as the number of visited locations in a journey, where a journey is defined as a simple cycle in the location graph that starts and ends at the home location following the definition of journeys from (Axhausen, 2007). We propose to approximate the journey length using a random walk in the graph: Starting at the home location node, we conduct a random walk of 5000 steps. Assuming that the current location is l_i , we select the next node j with probability $p(j) = \frac{w_{ij}}{\sum_k w_{ik}}$, i.e., proportional to the transition counts. When reaching a node that does not have any outgoing edges, or only an edge pointing at itself, we reset the random walk to home. The *journey length* is then defined as the number of steps between each consecutive encounter of the home location, excluding the resets.

Hub size

With the *hub size* feature we measure how many locations are visited on a regular basis and thereby account for a significant portion of the user's activity. It is therefore a measure of concentration of the mobility behavior (on few or many locations) and by that of the user's flexibility. In the graph, we measure the *hubbiness* as the number of nodes required to account for at least 80% of the total visits. The feature can be approximated from a random walk, similarly to the *journey length*. A random walk of 5000 steps yields a list of visited locations $L_{random}[l^1, l^2, \dots, l^{5000}]$. The locations are sorted by their occurrence count (c_l) in L_{random} , such that $c(\hat{l}_1) \geq c(\hat{l}_2) \geq \dots$. The (unnormalized) *hub size* h^* is the required number of locations such that their counts sum up to > 4000 occurrences (80% of 5000 steps), formally $h^* = \min_h : \sum_i^h c(\hat{l}_i) > 4000$. Since this number increases with the size of the graph, we normalize the feature by the square root of the total number

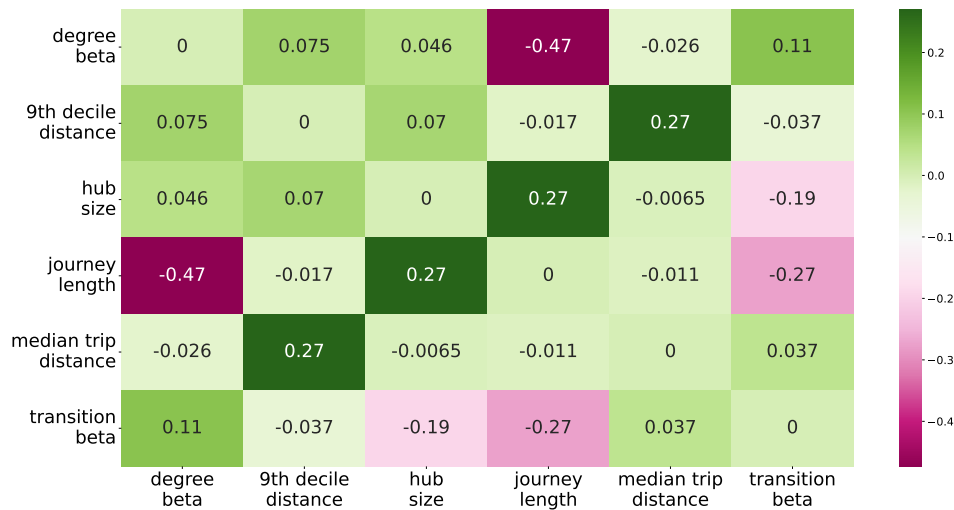


Fig. 4.3.: Correlation between features

of locations, $h = \frac{h^*}{\sqrt{|L|}}$. We chose a square root as it has been shown that the number of locations that are important to a person can be characterized with sub-linear exponential growth (Alessandretti et al., 2018). The importance values assigned to each node by this method correspond to the PageRank value of a node (Page et al., 1999; Schütze et al., 2008), with the slight variation that the implemented random walk always starts at the home node and restarts at the home node if it hits a dead end.

A visual summary of all features is given in Figure 4.1. In addition, Figure 4.3 shows the correlation matrix of all features of the users of six datasets combined as described in Section 4.1.5. It demonstrates that only few features are significantly correlated. Apart from the obvious correlation between *median* and *9th-decile trip distance*, we find that *journey length* is negatively correlated with *degree distribution* β (-0.47) and *transition* γ (-0.27), and positively correlated with *hub size* (0.27). Intuitively, if there are many nodes with high degree (=low *degree distribution* β) the probability to encounter longer cycles in a random walk (=high *journey length*) increases. Nevertheless, we decided to keep both features to further distinguish users with a flexible mobility behavior. Specifically, a high *average journey length* characterizes users that visit several locations in a row, independent of the locations' node degree. The *journey length* feature is also included due to its robustness to the tracking duration (cf. Section 4.1.6).

Identifying user groups

User group definition based on statistical testing

Given a clustering algorithm and m mobility features f_1, \dots, f_m , we aim to find meaningful user groups of distinct mobility behavior. Unsupervised machine learning methods can identify patterns in high-dimensional feature spaces based on a given distance metric. However, clustering methods are oftentimes sensitive to initialization or to their parameters,

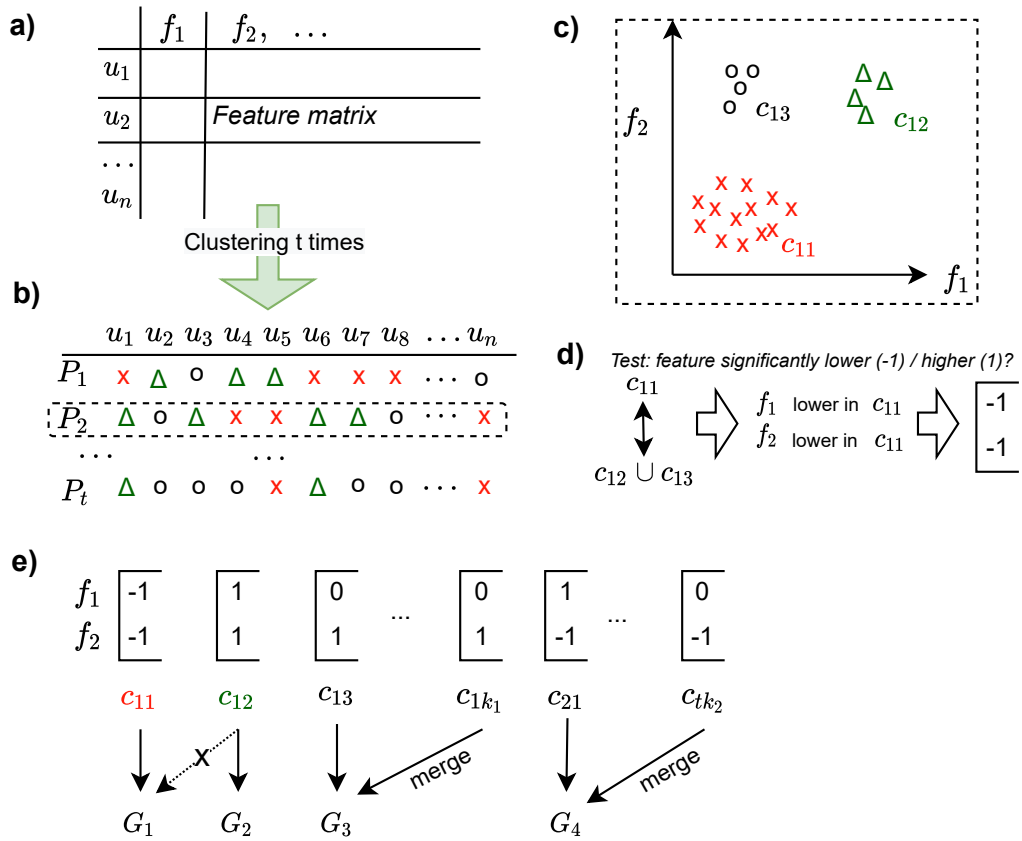


Fig. 4.4.: Workflow of user identification via clustering. a) Features f_i are extracted from the location graphs of each user u_i and form a feature matrix. b) A clustering algorithm is applied t times with different random initialization and parameters, yielding t different partitions of the users. In c), one such partition P is shown schematically for two features. d) By means of a statistical test, we determine for each feature whether the values in one cluster are significantly different from feature values in all other clusters of the partition. e) Based on the significant features, the clusters of all partitions can be merged to an existing group or define new groups. The result is a set of groups G with every cluster assigned to one group.

e.g., the number of desired clusters. To overcome these problems, we propose a method based on statistical testing that yields stable user groups. The method is explained visually in Figure 4.4.

Based on a fixed set of features and users, the clustering algorithm is run t times with different parameter choices or randomization, yielding multiple partitions P_1, P_2, \dots, P_t where each $P_i = \{c_{i1}, \dots, c_{is}\}$ is the set of clusters that define the i -th partition (see Figure 4.4a and b). One such partition is shown in Figure 4.4c. We then consider all clusters in each partition and apply a suitable statistical test to determine which features are significantly different from the other clusters (Figure 4.4d). Let $F_j(c_{ik})$ denote the distribution of feature f_j in cluster c_{ik} . We test the hypothesis that $F_j(c_{ik})$ does not differ from $F_j(P_i \setminus c_{ik})$, i.e., the distribution of f_j in all other clusters.

Next, the test results are combined in a m -tuple for each cluster, called $g(c_{ik})$, which defines a potential user group. The entries of $g(c_{ik})$ describe for each feature whether it is significantly lower (-1), higher (1) or not significantly different (0) from the other clusters. For example, given the features height, age and weight of a person, the tuple $g(c_{ik}) = (0, 1, -1)$ would denote that the people in cluster c_{ik} are not of significantly different height, but are significantly older and of lower weight than the people in other clusters.

Merging similar user groups

In the next step, we aggregate clusters to user groups (Figure 4.4e). In the aggregation step, the assignment of a cluster c_{ik} is independent of its partition P_i and we therefore drop the partition index to simplify notation in this section, i.e., a cluster is simply denoted as c_k . Two clusters are of the same user group if their significant features do not contradict each other. Formally, we define a valid merge of two clusters c_k, c_l as the following: Let $[g(c_k)g(c_l)]^-$ be the result of subtracting $g(c_k)$ from $g(c_l)$ element-wise. If any of the values in $[g(c_k)g(c_l)]^-$ is 2 or -2, then the clusters are not merged. The intuition is that if one or more features are opposite to each other (low in one but high in the other), the clusters are dissimilar. However, with this definition, it could still occur that, for example, $g(c_k) = (1, 1, 0)$ is merged with $g(c_l) = (0, 0, 1)$ because $[g(c_k)g(c_l)]^- = (1, 1, -1)$, even though not a single significant feature corresponds. Thus, as a second requirement, they are only merged if they share at least θ_{minf} significant features, i.e., at least two elements of $[g(c_k)g(c_l)]^-$ are zero. By definition of this merging process, the maximum number of resulting groups is 2^m .

Iterative group finding and assignment

In practice, we distinguish between an iterative group-finding phase and an assignment phase. In the group finding phase, we start with $c_{11} \in P_1$ and store $G_1 = g(c_{11})$ as the first user group. For c_{12} it is then checked whether a merge with G_1 is possible; otherwise, a new group $G_2 = g(c_{12})$ is added. All subsequent clusters are assigned to the existing groups or serve as new groups. Clusters with less than θ_{minf} significant features are skipped. Note that this rarely occurs in practice and the effect of θ_{minf} is therefore limited, as analyzed quantitatively in appendix A.3.4.

After the groups have been identified, we iterate over all clusters a second time and assign each cluster c_{ik} the group with best correspondence which is defined as the largest

Study	Tracking type	Users	Tracking days	Visited locations
Geolife	GPS tracker	66/177	301±392	168±187
GC1	GNSS via app	134/139	401±60	756±249
GC2	GNSS via app	47/50	321±63	718±318
Foursquare home	LBSN checkins	88/100	477±94	219±146
Foursquare random	LBSN checkins	82/100	413±116	88±43
yumuv	GNSS via app	653/871	99±27	184±86

Tab. 4.1.: Overview of the datasets used in this study. Column *users* shows the number of participants used in the study after filtering and the total number of available users. The columns *tracking days* and *visited locations* show the average and standard deviation over users.

overlap between characteristic tuples. Formally, let $g[j]$ denote the entry at the j -th position in the tuple g . Then the group of c_{ik} is assigned to the group with index l^* , $l^* = \arg \max_l \sum_j g(c_{ik})[j] \cdot G_l[j]$. Finally, each user is assigned to the group that occurred most often in her clusters. Two groups can occur equally often. However, as ties were not frequent (<5%), this case was disregarded, and ties were solved by randomly assigning a user to one of the groups.

4.1.4 Data and preprocessing

All preprocessing steps are performed using Python and the Trackintel movement data processing library (Martin et al., 2023c) which provides functionality to extract staypoints, triplegs, trips, and locations. *Triplegs* are defined as continuous movement trajectories, *staypoints* are periods of stationary behavior. Staypoints are defined as activity if their duration is longer than 25 minutes or if there exists a non-trivial purpose label (any purpose except *wait* or *unknown*). *Trips* are defined as the collection of all movement and idling between two activities (Axhausen, 2007). Trips with gaps longer than 25 minutes are considered to have an unknown destination as a person could have performed an activity in between, Locations are extracted using DBSCAN with 1 sample required for a cluster and a search radius of 30 meters for GNSS datasets.

yumuv

The yumuv dataset was recorded during the roll-out of a new MaaS offer in Zurich, Switzerland, called *yumuv*¹ to study the impact of mobility bundles on mobility behavior (Martin et al., 2021b). The total study duration was three months and participants were either part of the TG, which had access to the new MaaS offer available via the yumuv app after one month of pre-tracking, or part of the CG. Both groups had to install the app MyWay², a

¹<https://yumuv.ch/en>

²<https://play.google.com/store/apps/details?id=ch.sbb.myway>

GNSS based tracking app, on their phone to record their mobility behavior. The app already provided staypoints and triplegs. The participants labeled triplegs with the used mode of transportation and staypoints with an activity label. Additionally, all participants took part in an online survey before and after the study period and provided person and household specific data such as socio-demographic information or mobility tool ownership. The dataset contains a total of 871 users (161 TG, 710 CG) of which 498 (71 TG, 427 CG) finished the study.

We additionally separate the dataset into four weeks *before* and four weeks *after* getting access to the MaaS bundle via the yumuv app. The exact dates and durations of the before and after periods are slightly different for every user depending on when users started tracking and installed the yumuv app. For the users of the TG we use each individual's start and end date as defined above. As the users of the CG do not get access to a MaaS bundle, we use the average start and end date of all TG users as the start and end date for the CG.

Green Class

The Swiss Federal Railways (SBB) conducted two large-scale 1-year pilot studies to evaluate the use of a comprehensive *all you can travel* mobility package (Martin et al., 2019a). In the pilot studies the participants had access to a general public transport pass valid in Switzerland, access to popular car- and bike-sharing programs, and taxi vouchers. Additionally, participants of the first pilot study referred to as Green Class 1 (GC1) had access to a personal battery electric vehicle whereas participants of the second pilot study referred to as Green Class 2 (GC2) had access to a premium electric bike. Participants agreed to be tracked via the MyWay app and provided socio-demographic information in surveys.

Geolife

The Geolife GPS trajectory dataset was collected by Microsoft Research Asia over a span of three years (Zheng et al., 2009). Employees were provided with different global positioning system (GPS) loggers and GPS-phones that were used to passively track their everyday movement continuously. The dataset does not systematically provide additional labels or socio-demographic information; however, it is still one of the few publicly available large-scale tracking datasets and is included to allow an easy reproduction of the results of this study. Staypoints are generated with Trackintel using the staypoint detection algorithm from (Li et al., 2008). We used the parameters proposed by the authors for this dataset and we additionally added a threshold that excludes periods without trackpoints for more than 24 hours as gaps.

Foursquare

The global scale Foursquare-dataset³ presented in (Yang et al., 2015, 2016) is a vast collection of publicly available check-in data from the location-based social network Foursquare. We

³https://sites.google.com/site/yangdingqi/home/foursquare-dataset#h.p_ID_56

chose to include the Foursquare dataset to showcase the possibility of the graph-based approach for non-GPS datasets. The full dataset contains check-ins of 144'704 users all over the world collected over the course of 18 months. Users track their movements by checking in at venues (e.g., points of interests). Data quality varies highly between users and not all users provided check-ins at their home location. Especially the second issue is problematic, as the structure of the location graph with a missing home location would be systematically different. We therefore create two subsets of the Foursquare dataset. The *Foursquare Home* subset consists of the 100 users with the most home check-ins in the dataset. As these are some of the most active users in the dataset, we further create the *Foursquare Random* dataset where we randomly draw 100 users from the 27'227 users that have at least 81 check-ins reported (above the 25th percentile), that checked-in at least at 40 different locations (above the 25th percentile) and that have at least 24 check-ins at home (above the 75th percentile).

Graph generation

The location graph for each person is generated as described in 4.1.3. Following the definition of trips given in Section 4.1.4, every trip of a person increases the edge weight between the two activity locations (i.e. nodes) by 1. The graph creation for the Foursquare dataset is slightly different as it uses check-ins without stay duration instead of continuous tracking. Here, all sequential check-ins at locations are used to increase the edge weight between two venues (nodes). Before creating the graphs for the GNSS based datasets, we filter to include only users with at least 14 days with tracking coverage of more than 70 % of the time of the day.

4.1.5 Results and discussion

Given the location graphs for all users, we calculate the features as described in Section 4.1.3. We then combine the six datasets and classify users by their mobility behavior with K-means clustering and our group identification algorithm (Section 4.1.3). In the following, we first describe the obtained user groups, secondly, we present cross-sectional and longitudinal studies based on these groups and finally validate our feature set and the method.

Identification of user groups across datasets

The aim is to study *universal* differences in mobility behavior that appear in diverse tracking studies. Therefore, the six datasets (GC1, GC2, yumuv, Geolife, Foursquare-Random, and Foursquare-Home) are combined and processed in an analogous manner. We then exclude users if one or more of their feature values is more than four standard deviations apart from the mean value of that feature. By that 67 users or ~ 6% of the total users are excluded. The Foursquare-Random dataset has the highest outlier ratio with 18%.

Next, the features are normalized to z-scores. We proceed with the group identification algorithm introduced in Section 4.1.3. Here, the K-Means++ algorithm (Vassilvitskii and Arthur, 2006) is utilized for clustering, since it enforces compact clusters which are more likely to have significantly different feature values. Although the impact of the initialization of K-Means++ is significantly lower than for the original K-Means algorithm (Lloyd, 1982), we observed different outcomes depending on the initialization. Therefore, we vary both the random initialization as well as the number of clusters k . Specifically, we apply K-means three times for each $k \in [6, 7, 8, 9]$, resulting in $t = 4 * 3$ partitions P_1, \dots, P_{12} . We test for significant difference in the feature distribution with a Mann-Whitney U test (Mann and Whitney, 1947). Furthermore, we set $\theta_{minf} = 2$, such that clusters with less than two significant features are skipped. In the group finding phase (cf. Section 4.1.3) we identified six user groups. However, in the subsequent group assignment phase, the users were only assigned to five groups. In other words, one of the groups only appeared in few clustering runs and every user was assigned to one of the other five groups more often. The consistency of user-group assignments is analyzed further in Section 4.1.6.

Interpretation of user groups

To further analyze the identified groups, we inspect which feature dimensions are distinctive for a group. For that we calculate the deviation of the feature value of a single group from the mean of the distribution of this feature value from all other groups. Figure 4.5 shows the deviation for each feature by group. For example, the *median trip distance* of the first group is more than two standard deviations above the average of this feature in all other groups. On the basis of Figure 4.5 we interpret the clustering results and summarize each group's mobility behavior in one term. This group-naming is a subjective decision that is dataset- and context-dependent; however, it greatly facilitates the communication of results to decision-makers and the public. Here, we base our group description on the underlying mobility behavior that leads to a specific layout in the mobility graph as measured by the features described in Figure 4.1.

The first two groups are clearly related to trip distances, where we can safely assume that users with a high *median distance* cover much distance on a regular basis (as *commuters* do), whereas the *9th decile* is expected to be high only for users that regularly cover very long distances. The third group is characterized by their *flexibility*, because their activities are highly distributed (high *hub size*), they move from place to place (high *journey length*) and their graphs are less concentrated on single nodes or edges. In contrast, the fourth group's activity is more skewed towards one or few trips (high *transition* γ) and few nodes (low *hub size*) and takes place at a lower radius (low *median distance*). The fifth group uses a single node (or few nodes) with high degree (high *degree distribution* β) as a base and other activities are started from this center of the locations (low *journey length*). For a more detailed visualization of the groups, we refer to the scatterplot matrix of each pair of features in Figure A.2 in appendix A.3.1.

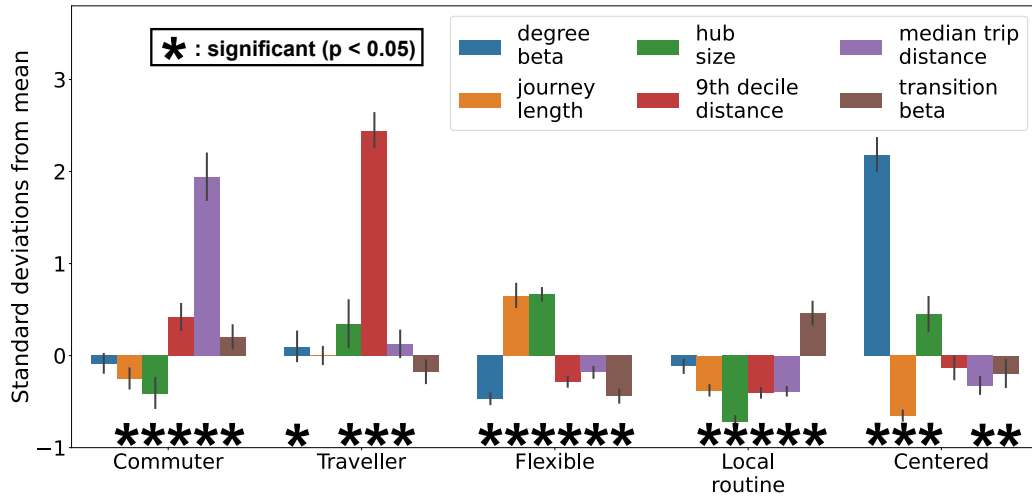


Fig. 4.5.: Feature properties for each user group: The user groups that are consistently found in the data are named based on significant differences in their feature values with respect to the other clusters.

Comparison of user groups across datasets

One of the main contributions of this work is the identification of user groups across several datasets that do not depend on technical data set properties such as tracking technology or the study duration. To verify this hypothesis, we analyze the distribution of user groups over the different studies in Figure 4.6. We observe that the groups are not study-specific as two groups appear in all studies and the other four groups appear in all studies but *Foursquare-Home*. This rules out the case where the clustering process identifies each dataset as its own cluster. In appendix A.3.6, we further use a logistic regression model to show that tracking duration and coverage have very little influence on the graph-based mobility profiles while they strongly influence the mobility profiles generated based on basic features from the literature (cf. Section 4.1.6).

The variation in group distribution over studies can be explained by actual differences between study populations: For example, the yumuv app attracted mainly young people living in urban areas while the Green Class studies had a focus on suburban professionals. An exception is the two subsets of the Foursquare dataset where only few user groups are present. In general, it is reasonable to assume that persons in the Foursquare datasets are rather young and live in cities, similar to users of the yumuv app. Therefore, it is not surprising that all three datasets share the same two majority classes *Flexible* and *Local routine*. One important difference between the datasets is certainly that many LBSN users do not check-in at their home location. However, even though the group distribution of Foursquare users who do check-in at home (Foursquare-Home) is closer to yumuv than the group distribution of Foursquare-Random, it is still significantly different. Further analysis revealed that Foursquare users rarely cover long distances and have a rather low *degree distribution* β . At this point, it is still unclear to which degree these differences can be attributed to a sample bias of the study participants or to a technical bias caused by the characteristics of check-ins compared to GNSS tracking.

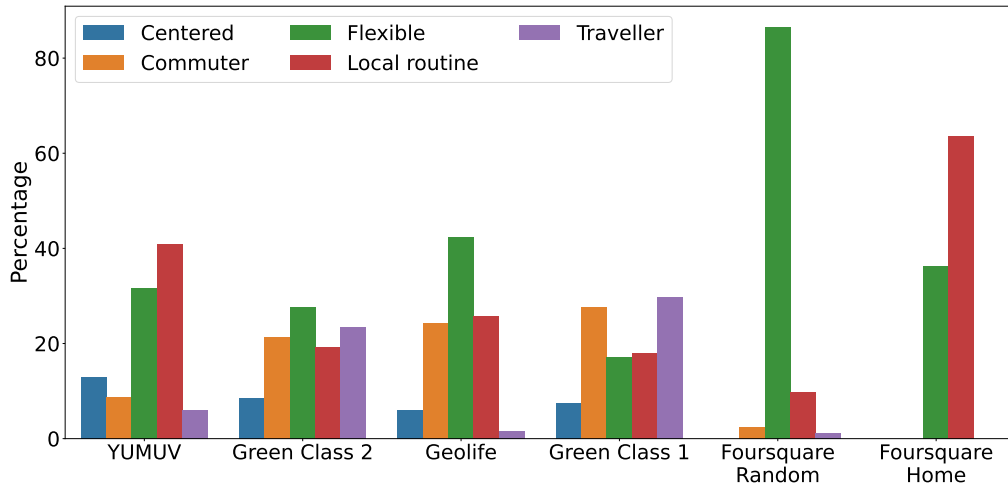


Fig. 4.6.: User groups by study. With the exception of the Foursquare dataset, the user groups are similarly distributed. Differences can be explained by variations in the study target group, e.g., the Green Class 1 offer attracted individuals that cover longer distances.

Overall, we conclude that the differences in the group distribution over datasets can mainly be explained by differences in the mobility behavior of study participants. The presented method is therefore robust to changes in tracking techniques and reflects actual differences in mobility behavior.

Use cases of mobility profiling in MaaS applications

After having established stable mobility profiles, we can use them to answer questions that are typically of interest when analyzing longitudinal or cross-sectional tracking studies. Here, we consider questions that arise around the introduction of a novel MaaS offer:

- What are the target groups for the MaaS offer? (Section 4.1.5)
- How does access to a specific MaaS offer change mobility behavior over time? (Section 4.1.5)

The yumuv dataset is very suitable for a case study due to the availability of distinct control group (CG) and treatment group (TG), and the availability of tracking data before and after access to the app. For the following analysis, we split the yumuv dataset into four parts: TG-Before, TG-After, CG-Before, and CG-After (cf. Section 4.1.4). For this analysis, we consider only participants who finished the study; after pre-processing and outlier-filtering the TG consists of 51 users and the CG of 372 users.

These four yumuv subsets were not part of the merged datasets D that were clustered (cf. Section 4.1.5). Consequently, the graphs of these subsets must be assigned to a user group first. However, the final user groups $G_{final} = G_1, \dots, G_5$ resulting from the iterative group finding and assignment procedure do not have a unique cluster center assigned to them,

as they are the result of merging different partitions P_1, \dots, P_{12} . We, therefore, chose the specific partition P_i out of all partitions P_1, \dots, P_{12} with the highest correspondence to the final user groups G_{final} , meaning that most users are assigned to the same group.

Here, in the best partition P_4 (with $k = 7$), 95% of the users were assigned to the same as their final user group. The graphs in subsets TG-Before, TG-After, CG-Before, and CG-After are now assigned to a user group by finding their closest cluster center in P_4 , as it is commonly done in K-Means clustering for new test data. These preprocessing steps yield a user group for all users per subset, such that the before- and after-groups of one user may differ.

Cross sectional comparison

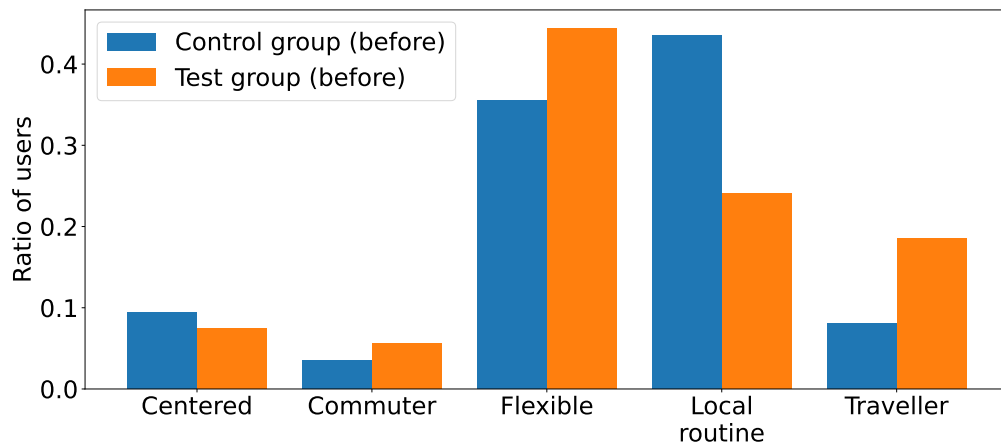


Fig. 4.7.: Characteristics of yumuv users compared to the control group. Users assigned to the groups *Flexible* and *Traveller* are more likely to be interested in the yumuv MaaS offer.

The assignment of user groups to each of the subsets (TG-Before, CG-Before) can now be used for a cross-sectional analysis. For mobility service providers and more generally the design of MaaS offers it is important to know the target group of their offer. For this purpose, we compare the distribution of groups in the TG with the group distribution in the CG. We do this comparison for the period *before* access to the app in order to exclude possible confounding factors from the app usage that might affect mobility behavior. The comparison in Figure 4.7 shows that persons who bought the yumuv offer (TG) are more often assigned to the groups *Flexible* and *Travellers*, whereas the *Local routine* group is more prevalent in the control group (distributions significantly different in χ^2 test with $p = 0.02$). These target and non-target groups can now be characterized using Figure 4.5 and using additional information such as demographics if available (cf. Section 4.1.5).

Longitudinal study

One of the main questions with regards to the introduction of MaaS offers is if and how they impact the mobility behavior of users (Hensher et al., 2021). The assignment of user

groups to participants before and after the intervention allows to analyze whether the group assignment of one user changes from the *before* period to the *after* period and allows to compare the changes between CG and TG.

Figure 4.8 shows the changes between user groups from the period before intervention to afterward. Each row is normalized to 1 and each cell shows the percentage of users that were assigned to a specific group before the intervention (row label) and moved to another group after the intervention (column label). We observe that more people in the treatment group switch towards the *Traveller* group than in the control group (cf. rightmost column). Furthermore, it seems that the *Flexible* group is more stable in the treatment group (cf. the values on the diagonal for the row *Flexible*). However, we compared the distributions row-wise with a χ^2 test, and due to the small size of the Treatment group (51) there are no significant differences (the lowest p-values are $p = 0.09$ for the changes of the former *Flexible* group and $p = 0.23$ for former *Traveller*). Despite the lack of statistical significance this analysis still serves as a show-case of how to use the identified user groups for a longitudinal analysis. A graph-based visualization of the user group changes can be seen in Figure A.4 in appendix A.3.3.



Fig. 4.8.: User group changes upon intervention (start of the yumuv offer).

Cluster analysis with respect to labels

The yumuv study also included surveys that collected socio-demographic and household information of the participants. In this section, we analyze which of these features are significantly different for a specific group with respect to all other groups and therefore characterize that group. Table 4.2 shows the replies for each user group for a selection of relevant questions. Many results confirm the assumptions about mobility behavior that determined our naming of the clusters as user groups. For example, *Commuters* are less satisfied with the public transport connections to their home; they oftentimes travel by car and are seldom in home office. Only 37% of the *Commuters* live in cities. *Travellers* and *Flexible* users in contrast are younger, oftentimes live in cities and spend more days in home office. Interestingly, the *Centered* group works from home significantly more often than others. The fact that working behavior such as home office is reflected in the user groups provides evidence for a strong influence of the home and work locations on the

graph features. Last, the naming of the group *Local routine* is reflected well in the users' self-reporting of their covered distance.

	Commuter (n=57)	Traveller (n=39)	Flexible (n=206)	Local routine (n=267)	Centered (n=84)
[pht] Age	39.11 p=0.46	37.72 p=0.351	37.86 p=0.022	40.7 p=0.059	40.65 p=0.217
Money spent on PT (in CHF per month)	66.5 p=0.023	78.89 p=0.127	94.19 p=0.243	94.04 p=0.39	107.0 p=0.311
Home office (in days per week)	0.86 p=0.034	1.81 p=0.007	1.31 p=0.168	0.93 p=0.003	2.19 p=0.002
Distance travelled by car yearly (in km)	11840 p=0.106	11844 p=0.097	11039 p=0.027	7855 p=0.001	9656 p=0.413
Satisfaction with PT reachability (%)	79.12 p=0.034	86.25 p=0.364	85.4 p=0.205	87.22 p=0.079	84.69 p=0.416

Tab. 4.2.: User group analysis with respect to demographic and mobility characteristics from study questionnaire. The mean values are given and compared to the other groups in a Mann-Whitney U test for continuous variables or a Chi-Squared test for categorical variables. Significant differences are marked bold, and PT denotes public transport. Note that all fields are self-reported in a questionnaire and not measured in the tracking study.

4.1.6 Validation

This section provides analyses to validate the method for extraction of generalizable user groups based on graph representations of individual mobility.

Feature robustness to study duration

An important factor in the feature selection process is their robustness to dataset properties. Here we investigate how much the feature values depend on the tracking duration. For this experiment, we split the Green Class 1 data into distinct bins of $t = 4, 8, 12, 16, 20, 24$ and $t = 28$ weeks. Since the participants in the study were tracked for 56 weeks, $t = 4$ yields 14 non-overlapping bins and $t = 28$ is the maximum duration with two distinct bins. Next, we construct the location graphs from the activity of each user in each time bin and compute the corresponding features. In Figure 4.9 the mean feature values for all users per time bin are shown. Our selected features (top two rows) are largely robust to the tracking period or converge after around $t = 12$ in the case of *degree* and *transition* γ . In contrast, other considered features show a strong time dependency, such as the mean Eigenvector centrality in a graph or the clustering coefficient as proposed by (Onnela et al., 2005). Similarly, features of human mobility that are commonly used in the literature usually depend on the number of nodes and therefore change significantly with the tracking duration.

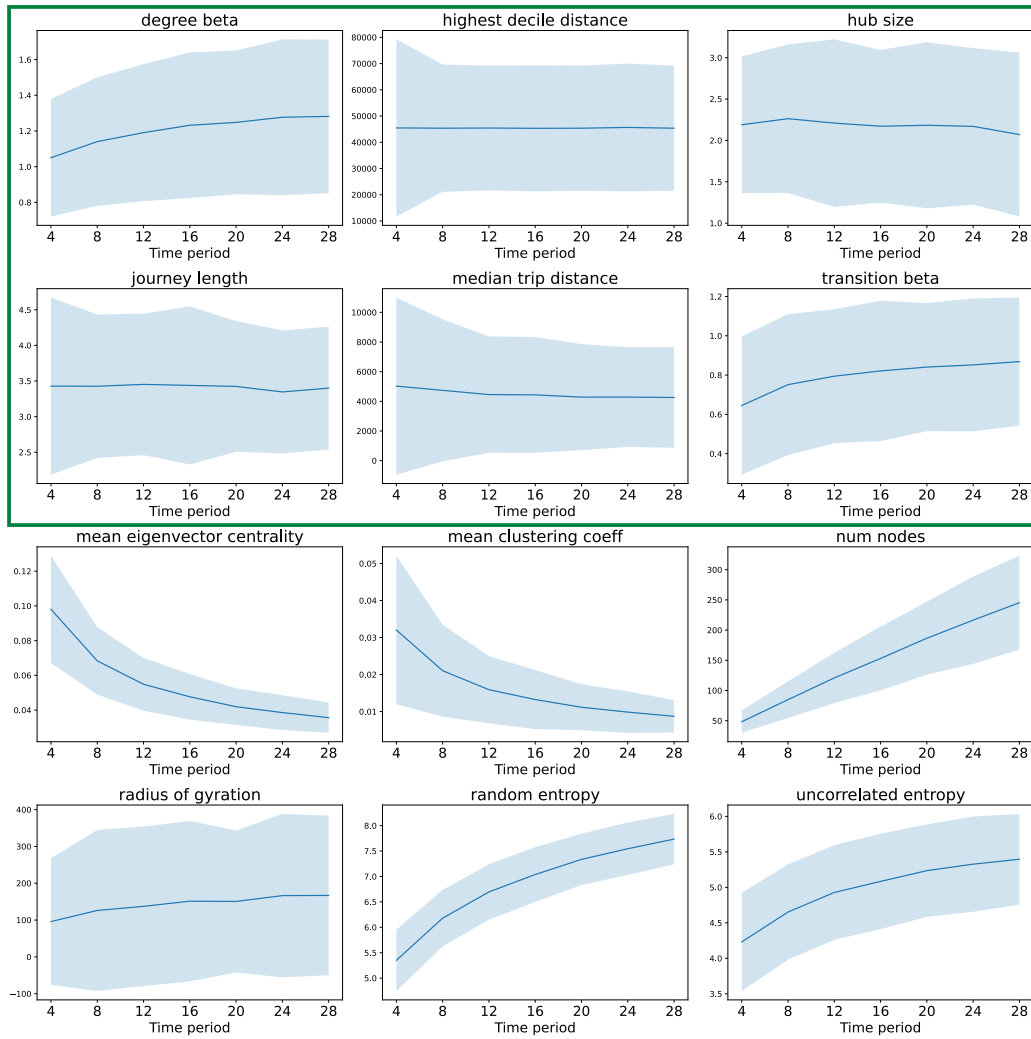


Fig. 4.9.: Mean and standard deviation of different features with respect to tracking period. The selected features are within the green box.

Relation of graph-based and classical mobility features

Furthermore, we investigate the relation of the proposed graph features to a selection of classical non graph-based mobility features that are commonly used to characterize mobility behavior. These features are directly calculated from the raw data and can therefore make use of information that is lost in the location graph, such as the order of the activities or their duration.

Concentrating on the most widely used measures, we consider the following features, which are termed *basic features* from here on:

- Number of visited locations
- Radius of gyration (González et al., 2008)
- Maximal distance from home
- Random, uncorrelated, and real entropy (Song et al., 2010a)
- Mean trip duration and distance

Except for trip distance and trip duration, we utilize the implementation in the `scikit-mobility` package (Pappalardo et al., 2022).

For this experiment, we use only the datasets Green Class 1 and 2, yumuv and Geolife, because there is no trip information available for the Foursquare datasets (only check-ins). First, we were interested in whether clustering based on the basic features results in similar clusters as our user groups. We cluster the basic features with $k = 5$ and compare the resulting clustering to our user groups with the Adjusted Rand Index (Hubert and Arabie, 1985; Rand, 1971). Intuitively, the Rand Index is proportional to the number of pairs that end up in the same cluster in both partitions or in different clusters in both partitions. The Adjusted Rand Index is its normalized version that yields a value between -1 and 1. An index of 0 means that there is no relation between two partitions while the same partitions would yield an index of 1. Here, the similarity to our user groups is 0.08., i.e., the user groups found with graph features are fairly distinct from clusters that are identified with the basic features.

Secondly, Figure 4.10 depicts the mean and standard deviation of the raw features in the proposed user groups. It can be observed that *Traveller* obtain significantly higher values also in these basic features, confirming the differences between groups also with respect to these basic measures.

We conclude that using graph features results in different user groups that can then be analyzed with respect to classical features.

Group robustness to cluster ordering

Last, although it was shown that all clusters of 12 partitions can be reduced to five user groups, we found that the resulting groups still depend on the *order* of considered clusters during the *iterative* group-finding phase. For example, the first cluster is always used as G_1 ,

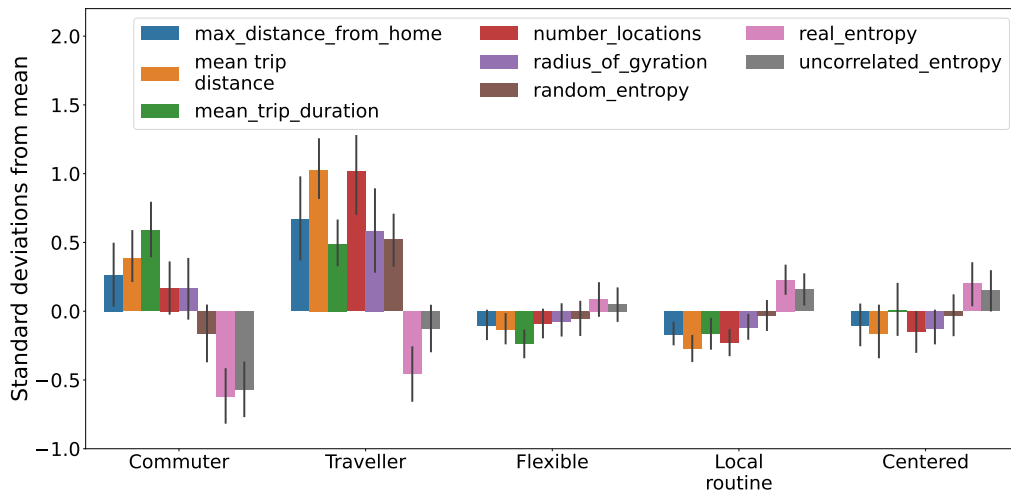


Fig. 4.10.: Distribution of basic features over the identified user groups. On the one hand, differences in the basic features are also reflected in our user groups. On the other hand, our groups seem to identify further differences in mobility behavior that are hardly reflected in the basic features (e.g., group *Centered*).

and other clusters may be merged with this particular cluster. The properties of G_1 thus depend on which cluster is considered first. Nevertheless, we qualitatively observed a strong stability of the resulting groups and their properties.

To quantify this stability, we perform 20 runs of the group-finding and assignment phases with different random seeds and compare the resulting groupings to the main grouping found in Section 4.1.5 on a per-user basis using the Adjusted Rand Index (Hubert and Arabie, 1985; Rand, 1971). In our analysis, the Adjusted Rand Index of the pair-wise comparison of the groupings with the main grouping is 0.91 on average. Another initialization would therefore yield a very similar output, where the resulting groups could be named in a similar manner and a large majority of users would be assigned to the same group.

Consistency of group assignment

In the method presented in (cf. Section 4.1.3) the clusters resulting from the different runs are assigned to groups. A user can therefore belong to clusters that are assigned to different groups and is finally assigned to the group that most of his clusters are assigned to (cf. Section 4.1.3).

We now analyze how consistent this assignment is by computing a *consistency score* that indicates how often a user belonged to its most dominant group. The score is calculated by counting the number of times a user was assigned to its majority group divided by the total number of assignments.

In our case study, the average consistency score over all users is 0.87, i.e., an average user is assigned in 87% of the runs to its most dominant group. 80% of the users obtain a consistency greater than 0.9.

4.1.7 Conclusion

Research on mobility behavior oftentimes suffers from a lack of reproducibility and transferability. Big tracking datasets are inherently noisy and usually unlabeled, and proposed methods do not generalize to other datasets. Here, we have presented an attempt to develop a generic clustering approach that yields stable mobility behavior groups on several diverse datasets. In contrast to previous work, we base our analysis on a compact graph representation of the tracking data, which 1) reduces memory resources needed to store long-term tracking data, 2) facilitates the comparability of different datasets, and 3) captures other aspects of mobility behavior than time-series based basic features. Based on six features that were shown to be particularly robust with respect to the time period, we apply a clustering algorithm multiple times and extract stable and interpretable user groups in an iterative fashion based on statistical testing.

Our analysis showed that five groups could consistently be found in the six datasets, which differ by the complexity, the role of home bases, and the geometric extent of their mobility behavior. All user groups can be found in all studies except Foursquare, despite significant differences in the tracking quality, duration, and user demography of the studies. These user groups were also shown to be robust to the clustering parameters, consistent, and seem to depict novel aspects of mobility behavior that are not contained in classical mobility features. It is still unclear to what degree the differences of the Foursquare data are due to a sample bias of highly active urban LBSN users or due to systematic differences between GNSS based and check-in based tracking data. However, the Foursquare datasets did not generate exclusive user groups and could still be described by our framework. Furthermore, it could be shown that differences of the distribution of user groups also reflect differences of the target groups of each study. Such analysis is of interest to providers of MaaS offers to direct their services to the right people. Similarly, it could be shown that the effect of a MaaS offer on mobility behavior can be viewed in the context of user group changes over time. While a detailed analysis of the changes of location-graphs over time is out of scope of this paper, it is an interesting endeavor for future research. While the cluster analysis can be used to describe the change in mobility behavior over time, we noticed that this description of mobility behavior exhibits a higher volatility than expected, i.e., up to 50% of users change their group from one slot to the next. A possible reason for this could be that the mobility behavior did not fully stabilize after the considered tracking duration or clusters are overlapping which may lead to a certain number of samples that lie between two clusters and thus easily switch clusters over time. Further work could explore the possibility to connect our clustering approach with soft assignments where each sample belongs to multiple groups with certain probability.

Finally, it should be stressed that much of the proposed methodology is by no means restricted to mobility research. Some of the proposed features could be relevant in other fields where data is represented in graph structures, such as molecules in biology or computer networks (e.g., hub size as a descriptor of network activity). More importantly, clustering is a popular method used in many fields, and the identification of stable, statistically valid groups is a common problem. Our algorithm is a simple yet effective method to make clustering results more generic and reproducible.

Acknowledgements

Funding: This work was supported by the Swiss Data Science Center [C17-14]; the Swiss Innovation Agency Innosuisse within the Swiss Centre for Competence in Energy Research on the Future Swiss Electrical Infrastructure (SCCER-FURIES); the Swiss Innovation Agency Innosuisse within the Swiss Centre for Competence in Energy Research - Efficient Technologies and Systems for Mobility (SCCER Mobility); the ETH Zurich Foundation [MI-01-19];

4.2 Graph convolutional neural networks for human activity purpose imputation from GPS-based trajectory data

The following section is a reprint of the publication:

Henry Martin*, Dominik Bucher*, Esra Suel, Pengxiang Zhao, Fernando Perez-Cruz, and Martin Raubal (2018). “Graph Convolutional Neural Networks for Human Activity Purpose Imputation”. *NIPS Spatiotemporal Workshop at the 32nd Annual Conference on Neural Information Processing Systems (NIPS 2018)*. DOI: 10.3929/ethz-b-000310251.

Conceptualization: HM, DB, ES; Methodology: HM, DB; Software: HM, DB; Writing – original draft: HM, DB, ES, PZ; Writing – review & editing: HM, DB, ES, PZ, FP, MR;

The content is reproduced ‘as is’, however, formatting changes and corrections of spelling have been applied.

Abstract

Automatic location tracking of people has recently become a viable source for mobility and movement data. Such data are used in a wide range of applications, from city and transport planning to individual recommendations and schedule optimization. For many of these uses, it is of high interest to know *why* a person visited at a given location at a certain point in time. We use multiple personalized graphs to model human mobility behavior and to embed a large variety of spatio-temporal information and structure in the graphs’ weights and connections. Taking these graphs as input for graph convolutional neural networks (GCNs) allows us to build models that can exploit the structural information inherent in human mobility. We use GPS travel survey data to build person specific mobility graphs and use GCNs to predict the purpose of a user’s visit at a certain location. Our results show that GCNs are suitable to exploit the structure embedded in the mobility graphs.

* Equal contribution

4.2.1 Introduction and related work

In recent years, human mobility analysis has become increasingly important since it is closely associated with opportunities (e.g., on-demand travel and shared mobility services, location-aware recommender systems) and challenges (e.g., traffic congestion, air and noise pollution, decisions on transportation infrastructure investments) in modern society. Advanced information and communication technology presents a unique opportunity to better understand highly complex mobility behavior (Urner et al., 2018). Specifically, the prevailing availability of Global Navigation Satellite System (GNSS) based movement trajectory data allows low-cost collection of trajectory data from very large numbers of participants over long periods of time. To take advantage of the available data, researchers both from domain sciences and machine learning are developing methods for cleaning raw data, identification of trips (i.e., movements) and activities (i.e., staypoints), and prediction of transport modes and activity types (Chen et al., 2016a; Zheng, 2015).

Previous studies have proposed a number of methods to identify activity types from various trajectory datasets (Chen et al., 2016a; Liao et al., 2006; Liu et al., 2013). For instance, Montini et al. (2015, 2014) trained a random forest model on GPS trace data to improve trip purpose identification and Chen et al. (2018) and Zhao et al. (2017) used Bayesian theory-based models to infer trip purposes using taxi trajectory data. However, these studies often neglect high regularity of human activity patterns while predicting activity types and mostly rely on extracted features of activities (e.g., average stay duration, time of day, points of interest in the vicinity, etc.). In this paper, we propose a new method based on graph convolutional neural networks (Defferrard et al., 2016) for the prediction of activity types (i.e., trip purposes) from GPS trajectory data generated by personal smartphones. This allows incorporating the high regularity of human mobility in terms of locations visited at specific times of day and the frequency of observed direct trips between two locations (Song et al., 2010a). For example, a person might prefer to go grocery shopping at their favorite supermarket after work every Friday; and such regularity in behavior is potentially useful for inferring activity types from location data. To the best of our knowledge, this is the first application of GCNs for the task of predicting activity purposes from trajectory data. Furthermore, we use a unique dataset that includes raw trajectories from 139 users over one year in Switzerland, in which the participants labeled each staypoint with the purpose of their visit.

4.2.2 Data and methods

Data

We use semantically enriched tracking data from the *SBB Green Class* pilot study⁴. Here, 139 Swiss users were tracked over the course of one year with an app on their smartphone. The tracking app⁵ segmented the movement data into staypoints (a user does not move out of a

⁴<https://www.sbb.ch/de/abos-billette/abonnemente/greenclass/ueber-sbb-green-class/pilotprojekte.html>

⁵<https://motion-tag.com/en/mobility/>

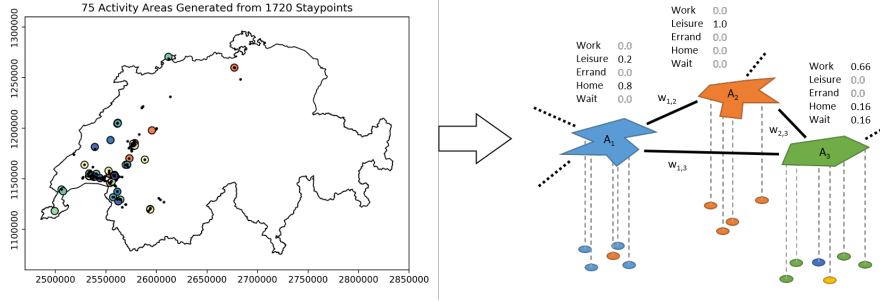


Fig. 4.11.: To create activity areas (A_i) from individual staypoints, we employ a DBSCAN clustering algorithm. Activity areas can be connected ($w_{i,j}$) in various ways, e.g., with their weights corresponding to the number of times a user directly traveled between two activity areas. The labels of the activity areas correspond to the distribution of the underlying staypoints.

certain area during a given period) and movement trajectories and presented it to the users for validation and labeling. The trajectories are labeled with the used mode of transport and the staypoints with an activity purpose $l_p \in \{\text{home, work, errand, leisure, wait}\}$. The study participants generate between 2 and 20 staypoints every day which results in between 2000 and 4000 staypoints per user over the course of the tracking period.

Methods

Activity areas

From prior work on human mobility analysis, we know that humans tend to visit a fixed set of spatial locations that evolves over time (Alessandretti et al., 2018; Song et al., 2010a). To exploit the spatio-temporal patterns and structure that lies within human mobility and the sequential visits of these places, we explicitly model the human mobility behavior between these locations. Based on the point-cloud like staypoint data, we create frequently visited places which we call activity areas using the DBSCAN clustering algorithm (Ester et al., 1996; Jonietz and Bucher, 2018).

Graph network construction

To exploit the spatial information encoded in the relation between activity areas and the spatio-temporal structure that lies within the sequential visit patterns, we construct a set of person-specific, undirected, and connected graphs $G_{u,j} = (V_u, E_{u,j})$. We define $G_{u,j}$ as the j^{th} mobility graph of user u with a user specific, finite ($|V_u| = n_u$) set of nodes $V_u = \{V_{u,i}\}$, $i \in [1, 2, \dots, n_u]$, where n_u is the number of nodes per user and $V_{u,i}$ denotes a specific node of that user. The nodes are connected with m sets of weighted edges $E_{u,j}$, $j \in [1, 2, \dots, m]$. For every graph $G_{u,j}$ we define $W_{u,j} \in \mathbb{R}^{n_u \times n_u}$ as its weighted adjacency matrix.

To construct the mobility graphs for one user, we define every activity area as a node $V_{u,i}$ in his graphs $G_{u,j}$ (cf. Figure 4.11). The creation of a simplified graph by clustering (spatially) similar staypoints to activity areas can be interpreted as a coarsening step comparable to the

graph coarsening step from (Defferrard et al., 2016). To showcase the ability of the GCNs to learn on different graphs (with the same set of nodes), we use two different definitions of spatio-temporal connectivity to create the edges between the nodes. Specifically, we use the transition frequency between nodes and the spatial (Euclidean) distance between all nodes as two different sets of weighted edges. The transition frequencies are calculated by counting how often a user directly traveled between two activity areas. Other suitable definitions for the edges between nodes could be based on the travel times (e.g., one graph per mode of transport), angles, ticket prices, or soft similarity measures such as cultural similarity.

Feature extraction

For each node (activity area) $V_{u,i}$, $i \in (1, 2, \dots, n_u)$, we extract 30 features⁶ based on the aggregated information of the associated staypoints. Similarly, we assign labels to the nodes based on the aggregated staypoints. This process is visualized in Figure 4.11 on the right. Staypoints associated with the same node can have different labels; we therefore assign a distribution over all classes to each node. The probability of label l_p at node i (short for $V_{u,i}$) is defined as $p_{i,l_p} = n_{i,l_p}/n_i$, where n_i is the number of all staypoints and n_{i,l_p} is the number of staypoints with label l_p at the node i . As the node labels are unbalanced, we introduce sample weights based on the inverse label weights $w_{l_p} = n_L/(|L| \cdot n_{l_p})$ (where n_L is the cumulative weight of all labels and n_{l_p} is the cumulative weight for a single label $l_p \in L$). These are later used to compute a weighted cross-entropy loss as the cost function. Out of the 136 users we select a subset of 82 users with similar labeling behavior (number of staypoints per class is over 100). During the experiment, we randomly select 41 out of the 82 users for training, 5 users for validation, and 36 completely unseen users for testing. Additionally, we standardize all training features by removing their mean and scaling them to unit variance and apply the standardization estimated on the training data to the test data.

Activity purpose imputation

The breakthrough success of convolutional neural networks (CNN) in many areas like image, speech, text, or video processing (Karpathy et al., 2014; Kim, 2014; Krizhevsky et al., 2012) is based on their ability to exploit the structure that lies within the data by using self-learned instead of handcrafted features. In recent publications, the concept of GCNs emerged to generalize this ability to arbitrary graph structures and manifolds (Bronstein et al., 2017; Defferrard et al., 2016). Applications of GCNs on traffic forecasting (Cui et al., 2019b) and human action classification (Yan et al., 2018) showed the potential of GCNs to use information embedded in the graph structure and to exploit spatio-temporal structure for their task.

⁶Mean stay duration, maximal duration, minimal duration, total number of staypoints within $V_{u,i}$, mean longitude and latitude, average distance to public transport stops (train, tram, bus), distribution of arrival and departure time at staypoints (classified into night, early morning, late morning, early afternoon, late afternoon, early evening, late evening), and the distribution over weekdays and weekends (Mon-Sun).

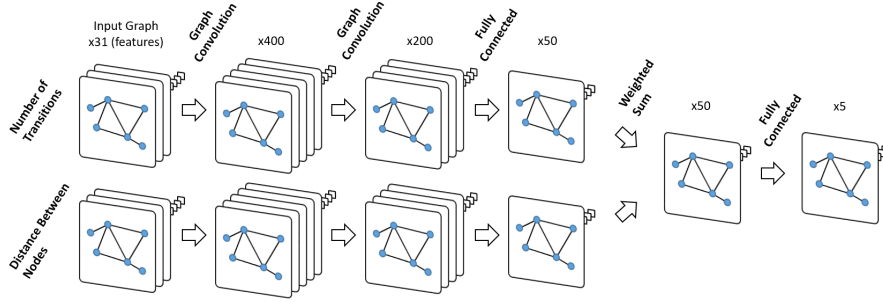


Fig. 4.12.: The computation graph used within this work: The transition and distance graphs are each convoluted twice (layer sizes of 400 and 200). Their outputs are then combined using a weighted sum (layer size 50). Finally, a fully connected layer produces the output label distributions.

We follow the approach of (Defferrard et al., 2016) to approximate graph convolutions formulated in the Fourier domain using a truncated expansion in terms of Chebyshev polynomials. As it was done in (Kipf and Welling, 2016), we only use the first-order approximation. For readability we now consider only the mobility graphs of a single user and omit the index u , however, all graph-related parameters stay user-specific. This results in the following propagation rule for graph convolution layers:

$$H_{G_j}^{(l+1)} = \sigma(D_j^{-\frac{1}{2}} W_j D_j^{-\frac{1}{2}} H_{G_j}^{(l)} M^{(l)}) \quad (4.3)$$

where l is the layer index, $H_{G_j}^l$ is the input from the previous layer with $H_{G_j}^0 = X_j$ (corresponding to the input feature matrix of graph G_j), D_j is the degree matrix corresponding to the weighted adjacency matrix W_j . D_j is used to normalize the weighted adjacency matrix. $M^{(l)}$ denotes the weight matrix for the l^{th} neural network layer and $\sigma(\cdot)$ is the ReLU activation function. Note that the network parameters $M^{(l)}$ are independent of a specific graph G_j (or user). Figure 4.12 shows the network used in this work. We employ two graph convolutional layers on each of the two graphs (G_1 and G_2 , one graph for every set of edges), whose outputs are then combined using a weighted sum:

$$H^{(S)} = H_{G_1}^{(2)} M_{G_1}^{(S)} + H_{G_2}^{(2)} M_{G_2}^{(S)} \quad (4.4)$$

Finally, a fully connected layer creates the label predictions $Y = H^{(S)} M^{(Y)} + b^{(Y)}$ (where $b^{(Y)}$ denotes a bias term).

4.2.3 Results

The here presented model predicts a distribution of labels for each activity area $V_{u,i}$. To measure the performance of our model we employ the earth mover's distance (EMD; also referred to as the Wasserstein metric) (Rubner et al., 1998). In essence, this metric reports the minimal shift of probability masses required to reach a given target distribution from

Tab. 4.3.: EMD comparison of different models.

Learning Method	EMD
GCN	0.076
MLP	0.097
RDF	0.146
RDF (regression)	0.146
Average Label Distr.	0.200

a predicted one. Table 4.3 shows the results for the here introduced GCN, for a multi-layer perceptron (MLP), a random forest (RDF) classifier, a multi-output random forest regressor and as a baseline, the average label distribution for each node⁷. All classifiers can outperform the trivial baseline and both neural network implementations outperform the random forests which might be because the random forests can not take advantage of the additional information provided by the soft class labels.

The GCN performs better than the MLP implementation, even though both are trained and tuned similarly. This indicates that the GCN can take advantage of the additional information embedded in the graphs (connections and weights) and use them to exploit local structure.

4.2.4 Conclusion and future work

In this work, we presented a GCN-based approach for imputing human activity purposes from GPS trajectory data. Our results show that we can use multiple personalized graphs to model human mobility behavior and embed a large variety of spatiotemporal information and structure in their weights and connections. We could also show that we can exploit this structure using GCNs.

For our future work on the exploitation of spatio-temporal structure using graph-based modeling in combination with GCNs, we plan to experiment with different graph-building methods and compare them to the clustering approach applied in this work. For example, the here presented study merely considers the features of staypoints for predicting activity types. Taking into account other contextual information (e.g., points of interest or the road network) could significantly influence the predictive powers of a GCN model. Finally, we would like to analyze the impact of using more complex GCN models (e.g., use a k^{th} -order approximation instead of a first-order approximation for the graph convolution).

⁷MLP: implementation in TensorFlow, cross-entropy-loss, parameters: 3 layers (400,200,50), dropout: p=(0.8,0.5) between layers; RDF classif.: implementation only supports one-hot label training, parameters: trees=500, balanced+scikit learn default; RDF reg.: output normalized that classes add up to 1, parameters: trees=500+scikit learn default. All classifiers except for the GCN only use the node features.

Acknowledgments

This research was supported by the Swiss Data Science Center (SDSC) and by the Swiss Innovation Agency Innosuisse within the Swiss Competence Center for Energy Research (SCCER) Mobility.

Supporting sustainability of personal vehicles

5.1 Using rooftop photovoltaic generation to cover individual electric vehicle demand - a detailed case study

The following section is a reprint of the publication:

Henry Martin, René Buffat, Dominik Bucher, Jannik Hamper, and Martin Raubal (2022). “Using rooftop photovoltaic generation to cover individual electric vehicle demand — A detailed case study”. *Renewable and Sustainable Energy Reviews* 157, p. 111969. DOI: 10.1016/j.rser.2021.111969.

Conceptualization: HM, RB, DB, JH (supporting); Methodology: HM, RB, DB; Software (Mobility data and scenarios): HM, JH (supporting); Software (PV Model): RB; Data curation: HM; Visualization: HM; Writing – original draft: HM, RB; Writing – review & editing: MR; Supervision: MR

The content is reproduced ‘as is’, however, formatting changes and corrections of spelling have been applied.

Abstract

The introduction of battery electric vehicles (BEV) and the expansion of rooftop photovoltaic (PV) power generation are both progressing at a fast pace to decarbonize the transport and the energy sector in Switzerland. These parallel developments have an enormous synergy potential as the actual decarbonization impact of BEVs depends heavily on the carbon footprint of the power source and the PV expansion requires local storage as a buffer to reduce negative impacts on the distribution grid. We present an empirical analysis based on a detailed 10-month data set of the charging and mobility behavior of 78 BEV users in Switzerland. It is combined with a fine-grained digital surface model of Switzerland to extract the detailed roof geometry and the corresponding rooftop PV generation capacity of each of the BEV owner’s houses.

We test four different smart charging strategies with a varying degree of complexity and find that when charging uncontrolled (the strategy used during the study), BEV owners can only cover 15 % of their BEV’s demand using PV generated from the roofs of their own houses. A simple controlled charging approach greatly increases the average coverage to 56 % and

up to 90 % or 99 % when using an optimized charging strategy without or with a home battery storage. All charging strategies ensure that the individual mobility behavior of the BEV owners is not affected.

We further show that using rooftop PV power generation for BEV charging has a large potential to further decrease the climate impact of BEVs and propose simple adjustments to consider in charging strategies that help to increase the owners' PV consumption.

* Equal contribution

5.1.1 Introduction

In light of the threatening progression of climate change, 196 countries have negotiated the Paris Agreement and committed to keeping the global temperature increase to well below 2°C compared to pre-industrial levels (UNFCCC, 2015). To reach this goal, the committing countries formulated national emission targets and measures to achieve decarbonization of their economies. For example, China plans to peak in carbon emissions in 2030 and to achieve carbon neutrality by 2060 (The Economist, 2020), the EU adopted the *European Green Deal* (a set of measures to achieve climate neutrality in 2050¹) and Switzerland passed the *Climate strategy 2050* that also targets climate neutrality in 2050.

These examples sound promising; however, we are currently losing the global fight against climate change. There is a significant gap between the emission reduction that is achieved if all countries implement their proposed measures and the emission reductions that are required to achieve the goals of the Paris Agreement. This is even true under the consideration of the impact that the COVID-19 global pandemic had on the world's economy (Olhoff and Christensen, 2020). As of 2021, only two larger countries (Morocco and Gambia) are considered to be on track to achieve the stricter 1.5°C target of the Paris Agreement, and of the eight countries that fulfill the 2°C goal India is the only one that is part of the G20².

Joint decarbonization of the transport and electricity sectors

This makes it clear that we must drastically increase our decarbonization efforts. A major contributor to climate change is the transport sector where the overwhelming majority of energy demand (e.g., 93.7 % in Switzerland (BFE, 2019b)) is covered by fossil fuels. From today's perspective, the most promising path to a fast decarbonization of the transport sector and especially of individual mobility is given by the introduction of BEVs (Haasz et al., 2018). However, the decarbonization potential of BEVs strongly depends on the emissions of the power used for charging (Casals et al., 2016; Hawkins et al., 2013). This makes the parallel decarbonization of the electricity sector a prerequisite for the decarbonization of the transport sector using BEVs.

The decarbonization of the electricity sector relies mainly on the large-scale integration of wind and solar power into the energy system while their relative importance varies depending on the geographic location (Lu et al., 2009; World Bank, 2020). Globally, installed photovoltaic (PV) generation grows faster (Breyer et al., 2017) and is expected to be the main driver of the expansion of renewable energy generation (IEA, 2020a). In contrast to wind power generation, small roof-top PV systems installed on the roofs of private homes play a major role in this growth. For example, in Germany, the share of non-utility scale systems (< 100 kWp) is close to 50 % of the total installed PV generation capacity (Philipps and Warmuth, 2020).

¹https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en

²<https://climateactiontracker.org/countries/>

Grid impact of PV and BEV growth, and impact mitigation strategies

A possible obstacle for the fast growth of PV generation and the quick introduction of BEVs is given by the stability of the distribution grid. Both phenomena, large-scale roof-top PV generation and the simultaneous charging of BEVs are possibly threatening the distribution grid stability as they introduce high and intermittent generation capacity and consumption into the grid (Clement-Nyns et al., 2010; Luthander et al., 2015).

One idea to mitigate the negative grid impact of rooftop PV generation and BEV charging at the same time is to charge the BEVs directly using power generated from rooftop PV systems installed on the BEV owners' home (Hoarau and Perez, 2018). In this case, the power is used directly where it is generated without entering and stressing the distribution grid. This idea is especially important for BEV charging as the majority of the BEV charging processes are undertaken at home (Mwasilu et al., 2014).

Another advantage of the use of residential rooftop PV for BEV charging lies in the potential for faster decarbonization of the transportation sector. The reason for this is that the self-consumption of rooftop PV has considerably lower GHG emissions than most other electricity sources (Fthenakis and Raugei, 2017; Fthenakis et al., 2008) and especially than the electricity generation mix of most countries.

Research gaps and contributions

In an optimistic, hypothetical future, a large share of the individual energy demand could be covered by PV cells that are installed directly on each person's house, as this could reduce the complexity and required capacity of the electric distribution grid and would increase the decarbonization potential of BEVs. Several studies consider the self-consumption potentials of individually generated renewable energy (Luthander et al., 2015) however, there is a lack of studies that analyze the combination of high-resolution BEV charging data with detailed PV generation models (Shepero et al., 2018). It is therefore still unknown how well the individual electricity demand for mobility can be covered when considering the intersection of when the BEV is at home and available for charging and the availability of PV generation at these times.

To fill this gap this study analyzes what share of the energy demand of a BEV can be covered by rooftop PV generation installed on the house of the BEV owner. This study has the following main contributions:

- We use a high-resolution data set of 78 BEVs that includes GPS position information, information about the battery's state of charge (SoC), and about the charging activity to calculate the energy demand of each of the BEVs over time.
- We estimate the potential rooftop PV generation based on a high-resolution digital surface model in combination with building footprints to extract the surface, tilt, and aspect of the roof of the actual house of every BEV owner. This allows us to create a PV generation profile with a half-hour resolution for each house.

- We match the potential PV energy generation with the recorded BEV usage of each of the 78 households and then analyze to what extent it is possible to fuel their BEVs with self-generated power.

Our study confirms findings of more theoretical work, e.g., by Munkhammar et al. (2013) that there is only a limited potential for increased self-consumption for BEVs if they charge uncontrolled. However, we can extend these results and demonstrate that this potential is greatly enhanced if an appropriate charging algorithm is used.

Paper outline

In the next section, we review previous research related to our case study. Section 5.1.3 explains the data used in detail and highlights strengths and weaknesses of the dataset. In Section 5.1.4 we explain how the data sources are integrated, in particular the different movement and mobility tracking data, as well as the data from the building and PV production models. Finally, section 5.1.5 presents the results. In the last sections we discuss our findings, potential generalizations and conclude the paper by providing an outlook on future research.

5.1.2 Background

Decarbonization paths for the transportation sector

The energy consumption of the transport sector is still mostly covered by fossil fuels. Globally, we use about 32 % of our energy for mobility and transport. Out of these 32 % only about 3.3 % are covered by renewable energy sources (REN21, 2020).

On a European level road transportation accounted for 26 % of the EU GHG emissions in 2018 and was the only energy sector that has increased its GHG emissions since 1990 (Pilzecker et al., 2020). The situation in Switzerland is similar where the transport sector is with 32.4 % of the GHG emissions the largest emitter (Schilt, 2020).

With regard to these numbers, it is clear that a quick decarbonization of the transport sector has to become a high priority in view of reaching the goals formulated in the Paris Agreement. Among the strategies to decarbonize the transport sector, we find soft incentives targeting mobility behavior (Bucher et al., 2019b; Cellina et al., 2019; Froehlich et al., 2009), policy interventions that use a top-down governing approach (Nash and Whitelegg, 2016), and technological innovations such as synthetic fuels (Çabukoglu et al., 2019; Küng et al., 2018) or battery improvements that foster quicker transitions to electric vehicles (Hu et al., 2017). While many of these proposals consider a wide range of different decarbonization measures, our study focuses only on BEVs as electrification is likely to be the main factor for the decarbonization of transport. Ruhnau et al. (2019) summarize their review of 22 scenarios from 12 independent studies about possible decarbonization paths as: “the more the emissions are to be reduced, the more road transport is expected to be electrified” (Ruhnau et al., 2019, p. 997). At this point, it is important to mention that the environmental impact

of BEVs will depend on the conditions under which the battery was produced (McManus, 2012; Saner et al., 2013) and on the emissions created to generate the energy necessary for charging the BEV (Faria et al., 2013).

It is estimated that by 2030 approximately 20 % of all vehicles in Switzerland will be electric (Haan and Bianchetti, 2016), a value that changes to around 6 % (or 125 million vehicles) globally (Bunsen et al., 2018). This is driven by the desire of many people to travel and behave more sustainably, but also because the costs for BEV batteries are falling rapidly and are expected to decrease further (Nykqvist and Nilsson, 2015).

Grid impact of battery electric vehicles

Apart from the mostly positive impact on CO₂ emissions, BEV charging has a potential impact on the stability of the distribution grid (Clement-Nyns et al., 2010; Hoarau and Perez, 2018; Richardson, 2013; Stiasny et al., 2020). Clement-Nyns et al. (2010) propose and evaluate coordinated charging strategies for BEVs to keep the impact of residential distribution grids within bounds. They find that BEV penetration levels of 10-30 % already significantly impact power losses and grid voltages, which can be counteracted to some degree by coordinated charging strategies. On the other hand, the smart chargers of BEVs could even be used to help stabilize the grid, especially considering that vehicles are typically unused 23 out of 24 hours per day (Zah and Haan, 2012).

Photovoltaic potentials and self-consumption of solar power

A potential solution to both problems, the slow decarbonization and potential grid impact of a BEV large-scale roll-out is utilizing the distributed nature of many renewable energy sources and charging the BEVs using locally generated power. As Luthander et al. (2015) note, problems arising from high penetration of distributed intermittent power generation such as the threat to exceed voltage limits can be mitigated with increased self-consumption of the distributed generation and Hoarau and Perez argue that BEVs in combination with PV systems can lead to higher self-consumption (Hoarau and Perez, 2018).

Of particular interest for distributed generation are PV installations on rooftops. Researchers explored the automatic extraction of rooftops from satellite or remote sensing imagery, e.g., Wiginton et al. (2010) analyze the PV potential in Ontario, Canada, and find that approx. 30 % of the region's energy demand could be met using on-house PV installations. Similarly, Ordóñez et al. (2010) classify satellite imagery by hand to estimate the potential of PV energy in Andalusia, Spain and find that around 80 % of the residential housing sector energy demand could be covered. Several studies computed the rooftop PV potential in Switzerland. In (Buffat et al., 2018a), the rooftop PV potential was modeled based on a detailed digital elevation model and building footprints. They estimated the PV energy potential in the range from 48.6 TWh to 58.8 TWh if the rooftops of all buildings within Switzerland would be covered by PV panels and assuming a conversion efficiency from solar irradiation to generated electricity of 10.33 %. Compared to the total Swiss electricity

consumption of 58.3 TWh in 2015, this is a considerable amount. Note that especially in regions like Ontario or Switzerland, solar irradiation is largely dependent on the seasons of the year³. As costs of PV panels have dwindled over the last decade, more and more PV installations are available. It is aimed that PV energy will contribute at least 11.4 TWh by 2030 and at least 34 TWh by 2050 to the total energy demand of Switzerland (BFE, 2020, 2019a).

With dropping prices for PV panels, it becomes increasingly interesting to add batteries to individual homes in order to further decrease the stress on the grid and bridge cloudy days. These batteries have the potential to reduce peak loads, especially in combination with smart meters and adaptive control systems (Mahmud et al., 2018). Regarding the direct influence on electric mobility, Buffat et al. (2018b) analyzed the effects of home-installed batteries on the provision of energy for BEVs used for commuting. The study shows that vehicles are concentrated during working hours in areas with a high workplace density while a significant share of the PV production is in suburban areas. Using home-installed batteries would enable using the produced PV energy locally without the grid. However, as this study used no real trajectories of BEV the assumed charging scenarios could not reflect the spatio-temporal characteristics of PV energy production and BEVs.

Effects of self-consumption on the decarbonization of the mobility sector

(Mohammadi and Taylor, 2017) show that human mobility and power consumption in buildings are strongly related. Using the battery from electric vehicles to increase the self-consumption of a household is seen as difficult because of the low coincidence between PV generation and the charging of electric vehicles (Munkhammar et al., 2013). It was also shown previously that changing mobility behaviors in combination with the provision of PV energy provides important steps to reach lower greenhouse gas emissions and thus fulfills the goals of various energy strategies (Bucher et al., 2019a; Buffat et al., 2018b). However, the question stays how much residential roof-top solar power may contribute to the decarbonization of the mobility sector by charging electric vehicles. Also, the multi-modality of transportation and flexible working concepts like home office are forecasted to increase in the future (the latter even received an unexpected boost due to the global COVID-19 pandemic), which could unlock new potentials for the usage of BEV charging to increase residential self-consumption.

³To help people plan individual PV installations, a number of online calculators and estimators are available such as www.sonnendach.ch from the Swiss Federal office for energy or www.solarpotenzial.ch from a local grid and power plant operator.

5.1.3 Data

Mobility data

We analyze the energy consumption from BEV usage, based on Swiss BEV users who participated in the SBB Green Class E-Car pilot study⁴ from Feb 1st to Dec 23rd, 2017. Participants of the study were equipped with a comprehensive Mobility as a Service (MaaS) package, containing (among others) a general public transport pass valid everywhere in Switzerland and a BEV for their personal use. Participation in this pilot study required paying a participation fee of 12'200 CHF, which led to a sample of participants with above-average income who travel more than the average Swiss person (Martin et al., 2019a). All study participants had access to a garage to install a home-charging station for the BEV, 83 % of the participants lived in single-family homes and all participants had access to more than 1 car at a household level (including the E-Car), however, participants had to pledge that at least 80 % of all drives of the BEV are done by themselves. The original Green Class E-Car pilot study had 144 participants of which we excluded participants who dropped out of the study early, who never used their BEV, users who could not be matched to houses (see Section 5.1.4), users who were matched to vacation homes or work locations (distinguished based on mobility patterns) and users with too large houses that resulted in PV system sizes with peak generation capacities over 30 kWp. For the remaining 78 users we also exclude the data between the 24st of December and the 31st of December because of low data coverage during these days. Furthermore, about 5 days of data are missing for most users in late September and early October (Martin et al., 2019a). These days are therefore also excluded from the study. Figure 5.1 shows the percentage of cars that are at the home location for every hour of the day over the full study period. The Figure allows to draw conclusions on the usage behavior of the study participants as it shows distinct usage patterns during the night and during the day as well as during the week and on week-ends.

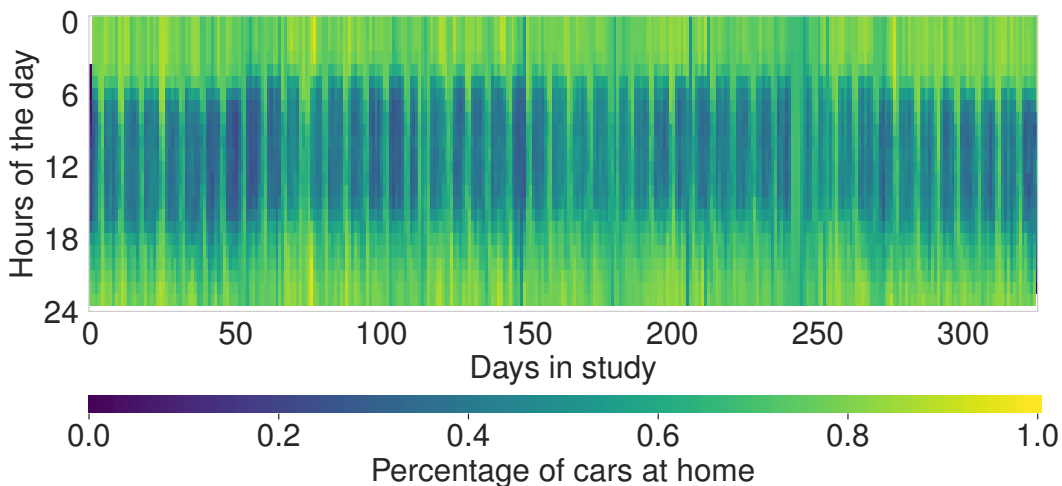


Fig. 5.1.: Share of BEVs at home or not at home

⁴More information about SBB Green Class can be found under www.sbb.ch/en/travelcards-and-tickets/railpasses/greenclass/pilotprojects.html.

The BEV of the participants records data event-driven and creates a new entry when the engine is started, the engine is stopped, the car starts charging or the car stops charging. For the start and the end of every entry, the following relevant information is stored: Timestamp, mileage, SoC, and geographic position. The entries are labeled as either *charging*, *driving*, *pausing* or *gap*. In total, the data set of the 78 used BEVs contains 189'446 entries with 77'065 drives and 19'867 charging processes and covers a total distance of 791'273 km. This corresponds to an average of about 10'145 km per car accumulated over 10 months.

Data used for solar irradiation estimation

The *Federal Register of Buildings and Dwellings*⁵ contains for each building in Switzerland associated properties such as a unique identification number, the coordinates of the building, the complete address, or the building type (e.g. if it is a single or multi-family house). In combination with the home addresses, we can match each participant to a building. However, the register does not contain the exact building footprint or rooftop shape (i.e., its slope and orientation), which is required for an accurate estimation of the PV potential. We retrieve building shapes from the swissBUILDINGS3D 2.0 dataset of the Federal Office of Topography⁶. This dataset contains the most current building data available for Switzerland and includes 3D vector data for each building. Furthermore, to model the shading effects of canopy and topography we use a digital surface model (DSM) with a resolution of 0.5 meters derived from LIDAR data (Buffat, 2016). To extract the PV potential in 2017 (when the mobility data was recorded), we additionally utilized solar irradiation data as provided by the Satellite Application Facility on Climate Monitoring (CM SAF) Surface Solar Radiation Data Set – Heliosat (SARAH) (Müller et al., 2015). This dataset contains direct as well as diffuse solar irradiation data derived from weather satellites and covers Switzerland with a spatial resolution of 0.05°, corresponding to rectangular grid cells of roughly 3.8 km × 5.6 km, and provides one data point for each cell every 30 minutes. Additionally, to estimate the efficiency of PV cell we use hourly ambient temperature data from MeteoSwiss weather stations extracted from IDAweb⁷.

Combining the building footprint, DSM and solar irradiation data (cf. (Buffat et al., 2018a)) lets us compute accurate energy production values for each building in intervals of 30 minutes.

⁵More information can be found under <https://www.bfs.admin.ch/bfs/en/home/registers/federal-register-buildings-dwellings.html>

⁶More information can be found under <https://shop.swisstopo.admin.ch/en/products/landscape/build3D2>

⁷More information can be found under <https://www.meteoschweiz.admin.ch/home/service-und-publikationen/beratung-und-service/datenportal-fuer-lehre-und-forschung.html>

5.1.4 Methods

Mobility data preparation

We preprocess the BEV tracking data in two steps. At first, we fill in missing tracking locations, then we aggregate all entries where the BEV is successively at the home location or not at the home location of the user.

Out of all available BEV entries 34 % have a missing start and/or end location. However, we can define rules to fill in trivial missing locations that leverage the available location information and the reading of the mileage counter. Given a BEV, we define the following rules about its starting location l_{start}^i and its stopping location l_{stop}^i for a tracking entry i :

$$\begin{aligned}l_{start}^i &= l_{stop}^i & \forall \Delta m_i = 0 \\l_{stop}^i &= l_{start}^i & \forall \Delta m_i = 0 \\l_{start}^i &= l_{stop}^{i-1} \\l_{stop}^i &= l_{start}^{i+1}\end{aligned}$$

with Δm_i being the difference of the mileage counter between the start and the end of the same segment. The first two rules state that the car is still at the same location when the mileage counter has not changed, the third rule states that a car starts at its last stopping location, and the last rule states that a car stops at its next starting location. If any of the right-side locations is known, we can use it to fill in the left counterpart. These rules are sufficient to fill 61% of the gaps which leaves 13% of the total tracking entries without complete location information.

In the next preprocessing step, all BEV entries of a user are aggregated into segments where the BEV was continuously either at the home location or not at the home location. For this we label the tracked locations as *at home* if the car is less than 500 meters away from the users' home location as it is defined in Section 5.1.4. The threshold of 500 meters is chosen relatively high because the BEVs are often parked in the garage which leads to a higher than usual GPS localization error. We then aggregate all information of all entries when a BEV was continuously *at home* or *not at home* into one new entry that we call segment. We especially keep the first starting time of an entry and the last ending time of an entry as the start and end time of the segment, the total energy consumption within the segment and the difference in the SoC of the first and last entry of the segment.

A one-week example of a single BEV user can be seen in Figure 5.2. Here the dark green patches correspond to segments where the BEV was at the home location and the light green patches correspond to the segments where the BEV was not close to the home location. The black line shows the SoC of the BEV over this period and the blue line the theoretical PV generation available at the house of the BEV owner (see Section 5.1.4 for information about PV generation).

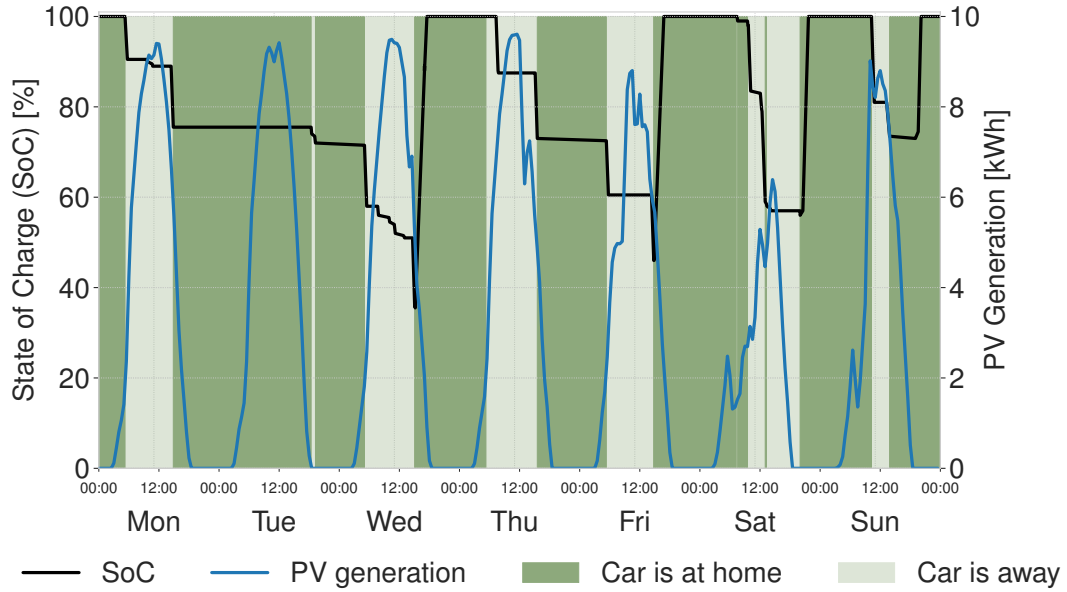


Fig. 5.2.: Exemplary week of a single user showing the usage of the BEV as well as the energy production by rooftop PV. In this recorded charging pattern (i.e., the baseline scenario, as explained below), it would be beneficial to not immediately recharge the car to the maximum, but instead wait for periods of increased PV energy generation.

Energy demand of the BEV

The BEV provides data about the SoC at the beginning of a drive and at the end of a drive, as well as the consumed energy. However, we do not know exactly how much energy is required to fully charge the BEV based on a specific SoC. We take advantage of the large amount of data we have access to and determine how much energy is required to fully reload the BEV as the solution to an ordinary least squares problem:

$$\hat{\alpha} = \arg \min_{\alpha} \|y - X\alpha\| \quad (5.1)$$

For the fit we exclude outliers where either the SoC or the consumed power are ≤ 0 and we enforce the intercept to be 0 to account for the physics of charging, meaning that there is no change in the SoC if no power was consumed. We chose a polynomial with a square root term in order to account for the small bend close to zero. Therefore, in our case y in Equation 5.1 is the consumed power during one drive and X is the data matrix that was combined using the vector of the change in the SoC and the vector of the square-root of the change in the SoC. Figure 5.3 shows a hex bin plot of the data together with the polynomial fit and a linear fit as a reference. The result of the least squares regression is:

$$\text{Consumed power} = 0.293 \text{ kWh} \cdot \Delta_{SoC} + 0.232 \text{ kWh} \cdot \sqrt{\Delta_{SoC}} \quad (5.2)$$

This method assumes that the battery capacity of the BEVs of all users is constant over the study period. All our study participants use the same (newly bought) type of BEV; however, we are neglecting ageing and temperature effects.

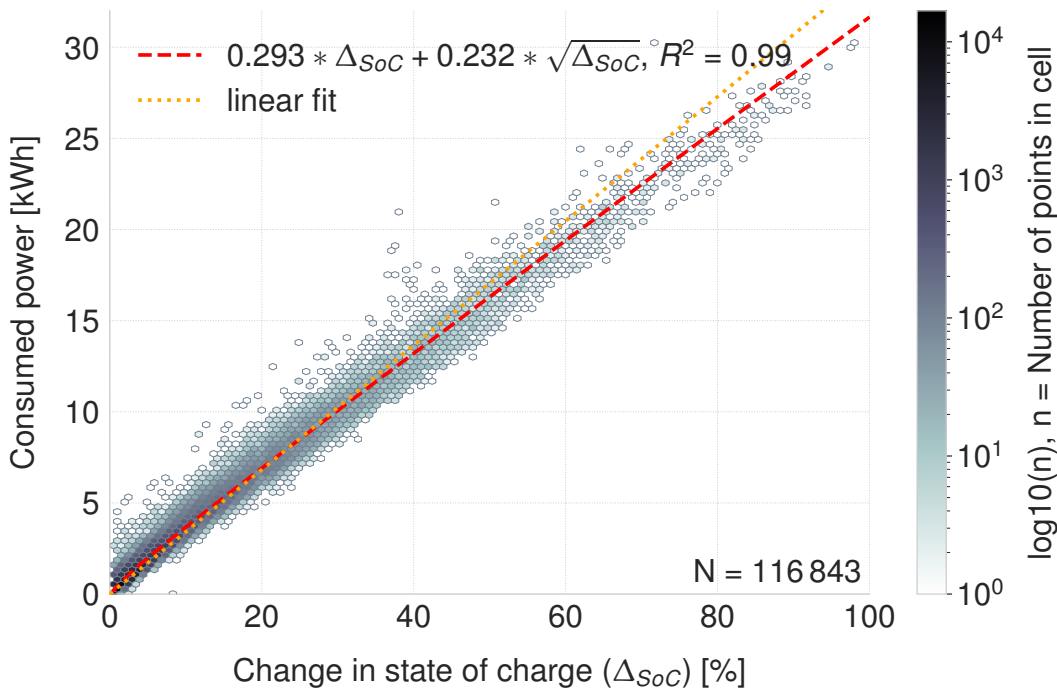


Fig. 5.3.: Scatterplot of the difference of the SoC before and after a drive and the consumed energy.

Over the whole study period the 78 BEVs consume and charge a total of 134.03 MWh of which 80 % is charged at home. Figure 5.4 (left) shows the average daily demand of all study participants and its 95 % confidence interval. This Figure shows that there is a slight seasonality in the consumption data such that participants consume more per week in cold months. Figure 5.4 (right) shows the daily energy demand for mobility over the course of the study and aggregated per weekday. It reveals the aggregated demand does not depend on particular weekdays which have median values between 6.28 kWh and 6.56 kWh with the exception of Sundays where the demand is considerably lower with a median value of 4.63 kWh.

PV generation

To estimate the PV energy production potential of rooftops, we use building footprints, a custom-created DSM, and satellite-based spatio-temporal solar irradiation data (at a spatial resolution of roughly 5 km and a temporal resolution of 30 min). We calculate the solar irradiation for each cell of the DSM within a building footprint for every 30 min during the case study period taking the shadowing effects of neighboring buildings and topography into account. Thereby we assumed that the complete roof can be used for PV generation. As the available technology impacts the overall electricity generation, we model the PV generation

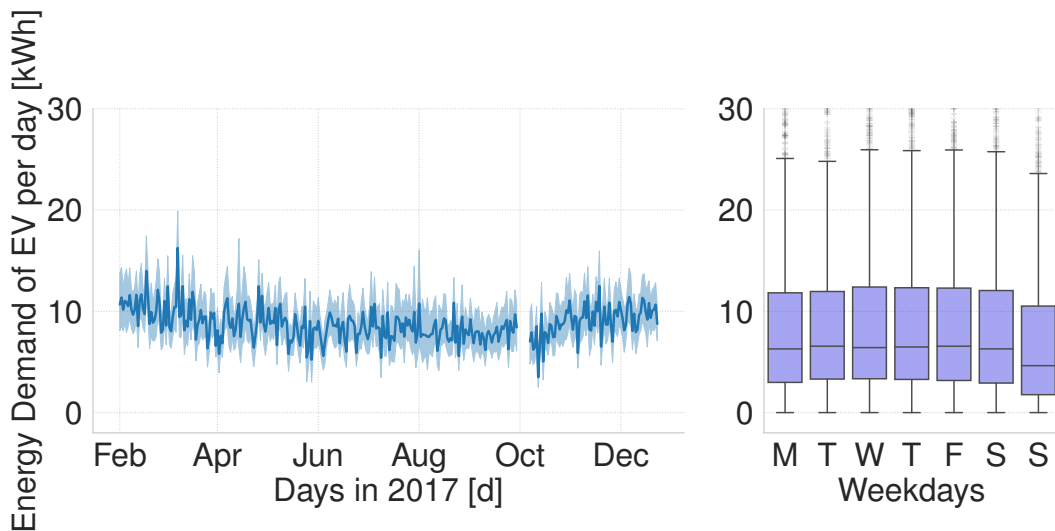


Fig. 5.4.: Daily energy demand of the electric vehicles of study participants. On the left, the average demand and its 95% confidence intervals are shown over the course of the study. The data shows a week of excluded data in the beginning of October. On the right, box plots of individual weekdays are shown. Outliers in the box plot go up until about 100 kWh which corresponds to the consumption of three full charges in a single day.

two folds. For all detailed analyses, we model the PV cell and inverter efficiencies for each 30 min based on the current conditions. The PV cell efficiency is estimated based on the PV cell properties (see A.4.1), the amount of solar irradiation hitting the cell, as well as the ambient temperature. Inverter efficiencies are modeled using the *PVWATT*⁸ model using the yearly peak production to dimension the maximum power of the inverter.

In a second experiment in Section 5.1.5, we investigate the sensitivity of the PV energy conversion and model only the effect of shadowing and assume otherwise constant efficiencies for PV panels and inverter. The applied workflow is based on (Buffat et al., 2018a), but differs in the following key areas:

- We modified the digital terrain model used in (Buffat et al., 2018a) by replacing all buildings using the *SwissBuildings 2.0*⁹ dataset. This is achieved by rasterization of the 3D vector shapes of the *SwissBuildings 2.0* with the same spatial resolution of 0.5 m of the existing DSM. Thereby, we ensure to use the most recent available rooftop shapes while still modeling the surrounding topography, such as terrain features or vegetation.
- In contrast to calculating long-term mean time series, we calculate the electricity generation for every 30 minutes for the year 2017 for each building assigned to a BEV user.
- For each building, the ambient temperature of the nearest available weather station is used in the process to model the PV cell efficiencies. The temperature is corrected

⁸<https://pvwatts.nrel.gov/downloads/pvwattsv5.pdf>

⁹<https://shop.swisstopo.admin.ch/en/products/landscape/build3D2>

based on the elevation difference between the building and the weather station and a factor of $-0.66\text{ }^{\circ}\text{C}$ per 100 m difference of elevation¹⁰.

Photovoltaic potential

By combining the solar irradiation data (cf. section 5.1.3) with the geographic data describing the rooftops of participants' houses (cf. sections 5.1.4 and 5.1.4) we compute accurate estimations of the energy that could potentially have been generated using photovoltaic cells on each house.

Figure 5.5 shows the average daily PV energy generation of all participants over the course of the study period. It can be seen that the production in winter is a fraction of the one during summer months, that there are occasional days on which the production on all houses drops to values close to zero (mostly days where all of Switzerland is covered by clouds).

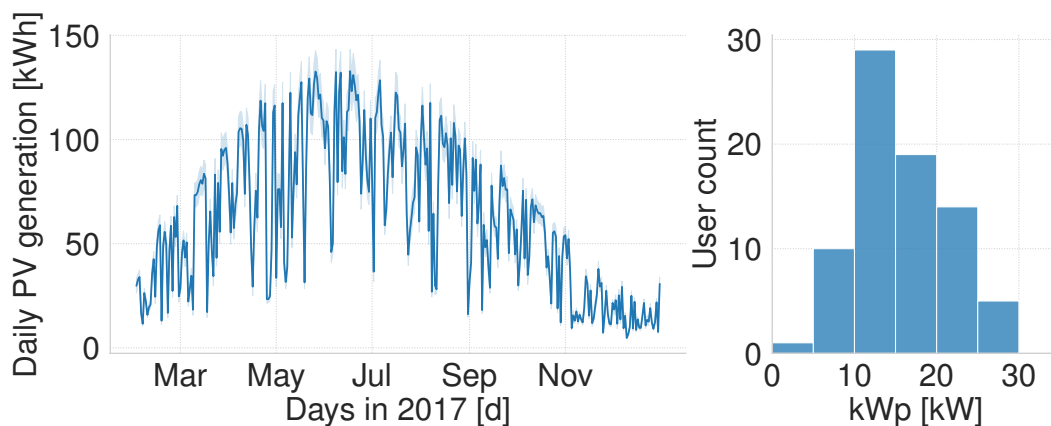


Fig. 5.5.: PV energy generation at homes of study participants. Left side shows the average daily PV generation per participant as well as the 95 % confidence interval over the course of the study period. The right side shows the distribution of the peak generation capacity per house.

Combination with PV model

We validate each match between home location and building footprint manually in a geographic information system (Paul A. Longley et al., 2011) by overlaying recent satellite imagery. 65 of the 78 persons live in single-family homes. We also include 10 users who live in two-family homes and 4 users who live in homes for multiple families with the condition that we could clearly identify the number of households living in this house. For houses with more than 1 household we later divide the PV system size and the generated power by the number of households as if the electricity would be equally shared among all households. The right side of Figure 5.5 shows the distribution of the resulting peak generation capacity for all users.

¹⁰This value was recommended by the Federal Office of Meteorology and Climatology MeteoSwiss as the vertical temperature gradient.

Smart charging scenarios

In this study we compare four different charging scenarios for the BEV. An overview of these charging scenarios is shown in Figure 5.6.

Scenario assumptions

All scenarios assume perfect knowledge of the PV generation as well as the mobility energy demand over the whole study period. These assumptions are justified because the goal of this work is not to present a novel operational smart charging algorithm but to analyze the potential to cover mobility energy demand using residential rooftop PV systems. An overview of existing smart charging algorithms for BEV charging using PV systems was presented in (Fachrizal et al., 2020). The temporal resolution of all scenarios with exception of the baseline is one segment in which the car was continuously at home or away from home. As described in Section 5.1.3 all users in the study have access to the same BEV with a battery capacity (cf. Section 5.1.4) of 31.61 kWh. As part of the project all participants installed a private charging station at home; we will assume that all participants have installed a 11 kW charger. In general, we do not explicitly model the plug-in behavior of a user, however, the plug-in behavior was recorded during the study and is implicitly considered in scenario 1. Scenarios 2 and 3 assume that the BEV is able to charge when it is at home.

Baseline scenario: recorded charging schedule

As a baseline we analyze the actual charging behavior of the BEV owners as it was recorded in the study. The charging of the BEVs in the tracking study is uncontrolled, i.e., if a user plugs in the BEV it immediately starts charging until it is fully charged or unplugged. An example of the recorded charging behavior is shown in Figure 5.2.

Scenario 1: segment-wise optimization

The first charging scenario, shown in the top left of Figure 5.6, optimizes the share of rooftop PV generation used for BEV charging by only shifting the charging schedule within the same segment (cf. Section 5.1.4 for the definition of segments). In this scenario, the SoC of the BEV at the start and the end of every segment is the same as the reference SoC (SoC_{ref}) that was recorded for this segment in the case study. We allow, however, to change the charging trajectory (e.g., by deferring charging) between these points. In the example given in Figure 5.2 this would mean that the SoC at the beginning and the end of every dark green segment is fixed but we can alter the time when charging takes place in between these points. The purpose of this scenario is to offer a charging schedule that optimizes the consumption of PV energy but remains as close as possible to the recorded data. By that, this scenario implicitly considers the plug-in behavior of a person on a per-segment level as we are only allowed to charge in this scenario when the user also charged in reality.

Scenario 2: cross-segment optimization

The second smart charging scenario, shown in the top right of Figure 5.6, implements a greedy approach to maximize the share of PV generation used for the charging of the BEV

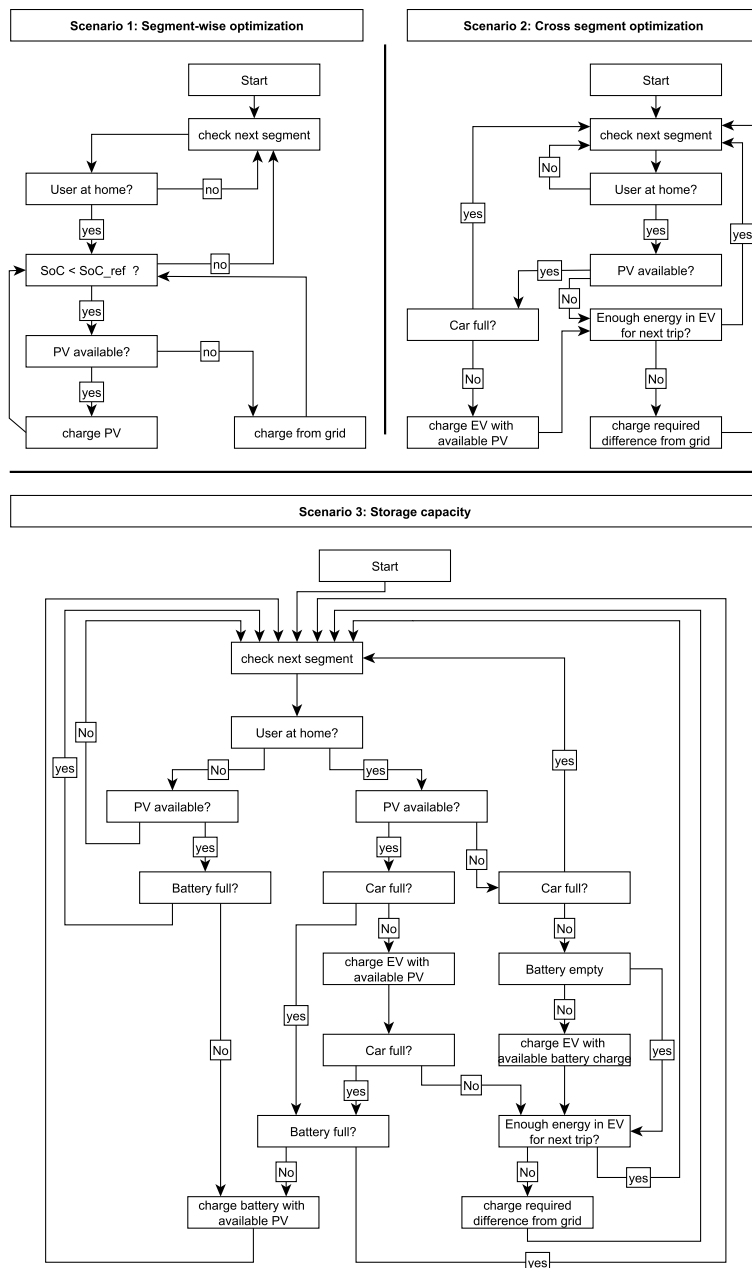


Fig. 5.6.: Flowcharts representing the charging logic of the different scenarios (top left: scenario 1 segment-wise optimization, top right: scenario 2 cross-segment optimization, bottom: scenario 3 cross-segment optimization with battery).

over the whole study period. Therefore, the BEV in scenario 2 always charges as much as possible using PV power and only takes a minimal amount of energy from the grid to be able to serve the mobility needs of the user. In the example given in Figure 5.2 we would be allowed to do any operation as long as we only charge when the user is at home, the maximum charging power is not surpassed and the SoC of the BEV is always greater than 0.

Scenario 3: storage capacity

In the third scenario we analyze the impact of a commercial home battery storage that is used to buffer the PV power generation. We set the capacity and the maximum symmetric power flow based on a commercial product to 13.5 kWh and 4.6 kW. An overview of the third scenario is shown in the bottom of Figure 5.6. The simulation runs as follows: We specify the battery size available. If there is energy generation available from PV, it gets charged to the electric vehicle if it is at home and not full. Otherwise, the generated power is stored in the battery.

The car is charged from either the PV power currently being generated or from the battery. Only if this is insufficient, energy from the power grid gets charged up to the minimum required level to fulfill the mobility requirements.

Calculation of CO₂ emissions

We evaluate the sustainability of the different scenarios by calculating their relative GHG emissions as CO₂ equivalent. For the comparison we include the GHG emissions of the power used for charging the BEVs, and the emissions generated by the production and the disposal of the rooftop PV and the battery system (scenario 3). We specifically exclude all emissions related to the production and the disposal of the BEV as we only use the calculated emissions to compare the scenarios against each other and the BEVs are the same in all scenarios. All used emission factors were determined in *cradle-to-grave* LCA studies with the exception of the emission factor for the German power mix which is only published based on marginal emissions.

Emission factor for grid charging

The emission factor for the Swiss household consumer mix is published by the Federal Office for the Environment (FOEN) and is 181.1 gCO₂-eq/kWh (Messmer and Frischknecht, 2016). While the power generated in Switzerland has considerably lower emissions, much of the Swiss energy is imported (often from Germany) and often stored in one of the many pumped storage power plants to be distributed at a later point in time. The emission factor for Germany is 401 gCO₂-eq/kWh (Icha and Kuhs, 2020) and is significantly higher even though it comprises only direct emissions.

Emission factor for PV charging

Even though PV power generation is considered a clean source of energy, it is not carbon free (Barros et al., 2020) and its cradle-to-grave emission factors are controversially debated

as they can strongly vary depending on the assumptions taken on production techniques (Nugent and Sovacool, 2014). The 2014 study from Nugent and Sovacool (2014) reported an average emission factor of 49.91 gCO₂-eq/kWh but noted that large-scale PV generation facilities are generally more efficient than rooftop PV systems. We assume an emission factor of 53.6 gCO₂-eq/kWh as this was reported by a recent study with a very similar use case to ours (Krebs et al., 2020). The factor is in line with the study of Nugent and Sovacool (Nugent and Sovacool, 2014) and another recent study (Jones and Gilbert, 2018) that reported average values ranging between 30 gCO₂-eq/kWh and 80 gCO₂-eq/kWh depending on the panel technology. We further omit emissions stemming from the hypothetical need of additional infrastructure (such as a distribution grid expansion) and system management losses as they are expected to only have a minor impact (Jones and Gilbert, 2018).

Emission factor for home storage

For the home storage we use a capacity specific emission factor of 76.1 kgCO₂-eq/kWh as it was reported for a recent lithium-ion battery (lithium cobalt phosphate) (Raugei and Winfield, 2019). We follow (Krebs et al., 2020) and assume a lifetime of 5000 cycles and an average usage of 1 cycle per day which results in an average lifetime of 13.7 years. To correctly account for the battery related emissions over the study duration, we calculate the average weekly emissions of the battery. A battery with a capacity of 13.5 kWh has then average emissions of 1.44 kgCO₂-eq per week.

5.1.5 Results

We calculate the detailed mobility energy demand and the rooftop PV generation potential for all users using the steps described in Section 5.1.4 and simulate a detailed charging schedule for each of the four charging scenarios presented in Section 5.1.4. In this section, we analyze the potential to cover the mobility energy demand using rooftop PV generation, the development of this coverage ratio over time, and the impact of the PV generation on the GHG emissions stemming from mobility. In the end of this section, we further analyze the sensitivity of the results to roof size, overall BEV usage and panel efficiency.

Coverage of mobility energy demand using rooftop PV generation

All scenarios were implemented as Python 3 scripts¹¹ following the procedures outlined in Section 5.1.4. The main result of this work is a detailed juxtaposition of mobility-induced electric energy requirements (stemming from individual motorized transport) and potential (local, i.e., on an individual's home's rooftop) generation using photovoltaics under the assumption of different smart charging scenarios.

Figure 5.7 shows a histogram per scenario for the overall coverage ratio per user for the full study duration. The coverage ratio is defined as the total energy charged from rooftop PV divided by the total energy charged by the BEV. Additionally, the average coverage ratio over

¹¹All code used for this project can be accessed via <https://github.com/mie-lab/rooftop-PV-EV-charging>

all users is given in the title of each scenario. As expected, the coverage ratio is increasing with increasing complexity of the scenarios. The baseline scenario represents the charging behavior of the users as it was recorded in the study. The evaluation of this scenario shows that uncontrolled charging results in a very low coverage of the mobility demand with rooftop PV generation of only 15 %. These results are lower than the reported self-consumption of the total household load in combination with a plug-in electric vehicle (Munkhammar et al., 2013) or the self-consumption reported in a case-study that looked at shared BEVs in a microgrid (Van Der Kam and Sark, 2015). However, already shifting the charging within individual at-home segments increases the coverage ratio by 41%. This is rather surprising given that the overlap between sunny hours and when the car is at home is thought to be comparatively low. Scenario 2 represents a theoretical upper boundary on how much of the individual's mobility behavior energy demand could be covered by PV generation under the assumption of perfect forecast of the upcoming PV generation and the mobility behavior. The results of scenario 2 show that 32 users could theoretically cover more than 95 % (height of the last bin) and the others substantially more than 50 % of their mobility energy demand by PV generation based on their own roofs. Scenario 3 follows the same charging strategy and assumptions as scenario 2 but uses an additional home energy storage to buffer energy generated by PV if possible. In this scenario almost all users can cover their demand by 100 %; while this is very high, the difference to scenario 2 is small.

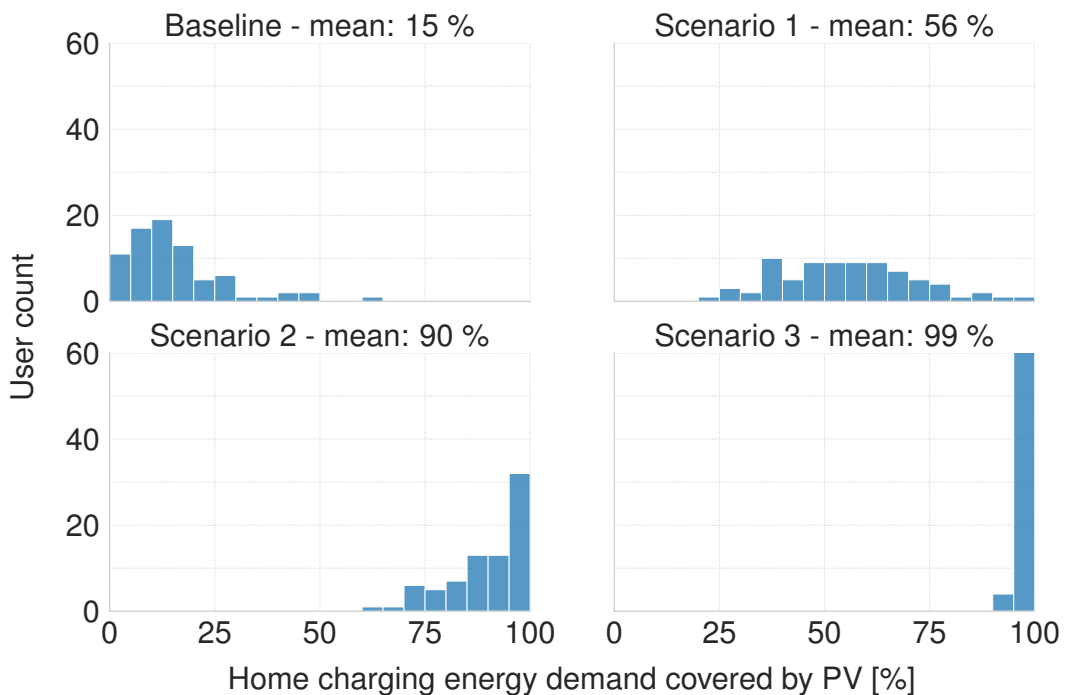


Fig. 5.7.: Share of energy (required for mobility) generated by local rooftop PV for each user in all scenarios over the full study duration.

Coverage ratio over time

Figure 5.8 shows a different view on the results. Here, we plot the cumulative sum of the energy that was used by the cars against the net energy generated (and utilized) by PV. In this graph the two lines through the origin $x = y$ and $x = -y$ represent the case when only PV generation or only energy from the grid was used to charge the BEV. The graph is based on the data of all study participants and was sorted by time before creating it. The total energy used, drawn on the x-axis, is therefore monotonic over time and can be seen as a proxy for temporal progression. This graph shows again that scenario 2 and scenario 3 are close to optimal over the whole study period, as opposed to the baseline charging strategy which is close to the worst-case scenario. Both, the baseline scenario and scenario 1 show a hump in the second half of the graph which is already a hint for the (expected) influence of seasonality on the PV generation and therefore on the results.

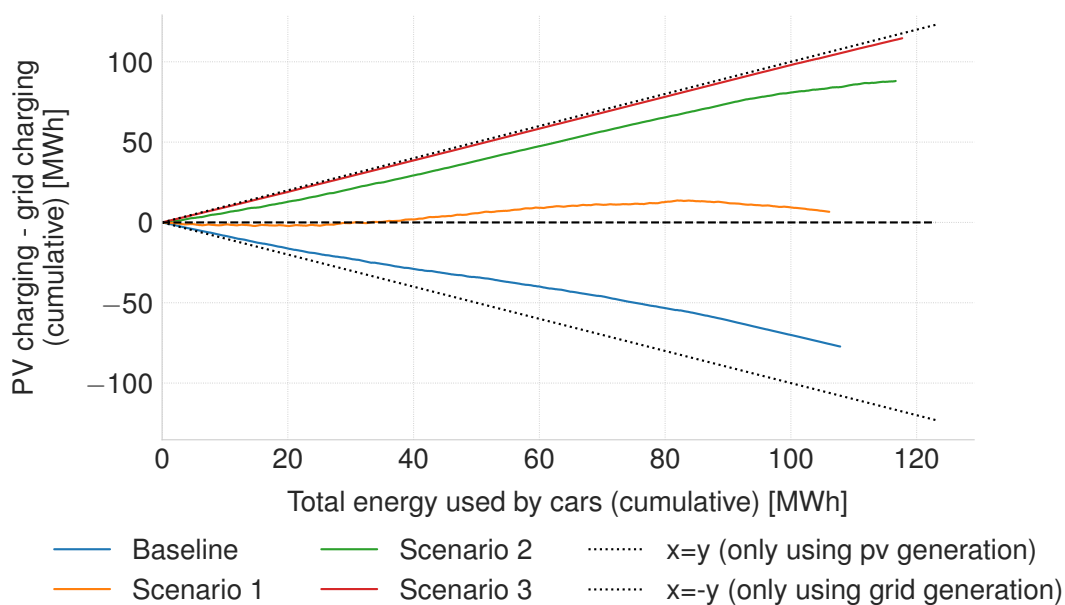


Fig. 5.8.: Energy sources (rooftop PV or the grid) for the (cumulative) energy demand of BEVs for each scenario.

Figure 5.9 shows the distribution of the coverage per week per user for every scenario. In this graph the seasonality effect is clearly visible. In all scenarios except for scenario 3 coverage is higher during the sunny months. Figure 5.9 could support the assumption that the coverage ratio is increased due to the higher generation during the summer months which could also lead to the assumption that the result mostly depends on the roof size and the installed generation capacity. This question is further investigated in section 5.1.5.

Effect on CO₂ Emissions

We calculate the average CO₂ emissions of each user on a weekly basis. Therefore, we multiply the sum of the weekly energy usage per user with the emission factors discussed

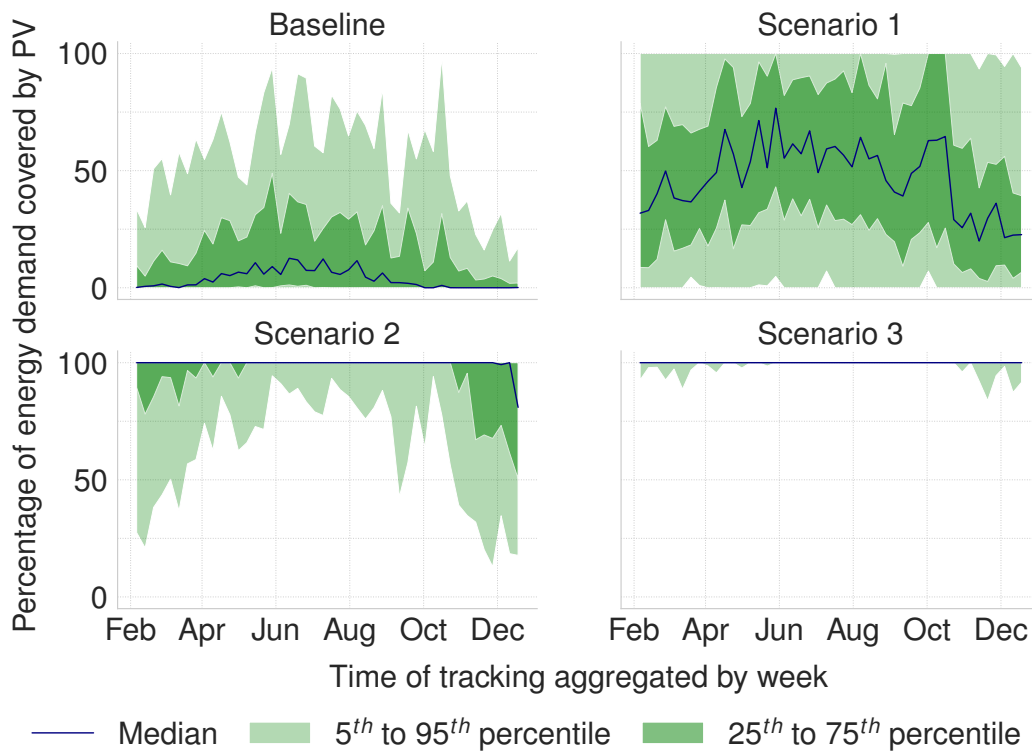


Fig. 5.9.: Shares of energy produced by rooftop PV for each user over the duration of the study period.

in Section 5.1.4. For the used power we distinguished between energy used from the grid and energy used from the rooftop PV system. For scenario 3 the emissions of the battery are added as a constant factor each week, PV generation that was temporarily stored in the battery is accounted for in the rooftop PV emission factor.

Figure 5.10 shows the average emissions over all users for the different scenarios. The left side of the Figure shows the results for the Swiss electricity mix. It demonstrates that there is a significant potential of saving CO₂ emissions. For the comparably clean Swiss power mix, every user would save on average 2.93 kgCO₂-eq per week for scenario 1, 4.11 kgCO₂-eq per week for scenario 2, and 3.13 kgCO₂-eq per week for scenario 3 in comparison to the baseline scenario. Further, the graph shows that already scenario 1 has a significant potential to save emissions when compared to the baseline. We can also see that the battery in scenario 3 is not able to amortize its initial emissions with higher PV usage which leads to weekly average emissions that are often higher than those of scenario 1.

The right side of Figure 5.10 shows the same analysis for the German power mix. The result follows a similar trend but the absolute difference in emissions is larger. For the comparably emission-heavy German power mix, every user would save on average 7.29 kgCO₂-eq per week for scenario 1, 10.78 kgCO₂-eq per week for scenario 2, and 10.63 kgCO₂-eq per week for scenario 3 in comparison to the baseline scenario. Furthermore, it shows that scenario 3 now achieves lower emissions than scenario 1 and also lower CO₂ emissions than scenario 2

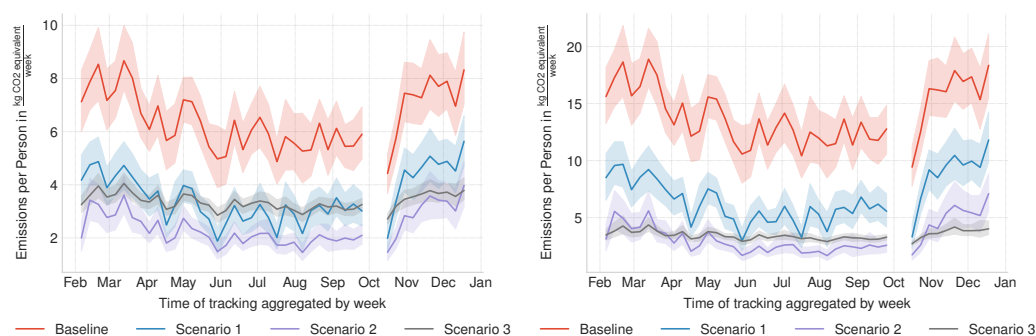


Fig. 5.10.: Average greenhouse gas emissions stemming from electric mobility for all users. On the left side, a typical Swiss power mix is assumed, on the right side a German power mix (only the direct emissions considered for Germany). Two weeks in October affected by missing data are excluded. Uncertainty intervals show the 95% confidence interval of the mean.

in early spring and late autumn. The higher difference in emissions between power from the grid and the rooftop PV generation allows for a better amortization.

Sensitivity

To provide insights into the reasons for the coverage variance between users, we analyze the influence of the roof size, the total energy demand, and the overall efficiency of the PV system.

Influence of PV system size and charging demand

Intuitively we would expect that it is easier to cover the mobility energy demand if a study participant has a large rooftop PV system or if she does not use the car very often. However, our analysis suggests that the influence of these factors is rather small.

Tab. 5.1.: Results of the regression analysis examining the correlations between total energy demand for mobility resp. peak PV power and PV coverage ratio.

	Demand		PV peak power		R ²
	Coef. $\frac{[\%]}{kWh}$	p	Coef. $\frac{[\%]}{kWp}$	p	
Baseline	-4.94E-03	2.40E-02	4.68E-01	4.73E-02	0.10
Scenario 1	-1.10E-02	2.20E-04	6.29E-01	4.42E-02	0.19
Scenario 2	-9.00E-03	3.83E-08	1.93E-01	2.16E-01	0.33
Scenario 3	-1.44E-03	2.55E-05	7.07E-02	4.04E-02	0.23

To test the influence of these factors we run a regression analysis for each scenario using the average yearly coverage ratio per person as dependent variable and the PV system's peak power and the total mobility energy consumption as independent variables.

The results of this analysis are shown in Table 5.1. Total energy demand and PV system size are below the 5 % significance level for all scenarios except for the PV peak power for

scenario 2. However, the rather low R^2 values suggest that these factors can only partially explain the result of the analysis. A more detailed look at the regression coefficients further reveals that the linear relationship might be significant, but for the baseline scenario as well as scenarios 2 and 4 its influence is rather low as a change of several hundreds of kWh in yearly demand or a change of several kWp in PV system size are necessary to change the coverage value by 1 %. This can be also seen in the scatter plots shown in Figure 5.11, where it is visible that neither the total yearly electricity demand of the BEV nor the size of the PV system can explain a substantial part of the variance.

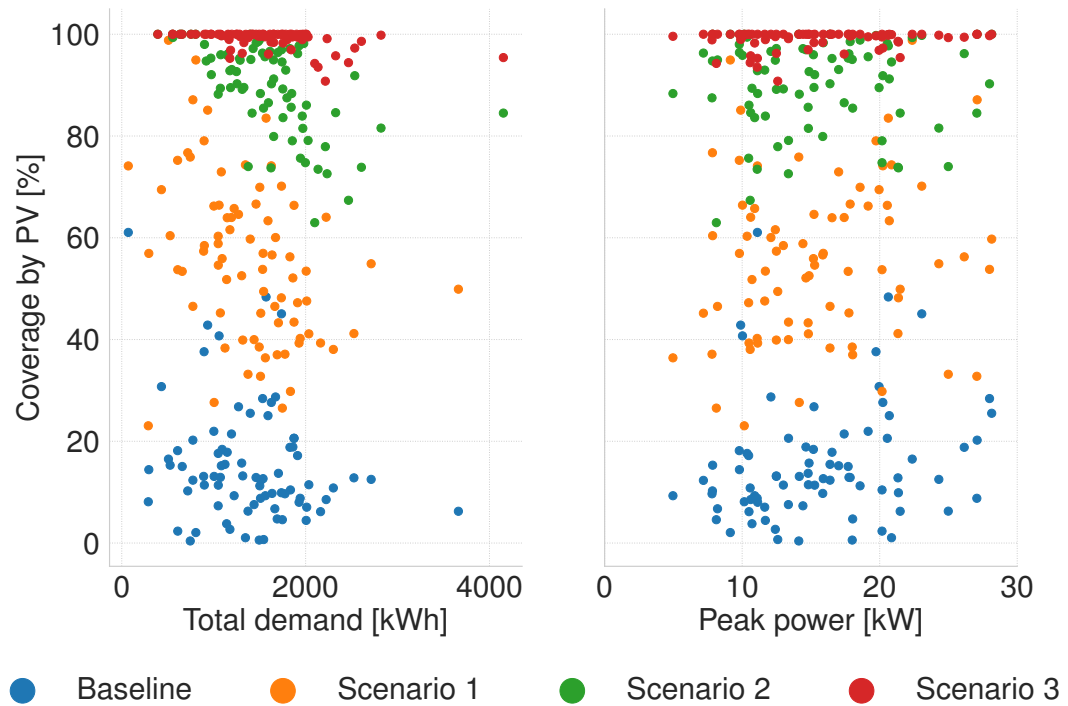


Fig. 5.11.: Analysis of correlation between total energy demand for mobility resp. peak power production on individual rooftops and energy coverage by PV power. In red, the baseline scenario is shown, in blue scenario 1, in purple scenario 2 and in grey scenario 3.

Influence of overall PV system efficiency

In addition to the sensitivity to PV system size and total energy demand, we tested the influence of varying the overall PV system efficiency. We therefore repeated the main experiment 24 times and calculated the coverage ratio for every user for all scenarios assuming a fixed PV system efficiency that varies its efficiency between 6 % and 29 %. The overall system efficiency here is defined as the ratio of generated power and incoming solar irradiation (see also Section 5.1.4). Figure 5.12 shows the average coverage value over all users and its 95 % confidence interval. The average coverage ratio grows quickly for low panel efficiency values and slowly saturates for efficiency values between 10 % and 16 % depending on the scenario.

A close look reveals that the uncertainty of the estimated mean slowly grows with growing panel efficiency for scenario 1 and 2 while it shrinks for scenario 3 and 4. This indicates that the scatter of the coverage values increases for the first two scenarios while it decreases for the others. Scenario 1 and 2 preserve the original mobility and plug-in behavior, the increase of the scatter in the data shows that there are some users who can better take advantage of the increased generation while others cannot, e.g., because they rarely plug-in their BEV. Scenario 3 and 4 optimize the PV usage with the mobility demand as the only boundary condition. In this case the only distinction between users are their usage patterns and an increased generation will lead to an increased coverage for all users that cannot yet fully cover their BEV's demand.

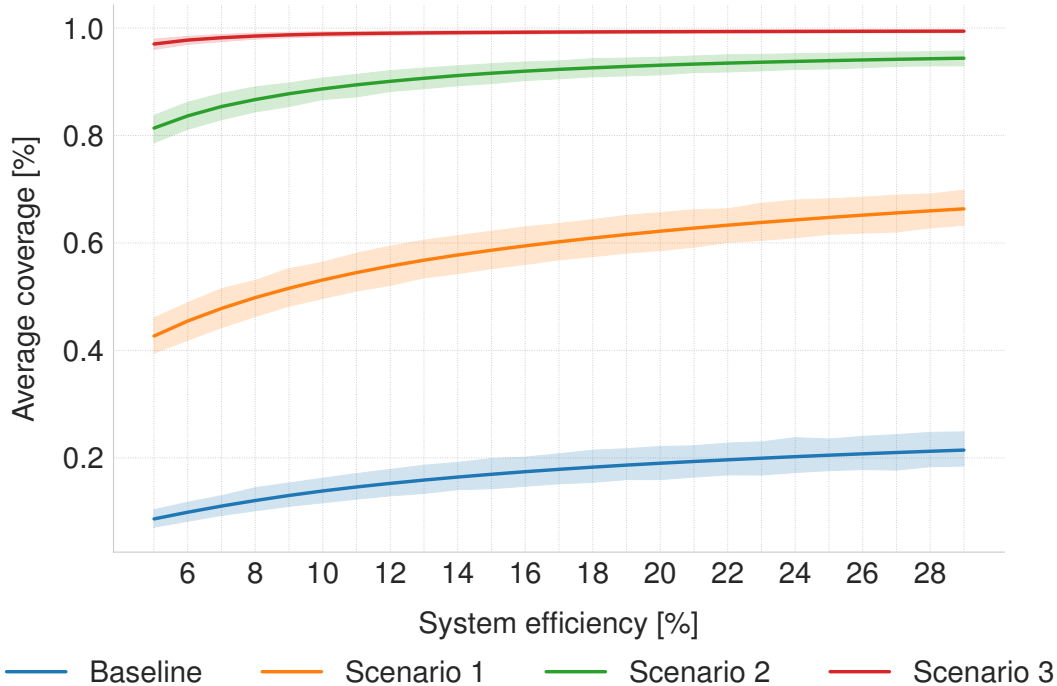


Fig. 5.12.: Average coverage value as a function of the overall PV system efficacy. The graph shows the average over all users per scenario and the 95 % confidence interval.

5.1.6 Discussion

Potential of home charging of EVs

The evaluation of the real-life charging schedules shows that there is a great potential to cover the mobility energy demand using rooftop PV generation. However, the results also show that uncontrolled greedy charging (as it is mostly the case at the moment) leads to almost worst-case results in terms of coverage (cf. Figure 5.9). The results from scenario 1 demonstrate that it is theoretically possible to increase the average coverage ratio by 41% without requiring any change in the behavior of the BEV owner (e.g., to always plug in the BEV). These results could be achieved by two simple adjustments to smart charging

algorithms: Restrict the maximum charge to the available PV generation as long as sunlight is available and do not fully charge the BEV during the night to have the possibility to take advantage of the sunlight in the morning hours.

Scenario 2 shows that almost all BEV owners could fully cover their mobility energy demand using rooftop PV generation. In practice it will be hard to achieve these values as it requires good forecasts of the PV generation and the individual mobility energy demand. In this study the results of scenario 2 are already close to optimal which could lead to the conclusion that the installation of an additional home battery storage is unnecessary. However, the design of a good charging strategy that relies on uncertain forecasts becomes significantly easier with the possibility of buffering the energy. This consideration could thus justify the installation of a home battery system.

Potential of rooftop PV to decarbonize transport

Even though BEV have lower GHG emissions than fossil fuel dependent ICEVs, their emissions depend strongly on the technologies used for power generation. The analysis of the average CO₂ emissions in the different scenarios reveals the large potential that the combination of rooftop PV generation and BEVs offers for the decarbonization of individual mobility. The reduction potential depends on the emission intensity of the energy mix available to the consumer, which means that the reduction potential becomes lower over time with the ongoing progression of the decarbonization of the energy sector. The impact of home-charging might be reduced if a BEV can be charged using PV at the workplace (Nunes et al., 2016) (e.g., if a person can charge her BEV using PV at the workplace, the BEV might be almost full when returning home).

Potential of rooftop PV and BEVs for increased grid stability

The stability of the distribution grid is an important factor that has to be considered during the expansion of PV generation as well as during the roll-out of BEVs. It is realistically feasible to cover a large portion of the mobility energy demand using the own rooftop PV generation and by that residential roof-top PV might be able to lower the grid impact of BEV charging. However, the inverse might not be true as for PV the critical point is the peak in the middle of the day where BEVs are usually not available at home.

Limitations

Due to the employed methods and because of the used datasets the results of this study are subject to several limitations.

The movement data used to estimate the BEV demand is entirely sourced from Switzerland, a country that is known for its excellent transportation and public transportation infrastructure. The transfer of the results to countries with differing infrastructure might be limited due to the impact that available infrastructure has on BEV demand.

Furthermore, the developed methodology cannot be used to directly derive operational smart charging algorithms due to the assumptions taken (see Section 5.1.4). Most notably, we are assuming the availability of perfect forecasts for mobility energy demand and the available solar generation. The results therefore present an upper limit of the achievable performance of an operational smart charging algorithm.

Lastly, we are not considering seasonal effects for the calculation of the CO₂ emission factor for grid charging in Switzerland, but simply use the yearly average. This might lead to an overestimation of the CO₂ emission factor for grid charging in summer and an underestimation in winter.

5.1.7 Conclusion

In this work we presented a detailed case study as a contribution to the open question to what extent BEV owners can be self-sustainable using rooftop PV generation. To answer this question, we combined the detailed mobility data records from 78 BEVs in combination with a detailed model of the BEV owners' roofs to estimate the potential rooftop PV generation.

Our results show that currently deployed uncontrolled greedy charging strategies (baseline scenario) lead to a very low coverage factor of only 15 % on average over all BEV users. Very simple adjustments without changing mobility or plug-in behavior (scenario 1) can already greatly increase this to an average coverage of 56 %. We argue that these values could be easily reproduced in practice by currently feasible smart charging algorithms if they charge slower and limit their charging intake to the currently available PV generation, and if they do not fully charge during the night in order to take advantage of the morning hours.

We further showed that the upper boundary of the coverage with and without additional home battery storage over all users is on average 90 % and 99 % respectively without restricting a person in their mobility. The degree to which these values can be achieved in practice will depend strongly on the quality of mobility demand and generation forecasts. Considering these uncertainties, it could be that the home battery storage will provide a more significant advantage than it does for the theoretical upper boundary.

No matter the charging schedule scenario, all of them greatly increase the consumption of rooftop generated PV which significantly reduces the emissions from driving the BEV. This highlights the large potential that rooftop PV power generation has to accelerate the BEV based decarbonization of the transport sector.

Future work should investigate smart charging algorithms that consider the goal of maximizing the usage of rooftop PV energy as it is cheaper and more sustainable for the user. Very high self-consumption rates for BEV charging will only be achievable if sufficiently good predictions of the energy demand of the BEV are available. However, reliable predictions of individual mobility behavior which induces the mobility demand are still an open problem and should be tackled in the future.

Acknowledgements

Funding: This work was supported by the Swiss Data Science Center [C17-14]; the Swiss Innovation Agency Innosuisse within the Swiss Centre for Competence in Energy Research on the Future Swiss Electrical Infrastructure (SCCER-FURIES); the Swiss Innovation Agency Innosuisse within the Swiss Centre for Competence in Energy Research - Efficient Technologies and Systems for Mobility (SCCER Mobility); the ETH Zurich Foundation [MI-01-19];

5.2 Graph-ResNets for short-term traffic forecasts in almost unknown cities

The following section is a reprint of the publication:

Henry Martin, Dominik Bucher, Ye Hong, René Buffat, Christian Rupprecht, and Martin Raubal (2020). “Graph-ResNets for short-term traffic forecasts in almost unknown cities”. *NeurIPS 2019 Competition and Demonstration Track*. Vol. 123. PMLR, pp. 153–163. DOI: 10.3929/ethz-b-000437682.

Conceptualization: HM, DB, YH, RB, CR; Methodology: HM; Software: HM; Writing: HM; Supervision: CR, MR

The content is reproduced ‘as is’, however, formatting changes and corrections of spelling have been applied.

Abstract

The 2019 IARAI *traffic4cast* competition is a traffic forecasting problem based on traffic data from three cities that are encoded as images. We developed a ResNet-inspired graph convolutional neural network (GCN) approach that uses street network-based subgraphs of the image lattice graphs as a prior. We train the Graph-ResNet together with GCN and convolutional neural network (CNN) benchmark models on Moscow traffic data and use them to first predict the traffic in Moscow and then to predict the traffic in Berlin and Istanbul. The results suggest that the graph-based models have superior generalization properties than CNN-based models for this application. We argue that in contrast to purely image-based approaches, formulating the prediction problem on a graph allows the neural network to learn properties given by the underlying street network. This facilitates the transfer of a learned network to predict the traffic status at unknown locations.

* Equal contribution

5.2.1 Introduction

Today 55 % of the world's population lives in urban areas. This number is expected to rise to 68 % by 2050 (UN, 2018). The growing urbanization in combination with population growth and private car ownership threatens to raise the already high levels of congestion which in turn increases pollution and economic costs and accelerates climate change (Reed, 2019). In order to facilitate the transition towards a more sustainable traffic system the growing traffic volume has to be managed in a smart way to reduce its negative impact.

A backbone of smart traffic management is short-term traffic flow predictions. They allow the detection of anomalies, such as accidents or obstacles (Wang et al., 2016), the smart routing of vehicles to optimally use the existing infrastructure (Ringhand and Vollrath, 2018) or the predictive control of the transportation system using traffic lights (Huang et al., 2018b). There is a vast body of literature concerning traffic forecasting available (Ermagun and Levinson, 2018; Vlahogianni et al., 2014), however, “*comparing the forecasting applications across studies is almost impossible*” (Ermagun and Levinson, 2018, p. 791) as studies use different spatio-temporal data resolution and report different error metrics on different aggregation levels.

Traffic4cast competition

In response to this lack of standardization, the Institute of Advanced Research in Artificial Intelligence (IARAI) published a novel, publicly available traffic forecasting benchmark dataset as part of the *traffic4cast* competition¹². The *Traffic4cast* dataset comprises traffic data from three cities (Berlin, Istanbul, and Moscow) and covers one year in 5-minute intervals. The data is given as three-channel images with normalized information about traffic volume, average speed, and average direction, which all range from 0 - 255. An example of the data can be seen in Figure 5.13. Here the logarithm of the per-city sum of all channels over all training images is shown as a proxy for the activity level per pixel. Only a few pixels are always zero over the whole dataset¹³, nevertheless, the street network is still clearly visible in each city.

The task in the *traffic4cast* competition is to predict the traffic of the next 15 minutes (3 images) based on the last hour (12 images). The results are then evaluated using a pixel-wise calculation of the mean-squared error (MSE) between the prediction and the ground truth.

Graph convolutional neural networks (GCNs)

The competition specifically encoded the traffic data as images to facilitate the usage of deep convolutional neural networks (CNN). However, the image-based representation omits explicit information about the street network and therefore disregards that the movement of cars is usually restricted to the road network. While a CNN-based approach will likely

¹²www.iarai.ac.at/traffic4cast

¹³29 % for Berlin, 11 % for Moscow and 23 % for Istanbul.

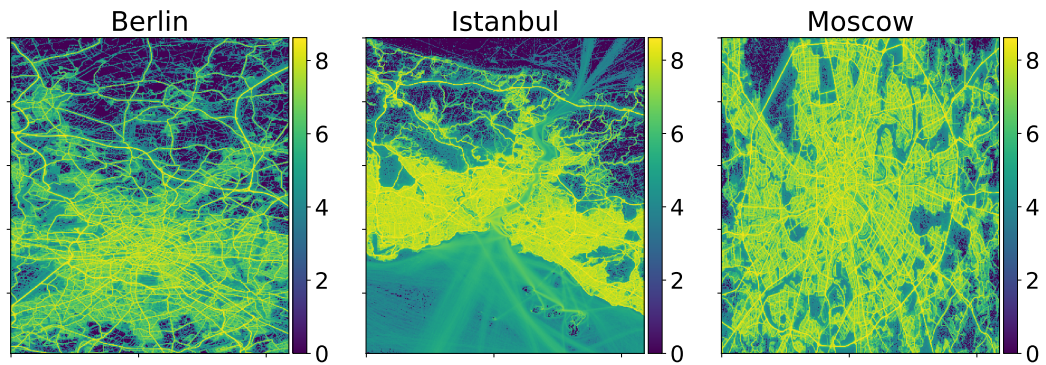


Fig. 5.13.: Logarithm of the per city sum of all training images of all channels.

be able to extract the street network and store it in its weights, the street network could be provided explicitly to greatly reduce the problem’s complexity.

GCNs have been developed over the past couple of years in an attempt to generalize the success of CNNs to irregularly structured domains that can often be described via graphs (Bronstein et al., 2017; Defferrard et al., 2016). By now, there are several propositions for graph (convolutional) neural network architectures that are described in the available review papers (Wu et al., 2019; Zhou et al., 2018). GCNs have already been applied to many different domains (Bronstein et al., 2017; Wu et al., 2019; Zhou et al., 2018) but there are only few examples that are related to the movement of people (Martin et al., 2018) or short-term traffic forecasting (Cui et al., 2019a; Yu et al., 2018; Zhang et al., 2019).

Contribution

In this work, we use the image-based U-Net approach of team MIE-Lab (Martin et al., 2019c)¹⁴ that won the second place in the *traffic4cast* competition and compare it with an alternative approach based on GCNs. We use well-known GCN architectures as well as suitable modifications such as the Graph Residual Network (Graph-ResNet) which is inspired by the residual learning network (ResNet) presented in (He et al., 2016a). We provide evidence that while the U-Net approach outperforms the GCN approach on known cities, the GCN approach generalizes better to unknown cities. The code to reproduce all experiments is publicly available¹⁵.

¹⁴Code and pretrained networks are available under <https://github.com/mie-lab/traffic4cast>.

¹⁵<https://github.com/mie-lab/traffic4cast-Graph-ResNet>

5.2.2 Graph-based traffic forecasting

Preprocessing

We introduce the ordered set of timestamps $T = \mathbb{N}_{12 \times 24 \times 365}$ to index all the 5-minute intervals throughout the year¹⁶. We denote all available training data as a 5-dimensional tensor \mathbf{P} with shape $(|U|, |T|, h, w, |C|)$, where an individual value $P_{u,t,i,j,c} \in \mathbb{N}_{255}$ denotes the integer value of a single pixel with $u \in U = \{B, I, M\}$ (for *Berlin, Istanbul and Moscow*), $t \in T$ the timestamp throughout the year, $i \in \mathbb{N}_{495}$ and $j \in \mathbb{N}_{436}$ the pixel coordinates, $c \in C = \{V, S, H\}$ the channel (corresponding to *Volume, Speed and Heading*), and $h = 495$ and $w = 436$ are the height and width of a single image.

We define a sample as $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ where $\mathbf{x}^{(i)}$ is a short movie of 12 consecutive traffic images with the timestamps $[t^{(i)}, t^{(i)} + 1, \dots, t^{(i)} + 12]$ and $\mathbf{y}^{(i)}$ consists of the images with the next three consecutive timestamps $[t^{(i)} + 13, t^{(i)} + 14, t^{(i)} + 15]$ which represent the prediction target. As it is described in (Martin et al., 2019c) we collapse the time dimension into the channel dimension. $\mathbf{x}^{(i)}$ is then of shape $(12 \cdot |C|, h, w)$ and $\mathbf{y}^{(i)}$ of shape $(3 \cdot |C|, h, w)$. The official test set in the *traffic4cast* competition contains only a small subset of timestamps for every day, namely *01:30, 04:45, 09:30, 14:30, 18:30*. In this work, all experiments are performed using only the subset of the data that corresponds to the available test timestamps (e.g. we use all training images corresponding to *01:30, 04:45, 09:30, 14:30, 18:30* but we omit training images corresponding to different time stamps). This approach leads to only a slight loss in performance while greatly decreasing training time.

Graph extraction

To enable the usage of GCNs we follow the workflow shown in Figure 5.14. (1) a city-specific mask is used to extract pixels that lie on the street network. The mask is generated based on the training images of the city. (2) the remaining pixels are transformed into a graph. In this graph representation, data is represented as a vector of node features and a sparse adjacency matrix to store their connectivity. (3) we use GCNs to generate node level predictions. (4) the data is transformed back from the graph domain into the image domain where the error is calculated based on the competition rules.

To create a graph from the traffic images, we define an image as a regular grid with pixels as nodes. The pixel-nodes are connected to all directly adjacent and diagonally adjacent pixel-nodes via undirected edges with weight 1. The pixel values of all available channels are stored as a vector of node features. For a given graph, the node feature vectors form a feature matrix with dimension $n \times 12 \cdot |C|$. We introduce sparsity in the above defined grid graph by deleting nodes corresponding to low activity pixels. A low-activity pixel $\hat{P}_{i,j}$ is defined as a pixel for which the sum over all available training images (of a single city u) does not exceed a certain user-defined threshold μ :

¹⁶We define $\mathbb{N}_w = \{1, 2, \dots, w\}$ as the set of natural numbers up until w .

$$P_{LA,u} = \{\hat{P}_{i,j} \mid \sum_{t \in T} \sum_{c \in C} P_{u,t,i,j,c} < \mu\} \quad (5.3)$$

A visualisation of the sum (before applying the threshold) is given in Figure 5.13.

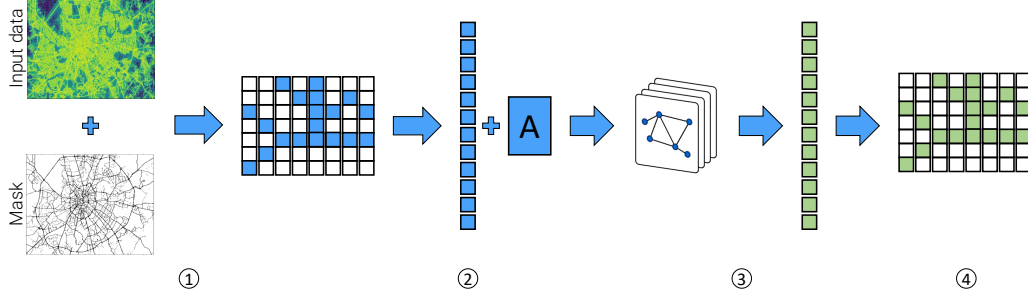


Fig. 5.14.: The graph-based traffic forecasting workflow. The input images are transformed into graphs by filtering out inactive pixels, after which the Graph ResNet learns to predict traffic on this graph.

GCN architectures

Based on the graphs extracted in Section 5.2.2, we use a GCN as proposed in (Kipf and Welling, 2016) as baseline and introduce three variations based on this baseline and on the ideas of the ResNet proposed by (He et al., 2016a).

Simple graph convolutional neural network

Similar to (Kipf and Welling, 2016), we define a graph convolutional processing block as the sequential application of a graph convolution and a ReLU activation function (we optionally consider dropout and batch normalization layers but do not explicitly denote them in the formalization below):

$$\mathbf{H}_{G_u}^{(l+1)} = \sigma(\mathbf{D}^{-\frac{1}{2}} \mathbf{A}^{(u)} \mathbf{D}^{-\frac{1}{2}} \mathbf{H}_{G_u}^{(l)} \mathbf{W}^{(l)}) \quad (5.4)$$

Here, l denotes the current layer in the network, $\mathbf{H}_{G_u}^{(l)}$ the input from the previous layer (where $\mathbf{H}_{G_u}^{(0)} = \mathbf{x}^{(i)}$ is the input, $\mathbf{A}^{(u)}$ is the adjacency matrix, \mathbf{D} is the diagonal node degree matrix of $\mathbf{A}^{(u)}$ (used to normalize the adjacency matrix), and $\mathbf{W}^{(l)}$ is a layer-specific (learned) weight matrix. $\sigma(\cdot)$ is an appropriate activation function; in this work, we use rectified linear units (ReLU) for all experiments. In addition to the network proposed by (Kipf and Welling, 2016), we add a skip connection that concatenates the input $\mathbf{x}^{(i)}$ with the output from the last GCN block before applying a last graph convolution (i.e., $\mathbf{y}^{(i)} = \mathbf{H}_{G_u}^{(m+1)} = \mathbf{D}^{-\frac{1}{2}} \mathbf{A}^{(u)} \mathbf{D}^{-\frac{1}{2}} (\mathbf{H}_{G_u}^{(m)} \oplus \mathbf{H}_{G_u}^{(0)}) \mathbf{W}^{(m)}$ with m being the number of GCN blocks). The combined SkipfNet has the advantage that it can very well learn functions resembling the identity function $f(x) = x$ due to the direct availability of the non-convoluted input (cf. (He et al., 2016b)). A central hyperparameter of the SkipfNet is the number of GCN blocks; Figure 5.15 shows exemplary networks with one and two blocks.

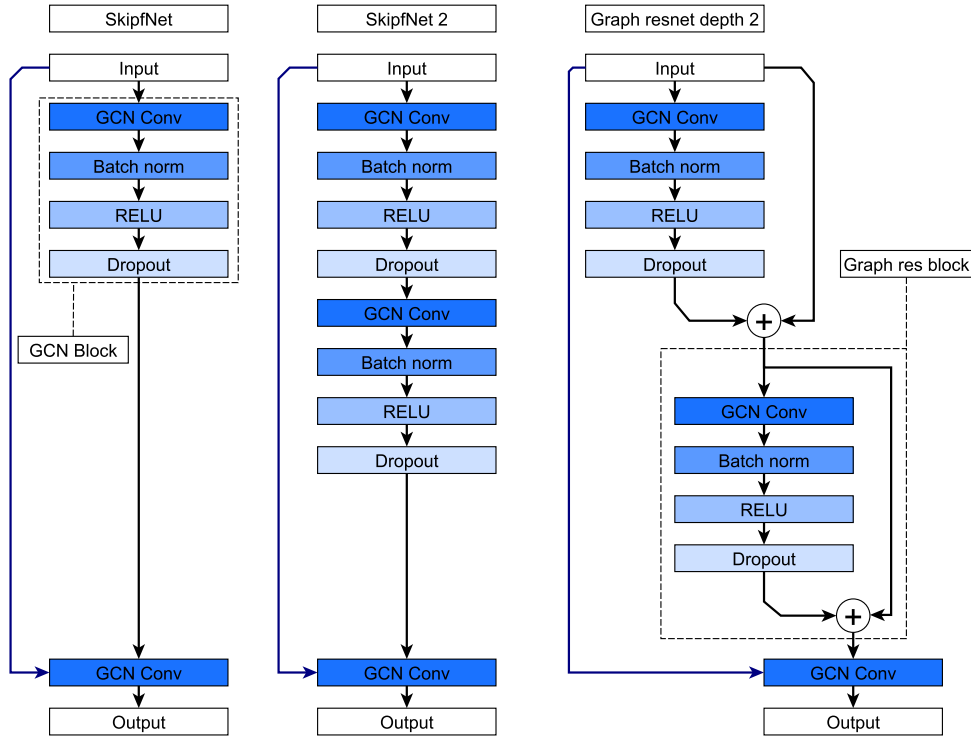


Fig. 5.15.: The different GCN architectures used and introduced in this paper.

Graph ResNet

Inspired by the popular ResNet architecture (He et al., 2016a), we introduce a Graph ResNet to further improve the predictive power of the SkipfNet. We define a graph residual block similar to the GCN block of the SkipfNet, but add input of the previous layer to the output of the current layer before passing it on to the next graph residual block (cf. Figure 5.15; again, we do not mathematically represent dropout and batch normalization below):

$$H_{G_j}^{(l+1)} = \sigma(D^{-\frac{1}{2}} A^{(u)} D^{-\frac{1}{2}} H_{G_j}^{(l)} W^{(l)}) + H_{G_j}^{(l)} \quad (5.5)$$

While in the original ResNet architecture, the layer size varied throughout the network using pooling. However, as pooling for graphs is more complex and still under active research (Lee et al., 2019; Ying et al., 2018), we keep the layer size constant.

5.2.3 Experiments and results

We perform two groups of experiments. At first, we train different graph networks to solve the *traffic4cast* prediction task in Moscow and compare the results to U-Nets of different depth. As a second set of experiments, we use the GCN and U-Nets that were trained only on Moscow and use them to predict the traffic in Berlin and Istanbul.

All U-Nets are trained using the parameters and training schedule from (Martin et al., 2019c) with a varying depth between 2 and 6 layers. Additionally, we include the original KipfNet with 16 hidden units as used in the original paper for the CORA dataset (Kipf and Welling, 2016) and with 128 hidden units, which is the maximum that our GPU allows. We fit 36 SkipfNets of depth 1, 29 SkipfNets of depth 2, and a total of 42 Graph-ResNets using random search on the hyperparameter space shown in Table 5.2. All Networks are fitted with the ADAM optimizer, learning rate 0.01, and weight decay of 0.0001 for a maximum of 10 epochs with early stopping if the validation error is not decreasing for two epochs in a row. The learning rate is divided by 10 after 5 epochs. We use a batch size of 2 and the mean squared error as a loss function. All graph networks receive the pixel coordinates of the nodes as two additional channels. These additional channels were omitted for the U-Nets as they degraded their performance. Training was performed using a Tesla P100 GPU with 12 GB of RAM. All experiments were implemented in Pytorch (Paszke et al., 2019) and Pytorch-Geometric (Fey and Lenssen, 2019).

Network	H 1	H 2 / depth	K	K mix	sc	bn	dropout
SkipfNet	8, 16, 32, 48, 64	-	2, 4, 6 , 8	1,2,4,6	0 , 1	0	0.5
SkipfNet2	8, 16, 32, 48, 64	8, 16, 32 , 48	2, 4, 6 , 8	1,2,4	0 , 1	0	0.5
Graph ResNet	16 - 100 (100)	2 - 60 (4)	2 - 6 (4)	1, 2 (2)	0, 1 (1)	0, 1	0, 0.5

Tab. 5.2.: Hyperparameter space explored during graph network training. H=Hidden layer; K and K mix=number of terms for the Chebyshev polynomial from (Defferrard et al., 2016) for the convolutional layers and the last mixing layer; sc=usage of optional skip connection (blue in Figure 5.15); bn=usage of batchnorm; dropout=used dropout probability. Hyperparameters of best model are shown in bold.

Traffic prediction capacity

The out-of-sample errors of the different models for the Moscow validation dataset are shown in the top graph of Figure 5.16 and in Table 5.3. The Figure shows a clear dependency between the number of model parameters and the performance. In general, there are two major ways to add complexity to the models: increasing the depth of the model or increasing the number of channels of the convolutional layers. The model complexity of the graph networks is limited because deeper networks lead to over-smoothing of the prediction (Li et al., 2018) and because the lack of suitable graph pooling operations forces us to add channels at the full-size graph which quickly drains GPU memory. In our work, the original KipfNet architecture reaches the memory limit of our GPU at 128 output channels of the convolution. The Skipfnet architecture with the variable K parameter and the skip connection allows adding additional parameters without over-smoothing. Finally, the Graph ResNet allows us to increase the performance even further. None of the graph networks reaches the performance of the deeper U-Net versions, however, all of the methods are below the official *traffic4cast* baseline¹⁷ that is shown as a dashed line.

¹⁷The baseline is a MSE of 1032 for Moscow, 789 for Istanbul and 582 for Berlin. It consists of predicting the per-channel mean of the last three images for the next three images. Due to the high

	KipfNet nh16	KipfNet nh128	Graph-ResNet	SkipfNet1	SkipfNet2
Nb. of params	$1.5 \cdot 10^3$	$1.2 \cdot 10^4$	$1.7 \cdot 10^5$	$1.8 \cdot 10^4$	$2.8 \cdot 10^4$
MSE Moscow	867	836	814	836	828
MSE Berlin	492	475	468	474	471
MSE Istanbul	671	645	633	644	639

	U-Net d2	U-Net d3	U-Net d4	U-Net d5	U-Net d6
Nb. of params	$4.2 \cdot 10^5$	$1.8 \cdot 10^6$	$7.7 \cdot 10^6$	$3.1 \cdot 10^7$	$1.2 \cdot 10^8$
MSE Moscow	813	797	791	794	794
MSE Berlin	478	500	520	519	501
MSE Istanbul	653	713	731	727	704

Tab. 5.3.: Number of parameters and the resulting mean squared error for all models in the different cities. The content of the table is visualized in Figure 5.16.

Generalization capacity

In a second step, we use the models trained on Moscow to predict the traffic in Berlin and Istanbul. To make up for the GCN’s additional knowledge of the street network we mask the U-Net predictions by multiplying them with the binary mask of the street network. The results of this experiment are shown in Figure 5.16.

The first result is that all methods generalize surprisingly well to unknown cities, as all methods can beat the baseline that is described in Footnote 17 and shown as a dashed line for each city. Next, the results show that the relationship between performance and the number of parameters does not hold anymore. For the U-Nets, we observe that the shallow U-Net has the best prediction results and that the U-Net loses performance the deeper it gets. This trend is not as clear for U-Nets for depths 5 and 6 which might be a sign that they have not been trained with enough data or not long enough. To support this claim, we repeated this experiment with the pre-trained U-Net of depth five used in the *traffic4cast* competition by Team MIE-Lab¹⁹. This competition network was tuned more carefully and trained with more data and indeed does fail miserably in the generalization task. It achieves a MSE of 774 for Moscow where it outperforms all other networks but a MSE of 798 for Berlin and an MSE of 3357 for Istanbul (both off the charts).

For the GCNs, the relationship between performance and the number of parameters is still intact with the Graph-ResNet now being the overall best performer. An explanation for these observations is that the more parameters the conventional CNNs have and the more data they get during training, the better they can extract the (non-transferable) street network and store it in their weights. Whereas with few parameters they will simply learn a baseline like the conditional mean of the input data, which is to a large extent generalizable to other cities. The graph-based networks however already know the street network as it is provided

regularity in traffic, this can be considered a strong baseline. The implementation of the baseline is available under <https://github.com/iarai/NeurIPS2019-traffic4cast>.

¹⁸See Footnote 17

¹⁹See Footnote 14.

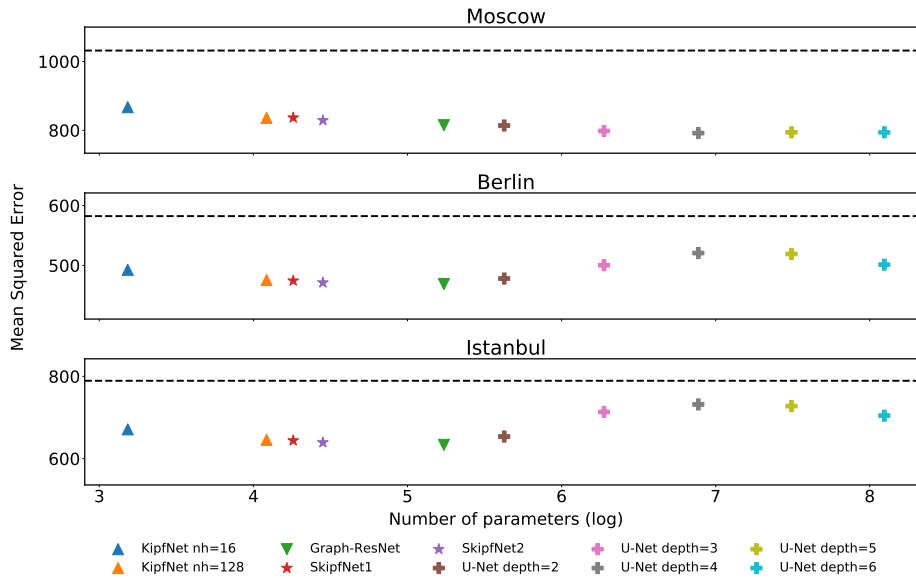


Fig. 5.16.: Traffic forecasting results. All models are trained on Moscow (top) and tested on Berlin (middle) and Istanbul (bottom). The competition baseline is shown as dashed line¹⁸.

as an explicit input and can therefore learn transferable rules based on the propagation of traffic through the network.

5.2.4 Conclusions and future work

In this work, we presented a graph-based approach for traffic forecasting and apply it to the publicly available dataset from the *traffic4cast* competition. We define ResNet-inspired GCN networks with skip connections that allow the training of deeper GCNs without over-smoothing. We train all models on traffic data from Moscow and validate them on data from Berlin and Istanbul. Our results suggest that the GCN-based models generalize better than the state-of-the-art CNN models from the *traffic4cast* competition. We think that the large performance difference between U-Nets and GCNs in the known city prediction is largely because the image-based competition design heavily favors conventional CNNs. Especially GPS positioning noise in combination with the MSE evaluation which favors blurry predictions (Lotter et al., 2016) is hard for the graph networks which can, by design, only predict values exactly on the graph. To still have a good performance under competition rules²⁰, the threshold for extracting the street network from Equation 5.3 was chosen rather low. However, a low threshold almost reproduces a grid graph where conventional CNNs are known to outperform GCNs. Therefore follow-up work should evaluate the performance of graph models under conditions that do not disadvantage graph-based approaches (e.g., by only evaluating pixels that lie on the graph). Furthermore, the relationship between parameters and model performance for graph models on unknown cities should be further explored by creating more complex graph models. A key role for this will be the successful integration of graph pooling methods in the model architecture. These experiments will

be important to validate the findings of this work that, as opposed to CNN-based models, GCN-based models continue to increase their performance with increasing model complexity when forecasting traffic in unknown cities.

Acknowledgement

This research was supported by the Swiss Data Science Center (SDSC) and by the Swiss Innovation Agency Innosuisse within the Swiss Competence Center for Energy Research (SCGER) Mobility. Christian Rupprecht is supported by ERC IDIU-638009.

²⁰In this context *competition rules* refers to the fact that all pixels are evaluated, not just pixels that are on the street graph; cf. Figure 5.13.

6.1 Summary and contributions

The goal of this dissertation was to develop computational methods to support sustainable individual mobility. It comprises two technical reports and seven peer-reviewed publications that cover the areas of empirical data collection, preprocessing, analyzing and interpreting individual mobility based on a graph representation, and the forecasting of mobility using machine learning. The following is a short summary of the content and the contributions of this dissertation.

6.1.1 Tracking data collection for the evaluation of mobility interventions

Chapter 2 summarizes three tracking studies that investigate the impact of different MaaS products on the mobility behavior of the study participants.

Section 2.1 summarizes the GC1 and GC2 pilot studies. The participants of both studies got access to a comprehensive mobility package comprising a general public transport pass for Switzerland, a BEV (GC1) or an e-bike (GC2), access to car and bike sharing, as well as taxi vouchers. The studies showed that the participants continued to travel multi-modal and combined their mobility tools with public transport rather than replacing them, which is an especially important result with regards to the BEV that was provided in the GC1 study. The provided BEV was primarily used to replace trips that were formerly served using ICEVs. This led to a strong and stable decline in the GHG emissions of the GC1 participants, a result that was not observed in the GC2 study where participants received the e-bike instead of the BEV.

Section 2.2 summarizes the EIM project that analyzed the usage and impact of a MaaS app called yumuv that was available in the city of Zurich. The app facilitated the booking of shared e-scooters and e-bikes of several mobility service providers and offered mobility subscriptions (bundles). Study participants were divided into a control group (CG) and a treatment group (TG) that had access to the yumuv app. The study showed that access to the mobility bundle jointly increases the usage of public transport and shared e-scooters. Furthermore, as part of the project, location graphs were developed as a representation of individual mobility (cf. Section 3.2) and as the basis of a novel method to analyze mobility behavior change over time (cf. Section 3.2).

6.1.2 Making tracking data processing reproducible

A major challenge when studying mobility behavior based on tracking data is the preprocessing of the data. On the one hand, the preprocessing is complex and time-consuming and on the other hand, it can have a significant impact on the results. Both problems are addressed in Section 3.1 that presents Trackintel, an open-source Python library for movement data processing. Trackintel is built on a standard data model for human mobility used in transport planning that is compatible with different types of tracking data. Based on this data model, Trackintel standardizes preprocessing steps by providing implementations for the most common preprocessing algorithms, as well as support for the analysis and visualization of tracking data. By that, Trackintel simplifies quantitative research based on tracking data and greatly increases its reproducibility and replicability. Section 3.1 introduces the data model for mobility data, its implementation in Trackintel, and finally showcases the capabilities of Trackintel in a comprehensive case study that jointly analyzes four different tracking datasets.

6.1.3 Analysis and prediction of individual mobility using location-based graphs

There are various ways to record the mobility behavior of individuals such as (paper-based) travel surveys, GNSS-based tracking, call detail records, location-based social network data, public transport check-in data, and others. These types of tracking data can additionally have a different spatiotemporal resolution (e.g., sampling rate), different spatiotemporal distribution (track point distribution, e.g., regular vs. burst patterns), and different context data associated with it. Section 3.2 formalizes individual location graphs, a graph-based representation for individual mobility data that uses unique visited locations as nodes, and the count of direct transitions between these nodes as weighted and directed edges. Location graphs only require information about sequential location visits and therefore can be created based on a wide range of different tracking datasets. Once transformed into the location graph representation, different datasets can be jointly analyzed based on the topology of the location graphs. This allows combining multiple small tracking datasets to increase the sample size or to transfer models fitted labeled datasets to unlabeled datasets. After the formal introduction of the graph representation, Section 3.2 offers a guide to analyze mobility behavior of individuals based on the in-degree distribution of the graph, the journey distribution, the connectedness, and the presence of hubs in the graph (hubbiness). Finally, the location graph representation is applied to four different datasets to show its ability to reproduce well-known statistical properties of mobility data (visitation frequency, distribution of displacements, and radius of gyration) and to show its limitations by comparing its ability to reproduce journey distributions compared to higher order Markov models.

Location graphs can represent the mobility behavior of an individual without geocoordinates and are therefore a more privacy-preserving alternative to a person's full profile. However, location graphs are not anonymous, as it was shown that users can be reidentified from a

pool of individuals based on past, non-overlapping tracking data (Manousakas et al., 2018). Section 3.3 further analyzes the privacy properties of location graphs with a focus on the influence of the tracking duration. Here, a scenario is constructed in which an attacker is in possession of a pool of graph-encoded tracking data and gets access to a novel set of graph-encoded tracking data that they want to match. The results of the experiment show that the top-1 re-identification accuracy is between 0.41% and 20.97% for different tracking duration combinations, meaning that graphs constructed from a dataset with longer tracking duration are easier to match. Therefore, we advise data collectors to limit the tracking duration or to reset user IDs regularly when storing long-term tracking data in order to protect the privacy of individuals.

Even though location graphs are not anonymous, they are a privacy-friendly alternative to commonly used mobility profiles and can still provide valuable insights into the mobility behavior of individuals. Section 4.1 presents a clustering-based method to identify user groups with similar mobility behavior that appear across several datasets and that can be used to analyze the adoption and the impact of mobility interventions such as access to a novel MaaS offer. We apply the method to six datasets with a total of 1070 users and identify five groups (*centered*, *commuter*, *flexible*, *local routine*, *traveler*) that are present in all datasets. Further analysis of the differences between CG and TG of the yumuv dataset introduced in Section 2.1 reveals that two groups (*flexible*, *traveler*) are overrepresented in the TG and one group (*local routine*) is underrepresented in the TG with respect to the CG. This target group analysis can be used by mobility service providers to optimize their mobility offers for the intended group of customers. Furthermore, we present a longitudinal analysis to evaluate the impact of mobility interventions on mobility behavior. Given a mobility intervention, the group assignment of study participants before and after the intervention reveals whether or not their behavior changed over the course of the study.

Finally, the short study in Section 4.2 presents an application of graph convolutional neural networks to location graphs for the imputation of missing activity labels. In transport planning, travel demand is commonly modeled based on activities (Castiglione et al., 2015; Jiang et al., 2017). A fine-grained understanding of our activities is therefore central to understanding and modeling human mobility behavior. However, while participants in tracking studies are usually asked to label their significant stays, they often only sporadically provide activity labels. Predicting missing activity labels given the existing labeled activities or labeling the locations of previously unknown users is therefore an important yet still open problem. The method proposed in Section 4.2 uses a GCN to predict the activity label distribution at activity locations. It is applied to undirected location graphs enriched with an additional pair of edges based on the Euclidean distances between locations. The proposed GCN outperforms all baselines, indicating that deep learning models can take advantage of the information encoded in the location graph.

6.1.4 Analysis of battery electric vehicle usage and charging behavior based on tracking data

Chapter 5 turns the attention away from location graphs toward the sustainability of personal vehicles. Section 5.1 presents a detailed analysis of the potential to charge BEVs at home using PV panels installed on the roof of the car owner's house. To answer this question, a subset of the GC1 participants that live in single-family homes ($n=78$) was selected. We then virtually equipped the roof of their houses with PV panels based on the true roof geometry extracted from a digital elevation model. Using the tracking and charging data recored by their BEVs in combination with historical weather data, we analyzed to what extent the energy needs for personal mobility can be covered using power generated by their rooftop PVs system. We simulated four different smart charging strategies with varying degrees of complexity and found that when charging uncontrolled (the strategy used by the participants during the study), BEV owners could only cover 15 % of their BEV's demand using PV generated from the roofs of their own houses. A simple controlled charging approach greatly increased the average coverage to 56 %, and up to 90 % or 99 % when using an optimized charging strategy without or with a home battery storage. This shows that it is possible to cover a large portion of the mobility energy demand of BEV owners using rooftop PV generation without restricting their mobility. However, it is necessary to use smart charging management with mobility demand and power generation forecasts.

6.1.5 Traffic prediction using graph neural networks

Chapter 5 presents a deep-learning method for forecasting the traffic speed and volume in a city. In the experiment, we compare our novel GCN based approach with more conventional image-based approaches that deploy convolutional neural network (CNN) models for the prediction of traffic in unseen cities. The results suggest that the graph-based models have superior generalization properties than CNN-based models for this application. We argue that in contrast to purely image-based approaches, formulating the prediction problem on a graph allows the neural network to learn properties given by the underlying street network. This facilitates the transfer of a trained model to predict traffic dynamics in cities with lower data coverage.

6.2 Research questions

Chapter 1 of this dissertation introduced two primary research questions that focused on how computational methods can support sustainable mobility and two secondary research questions that focus on the foundations of computational methods.

6.2.1 Research questions focused on sustainable mobility.

The primary research questions of this dissertation focus on the development of computational methods to support modal shift and how computational methods can be used to support the sustainability of personal vehicles.

How can the modal shift of individuals be supported using computational methods?

One of the main strategies to make individual mobility more sustainable is to cover our mobility needs using more sustainable modes of transport. The shift towards more sustainable modes is therefore one of the main interventions covered by the A-S-I framework mentioned in the introduction of this dissertation. It is also recognized as one of the two most efficient ways to reduce the carbon footprint of individual transport by the IPCC report on climate change (Jaramillo et al., 2022). However, while modal shift is widely recognized as an important and necessary measure, detailed knowledge about the impact of specific mobility offers is still to be collected. A key tool to collect empirical evidence on mobility behavior is to record the movement of individuals using travel diaries. In the past, this was done using paper-based surveys or phone interviews. Nowadays, travel diaries recorded passively using the mobile phone of a study participant provide higher spatiotemporal coverage and allow for tracking studies with a high number of participants and longer durations. In such case studies participants typically install an app on their phone which records and segments their movement. Each participant then has to provide or validate labels for transport modes and activities. Despite the advantages of mobile phone-based surveys, the provided labels are often incomplete or implausible. As these labels are essential for studies that evaluate the impact of mobility interventions, such as new incentives for modal shifts, completing and correcting these labels is of great importance. An important step to advance methods for the completion of activity labels is presented in Section 4.2, where a deep learning approach based on a GCN is used for the imputation of missing labels. Here, the imputation was performed based on location graphs and we showed that a GCN can take advantage of information encoded in the topology of the location graph for analysis and prediction tasks.

A challenge with regard to the introduction of new incentives for mobility behavior change is the detection and analysis of mobility behavior change. Section 4.1 of this dissertation introduces a novel method for the monitoring of mobility behavior change. The main purpose of the method is to identify groups with similar mobility behavior based on location graphs. The focus on location graphs allows combining multiple potentially smaller datasets to increase the sample size and offset potential biases in order to detect more robust and generalizable groups. An important application of these mobility behavior groups is to track the group assignment of users over the course of an intervention. This allows detecting and monitoring behavior change using only the less privacy-sensitive location graphs of an individual. This work can thereby support the evaluation of the effectiveness of mobility interventions.

Finally, another problem potentially preventing the modal shift of individuals is related to the mobility service providers that offer sustainable mobility options. Due to their novelty, there is little experience in how to design and manage these shared mobility products effectively. A key insight that is missing is information about the potential target groups for specific mobility offers. Section 4.1 presents how the target groups of a MaaS offer can be analyzed using the yumuv MaaS study introduced in Section 2.2 as an example.

How can the sustainability of personal vehicles be supported using computational methods?

The replacement of ICEVs with BEVs represents a highly effective way to lower the GHG emissions of individuals, as it could also be seen in the GC1 study summarized in Section 2.1. However, to unfold their full decarbonization potential, BEVs require clean energy for charging. Even though the energy transition is in progress globally (IEA, 2020b), electric power available to consumers is still partially generated from fossil fuels in most countries. As rooftop PV systems have relatively low specific GHG emissions (cf. Section 5.1.4), their installation might be an easy solution for homeowners to further decarbonize their mobility. However, there is skepticism regarding this solution, as personal vehicles are often not parked at home during the day when the sun shines. Our study in Section 5.1 found that even with a trivial smart charging strategy the participants could cover most of their mobility energy demand with power generated from their own PV system. The results seem counterintuitive at first, however, people often overestimate their daily driving distances or underestimate the range of BEV. Therefore, it is often enough if the car can be fully charged on weekends, on days when a person is working from home, or in the morning and evening hours. The study showed the high potential of rooftop PV systems for the decarbonization of personal mobility and the high importance of controlled charging for optimizing the self-consumption of solar power. This computational study thereby offered valuable decision support for individuals and policymakers to further decarbonize personal mobility.

Individual mobility and especially non-urban individual mobility will continue to rely on personal vehicles to a large extent. Even though using BEVs reduces the related GHG emissions significantly, individual motorized transportation will still block large areas of public space for parking and infrastructure, and the problem of increasing traffic is left unresolved. In light of further global urbanization (UN, 2018), it will therefore be essential to establish smart traffic management in order to maintain livable cities. A key component of smart traffic management systems is short-term traffic forecasts. Providing these forecasts is an important way how computational methods can support the sustainability of personal vehicles. Section 5.2 contributes to the existing literature in this field by exploring the capabilities of different deep learning models trained for traffic forecasting to generalize to unseen cities.

6.2.2 Research questions focused on computational methods

In addition to the primary research questions discussed above, the focus on computational methods for the support of sustainable mobility brought two further research questions into the focus of this doctoral dissertation.

How can individual mobility from different data sources be represented in a compact, and privacy-friendly way such that it allows the development of computational methods?

A core contribution of this dissertation is the formalization of location graphs as a representation of individual mobility in Section 3.2. The creation of a location graph requires only the sequence of visited locations without coordinates which are transformed using a lossy compression that retains unique locations and the count of 1st order transitions. This has two main advantages it allows the creation of location graphs based on many different types of tracking datasets and it is privacy-friendly because much of the sensitive information (e.g., multi-hop sequence of visits, visiting times, activity purpose at a location) is aggregated and thereby partially obscured on a location level. Despite its minimal requirements and the aggregation of personal information, Section 4.1 and Section 4.2 show that the location graph can still be used as a basis to develop computational methods to support the sustainability of individual mobility.

How can the reproducibility and generalizability of preprocessing and analysis methods for human mobility be improved?

The reproducibility and generalizability of computational methods related to human movement data are currently limited by several factors. The first concerns the preprocessing of tracking data directly. While tracking data are commonly used in various scientific disciplines, there is no common understanding of the unit of analysis. As an example, a *location* might refer to a single recorded raw trackpoint or to a place that was visited several times and can only be identified from data after several preprocessing steps. Furthermore, the preprocessing steps applied to tracking data are not standardized. Preprocessing and analysis algorithms are currently hard to compare, as there is no common understanding of the processed movement data and because they are often not available in a reliable open-source implementation. Both problems are addressed with the introduction of Trackintel in Section 3.1. This data model for movement data is used to standardize the definition and implementation of preprocessing steps. Furthermore, Trackintel provides a high-quality implementation of the most common preprocessing algorithms and provides functionalities for visualization and analysis. Trackintel thereby improves the reproducibility and generalizability of preprocessing and analysis methods for human mobility.

Another problem with the reproducibility and generalizability of computational methods is related to the properties of tracking datasets. Even though all tracking datasets record human

mobility, they are highly diverse as they use different tracking technologies and might be different in terms of spatiotemporal resolution, the spatiotemporal distribution of trackpoints, sample biases, available annotations, or the recorded context data. Additionally tracking datasets can usually not be shared due to privacy concerns. The result is that algorithms for the analysis of individual mobility are often highly specialized on a single dataset and not applicable to others even if an open-source implementation would be available. This dissertation addresses this issue by leveraging the location graph representation introduced in Section 3.2. Due to its minimal requirements on the dataset, it can be created based on many different tracking datasets and is yet flexible to accommodate optional context data as node or edge information. Section 4.1 and Section 4.2 present computational methods based on location graph data and Section 3.2 and Section 4.1 show how the location graph can be used to analyze datasets that are recorded using different technologies.

6.3 Outlook

The research presented in this dissertation was focused on the development of computational methods that support sustainable individual mobility. However, efforts on decarbonizing the energy and transport sector have to be further increased in the near future. The following is a short description of promising avenues for further research on computational methods to support sustainable mobility.

The first topic for future research is the continuation of research on location graphs. This dissertation described the advantages of these methods with respect to generalizability and the privacy-utility trade-off. However, there are very few methods for mobility prediction, labeling, and analysis of mobility data available, leaving plenty of opportunity for future research work. Furthermore, we have discussed the properties of location graphs and the relation of graph properties to mobility behavior. However, the influence of gaps in the tracking data on location graphs was not yet explored. Because location graphs are created by aggregation, they could be highly robust to tracking gaps. This would make them a suitable tool for tracking technologies with a low sampling frequency (eg., CDR data). This property could also be used to lower the sampling frequency of other tracking technologies which would benefit the privacy of individuals and the battery life of their devices. Similarly, the influence of tracking duration on location graphs was not yet explored in detail. Here it is important to know how long individuals need to be tracked until their location graph stabilizes.

A second avenue is to further reduce the workload for researchers associated with data preprocessing for tracking studies. Trackintel provides a standardized framework for preprocessing methods, however, there is currently no protocol that standardizes the preprocessing steps required to achieve a dataset with high data quality. This makes the data preparation process a highly manual and cumbersome process.

Finally, future research should continue to work on further decarbonizing BEVs. Given the insights on BEV charging from this dissertation, smart charging algorithms should be refined

to consider multiple objectives. These objectives are to fulfill the mobility needs of the vehicle owner while maximizing the usage of renewable energy generation for charging and supporting the stability of the distribution grid. This will require forecasts for the mobility demand of an individual and their combination with renewable generation forecasts.

A.1 Supporting information for Trackintel: an open-source Python library for human mobility analysis

A.1.1 Documentation score

Python

The documentation score reported in Table 3.1 for python libraries is based on the pyOpenSci package peer-review evaluation criteria¹

- Has an Open Software Initiative (OSI) approved license.
- Contains a README with instructions for installing the development version.
- Contains a vignette (notebook) with examples of its essential functions and uses.
- Has a test suite.
- Has continuous integration, such as Travis CI, AppVeyor, CircleCI, and/or others.
- Includes documentation with examples for all functions.

R

The documentation score reported in Table 3.1 for R libraries is based on the ROpenSci package peer-review evaluation criteria²

- Does the package have a CRAN accepted license?
- The package contains a reasonably complete readme with devtools install instructions.
- The package contains a vignette with examples of its essential functions.
- The package contains unit tests.
- The repository has continuous integration with Travis and/or another service.
- Package available on CRAN?

¹<https://www.pyopensci.org/contributing-guide/intro.html>

²<https://ropensci.org/>

A.2 Supporting information for Influence of tracking duration on the privacy of individual mobility graphs

A.2.1 Validation on Green Class 2

To validate our results for the Green Class 1 data, we compute the matching performance results on the Green Class 2 data accordingly. Figure A.1 visualizes the results corresponding to Figure 3.21. Due to the lower number of users in Green Class 2, the re-identification accuracy is generally higher, but the same patterns as for Green Class 2 can be observed: Both the pool and the test duration impact the matching performance, and the best results are obtained when pool and test duration are the same.

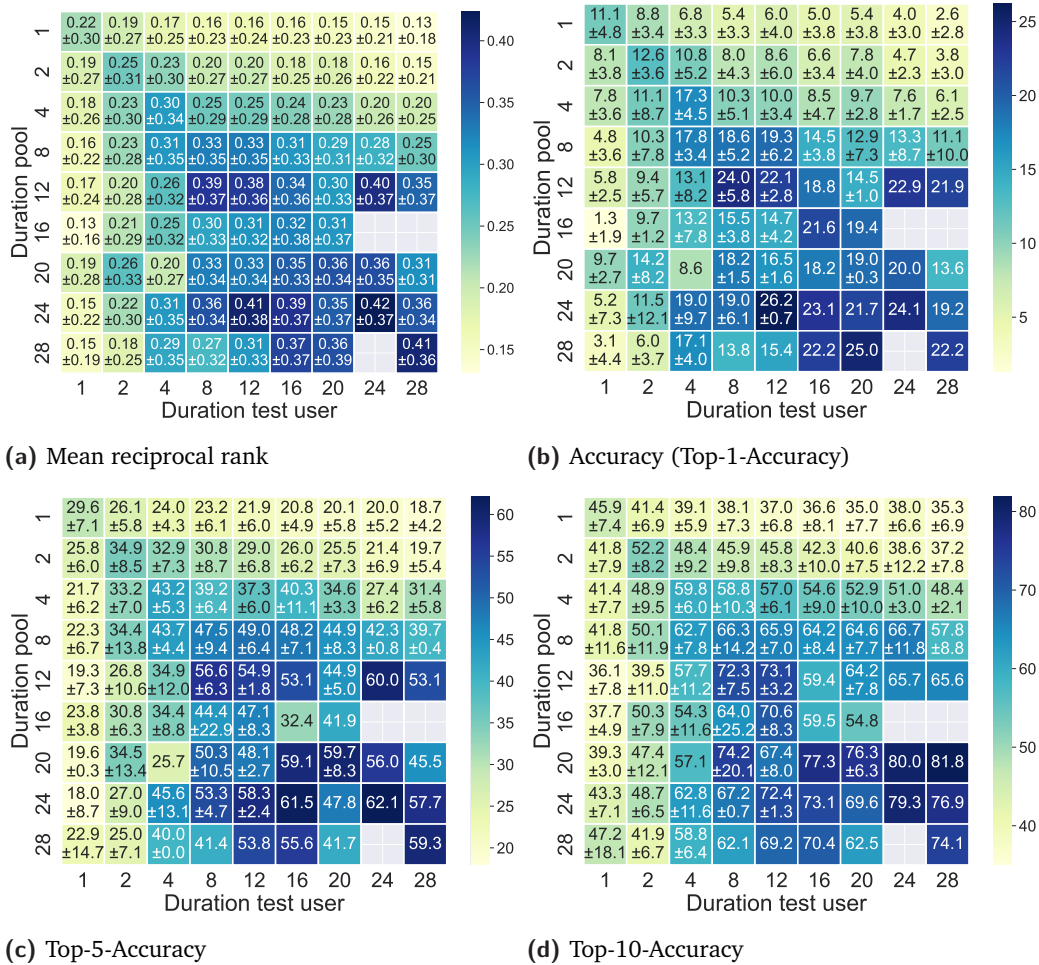


Fig. A.1.: Dependency of matching performance on tracking duration for the Green Class 2 data. Similarly to the results for Green Class 1, the top-k accuracy and MRR increase with both the tracking duration of the pool users as well as the test user. Due to the lower number of users, the re-identification performance is higher, reaching up to 82% top-10 accuracy.

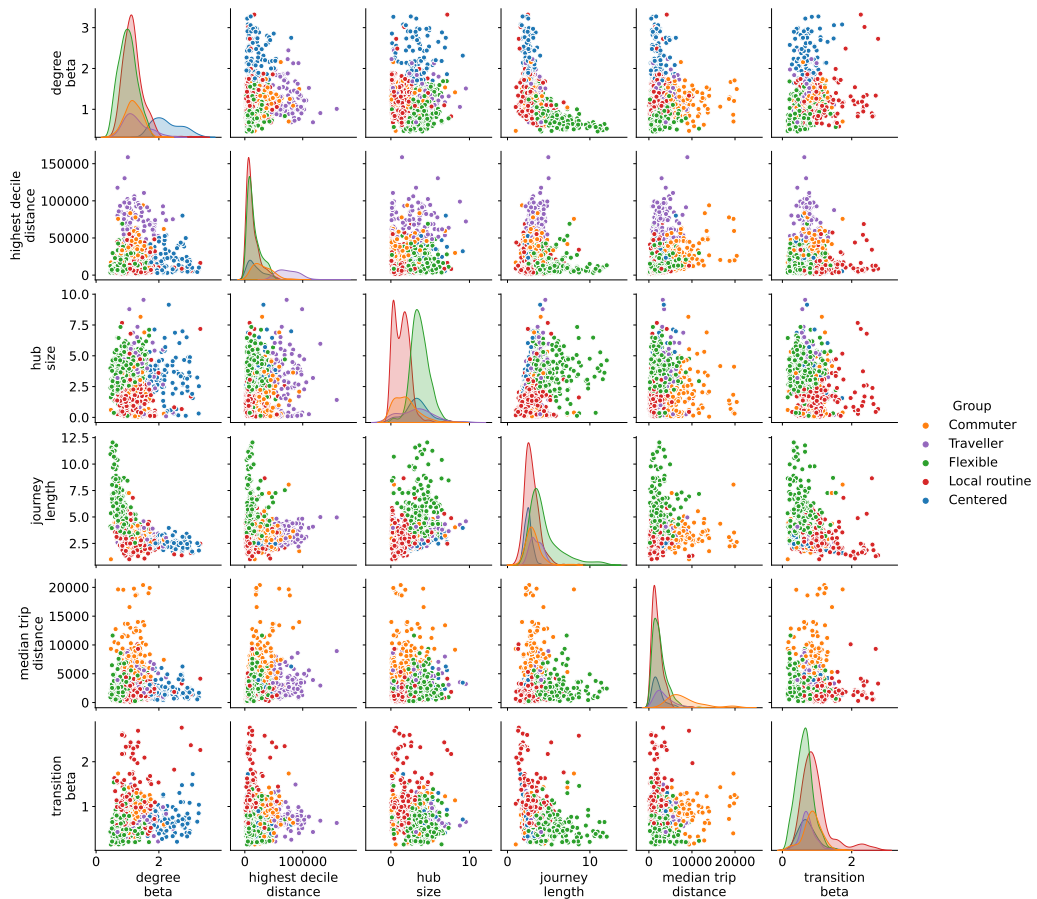


Fig. A.2.: Scatterplot matrix for the features of all datasets. Outliers were removed beforehand. The user groups can be clearly distinguished on certain features axes.

A.3 Supporting information for Graph-based mobility profiling

A.3.1 Feature exploration

For a more in-depth understanding of the distribution of features over user groups, we provide the scatterplot matrix in Figure A.2. The two largest groups are *Flexible* and *Local routine* which differ mostly in the *hub size* feature. The *Flexible* group also has a striking difference from the other groups with respect to the *journey length* feature. *Commuters* and *Travellers* are clearly distinguished by their high *median* and *9th-percentile trip distance* respectively. In contrast, the group *Local routine* has a particularly left-skewed distribution in the distance-based features. Last, the *Centered* profile is clearly characterized by the high node degree β in this group.

A.3.2 Cross-sectional study

The cross-sectional study on the yumuv dataset in Section 4.1.5 can be conducted in similar form for the Green Class dataset. Since no control group is available in the Green Class studies, we instead compare their user group distribution to the one in all other studies. Figure A.3 shows the group prevalence's. In Green Class 1 and 2, there is a significantly higher share of the *Commuter* and *Traveller* groups compared to the other studies, i.e., these two groups are above average attracted to the Green Class offer. In contrast, less users are part of the *Local routine* and *Flexible* groups. The differences are slightly weaker for Green Class 2. In both cases, these differences in the distribution of the user groups are significant (χ^2 test, $p < 0.01$). These target and non-target groups can be characterized using Figure 4.5 and it could be further analyzed with respect to additional information such as demographics (cf. Section 4.1.5).

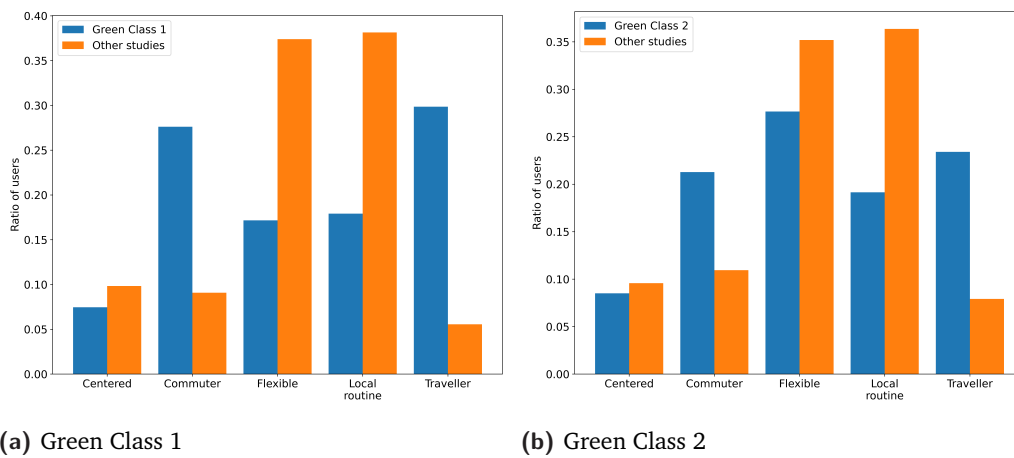


Fig. A.3.: Cross-sectional study for Green Class 1 and 2: There are more Travellers and Commuters taking part in the Green Class studies compared to the proportion in other datasets.

A.3.3 Longitudinal study

Analogous to Figure 4.8, Figure A.4 visualizes the movements of users between groups in a network. The width of the edges is proportional to the number of users that are assigned to group A before, but switch to group B during the trial period.

A.3.4 Dependence on clustering hyperparameters

Although our clustering approach does not depend on a major design choice such as the number of clusters as input, it still uses several hyperparameters. Those either provide more flexibility than before (e.g. the choices of k) or have minor influence on the result, such as the threshold θ_{minf} . The latter is demonstrated with an additional experiment reported in

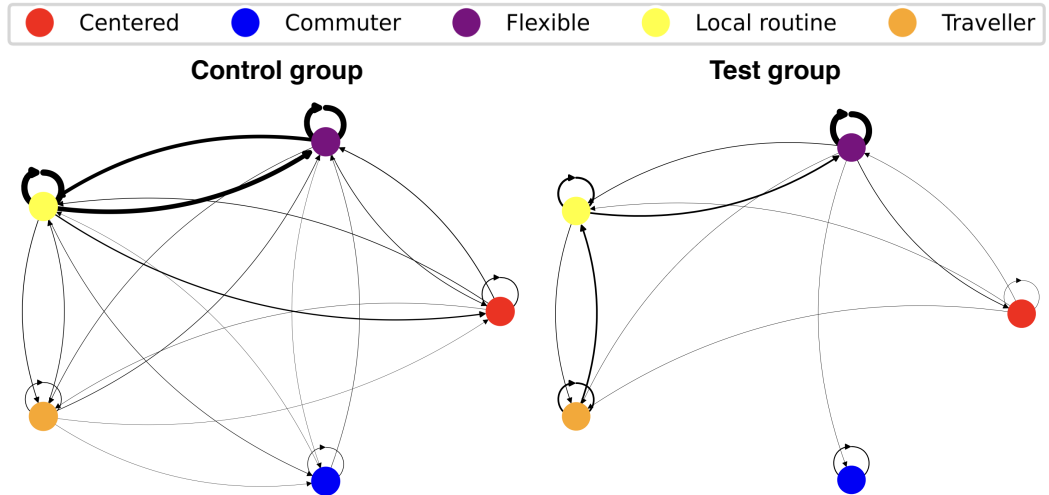


Fig. A.4.: Changes of user groups from the period before access to the yumuv offer to the period after intervention. The arrow width corresponds to the number of users that change from one group to another.

Table A.1. The initial partitions (cf. Figure 4.4a) with 3 runs for each of $k = [6, 7, 8, 9]$ yield 90 clusters in total. The cluster are merged (cf. Figure 4.4e) if no features are significant in contradicting directions and if there are more than θ_{minf} significant features corresponding. Increasing θ_{minf} leads to more unmerged user groups since more and more clusters have an insufficient number of significant features to be merged. Similarly, users can only be assigned to groups with a sufficient number of significant features. Therefore the number of unassigned users increases (see Table A.1). However, the resulting user groups that are consistently assigned (see section 4.1.3) are clearly stable and robust to θ_{minf} .

θ_{minf}	clusters	user groups (merged)	user groups (assigned consistently)	unassigned users (%)
1	90	6	5	0.00
2	90	6	5	0.00
3	90	8	5	0.00
4	90	18	6	0.03
5	90	35	7	0.14
6	90	80	2	0.82

Tab. A.1.: Effect of the θ_{minf} threshold, defining the minimum number of significant features to merge user groups. Only when θ_{minf} is set to a high value, the merging process is affected, i.e. clusters cannot be merged due to a small number of significant features.

As a rule of thumb, a user of the framework should set $\theta_{minf} \leq \frac{m}{2}$, i.e. not more than half of the number of features, and ensure that all users can be assigned to a group.

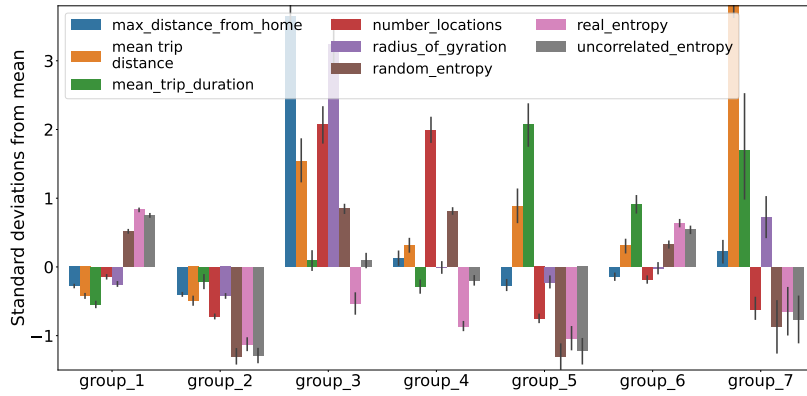


Fig. A.5.: Characteristics of the identified groups based on classical mobility features.

A.3.5 Mobility profiling based on classical mobility features

To demonstrate the generality of our clustering algorithm, we additionally show its results when applied on a set of classical mobility features, i.e. the *basic features* described in section 4.1.6. The same parameters are used. Note that we can not include the two Foursquare datasets, as the calculation of some of the features requires trajectory data. In the group finding phase, 10 groups are identified, but the users are only assigned consistently to 7 of them. The average consistency score (c.f. 4.1.6) is 0.9, meaning that on average a user is assigned to its most dominant group in 90% cases. This gives evidence that our algorithm is in general suitable to derive stable user groups. Figure A.5 shows the user groups based on classical features analogously to Figure 4.5 (we omit the step of naming the user groups as it is not the focus of this work). Figure A.6 depicts the groups per dataset and shows strong differences between the studies, indicating a strong influence of tracking period and other technical dataset properties on the user groups (c.f. appendix A.3.6 for further analysis). We argue that graph features are thus more robust to technical dataset properties and are therefore suitable for comparing mobility behavior of users in different studies.

A.3.6 Dependency of mobility profiles on technical dataset properties

In paragraph 4.1.5 the distribution of user groups over studies was discussed, and the occurrence of all groups in most datasets indicated a certain generality of the mobility profiles. Here, we provide further evidence that the user groups are robust to technical properties of the data. A multinomial logistic regression model is used to quantify the dependence on data properties, namely tracking duration and tracking coverage. We compare the resulting coefficients and p-values for the user groups derived from graph features (c.f. Section 4.1.5), given in Table A.2, to the ones for the basic features (c.f. appendix A.3.5), listed in Table A.3. The logistic regression model for graph-based user

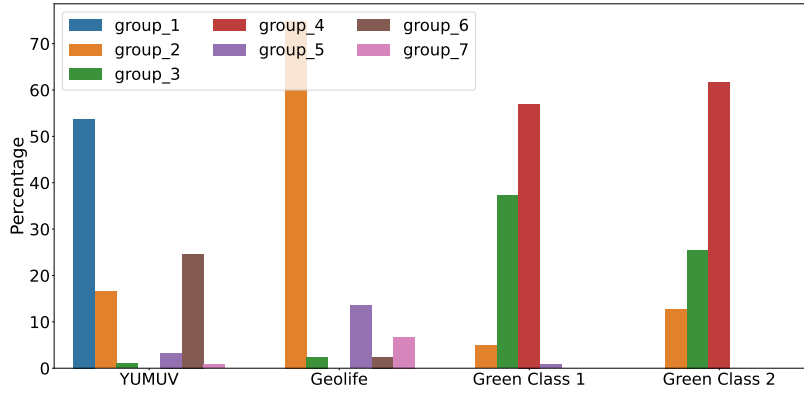


Fig. A.6.: Distribution of user groups (found based on classical mobility features) over datasets. Clearly, the groups are highly dependent on dataset properties.

groups is hardly better than random, with an accuracy of 0.375 (random is 0.35 for the 5 imbalanced classes) and an R-Squared value of 0.035. The coefficients in Table A.2 are also lower than for the basic features and mostly non-significant, with the exception of the Commuter group. Note that the coefficients of the first group (Centered) can not be computed since it serves as the reference group in the model.

	Commuter	Flexible	Local routine	Traveller
days tracked	5.48 (0.0)	0.13 (0.93)	-0.69 (0.62)	6.95 (0.0)
tracking coverage	-13.57 (0.02)	4.89 (0.37)	9.82 (0.07)	-5.94 (0.39)
Intercept	12.54 (0.02)	-3.78 (0.47)	-8.32 (0.11)	4.55 (0.49)

Tab. A.2.: Dependence of graph-based mobility profiles on tracking period and coverage. The coefficients with p-values in parentheses of a multinomial logistic regression model are shown. Significant coefficients are marked bold.

In contrast, the basic-features lead to user groups that are strongly influenced by the number of tracked days and the coverage, as shown by the large and significant coefficients in Table A.3. The accuracy is 0.59 (random: 0.47) and R-Squared is 0.279. In summary, while undesired dependencies on dataset-specific properties may still exist, the experiment shows the advantages of our approach with a comparably high robustness of the proposed graph-based feature set for mobility profiling.

	group 2	group 3	group 4	group 5	group 6	group 7
days tracked	-30.04 (0.0)	32.7 (0.0)	32.78 (0.0)	-18.13 (0.12)	-13.66 (0.0)	-38.51 (0.02)
tracking coverage	6.11 (0.3)	-40.0 (0.0)	-32.07 (0.0)	-21.11 (0.02)	-5.39 (0.22)	-6.51 (0.66)
Intercept	-5.25 (0.35)	31.98 (0.0)	24.59 (0.0)	17.87 (0.04)	5.5 (0.19)	4.52 (0.74)

Tab. A.3.: Dependence of mobility user groups, derived from basic features, on tracking coverage and duration. The coefficients with p-values in brackets of a multinomial logistic regression model are shown. Significant coefficients are marked bold.

A.4 Supporting information for Using rooftop photovoltaic generation to cover individual electric vehicle demand—a detailed case study

A.4.1 PV panel parameters

Manufacturer	Schüco
Model	SPV 170-SME-1
N_P	1
N_S	72
V_{OC0}	44 V
I_{SC0}	5.15 A
V_{MP0}	35 V
I_{MP0}	4.86 A
$\alpha_{I_{SC}}$	0.055
$\beta_{V_{OC}}$	-0.37
T_{ref}	45 °C
A_{cell}	125.0 * 125.0 mm
A_{panel}	1580.4 * 808.4 mm

Bibliography

- Ahas, Rein, Anto Aasa, Y Yuan, Martin Raubal, Zbigniew Smoreda, Yu Liu, Cezary Ziemlicki, Margus Tiru, and Matthew Zook (2015). “Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn”. *International Journal of Geographical Information Science* 29.11, pp. 2017–2039. DOI: 10.1080/13658816.2015.1063151 (cit. on p. 20).
- Akaike, Hirotogu (1998). “Information Theory and an Extension of the Maximum Likelihood Principle”. *Selected Papers of Hirotugu Akaike*. Ed. by Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa. New York, NY: Springer New York, pp. 199–213. DOI: 10.1007/978-1-4612-1694-0_15 (cit. on pp. 60, 65).
- Alessandretti, Laura, Piotr Sapiezynski, Vedran Sekara, Sune Lehmann, and Andrea Baronchelli (2018). “Evidence for a conserved quantity in human mobility”. *Nature Human Behaviour* 2.7, pp. 485–491. DOI: 10.1038/s41562-018-0364-x (cit. on pp. 20, 34, 47, 48, 88, 90, 94, 113).
- Alessandretti, Laura, Piotr Sapiezynski, Sune Lehmann, and Andrea Baronchelli (2017). “Multi-scale spatio-temporal analysis of human mobility”. *PLOS ONE* 12.2, e0171686. DOI: 10.1371/journal.pone.0171686 (cit. on p. 61).
- Alessandretti, Laura, Ulf Aslak, and Sune Lehmann (2020). “The scales of human mobility”. *Nature* 587.7834. Number: 7834 Publisher: Nature Publishing Group, pp. 402–407. DOI: 10.1038/s41586-020-2909-1 (cit. on p. 68).
- Alstott, Jeff, Ed Bullmore, and Dietmar Plenz (2014). “powerlaw: a Python package for analysis of heavy-tailed distributions”. *PloS one* 9.1, e85777. DOI: 10.1371/journal.pone.0085777 (cit. on pp. 54, 60).
- Andrés, Miguel E, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi (2013). “Geo-indistinguishability: Differential privacy for location-based systems”. *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pp. 901–914. DOI: 10.1145/2508859.2516735 (cit. on p. 72).
- Aslak, Ulf and Laura Alessandretti (2020). “Infostop: Scalable stop-location detection in multi-user mobility data”. *arXiv preprint arXiv:2003.14370* (cit. on p. 28).
- Axhausen, Kay W. and Tommy Gärling (1992). “Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems”. *Transport Reviews* 12.4, pp. 323–341. DOI: 10.1080/01441649208716826 (cit. on pp. 45, 87).
- Axhausen, Kay W. (2007). “Definition of movement and activity for transport modelling”. *Handbook of transport modelling*. Emerald Group Publishing Limited. DOI: 10.1108/9780857245670-016 (cit. on pp. 24–27, 29, 93, 97).

- Bachir, Danya, Ghazaleh Khodabandelou, Vincent Gauthier, Mounim El Yacoubi, and Eric Vachon (2018). “Combining bayesian inference and clustering for transport mode detection from sparse and noisy geolocation data”. *Joint European conference on machine learning and knowledge discovery in databases*. Springer, pp. 569–584. DOI: 10.1007/978-3-030-10997-4_35 (cit. on p. 33).
- BAFU (2023). *Kenngrossen zur Entwicklung der Treibhausgasemissionen in der Schweiz 1990–2021*. Tech. rep. Federal Office for the Environment (cit. on p. 2).
- Banerjee, Syagnik (2019). “Geosurveillance, location privacy, and personalization”. *Journal of Public Policy & Marketing* 38.4, pp. 484–499. DOI: 10.1177/0743915619860137 (cit. on p. 72).
- Banister, David (2008). “The sustainable mobility paradigm”. *Transport policy* 15.2, pp. 73–80. DOI: 10.1016/j.tranpol.2007.10.005 (cit. on p. 2).
- Barabási, Albert-László and Réka Albert (1999). “Emergence of scaling in random networks”. *science* 286.5439, pp. 509–512. DOI: 10.1515/9781400841356.349 (cit. on p. 55).
- Barabási, Albert-László (2016). *Network Science*. Third. Cambridge University Press (cit. on pp. 54, 55).
- Barros, Murillo Vetrone, Rodrigo Salvador, Cassiano Moro Piekarski, Antonio Carlos de Francisco, and Fausto Miguel Cereja Seixas Freire (2020). “Life cycle assessment of electricity generation: a review of the characteristics of existing literature”. *The International Journal of Life Cycle Assessment* 25.1, pp. 36–54. DOI: 10.1007/s11367-019-01652-4 (cit. on p. 135).
- Bassolas, Aleix, Hugo Barbosa-Filho, Brian Dickinson, Xerxes Dotiwalla, Paul Eastham, Riccardo Gallotti, Gourab Ghoshal, Bryant Gipson, Surendra A. Hazarie, Henry Kautz, Onur Kucuktunc, Allison Lieber, Adam Sadilek, and José J. Ramasco (Dec. 2019). “Hierarchical organization of urban mobility and its connection with city livability”. *Nature Communications* 10.1, p. 4817. DOI: 10.1038/s41467-019-12809-y (cit. on p. 20).
- Basterrech, Sebastián and Michał Woźniak (2022). “Tracking changes using Kullback-Leibler divergence for the continual learning”. *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, pp. 3279–3285. DOI: 10.1109/smc53654.2022.9945547 (cit. on p. 65).
- Ben-Gal, Irad, Shahar Weinstock, Gonen Singer, and Nicholas Bambos (2019). “Clustering users by their mobility behavioral patterns”. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13.4, pp. 1–28. DOI: 10.1145/3322126 (cit. on pp. 69, 86, 89).
- BFE (2020). *Energieperspektiven 2050+. Zusammenfassung der wichtigsten Ergebnisse*. Tech. rep. Bundesamt fuer Energie (cit. on p. 125).
- BFE (2019a). *Programmstrategie EnergieSchweiz: 2021 bis 2030*. Tech. rep. EnergieSchweiz and Bundesamt für Energie (cit. on p. 125).
- BFE (2019b). *Schweizerische Gesamtenergiestatistik 2019*. Tech. rep. Bern: Bundesamt fuer Energie (cit. on p. 121).

- BFS and ARE (2017). *Verkehrsverhalten der Bevölkerung - Ergebnisse des Mikrozensus Mobilität und Verkehr 2015*. Tech. rep. Bundesamt für Statistik und Bundesamt für Raumentwicklung (cit. on p. 10).
- Bhattacharya, Amiya and Sajal K. Das (2002). “LeZi-Update: An Information-Theoretic Framework for Personal Mobility Tracking in PCS Networks”. *Wireless Networks* 8.2, pp. 121–135. DOI: 10.1023/A:1013759724438 (cit. on pp. 46, 87).
- Boeing, Geoff (2017). “OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks”. *Computers, Environment and Urban Systems* 65, pp. 126–139. DOI: 10.1016/j.compenvurbsys.2017.05.004 (cit. on p. 35).
- Bongardt, Daniel, Lena Stiller, Anthea Swart, and Armin Wagner (2019). *Sustainable Urban Transport: Avoid-Shift-Improve (ASI)*. Tech. rep. Transformative Urban Mobility Initiative (cit. on p. 3).
- Boulouchos, Konstantinos, Christian Bach, Christian Bauer, Dominik Bucher, Davide Cerruti, Amin Dehdarian, Massimo Filippini, Maximilian Held, Stefan Hirschberg, Ramachandran Kannan, et al. (2021). *Pathways to a net zero CO2 Swiss mobility system: SCCER Mobility Whitepaper*. Tech. rep. ETH Zurich (cit. on pp. 2, 3).
- Breyer, Christian, Dmitrii Bogdanov, Ashish Gulagi, Arman Aghahosseini, Larissa S.N.S. Barbosa, Otto Koskinen, Maulidi Barasa, Upeksha Caldera, Svetlana Afanasyeva, Michael Child, Javier Farfan, and Pasi Vainikka (Mar. 2017). “On the role of solar photovoltaics in global energy transition scenarios”. *Progress in Photovoltaics: Research and Applications* 25.8, pp. 727–745. DOI: 10.1002/pip.2885 (cit. on p. 121).
- Brockmann, Dirk, Lars Hufnagel, and Theo Geisel (2006). “The scaling laws of human travel”. *Nature* 439.7075, pp. 462–465. DOI: 10.1038/nature04292 (cit. on pp. 20, 60, 93).
- Bronstein, Michael M, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst (2017). “Geometric Deep Learning: Going beyond Euclidean data”. *IEEE Signal Processing Magazine* 34.4, pp. 18–42. DOI: 10.1109/msp.2017.2693418 (cit. on pp. 4, 114, 149).
- Bucher, Dominik, René Buffat, Andreas Froemelt, and Martin Raubal (2019a). “Energy and greenhouse gas emission reduction potentials resulting from different commuter electric bicycle adoption scenarios in Switzerland”. *Renewable and Sustainable Energy Reviews* 114, p. 109298. DOI: 10.1016/j.rser.2019.109298 (cit. on p. 125).
- Bucher, Dominik, Francesca Mangili, Francesca Cellina, Claudio Bonesana, David Jonietz, and Martin Raubal (2019b). “From location tracking to personalized eco-feedback: A framework for geographic information collection, processing and visualization to promote sustainable mobility behaviors”. *Travel behaviour and society* 14, pp. 43–56. DOI: 10.1016/j.tbs.2018.09.005 (cit. on pp. 3, 24, 44, 123).
- Buffat, René, Stefano Grassi, and Martin Raubal (2018a). “A scalable method for estimating rooftop solar irradiation potential over large regions”. *Applied energy* 216, pp. 389–401. DOI: 10.1016/j.apenergy.2018.02.008 (cit. on pp. 124, 127, 131).
- Buffat, René (2016). “Feature-aware surface interpolation of rooftops using low-density Lidar data for photovoltaic applications”. *Geospatial Data in a Changing World*. Springer, pp. 337–350. DOI: 10.1007/978-3-319-33783-8_19 (cit. on p. 127).

- Buffat, René, Dominik Bucher, and Martin Raubal (2018b). “Using locally produced photovoltaic energy to charge electric vehicles”. *Computer Science-Research and Development* 33.1-2, pp. 37–47 (cit. on p. 125).
- Bunsen, Till, Pierpaolo Cazzola, Marine Gerner, Leonardo Paoli, Sacha Scheffer, Renske Schuitmaker, Jacopo Tattini, and Jacob Teter (2018). *Global EV Outlook 2018: Towards cross-modal electrification*. Tech. rep. International Energy Agency (cit. on p. 124).
- Burton, Elizabeth, Mike Jenks, and Katie Williams (2003). *The compact city: a sustainable urban form?* Routledge (cit. on p. 3).
- Çabukoglu, Emir, Gil Georges, Lukas Küng, Giacomo Pareschi, and Konstantinos Boulouchos (2019). “Fuel cell electric vehicles: An option to decarbonize heavy-duty transport? Results from a Swiss case-study”. *Transportation Research Part D: Transport and Environment* 70, pp. 35–48. DOI: 10.1016/j.trd.2019.03.004 (cit. on p. 123).
- Cai, H., Y. Xin, H. Martin, and M. Raubal (2022). “Optimizing Electric Vehicle Charging Schedules Based on Probabilistic Forecast of Individual Mobility”. *AGILE: GIScience Series* 3, p. 3. DOI: 10.5194/agile-giss-3-3-2022 (cit. on pp. 44, 61).
- Calafiore, Alessia, Gregory Palmer, Sam Comber, Daniel Arribas-Bel, and Alex Singleton (2021). “A geographic data science framework for the functional and contextual analysis of human dynamics within global cities”. *Computers, Environment and Urban Systems* 85, p. 101539. DOI: 10.1016/j.compenvurbsys.2020.101539 (cit. on p. 86).
- Calenge, Clement (2011). “Analysis of animal movements in R: the adehabitatLT package”. *R Foundation for Statistical Computing, Vienna* (cit. on p. 22).
- Calenge, Clement (2006). “The package “adehabitat” for the R software: a tool for the analysis of space and habitat use by animals”. *Ecological modelling* 197.3-4, pp. 516–519. DOI: 10.1016/j.ecolmodel.2006.03.017 (cit. on p. 21).
- Casals, Lluç Canals, Egoitz Martinez-Laserna, Beatriz Amante Garcia, and Nerea Nieto (2016). “Sustainability analysis of the electric vehicle use in Europe for CO2 emissions reduction”. *Journal of cleaner production* 127, pp. 425–437. DOI: 10.1016/j.jclepro.2016.03.120 (cit. on p. 121).
- Castiglione, Joe, Mark Bradley, and John Gliebe (2015). “Activity-Based Travel Demand Models: A Primer”. *SHRP 2 Report S2-C46-RR-1* (cit. on pp. 45, 87, 159).
- Cellina, Francesca, Dominik Bucher, Francesca Mangili, José Veiga Simão, Roman Rudel, and Martin Raubal (2019). “A Large Scale, App-Based Behaviour Change Experiment Persuading Sustainable Mobility Patterns: Methods, Results and Lessons Learnt”. *Sustainability* 11.9, p. 2674. DOI: 10.3390/su11092674 (cit. on p. 123).
- Chang, Serina, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec (2021). “Mobility network models of COVID-19 explain inequities and inform reopening”. *Nature* 589.7840, pp. 82–87. DOI: 10.1038/s41586-020-2923-3 (cit. on p. 20).
- Chapman, Lee (2007). “Transport and climate change: a review”. *Journal of transport geography* 15.5, pp. 354–367. DOI: 10.1016/j.jtrangeo.2006.11.008 (cit. on p. 86).

- Charleux, Laure and Katherine Schofield (2020). “True spatial k-anonymity: adaptive areal elimination vs. adaptive areal masking”. *Cartography and Geographic Information Science* 47.6, pp. 537–549. DOI: 10.1080/15230406.2020.1794975 (cit. on p. 72).
- Chen, Chao, Shuhai Jiao, Shu Zhang, Weichen Liu, Liang Feng, and Yasha Wang (2018). “TripImputor: real-time imputing taxi trip purpose leveraging multi-sourced urban data”. *IEEE Trans Intell Transp Syst* 99, pp. 1–13. DOI: 10.1109/tits.2017.2771231 (cit. on p. 112).
- Chen, Cynthia, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang (2016a). “The promises of big data and small data for travel behavior (aka human mobility) analysis”. *Transportation Research Part C: Emerging Technologies* 68, pp. 285–299. DOI: 10.1016/j.trc.2016.04.005 (cit. on pp. 5, 20, 24, 44, 45, 86, 87, 112).
- Chen, Meng, Yang Liu, and Xiaohui Yu (2014). “Nlpmm: A next location predictor with markov modeling”. *Advances in Knowledge Discovery and Data Mining: 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16, 2014. Proceedings, Part II 18*. Springer, pp. 186–197. DOI: 10.1007/978-3-319-06605-9_16 (cit. on p. 68).
- Chen, Qingqing and Ate Poorthuis (2021). “Identifying home locations in human mobility data: an open-source R package for comparison and reproducibility”. *International Journal of Geographical Information Science* 35.7, pp. 1425–1448. DOI: 10.1080/13658816.2021.1887489 (cit. on p. 34).
- Chen, Quanjun, Xuan Song, Harutoshi Yamada, and Ryosuke Shibasaki (2016b). “Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference”. *Thirtieth AAAI Conference on Artificial Intelligence*. DOI: 10.1609/aaai.v30i1.10011 (cit. on p. 4).
- Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman (2009). “Power-law distributions in empirical data”. *SIAM review* 51.4, pp. 661–703. DOI: 10.1137/070710111 (cit. on pp. 54, 60).
- Clement-Nyns, Kristien, Edwin Haesen, and Johan Driesen (2010). “The impact of charging plug-in hybrid electric vehicles on a residential distribution grid”. *IEEE Transactions on power systems* 25.1, pp. 371–380. DOI: 10.1109/tpwrs.2009.2036481 (cit. on pp. 122, 124).
- Cox, Brian, Christian Bauer, Angelica Mendoza Beltran, Detlef P van Vuuren, and Christopher L Mutel (2020). “Life cycle environmental and cost comparison of current and future passenger cars under different energy scenarios”. *Applied Energy* 269, p. 115021. DOI: 10.1016/j.apenergy.2020.115021 (cit. on p. 3).
- Craswell, Nick (2009). “Mean Reciprocal Rank.” *Encyclopedia of database systems* 1703. DOI: 10.1007/978-0-387-39940-9_488 (cit. on p. 76).
- Creutzig, Felix, Patrick Jochem, Oreane Y. Edelenbosch, Linus Mattauch, Detlef P. van Vuuren, David McCollum, and Jan Minx (2015). “Transport: A roadblock to climate change mitigation?” *Science* 350.6263, pp. 911–912. DOI: 10.1126/science.aac8033 (cit. on p. 86).

- Cui, Zhiyong, Kristian Henrickson, Ruimin Ke, and Yinhai Wang (2019a). “Traffic Graph Convolutional Recurrent Neural Network: A Deep Learning Framework for Network-Scale Traffic Learning and Forecasting”. *IEEE Transactions on Intelligent Transportation Systems*. Conference Name: IEEE Transactions on Intelligent Transportation Systems, pp. 1–12. DOI: 10.1109/tits.2019.2950416 (cit. on p. 149).
- Cui, Zhiyong, Kristian Henrickson, Ruimin Ke, and Yinhai Wang (2019b). “Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting”. *IEEE Transactions on Intelligent Transportation Systems* 21.11, pp. 4883–4894 (cit. on p. 114).
- Dabiri, Sina and Kevin Heaslip (2018). “Inferring transportation modes from GPS trajectories using a convolutional neural network”. *Transportation Research Part C: Emerging Technologies* 86, pp. 360–371. DOI: 10.1016/j.trc.2017.11.021 (cit. on p. 4).
- Damiani, Maria Luisa, Andrea Acquaviva, Fatima Hachem, and Matteo Rossini (2020). “Learning Behavioral Representations of Human Mobility”. *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*. SIGSPATIAL ’20. New York, NY, USA: Association for Computing Machinery, pp. 367–376. DOI: 10.1145/3397536.3422255 (cit. on p. 46).
- Das, Rahul Deb and Stephan Winter (2016). “A context-sensitive conceptual framework for activity modeling”. *Journal of Spatial Information Science* 2016.12, pp. 45–85. DOI: 10.5311/JOSIS.2016.12.260 (cit. on pp. 45, 48, 49).
- De Mulder, Yoni, George Danezis, Lejla Batina, and Bart Preneel (2008). “Identification via location-profiling in GSM networks”. *Proceedings of the 7th ACM workshop on Privacy in the electronic society*. WPES ’08. New York, NY, USA: Association for Computing Machinery, pp. 23–32. DOI: 10.1145/1456403.1456409 (cit. on pp. 46, 72).
- Defferrard, Michaël, Xavier Bresson, and Pierre Vandergheynst (2016). “Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering”. *Advances in Neural Information Processing Systems* 29. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, pp. 3844–3852 (cit. on pp. 112, 114, 115, 149, 153, 208).
- Dehmer, Matthias and Abbe Mowshowitz (2011). “A history of graph entropy measures”. *Information Sciences* 181.1, pp. 57–78. DOI: 10.1016/j.ins.2010.08.041 (cit. on p. 69).
- Dodge, Somayeh, Robert Weibel, Sean C. Ahearn, Maike Buchin, and Jennifer A. Miller (2016). “Analysis of movement data”. *International Journal of Geographical Information Science* 30.5, pp. 825–834. DOI: 10.1080/13658816.2015.1132424 (cit. on p. 44).
- Dodge, Somayeh, Song Gao, Martin Tomko, and Robert Weibel (2020). “Progress in computational movement analysis—towards movement data science”. *International Journal of Geographical Information Science* 34.12, pp. 2395–2400. DOI: 10.1080/13658816.2020.1784425 (cit. on p. 44).
- Douglas, David H and Thomas K Peucker (1973). “Algorithms for the reduction of the number of points required to represent a digitized line or its caricature”. *Cartographica: the international journal for geographic information and geovisualization* 10.2, pp. 112–122. DOI: 10.1002/9780470669488.ch2 (cit. on p. 33).

- Duckham, Matt and Lars Kulik (2005). “A Formal Model of Obfuscation and Negotiation for Location Privacy”. *Pervasive Computing*. Vol. 3468. Springer. Springer Berlin Heidelberg, pp. 152–170. DOI: 10.1007/11428572_10 (cit. on pp. 48, 72).
- Ebel, Holger, Lutz-Ingo Mielsch, and Stefan Bornholdt (2002). “Scale-free topology of e-mail networks”. *Physical review E* 66.3, p. 035103. DOI: 10.1103/physreve.66.035103 (cit. on p. 55).
- EEA (2023). *Annual European Union greenhouse gas inventory 1990–2021 and inventory report 2023*. Tech. rep. EEA/PUBL/2023/044. European Environment Agency (cit. on p. 2).
- Efstathiades, Hariton, Demetris Antoniadis, George Pallis, and Marios D Dikaiakos (2015). “Identification of key locations based on online social network activity”. *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, pp. 218–225. DOI: 10.1145/2808797.2808877 (cit. on pp. 34, 39).
- El Mahrsi, Mohamed K., Etienne Côme, Latifa Oukhellou, and Michel Verleysen (2016). “Clustering smart card data for urban mobility analysis”. *IEEE Transactions on intelligent transportation systems* 18.3, pp. 712–728. DOI: 10.1109/tits.2016.2600515 (cit. on p. 88).
- EPA (2023). *Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2021*. Tech. rep. EPA 430-R-23-002. U.S. Environmental Protection Agency (cit. on p. 2).
- Ermagun, Alireza and David Levinson (2018). “Spatiotemporal traffic forecasting: review and proposed directions”. *Transport Reviews* 38.6, pp. 786–814. DOI: 10.1080/01441647.2018.1442887 (cit. on p. 148).
- Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. (1996). “A density-based algorithm for discovering clusters in large spatial databases with noise.” *Kdd*. Vol. 96. 34, pp. 226–231 (cit. on pp. 53, 113).
- Evangelidou, Nikolaos, Henrik Grythe, Zbigniew Klimont, Chris Heyes, Sabine Eckhardt, Susana Lopez-Aparicio, and Andreas Stohl (2020). “Atmospheric transport is a major pathway of microplastics to remote regions”. *Nature communications* 11.1, pp. 1–11. DOI: 10.20944/preprints202003.0385.v1 (cit. on p. 86).
- Fachrizal, Reza, Mahmoud Shepero, Dennis van der Meer, Joakim Munkhammar, and Joakim Widén (2020). “Smart charging of electric vehicles considering photovoltaic power production and electricity consumption: A review”. *eTransportation* 4. DOI: 10.1016/j.etrans.2020.100056 (cit. on p. 133).
- Fagiolo, Giorgio (2007). “Clustering in complex directed networks”. *Physical Review E* 76.2, p. 026107. DOI: 10.1103/physreve.76.026107 (cit. on p. 58).
- Fan, Zipei, Xuan Song, Tianqi Xia, Renhe Jiang, Ryosuke Shibasaki, and Ritsu Sakuramachi (2018). “Online Deep Ensemble Learning for Predicting Citywide Human Mobility”. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2.3, pp. 1–21. DOI: 10.1145/3264915 (cit. on p. 4).

- Faria, Ricardo, Pedro Marques, Pedro Moura, Fausto Freire, Joaquim Delgado, and Anibal T. de Almeida (2013). “Impact of the electricity mix and use profile in the life-cycle assessment of electric vehicles”. *Renewable and Sustainable Energy Reviews* 24, pp. 271–287. DOI: 10.1016/j.rser.2013.03.063 (cit. on p. 124).
- Feng, Jie, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin (2018). “DeepMove: Predicting Human Mobility with Attentional Recurrent Networks”. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. the 2018 World Wide Web Conference. Lyon, France: ACM Press, pp. 1459–1468. DOI: 10.1145/3178876.3186058 (cit. on p. 20).
- Fey, Matthias and Jan Eric Lenssen (2019). “Fast Graph Representation Learning with PyTorch Geometric”. arXiv: 1903.02428 (cit. on p. 153).
- Fiore, Marco, Panagiota Katsikouli, Elli Zavou, Mathieu Cunche, Françoise Fessant, Dominique Le Hello, Ulrich Matchi Aivodji, Baptiste Olivier, Tony Quertier, and Razvan Stanica (2020). *Privacy in trajectory micro-data publishing : a survey* (cit. on p. 72).
- Floyd, Robert W (1962). “Algorithm 97: shortest path”. *Communications of the ACM* 5.6, p. 345. DOI: 10.1145/367766.368168 (cit. on p. 74).
- Freeman, Linton C (1977). “A set of measures of centrality based on betweenness”. *Sociometry*, pp. 35–41. DOI: 10.2307/3033543 (cit. on p. 74).
- Frick, Hannah and Ioannis Kosmidis (2017). “trackerR: Infrastructure for running and cycling data from GPS-enabled tracking devices in R”. *Journal of Statistical Software* 82, pp. 1–29 (cit. on p. 21).
- Froehlich, Jon, Tawanna Dillahunt, Predrag Klasnja, Jennifer Mankoff, Sunny Consolvo, Beverly Harrison, and James A Landay (2009). “UbiGreen: investigating a mobile tool for tracking and supporting green transportation habits”. *Proceedings of the sigchi conference on human factors in computing systems*. ACM, pp. 1043–1052. DOI: 10.1145/1518701.1518861 (cit. on p. 123).
- Fthenakis, V and M Rauegi (2017). “Environmental life-cycle assessment of photovoltaic systems”. *The Performance of Photovoltaic (PV) System*. Elsevier, pp. 209–232. DOI: 10.1016/b978-1-78242-336-2.00007-0 (cit. on p. 122).
- Fthenakis, Vasilis M, Hyung Chul Kim, and Erik Alsema (2008). “Emissions from photovoltaic life cycles”. *Environmental science & technology* 42.6, pp. 2168–2174. DOI: 10.1021/es071763q (cit. on p. 122).
- Gambs, Sébastien, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez (2014). “De-anonymization attack on geolocated data”. *Journal of Computer and System Sciences* 80.8, pp. 1597–1614. DOI: 10.1016/j.jcss.2014.04.024 (cit. on p. 72).
- Gambs, Sébastien, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez (2012). “Next Place Prediction Using Mobility Markov Chains”. *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*. MPM '12. New York, NY, USA: ACM, pp. 1–6. DOI: 10.1145/2181196.2181199 (cit. on p. 68).

- Ghosh, Shreya, Soumya K Ghosh, Rahul Deb Das, and Stephan Winter (2018). “Activity-based mobility profiling: A purely temporal modeling approach”. *Companion Proceedings of the The Web Conference 2018*, pp. 409–416. DOI: 10.1145/3184558.3186356 (cit. on p. 45).
- Gillies, Sean (2013). “The shapely user manual” (cit. on p. 23).
- Golle, Philippe and Kurt Partridge (2009). “On the anonymity of home/work location pairs”. *International Conference on Pervasive Computing*. Springer, pp. 390–397. DOI: 10.1007/978-3-642-01516-8_26 (cit. on p. 72).
- Gong, Lei, Toshiyuki Yamamoto, and Takayuki Morikawa (2016). “Comparison of Activity Type Identification from Mobile Phone GPS Data Using Various Machine Learning Methods”. *Asian Transport Studies* 4.1, pp. 114–128. DOI: 10.11175/eastsats.4.114 (cit. on p. 44).
- González, Marta C., César A. Hidalgo, and Albert-László Barabási (2008). “Understanding individual human mobility patterns”. *Nature* 453.7196, pp. 779–782. DOI: 10.1038/nature06958 (cit. on pp. 20, 60, 61, 90, 92, 107).
- Gössling, Stefan (2020). “Why cities need to take road space from cars - and how this could be done”. *Journal of Urban Design* 25.4, pp. 443–448. DOI: 10.1080/13574809.2020.1727318 (cit. on p. 86).
- Graser, Anita (2019). “MovingPandas: Efficient Structures for Movement Data in Python”. *GI_Forum* 1, pp. 54–68. DOI: 10.1553/giscience2019_01_s54 (cit. on pp. 22, 23).
- Gruteser, Marco and Dirk Grunwald (2003). “Anonymous usage of location-based services through spatial and temporal cloaking”. *Proceedings of the 1st international conference on Mobile systems, applications and services*, pp. 31–42. DOI: 10.1145/1066116.1189037 (cit. on p. 72).
- Haan, P and R Bianchetti (2016). *Szenarien der Elektromobilität in der Schweiz—Update 2016*. Tech. rep. EPB (cit. on p. 124).
- Haasz, Thomas, Jonatan J Gómez Vilchez, Robert Kunze, Paul Deane, David Fraboulet, Ulrich Fahl, and Eamonn Mulholland (2018). “Perspectives on decarbonizing the transport sector in the EU-28”. *Energy strategy reviews* 20, pp. 124–132. DOI: 10.1016/j.esr.2017.12.007 (cit. on pp. 3, 121).
- Haidri, Salman, Yaksh J Haranwala, Vania Bogorny, Chiara Renso, Vinicius Prado da Fonseca, and Amilcar Soares (2021). “PTRAIL—A python package for parallel trajectory data preprocessing”. *arXiv preprint arXiv:2108.13202*. DOI: 10.1016/j.softx.2022.101176 (cit. on pp. 22, 23).
- Hariharan, Ramaswamy and Kentaro Toyama (2004). “Project Lachesis: parsing and modeling location histories”. *International Conference on Geographic Information Science*. Springer, pp. 106–124. DOI: 10.1007/978-3-540-30231-5_8 (cit. on p. 28).
- Haussler, David (1999). *Convolution kernels on discrete structures*. Tech. rep. Department of Computer Science, University of California (cit. on p. 74).

- Hawkins, Troy R, Bhawna Singh, Guillaume Majeau-Bettez, and Anders Hammer Strømman (2013). “Comparative environmental life cycle assessment of conventional and electric vehicles”. *Journal of Industrial Ecology* 17.1, pp. 53–64 (cit. on p. 121).
- Haydari, Ammar, Michael Zhang, Chen-Nee Chuah, Jane Macfarlane, and Sean Peisert (2021). “Adaptive Differential Privacy Mechanism for Aggregated Mobility Dataset”. *arXiv:2112.08487 [cs]* (cit. on p. 72).
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016a). “Deep residual learning for image recognition”. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778. DOI: 10.1109/cvpr.2016.90 (cit. on pp. 149, 151, 152).
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016b). “Identity mappings in deep residual networks”. *European conference on computer vision*. Springer, pp. 630–645. DOI: 10.1007/978-3-319-46493-0_38 (cit. on p. 151).
- Hensher, David A., Chinh Q. Ho, and Daniel J. Reck (2021). “Mobility as a service and private car use: Evidence from the Sydney MaaS trial”. *Transportation Research Part A: Policy and Practice* 145, pp. 17–33. DOI: 10.1016/j.tra.2020.12.015 (cit. on p. 103).
- Hernandez, Diego (2018). “Uneven mobilities, uneven opportunities: Social distribution of public transport accessibility to jobs and education in Montevideo”. *Journal of Transport Geography* 67, pp. 119–125. DOI: 10.1016/j.jtrangeo.2017.08.017 (cit. on p. 2).
- Hinton, Geoffrey, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, and Tara Sainath (2012). “Deep Neural Networks for Acoustic Modeling in Speech Recognition”. *IEEE Signal Processing Magazine* 29 (cit. on p. 4).
- Hoarau, Quentin and Yannick Perez (2018). “Interactions between electric mobility and photovoltaic generation: A review”. *Renewable and Sustainable Energy Reviews* 94, pp. 510–522. DOI: 10.1016/j.rser.2018.06.039 (cit. on pp. 122, 124).
- Holden, Erling, David Banister, Stefan Gössling, Geoffrey Gilpin, and Kristin Linnerud (2020). “Grand Narratives for sustainable mobility: A conceptual review”. *Energy Research & Social Science* 65, p. 101454. DOI: 10.1016/j.erss.2020.101454 (cit. on p. 2).
- Holme, Petter and Jari Saramäki (2012). “Temporal networks”. *Physics Reports. Temporal Networks* 519.3, pp. 97–125. DOI: 10.1016/j.physrep.2012.03.001 (cit. on pp. 48, 52).
- Hong, Ye, Yanan Xin, Henry Martin, Dominik Bucher, and Martin Raubal (2021). “A Clustering-Based Framework for Individual Travel Behaviour Change Detection”. *11th International Conference on Geographic Information Science (GIScience 2021)-Part II*. Vol. 208. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, p. 4 (cit. on pp. 27, 86).
- Hong, Ye, Henry Martin, and Martin Raubal (2022). “How do you go where? Improving next location prediction by learning travel mode information using transformers”. *Proceedings of the 30th International Conference on Advances in Geographic Information Systems. SIGSPATIAL '22*. Seattle, Washington: Association for Computing Machinery, pp. 1–10. DOI: 10.1145/3557915.3560996. eprint: 2210.04095 (cit. on p. 46).
- Horni, Andreas, Kai Nagel, and Kay W. Axhausen (2016). *The Multi-Agent Transport Simulation MATSim*. Ubiquity Press. DOI: 10.5334/baw (cit. on p. 16).

- Hornsby, Kathleen and Max J. Egenhofer (2002). “Modeling Moving Objects over Multiple Granularities”. *Annals of Mathematics and Artificial Intelligence* 36.1-2, pp. 177–194. DOI: 10.1023/A:1015812206586 (cit. on pp. 46, 87).
- Hu, Xiaosong, Changfu Zou, Caiping Zhang, and Yang Li (2017). “Technological developments in batteries: a survey of principal roles, types, and management needs”. *IEEE Power and Energy Magazine* 15.5, pp. 20–31. DOI: 10.1109/mpe.2017.2708812 (cit. on p. 123).
- Huang, Haosheng, Georg Gartner, Jukka M. Krisp, Martin Raubal, and Nico Van de Weghe (2018a). “Location based services: ongoing evolution and research agenda”. *Journal of Location Based Services* 12.2, pp. 63–93. DOI: 10/ghx2v9 (cit. on pp. 20, 44).
- Huang, Haosheng, Yi Cheng, and Robert Weibel (2019). “Transport mode detection based on mobile phone network data: A systematic review”. *Transportation Research Part C: Emerging Technologies* 101, pp. 297–312. DOI: 10.1016/j.trc.2019.02.008 (cit. on p. 33).
- Huang, Wei, Lubing Li, and Hong K. Lo (2018b). “Adaptive traffic signal control with equilibrium constraints under stochastic demand”. *Transportation Research Part C: Emerging Technologies* 95. DOI: 10.1016/j.trc.2018.07.018 (cit. on p. 148).
- Hubert, Lawrence and Phipps Arabie (1985). “Comparing partitions”. *Journal of classification* 2.1, pp. 193–218. DOI: 10.1007/bf01908075 (cit. on pp. 107, 108).
- Icha, Petra and Gunter Kuhs (2020). *Entwicklung der spezifischen Kohlendioxid-Emissionen des deutschen Strommix in den Jahren 1990–2019*. Tech. rep. Umweltbundesamt (cit. on p. 135).
- IEA (2020a). *World Energy Outlook 2020*. Tech. rep. Paris: International Energy Agency. DOI: 10.1787/557a761b-en (cit. on p. 121).
- IEA (2020b). *World Energy Outlook 2022*. Tech. rep. Paris: International Energy Agency. DOI: 10.1787/3a469970-en (cit. on pp. 2, 162).
- ITF (2021). *Decarbonising Transport in Europe - the way forward*. Tech. rep. International Transport Forum (cit. on p. 4).
- Jaramillo, Paul, Suzana Kahn Ribeiro, Peter Newman, Subash Dhar, Obinna E. Diemuodeke, Toshihiko Kajino, Dongsoo S. Lee, Seiji B. Nugroho, Xiaoyu Ou, Anders Hammer Strømman, and Jake Whitehead (2022). “Transport”. *IPCC, 2022: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [P.R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, J. Malley (eds.)]* Cambridge, UK and New York, NY, USA: Cambridge University Press. DOI: 10.1017/9781009157926.012 (cit. on pp. 2, 161).
- Jiang, Shan, Joseph Ferreira, and Marta C Gonzalez (2017). “Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore”. *IEEE Transactions on Big Data* 3.2, pp. 208–219. DOI: 10.1109/tbdata.2016.2631141 (cit. on pp. 45, 87, 159).

- Jones, Christopher and Paul Gilbert (2018). “Determining the consequential life cycle greenhouse gas emissions of increased rooftop photovoltaic deployment”. *Journal of Cleaner Production* 184, pp. 211–219. DOI: 10.1016/j.jclepro.2018.02.140 (cit. on p. 136).
- Jonietz, David and Dominik Bucher (2018). “Continuous trajectory pattern mining for mobility behaviour change detection”. *LBS 2018: 14th International Conference on Location Based Services*. Springer. Zurich (cit. on pp. 25, 28, 113, 201).
- Jonietz, David, Dominik Bucher, Henry Martin, and Martin Raubal (2018). “Identifying and interpreting clusters of persons with similar mobility behaviour change processes”. *The Annual International Conference on Geographic Information Science*. Springer, pp. 291–307. DOI: 10.1007/978-3-319-78208-9_15 (cit. on p. 86).
- Joo, Rocío, Matthew E. Boone, Thomas A. Clay, Samantha C. Patrick, Susana Clusella-Trullas, and Mathieu Basille (2020). “Navigating through the R packages for movement”. *Journal of Animal Ecology* 89.1, pp. 248–267. DOI: 10.1111/1365-2656.13116 (cit. on p. 21).
- Jordahl, Kelsey, Joris Van den Bossche, Martin Fleischmann, James McBride, Jacob Wasserman, Adrian Garcia Badaracco, Jeffrey Gerard, Alan D. Snow, Jeff Tratner, Matthew Perry, Carson Farmer, Geir Arne Hjelle, Micah Cochran, Sean Gillies, Lucas Culbertson, Matt Bartos, Brendan Ward, Giacomo Caria, Mike Taves, Nick Eubank, sangarshanan, John Flavin, Matt Richards, Sergio Rey, maxalbert, Aleksey Bilogur, Christopher Ren, Dani Arribas-Bel, Daniel Mesejo-León, and Leah Wasser (2021). *geopandas/geopandas: v0.10.2*. Version v0.10.2. DOI: 10.5281/zenodo.5573592 (cit. on pp. 23, 24).
- Karpathy, Andrej, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei (2014). “Large-scale video classification with convolutional neural networks”. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732. DOI: 10.1109/cvpr.2014.223 (cit. on p. 114).
- Keßler, Carsten and Grant McKenzie (2018). “A geoprivacy manifesto”. *Transactions in GIS* 22.1, pp. 3–19 (cit. on pp. 5, 20, 44, 72).
- Kim, Jinsoo, Jae Hun Kim, and Gunwoo Lee (2022). “GPS data-based mobility mode inference model using long-term recurrent convolutional networks”. *Transportation Research Part C: Emerging Technologies* 135, p. 103523. DOI: 10.1016/j.trc.2021.103523 (cit. on p. 33).
- Kim, Yoon (2014). “Convolutional neural networks for sentence classification”. *arXiv preprint arXiv:1408.5882*. DOI: 10.3115/v1/d14-1181 (cit. on p. 114).
- Kipf, Thomas N and Max Welling (2016). “Semi-supervised classification with graph convolutional networks”. *arXiv preprint arXiv:1609.02907* (cit. on pp. 115, 151, 153).
- Kivelä, Mikko, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter (2014). “Multilayer networks”. *Journal of Complex Networks* 2.3, pp. 203–271. DOI: 10.1093/comnet/cnu016 (cit. on pp. 48, 49, 52).
- Konkol, Markus, Christian Kray, and Max Pfeiffer (2019). “Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study”. *International Journal of Geographical Information Science* 33.2, pp. 408–429 (cit. on p. 40).

- Kounadi, Ourania, Bernd Resch, and Andreas Petutschnig (2018). “Privacy Threats and Protection Recommendations for the Use of Geosocial Network Data in Research”. *Social Sciences* 7.10. DOI: 10.3390/socsci7100191 (cit. on p. 72).
- Krebs, Luana, Rolf Frischknecht, and Philippe Stolz (2020). *Environmental Life Cycle Assessment of Residential PV and Battery Storage Systems*. Tech. rep. ISBN 978-3-906042-97-8. International Energy Agency (IEA) PVPS Task 12 (cit. on p. 136).
- Kreil, David P., Michael K. Kopp, David Jonietz, Moritz Neun, Aleksandra Gruca, Pedro Herruzo, Henry Martin, Ali Soleymani, and Sepp Hochreiter (2020). “The surprising efficiency of framing geo-spatial time series forecasting as a video prediction task – Insights from the IARAI T4C Competition at NeurIPS 2019”. *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*. Ed. by Hugo Jair Escalante and Raia Hadsell. Vol. 123. Proceedings of Machine Learning Research. PMLR, pp. 232–241 (cit. on p. 86).
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). “ImageNet Classification with Deep Convolutional Neural Networks”. *Advances in Neural Information Processing Systems* 25. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., pp. 1097–1105. DOI: 10.1145/3065386 (cit. on pp. 4, 114).
- Krzyzanowski, Michal, Birgit Kuna-Dibbert, and Jürgen Schneider (2005). *Health effects of transport-related air pollution*. WHO Regional Office Europe (cit. on p. 2).
- Kulkarni, Vaibhav, Abhijit Mahalunkar, Benoit Garbinato, and John D. Kelleher (2019). “On the Inability of Markov Models to Capture Criticality in Human Mobility”. *Artificial Neural Networks and Machine Learning – ICANN 2019: Image Processing*. Ed. by Igor V. Tetko, Věra Kůrková, Pavel Karpov, and Fabian Theis. Cham: Springer International Publishing, pp. 484–497. DOI: 10.1007/978-3-030-30508-6_39 (cit. on p. 46).
- Kumar, Nishant and Martin Raubal (2021). “Applications of deep learning in congestion detection, prediction and alleviation: A survey”. *Transportation Research Part C: Emerging Technologies* 133, p. 103432. DOI: 10.1016/j.trc.2021.103432 (cit. on pp. 4, 44, 86).
- Küng, Lukas, Thomas Bütler, Gil Georges, and Konstantinos Boulouchos (2018). “Decarbonizing passenger cars using different powertrain technologies: Optimal fleet composition under evolving electricity supply”. *Transportation Research Part C: Emerging Technologies* 95, pp. 785–801. DOI: 10.1016/j.trc.2018.09.003 (cit. on p. 123).
- Lawrence, Martha and Richard Bullock (2022). *The Role of Rail in Decarbonizing Transport in Developing Countries*. Tech. rep. Mobility and Transport Connectivity Series. World Bank. DOI: 10.1596/38214 (cit. on p. 3).
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). “Deep learning”. *nature* 521.7553, pp. 436–444. DOI: 10.1038/nature14539 (cit. on pp. 4, 86).
- Lee, Junhyun, Inyeop Lee, and Jaewoo Kang (2019). “Self-Attention Graph Pooling”. arXiv: 1904.08082 (cit. on p. 152).
- Levin, Lena and Charlotta Faith-Ell (2019). “How to apply gender equality goals in transport and infrastructure planning”. *Integrating gender into transport planning: From one to many tracks*, pp. 89–118. DOI: 10.1007/978-3-030-05042-9_5 (cit. on p. 2).

- Li, Qimai, Zhichao Han, and Xiao-Ming Wu (2018). “Deeper insights into graph convolutional networks for semi-supervised learning”. *Thirty-Second AAAI Conference on Artificial Intelligence*. DOI: 10.1609/aaai.v32i1.11604 (cit. on p. 153).
- Li, Quannan, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma (2008). “Mining user similarity based on location history”. *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, pp. 1–10. DOI: 10.1145/1463434.1463477 (cit. on pp. 27, 28, 37, 53, 98).
- Liao, Lin, Dieter Fox, and Henry Kautz (2006). “Location-based activity recognition”. *Advances in Neural Information Processing Systems*, pp. 787–794 (cit. on p. 112).
- Lin, Miao, Hong Cao, Vincent Zheng, Kevin Chen-Chuan Chang, and Shonali Krishnaswamy (2015). “Mobility profiling for user verification with anonymized location data”. *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI’15. Buenos Aires, Argentina: AAAI Press, pp. 960–966 (cit. on p. 46).
- Liu, Feng, Davy Janssens, Geert Wets, and Mario Cools (2013). “Annotating mobile phone location data with activity purposes using machine learning algorithms”. *Expert Systems with Applications* 40.8, pp. 3299–3311. DOI: 10.1016/j.eswa.2012.12.100 (cit. on p. 112).
- Lloyd, Stuart (1982). “Least squares quantization in PCM”. *IEEE transactions on information theory* 28.2, pp. 129–137. DOI: 10.1109/tit.1982.1056489 (cit. on p. 100).
- Long, Jed A., Robert Weibel, Somayeh Dodge, and Patrick Laube (2018). “Moving ahead with computational movement analysis”. *International Journal of Geographical Information Science* 32.7, pp. 1275–1281. DOI: 10.1080/13658816.2018.1442974 (cit. on p. 86).
- Lotter, William, Gabriel Kreiman, and David Cox (2016). “Deep predictive coding networks for video prediction and unsupervised learning”. arXiv: 1605.08104 (cit. on p. 155).
- Lovelace, Robin and Richard Ellison (2018). “stplanr: A package for transport planning”. *The R Journal* 10.2, pp. 7–23. DOI: 10.32614/rj-2018-053 (cit. on pp. 22, 23).
- Lu, Xi, Michael B. McElroy, and Juha Kiviluoma (2009). “Global potential for wind-generated electricity”. *Proceedings of the National Academy of Sciences* 106.27. Publisher: National Academy of Sciences Section: Physical Sciences, pp. 10933–10938. DOI: 10.1073/pnas.0904101106 (cit. on p. 121).
- Lu, Xin, Erik Wetter, Nita Bharti, Andrew J. Tatem, and Linus Bengtsson (2013). “Approaching the Limit of Predictability in Human Mobility”. *Scientific Reports* 3, p. 2923. DOI: 10.1038/srep02923 (cit. on p. 68).
- Luca, Massimiliano, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo (2021). “A Survey on Deep Learning for Human Mobility”. *ACM Computing Surveys* 55.1, pp. 1–44. DOI: 10.1145/3485125 (cit. on pp. 20, 23, 44, 86).
- Luo, Ting, Xinwei Zheng, Guangluan Xu, Kun Fu, and Wenjuan Ren (2017). “An Improved DBSCAN Algorithm to Detect Stops in Individual Trajectories”. *ISPRS International Journal of Geo-Information* 6.3, p. 63. DOI: 10.3390/ijgi6030063 (cit. on p. 28).

- Luthander, Rasmus, Joakim Widén, Daniel Nilsson, and Jenny Palm (2015). “Photovoltaic self-consumption in buildings: A review”. *Applied Energy* 142, pp. 80–94. DOI: 10.1016/j.apenergy.2014.12.028 (cit. on pp. 122, 124).
- Ma, Guixiang, Nesreen K Ahmed, Theodore L Willke, and Philip S Yu (2021). “Deep graph similarity learning: A survey”. *Data Mining and Knowledge Discovery* 35, pp. 688–725. DOI: 10.1007/s10618-020-00733-5 (cit. on p. 84).
- Mahmud, Khizir, M Jahangir Hossain, and Graham E Town (2018). “Peak-load reduction by coordinated response of photovoltaics, battery storage, and electric vehicles”. *IEEE Access* 6, pp. 29353–29365. DOI: 10.1109/access.2018.2837144 (cit. on p. 125).
- Mann, Henry B. and Donald R. Whitney (1947). “On a test of whether one of two random variables is stochastically larger than the other”. *The annals of mathematical statistics*, pp. 50–60. DOI: 10.1214/aoms/1177730491 (cit. on p. 100).
- Manousakas, Dionysis, Cecilia Mascolo, Alastair R. Beresford, Dennis Chan, and Nikhil Sharma (2018). “Quantifying privacy loss of human mobility graph topology”. *Proceedings on Privacy Enhancing Technologies* 2018.3, pp. 5–21. DOI: 10.1515/popets-2018-0018 (cit. on pp. 47, 67, 72, 73, 77, 80, 83, 90, 159, 202).
- Martin, Henry, Henrik Becker, Dominik Bucher, David Jonietz, Martin Raubal, and Kay W Axhausen (2019a). “Begleitstudie SBB Green Class - Abschlussbericht”. *Arbeitsberichte Verkehrs-und Raumplanung* 1439. DOI: 10.3929/ethz-b-000353337 (cit. on pp. 36, 52, 73, 98, 126).
- Martin, Henry, Daniel J. Reck, Kay W. Axhausen, and Martin Raubal (2021b). *ETH Mobility Initiative Project MI-01-19 Empirical use and Impact analysis of MaaS: Ergebnisse*. Tech. rep. ETH Zurich. DOI: 10.3929/ethz-b-000521380 (cit. on pp. 36, 52, 97).
- Martin, Henry, Dominik Bucher, Esra Suel, Pengxiang Zhao, Fernando Perez-Cruz, and Martin Raubal (2018). “Graph Convolutional Neural Networks for Human Activity Purpose Imputation”. *NIPS Spatiotemporal Workshop at the 32nd Annual Conference on Neural Information Processing Systems (NIPS 2018)*. DOI: 10.3929/ethz-b-000310251 (cit. on pp. 4, 48, 69, 72, 88, 149).
- Martin, Henry, Nina Wiedemann, Daniel J Reck, and Martin Raubal (2023b). “Graph-based mobility profiling”. *Computers, Environment and Urban Systems* 100, p. 101910. DOI: 10.1016/j.compenvurbsys.2022.101910 (cit. on pp. 4, 69, 72, 81).
- Martin, Henry, Ye Hong, Nina Wiedemann, Dominik Bucher, and Martin Raubal (2023c). “Trackintel: An open-source Python library for human mobility analysis”. *Computers, Environment and Urban Systems* 101, p. 101938. DOI: 10.1016/j.compenvurbsys.2023.101938 (cit. on pp. 17, 53, 73, 97).
- Martin, Henry, Ye Hong, Dominik Bucher, Christian Rupprecht, and René Buffat (2019c). “Traffic4cast-Traffic Map Movie Forecasting–Team MIE-Lab”. *arXiv preprint arXiv:1910.13824*. arXiv: 1910.13824. DOI: 10.48550/arXiv.1910.13824 (cit. on pp. 149, 150, 153).
- Martin, Henry, Daniel Jan Reck, and Martin Raubal (2021c). “Using Information and Communication Technologies to facilitate mobility behaviour change and enable Mobility as a Service”. Vol. 9. 1. Austrian Academy of Sciences Press, pp. 187–193. DOI: 10.1553/giscience2021_01_s187 (cit. on pp. 3, 52).

- McKenzie, Grant and Daniel Romm (2021). “Measuring urban regional similarity through mobility signatures”. *Computers, Environment and Urban Systems* 89, p. 101684. DOI: 10.1016/j.compenvurbsys.2021.101684 (cit. on p. 86).
- McKinney, Wes (2010). “Data structures for statistical computing in python”. *Proceedings of the 9th Python in Science Conference*, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a (cit. on p. 24).
- McManus, Marcelle C (2012). “Environmental consequences of the use of batteries in low carbon systems: The impact of battery production”. *Applied Energy* 93, pp. 288–295. DOI: 10.1016/j.apenergy.2011.12.062 (cit. on p. 124).
- Melendez, Steven and Alex Pasternack (Mar. 2, 2019). “Here are the data brokers quietly buying and selling your personal information”. *The Fast Company* (cit. on p. 72).
- Messmer, Annika and Rolf Frischknecht (2016). *Umweltbilanz Strommix Schweiz 2014*. Tech. rep. Bundesamt für Umwelt (cit. on p. 135).
- Milo, Ron, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon (2002). “Network motifs: simple building blocks of complex networks”. *Science* 298.5594, pp. 824–827. DOI: 10.1126/science.298.5594.824 (cit. on p. 57).
- Mohammadi, Neda and John E Taylor (2017). “Urban energy flux: Spatiotemporal fluctuations of building energy consumption and human mobility-driven prediction”. *Applied energy* 195, pp. 810–818. DOI: 10.1016/j.apenergy.2017.03.044 (cit. on p. 125).
- Mokhtarian, Patricia L. and Ilan Salomon (2001). “How derived is the demand for travel? Some conceptual and measurement considerations”. *Transportation Research Part A: Policy and Practice* 35.8, pp. 695–719. DOI: 10.1016/S0965-8564(00)00013-6 (cit. on p. 45).
- Montini, Lara, Sebastian Probst, Johann Schrammel, Nadine Rieser-Schüssler, and Kay W Axhausen (2015). “Comparison of travel diaries generated from smartphone data and dedicated GPS devices”. *Transportation Research Procedia* 11, pp. 227–241. DOI: 10.1016/j.trpro.2015.12.020 (cit. on p. 112).
- Montini, Lara, Nadine Rieser-Schüssler, Andreas Horni, and Kay W Axhausen (2014). “Trip purpose identification from GPS tracks”. *Transportation Research Record* 2405.1, pp. 16–23. DOI: 10.3141/2405-03 (cit. on p. 112).
- Montjoye, Yves-Alexandre de, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel (2013). “Unique in the Crowd: The privacy bounds of human mobility”. *Scientific Reports* 3.1, p. 1376. DOI: 10.1038/srep01376 (cit. on pp. 44, 46, 72).
- Moro, Esteban, Dan Calacci, Xiaowen Dong, and Alex Pentland (2021). “Mobility patterns are associated with experienced income segregation in large US cities”. *Nature Communications* 12.1, p. 4633. DOI: 10.1038/s41467-021-24899-8 (cit. on p. 20).
- Müller, Richard, Uwe Pfeifroth, Christine Träger-Chatterjee, Roswitha Cremer, Jörg Trentmann, and Rainer Hollmann (2015). “Surface solar radiation data set - Heliosat (SARAH) - edition 1”. *Climate Monitoring Satellite Application Facility: Darmstadt, Germany* (cit. on p. 127).

- Munkhammar, Joakim, Pia Grahn, and Joakim Widén (2013). “Quantifying self-consumption of on-site photovoltaic power generation in households with electric vehicle home charging”. *Solar energy* 97, pp. 208–216. DOI: 10.1016/j.solener.2013.08.015 (cit. on pp. 123, 125, 137).
- Murphy, Kevin P. (2012). *Machine Learning: a Probabilistic Perspective*. MIT Press (cit. on p. 64).
- Mwasilu, Francis, Jackson John Justo, Eun-Kyung Kim, Ton Duc Do, and Jin-Woo Jung (2014). “Electric vehicles and smart grid interaction: A review on vehicle to grid and renewable energy sources integration”. *Renewable and Sustainable Energy Reviews* 34, pp. 501–516. DOI: 10.1016/j.rser.2014.03.031 (cit. on p. 122).
- Nash, Chris and John Whitelegg (2016). “Key research themes on regulation, pricing, and sustainable urban mobility”. *International Journal of Sustainable Transportation* 10.1, pp. 33–39 (cit. on p. 123).
- Nugent, Daniel and Benjamin K. Sovacool (2014). “Assessing the lifecycle greenhouse gas emissions from solar PV and wind energy: A critical meta-survey”. *Energy Policy* 65, pp. 229–244. DOI: 10.1016/j.enpol.2013.10.048 (cit. on p. 136).
- Nunes, Pedro, Raquel Figueiredo, and Miguel C Brito (2016). “The use of parking lots to solar-charge electric vehicles”. *Renewable and Sustainable Energy Reviews* 66, pp. 679–693. DOI: 10.1016/j.rser.2016.08.015 (cit. on p. 143).
- Nykqvist, Björn and Måns Nilsson (2015). “Rapidly falling costs of battery packs for electric vehicles”. *Nature Climate Change* 5.4, pp. 329–332. DOI: 10.1038/nclimate2564 (cit. on p. 124).
- Emissions Gap Report 2020* (2020). Tech. rep. UNEP DTU Partnership. DOI: 10.18356/9789280738124 (cit. on p. 121).
- Onnela, Jukka-Pekka, Jari Saramäki, János Kertész, and Kimmo Kaski (2005). “Intensity and coherence of motifs in weighted complex networks”. *Physical Review E* 71.6, p. 065103. DOI: 10.1103/physreve.71.065103 (cit. on pp. 58, 105).
- Ordóñez, J, E Jadraque, J Alegre, and G Martinez (2010). “Analysis of the photovoltaic solar energy capacity of residential rooftops in Andalusia (Spain)”. *Renewable and Sustainable Energy Reviews* 14.7, pp. 2122–2130. DOI: 10.1016/j.rser.2010.01.001 (cit. on p. 124).
- Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd (1999). *The PageRank citation ranking: Bringing order to the web*. Tech. rep. Stanford InfoLab (cit. on pp. 59, 94).
- Pappalardo, Luca, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási (2015). “Returners and explorers dichotomy in human mobility”. *Nature Communications* 6.1, p. 8166. DOI: 10.1038/ncomms9166 (cit. on pp. 4, 20, 48, 61, 74, 86, 88).
- Pappalardo, Luca, Filippo Simini, Gianni Barlacchi, Roberto Pellungrini, et al. (2022). “scikit-mobility: a Python library for the analysis, generation, and risk assessment of mobility data”. *JOURNAL OF STATISTICAL SOFTWARE* 103.4 (cit. on pp. 22, 23, 107).

- Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala (2019). “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. *Advances in Neural Information Processing Systems* 32. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., pp. 8026–8037 (cit. on p. 153).
- Paul A. Longley, Michael F. Goodchild, David J. Maguire, and David W. Rhind (2011). *Geographic Information Systems and Science (3rd ed.)* Chichester, England: John Wiley and Sons. (cit. on p. 132).
- Philipps, Simon and Werner Warmuth (2020). *Photovoltaics Report*. Tech. rep. Freiburg: Fraunhofer ISE (cit. on p. 121).
- Pilzecker, Andreas, Ricardo Fernandez, Nicole Mandl, and Elisabeth Rigler (2020). *Annual European Union greenhouse gas inventory 1990–2018 and inventory report 2020*. Tech. rep. European Environment Agency (cit. on p. 123).
- Prati, Gabriele (2018). “Gender equality and women’s participation in transport cycling”. *Journal of transport geography* 66, pp. 369–375. DOI: 10.31234/osf.io/d6z2t (cit. on p. 2).
- Prelicpean, Adrian C., Gyöző Gidófalvi, and Yusak O. Susilo (2017). “Transportation mode detection—an in-depth review of applicability and reliability”. *Transport reviews* 37.4, pp. 442–464 (cit. on p. 33).
- Primault, Vincent, Antoine Boutet, Sonia Ben Mokhtar, and Lionel Brunie (2018). “The long road to computational location privacy: A survey”. *IEEE Communications Surveys & Tutorials* 21.3, pp. 2772–2793. DOI: 10.1109/comst.2018.2873950 (cit. on p. 72).
- Rand, William M (1971). “Objective criteria for the evaluation of clustering methods”. *Journal of the American Statistical association* 66.336, pp. 846–850. DOI: 10.1080/01621459.1971.10482356 (cit. on pp. 107, 108).
- Ratti, Carlo, Dennis Frenchman, Riccardo Maria Pulselli, and Sarah Williams (2006). “Mobile landscapes: using location data from cell phones for urban analysis”. *Environment and planning B: Planning and design* 33.5, pp. 727–748. DOI: 10.1068/b32047 (cit. on p. 88).
- Raubal, Martin, Dominik Bucher, and Henry Martin (2021). “Geosmartness for Personalized and Sustainable Future Urban Mobility”. *Urban Informatics*. Ed. by Wenzhong Shi, Michael F. Goodchild, Michael Batty, Mei-Po Kwan, and Anshu Zhang. Singapore: Springer Singapore, pp. 59–83. DOI: 10.1007/978-981-15-8983-6_6 (cit. on pp. 3, 72).
- Raugei, Marco and Patricia Winfield (2019). “Prospective LCA of the production and EoL recycling of a novel type of Li-ion battery for electric vehicles”. *Journal of Cleaner Production* 213, pp. 926–932. DOI: 10.1016/j.jclepro.2018.12.237 (cit. on p. 136).
- Reck, Daniel J., David A. Hensher, and Chinh Q. Ho (2020). “MaaS bundle design”. *Transportation Research Part A: Policy and Practice* 141, pp. 485–501. DOI: 10.1016/j.tra.2020.09.021 (cit. on p. 3).
- Reed, Trevor (2019). “INRIX Global Traffic Scorecard” (cit. on pp. 86, 148).

- REN21 (2020). *Renewables 2020 Global Status Report*. Tech. rep. Paris: Renewable Energy Policy Network for the 21st Century (cit. on p. 123).
- Rhee, Injong, Minsu Shin, Seongik Hong, Kyunghan Lee, Seong Joon Kim, and Song Chong (2011). “On the Levy-Walk Nature of Human Mobility”. *IEEE/ACM Transactions on Networking* 19.3, pp. 630–643. DOI: 10.1109/TNET.2011.2120618 (cit. on p. 20).
- Richardson, David B (2013). “Electric vehicles and the electric grid: A review of modeling approaches, Impacts, and renewable energy integration”. *Renewable and Sustainable Energy Reviews* 19, pp. 247–254. DOI: 10.1016/j.rser.2012.11.042 (cit. on p. 124).
- Riesen, Kaspar, Xiaoyi Jiang, and Horst Bunke (2010). “Exact and inexact graph matching: Methodology and applications”. *Managing and Mining Graph Data*. Springer, pp. 217–247. DOI: 10.1007/978-1-4419-6045-0_7 (cit. on p. 74).
- Ringhand, Madlen and Mark Vollrath (2018). “Make this detour and be unselfish! Influencing urban route choice by explaining traffic management”. *Transportation Research Part F: Traffic Psychology and Behaviour* 53, pp. 99–116. DOI: 10.1016/j.trf.2017.12.010 (cit. on p. 148).
- Rinzivillo, Salvatore, Lorenzo Gabrielli, Mirco Nanni, Luca Pappalardo, Dino Pedreschi, and Fosca Giannotti (2014). “The purpose of motion: Learning activities from individual mobility networks”. *International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE. IEEE, pp. 312–318. DOI: 10.1109/dsaa.2014.7058090 (cit. on pp. 4, 47, 48, 69, 72, 88).
- Ritchie, Hannah, Max Roser, and Pablo Rosado (2020). “CO2 and Greenhouse Gas Emissions”. *Our World in Data* (cit. on pp. 2, 44).
- Rodrigue, Jean-Paul (2020). “Chapter 3 – Transportation, Economy and Society”. *The Geography of Transport Systems*. 5th ed. Routledge. DOI: 10.4324/9780429346323 (cit. on p. 2).
- Rossi, Luca, James Walker, and Mirco Musolesi (2015). “Spatio-temporal techniques for user identification by means of GPS mobility data”. *EPJ Data Science* 4.1, p. 11. DOI: 10.1140/epjds/s13688-015-0049-x (cit. on p. 72).
- Rossi, Ryan A and Nesreen K Ahmed (2014). “Role discovery in networks”. *IEEE Transactions on Knowledge and Data Engineering* 27.4, pp. 1112–1131. DOI: 10.1109/tkde.2014.2349913 (cit. on p. 68).
- Rout, Angela, Sophie Nitoslawski, Andrew Ladle, and Paul Galpern (2021). “Using smartphone-GPS data to understand pedestrian-scale behavior in urban settings: A review of themes and approaches”. *Computers, Environment and Urban Systems* 90, p. 101705. DOI: 10.1016/j.compenvurbsys.2021.101705 (cit. on p. 24).
- Rubner, Y., C. Tomasi, and L.J. Guibas (1998). “A metric for distributions with applications to image databases”. *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*. IEEE. Narosa Publishing House, pp. 59–66. DOI: 10.1109/iccv.1998.710701 (cit. on p. 115).

- Ruhnau, Oliver, Sergej Bannik, Sydney Otten, Aaron Praktijnjo, and Martin Robinius (2019). “Direct or indirect electrification? A review of heat generation and road transport decarbonisation scenarios for Germany 2050”. *Energy* 166, pp. 989–999. DOI: 10.1016/j.energy.2018.10.114 (cit. on pp. 3, 123).
- Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo (2021). ““Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI”. *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. DOI: 10.1145/3411764.3445518 (cit. on p. 40).
- Saner, Dominik, Niko Heeren, Boris Jäggi, Rashid A Waraich, and Stefanie Hellweg (2013). “Housing and mobility demands of individual households and their life cycle assessment”. *Environmental science & technology* 47.11, pp. 5988–5997. DOI: 10.1021/es304084p (cit. on p. 124).
- Schilt, Adrian (2020). *Emissionen von Treibhausgasen nach revidiertem CO2-Gesetz und Kyoto-Protokoll, 2. Verpflichtungsperiode (2013–2020)*. Tech. rep. Bern: Bundesamt für Umwelt BAFU (cit. on p. 123).
- Schneider, Christian M., Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C. González (2013). “Unravelling daily human mobility motifs”. *Journal of The Royal Society Interface* 10.84, p. 20130246. DOI: 10.1098/rsif.2013.0246 (cit. on pp. 20, 47, 48, 57, 88, 90).
- Schönfelder, Stefan and Kay W. Axhausen (2016). *Urban rhythms and travel behaviour: spatial and temporal phenomena of daily travel*. Routledge (cit. on p. 25).
- Schütze, Hinrich, Christopher D Manning, and Prabhakar Raghavan (2008). *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge. DOI: 10.1017/cbo9780511809071 (cit. on p. 94).
- Shenk, Justin, Wolf Byttner, Saranraj Nambusubramanian, and Alexander Zoeller (2021). “Traja: A Python toolbox for animal trajectory analysis”. *Journal of Open Source Software* 6.63, p. 3202. DOI: 10.21105/joss.03202 (cit. on pp. 22, 23).
- Shepero, Mahmoud, Joakim Munkhammar, Joakim Widén, Justin D. K. Bishop, and Tobias Boström (2018). “Modeling of photovoltaic power generation and electric vehicles charging on city-scale: A review”. *Renewable and Sustainable Energy Reviews* 89, pp. 61–71. DOI: 10.1016/j.rser.2018.02.034 (cit. on p. 122).
- Shokri, Reza, Carmela Troncoso, Claudia Diaz, Julien Freudiger, and Jean-Pierre Hubaux (2010). “Unraveling an old cloak: k-anonymity for location privacy”. *Proceedings of the 9th annual ACM workshop on Privacy in the electronic society*, pp. 115–118. DOI: 10.1145/1866919.1866936 (cit. on p. 72).
- Singer, Philipp, Denis Helic, Behnam Taraghi, and Markus Strohmaier (2014). “Detecting memory and structure in human navigation patterns using Markov chain models of varying order”. *PloS one* 9.7, e102070. DOI: 10.1371/journal.pone.0102070 (cit. on pp. 64, 65).
- SLOCAT (2023). *Transport and Climate Change Global Status Report: First insights*. Tech. rep. 3rd edition. Partnership on Sustainable, Low Carbon Transport (cit. on p. 2).

- SLOCAT (2021). *Transport and Climate Change Global Status Report: Tracking Trends in a Time of Change: The Need for Radical Action Towards Sustainable Transport Decarbonisation*. Tech. rep. 2nd edition. Partnership on Sustainable, Low Carbon Transport (cit. on p. 2).
- Smolak, Kamil, Katarzyna Siła-Nowicka, Jean-Charles Delvenne, Michał Wierzbiński, and Witold Rohm (2021). “The impact of human mobility data scales and processing on movement predictability”. *Scientific Reports* 11.1, pp. 1–10. DOI: 10.1038/s41598-021-94102-x (cit. on pp. 22, 23).
- Sofiev, Mikhail, James J Winebrake, Lasse Johansson, Edward W Carr, Marje Prank, Joana Soares, Julius Vira, Rostislav Kouznetsov, Jukka-Pekka Jalkanen, and James J Corbett (2018). “Cleaner fuels for ships provide public health benefits with climate tradeoffs”. *Nature communications* 9.1, p. 406. DOI: 10.1038/s41467-017-02774-9 (cit. on p. 2).
- Solomon, Adir, Amit Livne, Gilad Katz, Bracha Shapira, and Lior Rokach (May 2021). “Analyzing movement predictability using human attributes and behavioral patterns”. *Computers, Environment and Urban Systems* 87, p. 101596. DOI: 10.1016/j.compenvurbsys.2021.101596 (cit. on pp. 20, 28).
- Song, Chaoming, Zehui Qu, Nicholas Blumm, and Albert-László Barabási (2010a). “Limits of Predictability in Human Mobility”. *Science* 327.5968. Publisher: American Association for the Advancement of Science Section: Report, pp. 1018–1021. DOI: 10.1126/science.1177170 (cit. on pp. 5, 20, 81, 107, 112, 113).
- Song, Chaoming, Tal Koren, Pu Wang, and Albert-László Barabási (2010b). “Modelling the scaling properties of human mobility”. *Nature Physics* 6.10, pp. 818–823. DOI: 10.1038/nphys1760 (cit. on pp. 20, 60, 61).
- Stiasny, Jochen, Thierry Zufferey, Giacomo Pareschi, Damiano Toffanin, Gabriela Hug, and Konstantinos Boulouchos (2020). “Sensitivity Analysis of Electric Vehicle Impact on Low-Voltage Distribution Grids”. *21st Power Systems Computation Conference*. DOI: 10.1016/j.epsr.2020.106696 (cit. on p. 124).
- Sulis, Patrizia and Ed Manley (2018). “Exploring similarities and variations of human mobility patterns in the city of London”. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives*. Vol. 42. 4/W11. ISPRS, pp. 51–58. DOI: 10.5194/isprs-archives-xlii-4-w11-51-2018 (cit. on p. 88).
- Sweeney, Latanya (2002). “k-anonymity: A model for protecting privacy”. *International journal of uncertainty, fuzziness and knowledge-based systems* 10.05, pp. 557–570. DOI: 10.1142/s0218488502001648 (cit. on p. 72).
- The Economist (2020). “China aims to cut its net carbon-dioxide emissions to zero by 2060”. *The Economist* (cit. on p. 121).
- The pandas development team (2020). *pandas-dev/pandas: Pandas*. Version latest. DOI: 10.5281/zenodo.3509134 (cit. on p. 23).
- Thompson, Stuart A and Charlie Warzel (Dec. 19, 2019). “The Privacy Project: Twelve Million Phones, One Dataset, Zero Privacy”. *The New York Times* (cit. on p. 72).

- Toch, Eran, Boaz Lerner, Eyal Ben-Zion, and Irad Ben-Gal (2018). “Analyzing large-scale human mobility data: a survey of machine learning methods and applications”. *Knowledge and Information Systems* 58.3, pp. 501–523. DOI: 10.1007/s10115-018-1186-x (cit. on pp. 4, 23, 44, 86).
- Tong, Howell (1975). “Determination of the order of a Markov chain by Akaike’s information criterion”. *Journal of applied probability* 12.3, pp. 488–497. DOI: 10.1017/s0021900200048294 (cit. on p. 65).
- Tong, Wei, Yinggang Tong, Chang Xia, Jingyu Hua, Qun Li, and Sheng Zhong (2022). “Understanding Location Privacy of the Point-of-Interest Aggregate Data via Practical Attacks and Defenses”. *IEEE Transactions on Dependable and Secure Computing*. DOI: 10.1109/tdsc.2022.3184279 (cit. on p. 72).
- Trasarti, R., R. Guidotti, A. Monreale, and F. Giannotti (2017). “MyWay: Location prediction via mobility profiling”. *Information Systems* 64, pp. 350–367. DOI: 10.1016/j.is.2015.11.002 (cit. on p. 47).
- Tu, Zhen, Kai Zhao, Fengli Xu, Yong Li, Li Su, and Depeng Jin (2019). “Protecting Trajectory From Semantic Attack Considering k -Anonymity, l -Diversity, and t -Closeness”. *IEEE Transactions on Network and Service Management* 16.1, pp. 264–278 (cit. on p. 72).
- UN (2018). *2018 revision of world urbanization prospects*. Tech. rep. United Nations (cit. on pp. 4, 148, 162).
- UN (2015). *Transforming our world: the 2030 Agenda for Sustainable Development*. Tech. rep. A/RES/70/1. United Nations (cit. on p. 2).
- UNEP (2022). *Emissions Gap Report 2022: The Closing Window — Climate crisis calls for rapid transformation of societies*. Tech. rep. United Nations Environment Programme. DOI: 10.18356/9789210023993 (cit. on p. 2).
- UNFCCC (2015). *Paris Agreement*. Tech. rep. COP Report No. 21. United Nations Framework Convention on Climate Change (cit. on pp. 2, 121).
- Urner, Jorim, Dominik Bucher, Jing Yang, and David Jonietz (2018). “Assessing the Influence of Spatio-Temporal Context for Next Place Prediction using Different Machine Learning Approaches”. *ISPRS International Journal of Geo-Information* 7.5, p. 166. DOI: 10.3390/ijgi7050166 (cit. on pp. 20, 44, 112).
- Van Der Kam, Mart and Wilfried van Sark (2015). “Smart charging of electric vehicles with photovoltaic power and vehicle-to-grid technology in a microgrid; a case study”. *Applied energy* 152, pp. 20–30. DOI: 10.1016/j.apenergy.2015.04.092 (cit. on p. 137).
- Vassilvitskii, Sergei and David Arthur (2006). “ k -means++: The advantages of careful seeding”. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035 (cit. on p. 100).
- Vlahogianni, Eleni I., Matthew G. Karlaftis, and John C. Golias (2014). “Short-term traffic forecasting: Where we are and where we’re going”. *Transportation Research Part C: Emerging Technologies*. Special Issue on Short-term Traffic Flow Forecasting 43, pp. 3–19. DOI: 10.1016/j.trc.2014.01.005 (cit. on p. 148).

- Voorhees, Ellen M (1999). “The trec-8 question answering track report.” *Trec 99*, pp. 77–82 (cit. on p. 75).
- Vuchic, Vukan (1999). *Transportation for Livable Cities*. 1st. Routledge. DOI: 10.4324/9781351318167 (cit. on p. 2).
- Wang, Leye, Gehua Qin, Dingqi Yang, Xiao Han, and Xiaojuan Ma (2017). “Geographic Differential Privacy for Mobile Crowd Coverage Maximization”. *arXiv:1710.10477 [cs]*. DOI: 10.1609/aaai.v32i1.11285 (cit. on p. 72).
- Wang, Pengyang, Yanjie Fu, Hui Xiong, and Xiaolin Li (2019). “Adversarial substructured representation learning for mobile user profiling”. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 130–138. DOI: 10.1145/3292500.3330869 (cit. on p. 47).
- Wang, Ruoxi, Nan Li, and Yan Wang (2021). “Does the returners and explorers dichotomy in urban human mobility depend on the observation duration? An empirical study in Guangzhou, China”. *Sustainable Cities and Society* 69, p. 102862. DOI: 10.1016/j.scs.2021.102862 (cit. on p. 89).
- Wang, Shen, Soufiene Djahel, Zonghua Zhang, and Jennifer McManis (2016). “Next Road Rerouting: A Multiagent System for Mitigating Unexpected Urban Traffic Congestion”. *IEEE Transactions on Intelligent Transportation Systems* 17.10, pp. 2888–2899. DOI: 10.1109/tits.2016.2531425 (cit. on p. 148).
- Wang, Yingzi, Nicholas Jing Yuan, Defu Lian, Linli Xu, Xing Xie, Enhong Chen, and Yong Rui (2015). “Regularity and Conformity: Location Prediction Using Heterogeneous Mobility Data”. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’15. event-place: Sydney, NSW, Australia. New York, NY, USA: ACM, pp. 1275–1284. DOI: 10.1145/2783258.2783350 (cit. on p. 44).
- Warshall, Stephen (1962). “A theorem on boolean matrices”. *Journal of the ACM (JACM)* 9.1, pp. 11–12. DOI: 10.1145/321105.321107 (cit. on p. 74).
- WCED (1987). *Our Common Future*. Tech. rep. World Commission on Environment and Development (cit. on p. 2).
- Weiser, Paul, Simon Scheider, Dominik Bucher, Peter Kiefer, and Martin Raubal (2016). “Towards sustainable mobility behavior: Research challenges for location-aware information and communication technology”. *GeoInformatica* 20.2, pp. 213–239. DOI: 10.1007/s10707-015-0242-x (cit. on pp. 3, 44).
- Widhalm, Peter, Philippe Nitsche, and Norbert Brändie (2012). “Transport mode detection with realistic smartphone sensor data”. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, pp. 573–576 (cit. on p. 33).
- Wiedemann, Nina, Martin Henry, Esra Suel, Ye Hong, and Yanan Xin (2023). “Influence of tracking duration on the privacy of individual mobility graphs”. *Journal of Location Based Services (accepted)* (cit. on pp. 47, 52, 67).
- Wiedemann, Nina, Henry Martin, and Martin Raubal (2022). “Unlocking social network analysis methods for studying human mobility”. Vol. 3. Copernicus GmbH, pp. 1–12. DOI: 10.5194/agile-giss-3-19-2022 (cit. on pp. 4, 72).

- Wiginton, LK, Ha T Nguyen, and Joshua M Pearce (2010). “Quantifying rooftop solar photovoltaic potential for regional renewable energy policy”. *Computers, Environment and Urban Systems* 34.4, pp. 345–357. DOI: 10.1016/j.compenvurbsys.2010.01.001 (cit. on p. 124).
- Wilson, Andrew T (2014). *TrackTable Trajectory Analysis*. Tech. rep. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States) (cit. on pp. 21, 22).
- WMO (2021). *Climate Indicators and Sustainable Development Demonstrating the Interconnections*. Tech. rep. WMO-No. 1271. World Meteorological Organization (cit. on p. 2).
- World Bank (2020). *Global Photovoltaic Power Potential by Country*. Tech. rep. World Bank. DOI: 10.1596/34102 (cit. on p. 121).
- Wu, Zonghan, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu (2019). “A Comprehensive Survey on Graph Neural Networks”. *arXiv:1901.00596 [cs, stat]*. arXiv: 1901.00596. DOI: 10.1109/tnnls.2020.2978386 (cit. on p. 149).
- Xin, Yanan and Alan M MacEachren (2020). “Characterizing traveling fans: a workflow for event-oriented travel pattern analysis using Twitter data”. *International Journal of Geographical Information Science* 34.12, pp. 2497–2516. DOI: 10.1080/13658816.2020.1770259 (cit. on p. 88).
- Xu, Yanyan, Serdar Çolak, Emre C. Kara, Scott J. Moura, and Marta C. González (2018). “Planning for electric vehicle needs by coupling charging profiles with urban mobility”. *Nature Energy* 3.6, pp. 484–493. DOI: 10.1038/s41560-018-0136-x (cit. on p. 20).
- Yan, Sijie, Yuanjun Xiong, and Dahua Lin (2018). “Spatial temporal graph convolutional networks for skeleton-based action recognition”. *arXiv preprint arXiv:1801.07455*. DOI: 10.1609/aaai.v32i1.12328 (cit. on p. 114).
- Yan, Xiao-Yong, Wen-Xu Wang, Zi-You Gao, and Ying-Cheng Lai (2017). “Universal model of individual and population mobility on diverse spatial scales”. *Nature Communications* 8.1, p. 1639. DOI: 10.1038/s41467-017-01892-8 (cit. on p. 88).
- Yanardag, Pinar and S. V. N. Vishwanathan (2015). “Deep graph kernels”. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1365–1374. DOI: 10.1145/2783258.2783417 (cit. on p. 79).
- Yang, Dingqi, Daqing Zhang, Longbiao Chen, and Bingqing Qu (2015). “Nationtelescope: Monitoring and visualizing large-scale collective behavior in lbsns”. *Journal of Network and Computer Applications* 55, pp. 170–180. DOI: 10.1016/j.jnca.2015.05.010 (cit. on pp. 53, 98).
- Yang, Dingqi, Daqing Zhang, and Bingqing Qu (2016). “Participatory cultural mapping based on collective behavior data in location-based social networks”. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7.3, pp. 1–23. DOI: 10.1145/2814575 (cit. on pp. 53, 98).
- Ying, Zhitao, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec (2018). “Hierarchical Graph Representation Learning with Differentiable Pooling”. *Advances in Neural Information Processing Systems* 31. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., pp. 4800–4810 (cit. on p. 152).

- Yu, Bing, Haoteng Yin, and Zhanxing Zhu (2018). "Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting". *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. arXiv: 1709.04875, pp. 3634–3640. DOI: 10.24963/ijcai.2018/505 (cit. on p. 149).
- Yu, Qing and Jian Yuan (2022). "TransBigData: A Python package for transportation spatio-temporal big data processing, analysis and visualization". *Journal of Open Source Software* 7.71, p. 4021. DOI: 10.21105/joss.04021 (cit. on pp. 22, 23).
- Yuan, Nicholas Jing, Yu Zheng, Liuhan Zhang, and Xing Xie (2013). "T-Finder: A Recommender System for Finding Passengers and Vacant Taxis". *IEEE Transactions on Knowledge and Data Engineering* 25.10, pp. 2390–2403. DOI: 10.1109/TKDE.2012.153 (cit. on p. 28).
- Yuan, Yihong, Martin Raubal, and Yu Liu (2012). "Correlating mobile phone usage and travel behavior—A case study of Harbin, China". *Computers, Environment and Urban Systems* 36.2, pp. 118–130. DOI: 10.1016/j.compenvurbsys.2011.07.003 (cit. on p. 44).
- Yuan, Yihong and Martin Raubal (2012). "Extracting dynamic urban mobility patterns from mobile phone data". *International conference on geographic information science*. Springer, pp. 354–367. DOI: 10.1007/978-3-642-33024-7_26 (cit. on pp. 20, 88).
- Zah, Rainer and Peter de Haan (2012). *Chancen und Risiken der Elektromobilität in der Schweiz*. Vol. 59. vdf Hochschulverlag AG. DOI: 10.3218/3488-2 (cit. on p. 124).
- Zang, Hui and Jean Bolot (2011). "Anonymization of location data does not work: A large-scale measurement study". *Proceedings of the 17th annual international conference on Mobile computing and networking*, pp. 145–156. DOI: 10.1145/2030613.2030630 (cit. on p. 72).
- Zhang, Yang, Tao Cheng, Yibin Ren, and Kun Xie (2019). "A novel residual graph convolution deep learning model for short-term network-based traffic forecasting". *International Journal of Geographical Information Science* 0.0, pp. 1–27. DOI: 10.1080/13658816.2019.1697879 (cit. on p. 149).
- Zhang, Yatao and Martin Raubal (2022). "Street-level traffic flow and context sensing analysis through semantic integration of multisource geospatial data". *Transactions in GIS*. DOI: 10.1111/tgis.13005 (cit. on p. 69).
- Zhao, Pengxiang, David Jonietz, and Martin Raubal (2021). "Applying frequent-pattern mining and time geography to impute gaps in smartphone-based human-movement data". *International Journal of Geographical Information Science*, pp. 1–29. DOI: 10.1080/13658816.2020.1862126 (cit. on p. 29).
- Zhao, Pengxiang, Mei-Po Kwan, and Kun Qin (2017). "Uncovering the spatiotemporal patterns of CO2 emissions by taxis based on Individuals' daily travel". *Journal of Transport Geography* 62, pp. 122–135. DOI: 10.1016/j.jtrangeo.2017.05.001 (cit. on p. 112).
- Zhao, Zhi-Dan, Shi-Min Cai, and Yang Lu (2015). "Non-Markovian character in human mobility: Online and offline". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25.6, p. 063106. DOI: 10.1063/1.4922302 (cit. on p. 46).

- Zheng, Yu, Xing Xie, Wei-Ying Ma, et al. (2010). “Geolife: A collaborative social networking service among user, location and trajectory.” *IEEE Data Eng. Bull.* 33.2, pp. 32–39 (cit. on pp. 27, 33).
- Zheng, Yu, Lizhu Zhang, Xing Xie, and Wei-Ying Ma (2009). “Mining interesting locations and travel sequences from GPS trajectories”. *Proceedings of the 18th international conference on World wide web*, pp. 791–800. DOI: 10.1145/1526709.1526816 (cit. on pp. 36, 46, 52, 87, 98).
- Zheng, Yu (2015). “Trajectory data mining: an overview”. *ACM Transactions on Intelligent Systems and Technology (TIST)* 6.3, pp. 1–41. DOI: 10.1145/2743025 (cit. on pp. 23, 28, 112).
- Zheng, Yu, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma (2008). “Understanding Mobility Based on GPS Data”. *Proceedings of the 10th International Conference on Ubiquitous Computing. UbiComp '08*. event-place: Seoul, Korea. New York, NY, USA: ACM, pp. 312–321. DOI: 10.1145/1409635.1409677 (cit. on pp. 3, 4, 48, 88).
- Zhou, Jie, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun (2018). “Graph Neural Networks: A Review of Methods and Applications”. *arXiv:1812.08434 [cs, stat]*. arXiv: 1812.08434. DOI: 10.1016/j.aiopen.2021.01.001 (cit. on p. 149).
- Zimányi, Esteban, Mahmoud Sakr, and Arthur Lesuisse (2020). “MobilityDB: A mobility database based on PostgreSQL and PostGIS”. *ACM Transactions on Database Systems (TODS)* 45.4, pp. 1–42. DOI: 10.1145/3406534 (cit. on pp. 22, 23).

Webpages

- Graser, Anita (2020). *Tools for the analysis of movement data*. URL: <https://github.com/anitagraser/movement-analysis-tools> (cit. on p. 21).
- Sasidharan, Manu (2017). *Better Rural Transport is Key to Food Security and Zero Hunger*. URL: <https://slocat.net/1901-2/> (visited on May 22, 2023) (cit. on p. 2).

List of Figures

1.1	Organization of the dissertation.	6
2.1	Data preprocessing workflow deployed for the SBB Green Class pilot studies.	11
2.2	Distance-based modal split of GC1 participants over the course of the project.	13
2.3	Distance-based modal split of GC2 participants over the course of the project.	13
2.4	Overview of study design of the EIM project and the number of participants of the treatment group (TG) and the control group (CG) that successfully completed each phase of the study.	16
2.5	Data preprocessing workflow deployed for the EIM project.	17
3.1	Overview of the Trackintel framework.	24
3.2	Semantic visualization of the Trackintel data models and their UML diagram, with mandatory and optional attributes (shown in square brackets). The relations between the different classes are shown in the connecting lines. Figure adapted from (Jonietz and Bucher, 2018)	25
3.3	Semantic visualization of the relations between positionfix, staypoint and locations. Staypoints are groups of positionfixes where the users are stationary, and locations are aggregations of staypoints that the user visits multiple times. Locations can be generated across users (left) or for each user individually (right). Map data ©2022 Google.	29
3.4	The algorithm of tour generation implemented in Trackintel. A list of start candidates is maintained and iteratively checked for tour-closing trips.	32
3.5	The Trackintel framework offers functions to plot positionfixes (a), staypoints (b), and triplegs (c) together with the road network acquired from OpenStreetMaps. This example maps the movements of one Geolife participant.	35
3.6	The visualization result of the Trackintel <code>plot_modal_split()</code> function of the triplegs recorded from one Geolife participant. Major differences can be observed between the aggregation by count (number of triplegs) (a) and distance traveled (b).	35
3.7	Distribution of the individual temporal tracking quality for the considered datasets.	38
3.8	Comparison of modal split between datasets. The users of different studies differ considerably in terms of their usage of slow, motorized or fast transport.	39
3.9	Distribution of activities over time.	40

3.10	Location graphs are generated based on the activity sequence of a person in two aggregation steps. In the first step, activities represented by nodes are aggregated if they were performed at the same location (l_i) and share the same activity label (s_i), which leads to the activity graph. In the second step, all nodes of the activity graph are aggregated if they share the same location (l_i).	49
3.11	a: example of a typical location graph with a relatively low in-degree distribution β ; b: example of a typical location graph with a relatively high in-degree distribution β ; c: degree distribution of the graphs shown in a) (blue) and b) (orange); d: distribution of in-degree distribution β for different datasets. . .	55
3.12	a: Location graph of an individual with low $n_{\phi_{50}}$; b: Location graph of an individual with high $n_{\phi_{50}}$; c: Histogram of frequency of top 10 journeys for the individual in a); d: Histogram of frequency of top 10 journeys for the individual in b); e: distribution of $n_{\phi_{50}}$ over individuals for different datasets.	57
3.13	Examples of (unweighted and undirected) location graphs with low and high unweighted and undirected clustering coefficient and the distribution of unweighted and undirected clustering coefficients over users for different datasets.	59
3.14	Examples of (unweighted and undirected) location graphs with low and high $PR_{0.5}$ and the distribution of $PR_{0.5}$ over users for different datasets. Graphs are drawn unweighted and undirected for better visibility	59
3.15	Distribution of displacements and best-fit comparison between original data and graph-based approximation for different datasets.	62
3.16	Box plot of the difference between the radius of gyration and the graph-based radius of gyration over users for different datasets (absolute average e_{rel} shown in parentheses).	63
3.17	KL divergence for journey distributions generated from Markov models of different orders.	66
3.18	AIC for different k normalized by division by AIC of $k = 0$. 66	
3.19	Log likelihood evaluated for Markov models with different order evaluated for unseen test data. Impossible transitions (as defined by the mm) return no log-likelihood.	66
3.20	Experimental setup: The tracking data, comprising 56 weeks, are split into non-overlapping bins of varying duration. In the attack scenario, new tracking data from one period is matched to a pool of users at a previous time period. In example 1) the test data of four weeks length can be compared to the pool in the preceding 1, 2, 4 and 8 weeks. In the second example (marked as 2), a test user with tracking data from the second 24-week period is matched to users from all directly preceding tracking data, which includes one from each tracking duration except for 28 weeks.	76
3.21	Dependency of matching performance on tracking duration. Top-k accuracy and MRR increase with both the tracking duration of the pool users as well as the test user.	78
3.22	Evaluation of rank distribution and privacy loss as proposed by Manousakas et al. (2018).	80

3.23	Inter vs intra person variability of matching performance. The variance over users is higher than the variance over time bins. Intra-user variance decreases with growing tracking duration.	81
3.24	The re-identification accuracy decreases when there is a larger temporal gap between pool-bin and user-bin. However, the accuracy converges slowly and retains more than half of its former value even after one year.	83
4.1	Overview of graph features with an example graph from the Green Class 1 dataset for which the feature is rather low or rather high.	91
4.2	Power law fit for location graphs. The degrees are normalized by the highest degree found and ranked, and a power law is fitted.	92
4.3	Correlation between features	94
4.4	Workflow of user identification via clustering. a) Features f_i are extracted from the location graphs of each user u_i and form a feature matrix. b) A clustering algorithm is applied t times with different random initialization and parameters, yielding t different partitions of the users. In c), one such partition P is shown schematically for two features. d) By means of a statistical test, we determine for each feature whether the values in one cluster are significantly different from feature values in all other clusters of the partition. e) Based on the significant features, the clusters of all partitions can be merged to an existing group or define new groups. The result is a set of groups G with every cluster assigned to one group.	95
4.5	Feature properties for each user group: The user groups that are consistently found in the data are named based on significant differences in their feature values with respect to the other clusters.	101
4.6	User groups by study. With the exception of the Foursquare dataset, the user groups are similarly distributed. Differences can be explained by variations in the study target group, e.g., the Green Class 1 offer attracted individuals that cover longer distances.	102
4.7	Characteristics of yumuv users compared to the control group. Users assigned to the groups <i>Flexible</i> and <i>Traveller</i> are more likely to be interested in the yumuv MaaS offer.	103
4.8	User group changes upon intervention (start of the yumuv offer).	104
4.9	Mean and standard deviation of different features with respect to tracking period. The selected features are within the green box.	106
4.10	Distribution of basic features over the identified user groups. On the one hand, differences in the basic features are also reflected in our user groups. On the other hand, our groups seem to identify further differences in mobility behavior that are hardly reflected in the basic features (e.g., group <i>Centered</i>).	108
4.11	To create activity areas (A_i) from individual staypoints, we employ a DBSCAN clustering algorithm. Activity areas can be connected ($w_{i,j}$) in various ways, e.g., with their weights corresponding to the number of times a user directly traveled between two activity areas. The labels of the activity areas correspond to the distribution of the underlying staypoints.	113

4.12	The computation graph used within this work: The transition and distance graphs are each convoluted twice (layer sizes of 400 and 200). Their outputs are then combined using a weighted sum (layer size 50). Finally, a fully connected layer produces the output label distributions.	115
5.1	Share of BEVs at home or not at home	126
5.2	Exemplary week of a single user showing the usage of the BEV as well as the energy production by rooftop PV. In this recorded charging pattern (i.e., the baseline scenario, as explained below), it would be beneficial to not immediately recharge the car to the maximum, but instead wait for periods of increased PV energy generation.	129
5.3	Scatterplot of the difference of the SoC before and after a drive and the consumed energy.	130
5.4	Daily energy demand of the electric vehicles of study participants. On the left, the average demand and its 95% confidence intervals are shown over the course of the study. The data shows a week of excluded data in the beginning of October. On the right, box plots of individual weekdays are shown. Outliers in the box plot go up until about 100 kWh which corresponds to the consumption of three full charges in a single day.	131
5.5	PV energy generation at homes of study participants. Left side shows the average daily PV generation per participant as well as the 95 % confidence interval over the course of the study period. The right side shows the distribution of the peak generation capacity per house.	132
5.6	Flowcharts representing the charging logic of the different scenarios (top left: scenario 1 segment-wise optimization, top right: scenario 2 cross-segment optimization, bottom: scenario 3 cross-segment optimization with battery).	134
5.7	Share of energy (required for mobility) generated by local rooftop PV for each user in all scenarios over the full study duration.	137
5.8	Energy sources (rooftop PV or the grid) for the (cumulative) energy demand of BEVs for each scenario.	138
5.9	Shares of energy produced by rooftop PV for each user over the duration of the study period.	139
5.10	Average greenhouse gas emissions stemming from electric mobility for all users. On the left side, a typical Swiss power mix is assumed, on the right side a German power mix (only the direct emissions considered for Germany). Two weeks in October affected by missing data are excluded. Uncertainty intervals show the 95% confidence interval of the mean.	140
5.11	Analysis of correlation between total energy demand for mobility resp. peak power production on individual rooftops and energy coverage by PV power. In red, the baseline scenario is shown, in blue scenario 1, in purple scenario 2 and in grey scenario 3.	141
5.12	Average coverage value as a function of the overall PV system efficacy. The graph shows the average over all users per scenario and the 95 % confidence interval.	142

5.13	Logarithm of the per city sum of all training images of all channels.	149
5.14	The graph-based traffic forecasting workflow. The input images are transformed into graphs by filtering out inactive pixels, after which the Graph ResNet learns to predict traffic on this graph.	151
5.15	The different GCN architectures used and introduced in this paper.	152
5.16	Traffic forecasting results. All models are trained on Moscow (top) and tested on Berlin (middle) and Istanbul (bottom). The competition baseline is shown as dashed line ¹	155
A.1	Dependency of matching performance on tracking duration for the Green Class 2 data. Similarly to the results for Green Class 1, the top-k accuracy and MRR increase with both the tracking duration of the pool users as well as the test user. Due to the lower number of users, the re-identification performance is higher, reaching up to 82% top-10 accuracy.	168
A.2	Scatterplot matrix for the features of all datasets. Outliers were removed beforehand. The user groups can be clearly distinguished on certain features axes.	169
A.3	Cross-sectional study for Green Class 1 and 2: There are more Travellers and Commuters taking part in the Green Class studies compared to the proportion in other datasets.	170
A.4	Changes of user groups from the period before access to the yumuv offer to the period after intervention. The arrow width corresponds to the number of users that change from one group to another.	171
A.5	Characteristics of the identified groups based on classical mobility features.	172
A.6	Distribution of user groups (found based on classical mobility features) over datasets. Clearly, the groups are highly dependent on dataset properties.	173

List of Tables

3.1	Comparison of movement data libraries.	22
3.2	Description of the mandatory and optional columns for Trackintel data models.	26
3.3	Overview of basic features of the considered tracking studies. Locations, staypoints, triplets, trips and tours are given in multiples of a thousand.	36
3.4	Overview of the mobility statistics for the considered tracking datasets.	38
3.5	Used datasets with properties. Column <i>users</i> shows participants after filtering and the total participants in the dataset. For <i>Number of nodes</i> and <i>Average daily activities</i> the average and the standard deviation over users is shown.	53
3.6	Average and standard deviation of power law fitting parameters over users for the distribution of visitation frequency. The asterisk indicates a significant difference on a $p = 0.05$ level.	61
3.7	Regression analysis of the effect of the pool- and test-user tracking duration on the matching performance. Both positively affect the re-identification performance (=negative impact on privacy); however, the effect of the test duration is slightly higher. The matching performance is higher if the absolute difference between the pool and test user duration is low. All results are significant (p-values $\ll 0.01$).	78
3.8	Matching performance of different combinations of features, distance functions, and evaluation metrics. The highest matching accuracy is achieved with an R-convolution kernel that computes the MSE between all graph-features distributions combined.	79
3.9	Effect of mobility behavior and socio-demographics on re-identification accuracy, i.e., the rank of a user. A linear model is fitted and the coefficients are reported. Significant coefficients (p-value below 0.05) are marked with (*).	82
4.1	Overview of the datasets used in this study. Column <i>users</i> shows the number of participants used in the study after filtering and the total number of available users. The columns <i>tracking days</i> and <i>visited locations</i> show the average and standard deviation over users.	97
4.2	User group analysis with respect to demographic and mobility characteristics from study questionnaire. The mean values are given and compared to the other groups in a Mann-Whitney U test for continuous variables or a Chi-Squared test for categorical variables. Significant differences are marked bold, and PT denotes public transport. Note that all fields are self-reported in a questionnaire and not measured in the tracking study.	105
4.3	EMD comparison of different models.	116

5.1	Results of the regression analysis examining the correlations between total energy demand for mobility resp. peak PV power and PV coverage ratio. . . .	140
5.2	Hyperparameter space explored during graph network training. H=Hidden layer; K and K mix=number of terms for the Chebyshev polynomial from (Deferrard et al., 2016) for the convolutional layers and the last mixing layer; sc=usage of optional skip connection (blue in Figure 5.15); bn=usage of batch-norm; dropout=used dropout probability. Hyperparameters of best model are shown in bold.	153
5.3	Number of parameters and the resulting mean squared error for all models in the different cities. The content of the table is visualized in Figure 5.16. . . .	154
A.1	Effect of the θ_{minf} threshold, defining the minimum number of significant features to merge user groups. Only when θ_{minf} is set to a high value, the merging process is affected, i.e. clusters cannot be merged due to a small number of significant features.	171
A.2	Dependence of graph-based mobility profiles on tracking period and coverage. The coefficients with p-values in parentheses of a multinomial logistic regression model are shown. Significant coefficients are marked bold. . . .	173
A.3	Dependence of mobility user groups, derived from basic features, on tracking coverage and duration. The coefficients with p-values in brackets of a multinomial logistic regression model are shown. Significant coefficients are marked bold.	173

Acronyms

- AIC** Akaike information criterion. 60, 65
- BEV** battery electric vehicle. 3, 12, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 133, 135, 136, 138, 141, 142, 143, 144, 157, 160, 162, 164, 204
- CDR** call detail record. 44, 67, 88, 90
- CG** control group. 15, 16, 17, 97, 98, 157, 159, 201
- CNN** convolutional neural network. 160
- DSM** digital surface model. 127, 130, 131
- EIM** Empirical use and Impact analysis of MaaS. 9, 15, 16, 17, 157, 201
- EU** European Union. 2, 121, 123
- GC1** Green Class e-car. 9, 10, 11, 12, 13, 14, 157, 160, 162, 201
- GC2** Green Class e-bike. 9, 10, 11, 12, 13, 157, 201
- GCN** graph convolutional neural network. 112, 114, 116, 159, 160, 161
- GHG** greenhouse gas. 2, 3, 86, 122, 123, 135, 136, 143, 157, 162
- GNSS** global navigation satellite system. 44, 53, 67, 90, 97, 98, 99, 101, 109, 158
- GPS** global positioning system. 98
- ICEV** internal combustion engine vehicles. 3, 157, 162
- ICT** Information and Communication Technologies. 4
- KL** Kullback–Leibler. 65, 66, 68
- LBSN** location based social network. 67, 101, 109
- MaaS** Mobility as a Service. 3, 5, 9, 12, 15, 18, 86, 97, 98, 102, 103, 109, 157, 159, 162, 203
- MC** Markov chain. 63, 65, 66, 67, 68
- MTMC** Mobility and Transport Microcensus. 10, 11, 12

PV photovoltaic. 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 160, 162, 204, 208

SBB Swiss Federal Railways. 9, 10, 11, 15, 17, 201

SDG Sustainable Development Goal. 2

SoC state of charge. 122, 127, 128, 129, 130, 133, 135, 204

TG treatment group. 15, 16, 17, 97, 98, 157, 159, 201

Declaration

I, Henry Martin, declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute.

Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given in the bibliography.

Zürich, 2023



Henry Martin

