TRAFFIC SPEED MODELLING TO IMPROVE TRAVEL TIME ESTIMATION IN OPENROUTESERVICE

C. Ludwig [1,2]*, J. Psotta [2], A. Buch [1], N. Kolaxidis [2], S. Fendrich [2], M. Zia [2], J. Fürle [1], A. Rousell [2], A. Zipf [1,2]

[1] GIScience Research Group, Heidelberg University, Heidelberg, Germany
(ludwig,fuerle,buch,zipf)@uni-heidelberg.de
[2] HeiGIT gGmbH (Heidelberg Institute for Geoinformation Technology), Heidelberg, Germany
(psotta,kolaxidis,fendrich,zia,rousell)@heigit.org

KEY WORDS: Routing, Traffic speed, OSM, Twitter, Centrality.

ABSTRACT:

Time-dependent traffic speed information at a street level is important for routing services to estimate accurate travel times and to recommend routes which avoid traffic congestion. Still, most open-source routing machines that use OpenStreetMap (OSM) as the primary data source rely on static driving speeds derived from OSM tags, since comprehensive traffic speed data is not openly available. In this study, a method was developed to model traffic speed by hour of day at a street level using open data from OpenStreetMap, Twitter and population data. The modelled traffic speed data was subsequently integrated into the open-source routing engine openrouteservice to improve travel time estimation in route planning. Machine learning models were trained for ten cities worldwide using traffic speed data from Uber Movement as reference data. Different indicators based on geolocation and timestamp of Twitter data as well as a geographically adapted betweeness centrality indicator were evaluated for their potential to improve prediction accuracy. In all cities, the Twitter indicators improved the model, although this effect was only visible for certain road types. The centrality indicator improved the model as well but to a lesser extent. The Google Routing API was used as reference to evaluate the accuracy in travel time estimation. Deviations in travel times were regionally different and were partly alleviated by including the raw traffic data by Uber or the modelled traffic speed data in openrouteservice.

1. INTRODUCTION

Time-dependent traffic speed information at a street level is important for route planning to accurately estimate travel times and to recommend routes which avoid traffic congestion. However, most open-source routing engines that rely on OpenStreetMap (OSM) as their primary data source do not extensively support the integration of real-time or historic traffic speed data. Some engines offer prototypical implementations, but driving speeds are primarily estimated based on the OSM tags assigned to road features. This approach is also followed by openrouteservice, an open-source routing engine based on OSM data (openrouteservice, 2023).

One reason for the limited development in this area is that comprehensive global traffic speed data is only available from commercial providers such as Google or Here. Some cities publish traffic related data within their metropolitan area (e.g. GraphHopper Open Traffic Collection (2023)), but combining multiple of these data sets to increase coverage is not feasible, since they are all structured very differently which makes data fusion hard to automatise and therefore labour-intensive. Furthermore, municipal data sets are often not based on the OSM street network, requiring extensive map matching procedures to transfer traffic speed information to the corresponding OSM road features.

Currently, the most promising open dataset suitable for usage in open-source routing engines is provided by Uber Movement. It contains hourly traffic speed data based on the OSM road network for 51 cities worldwide (Uber Technologies Inc., 2023). However, this data is limited to the time period from 2015 to 2020 and covers only a fraction of the roads within these cities. To address this gap, several studies have proposed methods for modelling traffic speed using various open data sources. Many of these studies utilized machine learning techniques with different indicators such as OSM tags (e.g., highway=*), points-of-interest (Camargo et al., 2020), centrality indicators (Zhao et al., 2017), or social media data (Pandhare and Shah, 2017). These indicators have demonstrated their suitability for modelling traffic flow, but none of these studies have specifically evaluated the impact of using modelled traffic speed data on travel time estimation in route planning.

The aim of this study is to model historic traffic speed at a street level and by the hour of the day based on open and globally available data sources. Additionally, it seeks to evaluate the potential of using this data to improve travel time estimation in openrouteservice (Figure 1). The study specifically focuses on assessing the benefits of incorporating geolocation and timestamp information from Twitter data as well as a geographically adapted betweenness centrality indicator into traffic speed modeling using machine learning models in ten cities worldwide. More specifically, the study addresses the following research questions:

1. How can geolocation and timestmap information from Twitter data and the geographically adapted betweenness centrality indicator improve the traffic speed model?
2. What impact does the incorporation of traffic speed data have on the accuracy of travel time estimation in openrouteservice?

The source code and data used in this study are available at https://zenodo.org/record/7857038.ZEWubHbP0qw.

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-4-W7-2023-109-2023 | © Author(s) 2023. CC BY 4.0 License.

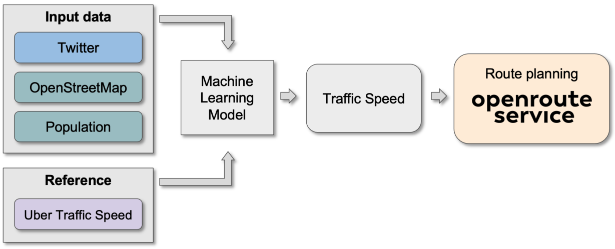

Figure 1. Conceptual workflow

2. RELATED WORK

At the time of writing, most open-source routing engines offer only experimental implementations for integrating time-dependent traffic speed data into route planning. The routing engine Valhalla has a proof-of-concept implementation for traffic-influenced routing that supports real-time traffic data (Valhalla, 2023). However, only a single traffic speed value can be provided for each OSM road feature, which might be an issue especially for long features as the actual traffic speed may vary along their length. The Open Source Routing Machine (OSRM) also has an experimental integration of traffic data which allows for a single traffic speed value to be provided for sections of OSM road features in both driving directions (Open Source Routing Machine, 2023). Graphhopper published an experimental implementation of traffic data in 2015, which relied on real-time traffic speed data from the City of Cologne (Graphhopper, 2019). It is however no longer maintained.

In the past, various methods have been developed to model traffic speed information using different kinds of openly available data sources. Studies that utilize social media data to model traffic speed or traffic volume primarily focused on analyzing the text content of the posts (Wang et al., 2018; Das and Purves, 2019; Pereira et al., 2021). Alternatively, Fan and Stewart (2021) and Liao et al. (2022) utilized only the geolocation of the tweets, without considering their content, to model regional mobility behavior. Traffic delays have also been estimated using the frequency of points-of-interest (POIs) from OSM (Camargo et al., 2020; Hu and Jin, 2017; Ge et al., 2020). Recently, network indicators such as betweenness centrality have gained popularity for modeling traffic flows in transportation networks (Gao et al., 2013; Pedreira Junior et al., 2021; Zhao et al., 2017; Jayasinghe and Sano, 2017).

In most studies, machine learning models, such as decision tree-based models, have been utilized for modeling traffic speed (Pandhare and Shah, 2017; Pedreira Junior et al., 2021; Integrating Household Travel Survey and Social Media Data to Improve the Quality of OD Matrix: A Comparative Case Study, n.d.). Recent studies have focused on evaluating the potential of deep learning in modeling traffic speed (Ge et al., 2020; Fang et al., 2019; Ren et al., 2022; Cui et al., 2020). However, many of these studies were restricted to small study areas due to the high computational effort to train these models. Nevertheless, some studies have explored the scalability of these models to larger regions Derrow-Pinion et al. (2021); Fang et al. (2020).

Previous studies have predominantly utilized datasets from official agencies for training and evaluating their models focusing on single cities only. The applicability of these models to multiple cities has rarely been investigated, partly due to the challenge of obtaining suitable reference data sets from different regions which share the same data format (Camargo et al., 2020).

3. DATA

OpenStreetMap ’OSM is a global digital map of the world that contains information about roads, land use, and points of interest (POI). As a community project similar to Wikipedia, the OSM data is primarily created and maintained by volunteer members.

Objects in OSM can be represented as nodes (points), ways (lines, polygons), or relations (groups of nodes or ways). The properties of these objects can be described using tags, which consist of a key and a value. For example, a bench can be mapped as a point geometry with the tag amenity=bench. Roads in OSM are labeled with the key highway=*, with the value depending on the type of road. For instance, highway=motorway st[tyrepresents a motorway, while highway=residential represents a residential street. An OSM object can have multiple tags assigned to it to describe its properties. OSM users have the flexibility to create and assign tags according to their needs. To maintain semantic consistency in the data, the OSM wiki provides guidelines for tag usage. Additionally, OSM members can propose new tags in the OSM forum. After discussion within the OSM community, it is decided whether the tag will be added to the wiki.

In this study, OSM data was utilized for traffic speed modeling to calculate the centrality indicator (see section 4.4) and for route planning in openrouteservice (see section 5). The OSM data used for all cities corresponds to March 31, 2020, aligning with the time period of the Twitter and Uber traffic speed data. The OSM data was downloaded from Geofabrik (2023).’

Twitter Twitter is a social media platform where users can publicly post text messages, called tweets. Optionally, the current geolocation at the time of sending the tweet is stored as well, provided that the user has activated this feature. Twitter data can be downloaded free of charge using the public Twitter API, but only a fraction of all tweets is available for download. In this study, roughly 10 million tweets were downloaded for the ten cities investigated, spanning the period from January 2018 to March 2020 (Table 1).

City	Twitter Users	Tweets	Population
Barcelona	44,758	542,076	5,687,356
Berlin	26,070	418,932	3,573,938
Cincinnati	11,445	157,754	296,943
Kyiv	5,398	73,046	3,016,789
London	151,509	1,543,018	9,648,110
Madrid	58,505	552,925	6,751,374
Nairobi	12,750	130,681	5,325,160
New York City	198,144	3,981,137	7,888,121
Sao Paulo	89,599	1,263,890	22,619,736
Seattle	34,694	518,950	737,015

Table 1. Number of Twitter users and number of Tweets (January 2018 - March 2020) and population in 2023 provided by World Population Review (2023) for all cities.

Population data The population distribution was considered in the calculation of the geographically adapted betweenness centrality of the road network (see section 4.4). To accomplish this, the study utilized the Global Human Settlement Population Layer dataset, provided free of charge by the Joint Research Center (JRC). This dataset offers global coverage at a resolution of 250 meters (Schiavina et al., 2022).

Uber Movement traffic speed data Uber Movement is a platform that offers open data on traffic flow for 51 cities worldwide (Uber Technologies Inc., 2023). The traffic speed data is derived based on the OSM road network using user data from the Uber app. Hourly mean, median, and 85th percentile traffic speeds are provided for road segments with at least 5 valid measurements by Uber users. The data is publicly available from January 2015 to March 2020. In this study, the quarterly speed statistics by hour of the day for the first quarter of 2020 were used.

4. TRAFFIC SPEED MODEL

4.1 Assumptions

The model is based on the fundamental assumption that publicly available geocoded data from social media platforms can serve as indicators of spatio-temporal human mobility in the real world. The presence of a higher number of social media messages at a specific location and time suggests a larger crowd of people in that area. Consequently, it implies increased traffic flow in the nearby road network, as individuals would have traveled to that location using various means of transportation, such as cars or public transport.

The approach presented in this study does not rely on the content of social media posts or personal user information. Instead, it solely utilizes the geolocation and timestamp of Twitter messages to infer traffic conditions. It does not assume that individuals create social media messages while driving a car, and therefore, it does not employ a telemetric approach to estimate user speed based on the spatio-temporal difference between two messages.

While this study tests the approach using Twitter data, it is worth noting that any datasets containing geolocation and timestamp information, such as those from other social media platforms or mobile phone data, could potentially be used as well. This expands the model's potential for application in other regions and reduces its dependence on a single data source.

4.2 Model setup

Traffic speed prediction is performed for individual street segments, considering the hour of the day, using supervised machine learning techniques across ten cities worldwide. These cities include Berlin (Germany), London (UK), Barcelona (Spain), Madrid (Spain), Nairobi (Kenya), Kyiv (Ukraine), Sao Paulo (Brazil), Seattle (USA), San Francisco (USA), New York City (USA), and Cincinnati (USA).

The street segments are derived from the OpenStreetMap (OSM) road network, which is transformed into a directed graph using the Python package omsnx (Boeing, 2017). For each street segment, various indicators are calculated and utilized as predictors for traffic speed in the model. These indicators encompass OSM tags such as "highway" and "max_speed," a geographically adapted betweenness centrality indicator, and several Twitter indicators. Twitter indicators. The input variables are standardized based on the training data prior to model training. To avoid model leakage the standardization is performed separately for training and testing data. Missing values in the max_speed tag are filled using the mean max_speed value for the respective highway tag. The highway tags are encoded to a numeric representation using a One-Hot-Encoding.

Separate models are trained for each city using the gradient boosting method implemented in the Python package XGboost (Chen and Guestrin, 2016). Different combinations of features are utilized to assess their impact on model performance. Uber data provides information on the mean, median, and 85th percentile of traffic speed. Hence, three different models with different target variables are trained for each city to evaluate the influence of these metrics on the accuracy of travel time estimation (see section 6.2).

To train the models, 1000 samples for each highway tag are randomly selected for each city. If there are fewer than 1000 samples available for a specific highway tag, all available features are used as samples. The training data comprises 70% of the samples, while the remaining samples are used for model evaluation. The quality of the different models is assessed using the coefficient of determination (R^2) and the root mean square error (RMSE).

4.3 Twitter indicators

Using the Twitter data, eight different indicators are computed for each street segment. To determine the optimal spatial aggregation of Twitter messages, the number of tweets in the vicinity of each street segment is calculated using four different buffer distances: 50 meters, 100 meters, 250 meters, and 500 meters. In assessing the temporal aggregation, the Twitter messages within each buffer are aggregated both by the hour of the day and by the total tweet count, disregarding the specific hour of the day.

4.4 Geographically adapted betweenness centrality

Betweenness centrality is an indicator used to identify significant nodes or edges in a network. It is computed by generating the shortest paths between all possible pairs of nodes in the graph and subsequently calculating the number of times each node or edge has been traversed. Although initially developed for social network analysis, betweenness centrality can be applied to other graph-like structures such as road networks, where it reveals the relative importance of road segments within the network.

Calculating the betweenness centrality of road networks requires considering their geographical context. People tend to choose faster routes rather than shorter ones, so road type becomes an important consideration. Additionally, human mobility is often purpose-driven, resulting in varying levels of travel to different locations. Taking these factors into account increases the complexity and computational effort, especially for large road networks, necessitating the calculation of a randomized sample of routes due to resource constraints.

Therefore, a geographically adapted version of the betweenness centrality indicator was calculated instead of using the original form. In each city, 20,000 car trips were generated using the openrouteservice car profile. For each trip, a random starting point, weighted by the population distribution, and a randomly selected destination point represented by a POI, were chosen. POIs were obtained from OSM based on predefined OSM tags, categorized into work, education, shopping, or leisure. To simulate more realistic trips, a distance decay function was used, favoring POIs closer to the starting point, as described in Gao et al. (2013). The generated routes were then matched to the corresponding OSM road segments using fast map matching (Yang and Gidófalvi, 2018), and aggregated to derive a centrality score for each road segment.

5. TRAFFIC SPEED INTEGRATION IN OPENROUTESERVICE

Openrouteservice is a freely available, open-source routing service that based on OSM data. The provided services include route planning, distance and travel time matrices and isolines for different profiles such as pedestrian, bike, car or wheelchair as well as geocoding and map-matching functions. Currently, openrouteservice uses a heavily modified fork of Graphhopper for routing.

The original version of the openrouteservice routing engine does not utilize traffic speed data as input. Instead, it assumes static driving speeds based on factors such as the highway class, road surface, and country-specific rules. Additionally, an acceleration heuristic is implemented to account for reduced speeds at junctions.

In this study, a traffic speed integration was implemented in openrouteservice to consider historic traffic speed data during route planning. It allows the provision of traffic speed data for each hour of the day, irrespective of the date or day of the week. This information is parsed and stored by openrouteservice as supplementary data for each edge in the routing graph.

The traffic speed data should be provided in a CSV file format. Each row in the dataset represents the traffic speed value for a specific hour of the day for an OSM street segment, identified by the OSM way ID it belongs to as well as the start and end node IDs. This structure enables the specification of driving speeds for both directions throughout a typical day. The data format aligns with the quarterly traffic speed datasets offered by Uber Movement, enabling their direct usage with openrouteservice without requiring map matching or other preprocessing steps.

To evaluate the potential improvement in travel time estimation by incorporating traffic speed data, comparisons were made between the travel times estimated by openrouteservice and those estimated by the Google Routing API. Google utilizes data from smartphones using its services to derive real-time and historic traffic speed information. Since the traffic speed model in this study only provides historic traffic speed, routes from Google were requested for a date several weeks in the future to ensure the utilization of historic data rather than real-time data. For each city, 50 random routes for each hour of the day were requested from the Google Routing API for June 23rd, 2023. To increase the number of routes for comparison, alternative routes were also included.

For each Google route, five routes were generated using openrouteservice with different traffic data sets: openrouteservice without traffic speed data, with modeled mean traffic speed, with modeled median traffic speed, with modeled 85th percentile traffic speed, and with raw 85th percentile traffic speed data from Uber. To replicate the Google routes as closely as possible, 10 waypoints distributed along the route were passed to openrouteservice. However, deviations between the routes generated by Google and openrouteservice still occurred. To exclude these form the analysis, routes with a deviation in geometry larger than 3% and differences in distance greater than 1% were excluded from the analysis.

This analysis was conducted for three cities located in different geographic regions and with varying spatio-temporal densities of Twitter data relative to population numbers: Nairobi (low density), Berlin (medium density), and Seattle (high density) (Table 1).

6. RESULTS

6.1 Traffic speed model

The variation in model performance based on different model features exhibits a consistent pattern across all cities (Figure 2). The models utilizing only the OSM tags ”highway” and ”max_speed” demonstrate the lowest accuracy. The inclusion of the hour of the day as an additional indicator does not considerably enhance model performance; in some cities, it even shows a slight decline. On the other hand, substantial improvements are observed when incorporating the total number of Tweets within a 250-meter distance of streets into the model. However, this improvement is not evident when aggregating the Tweets by hour of the day. In cities with limited Twitter data, such as Berlin or Nairobi, the addition of the betweenness centrality indicator further enhances the model. However, for cities with a substantial amount of Twitter data, such as New York City or London, the improvement is relatively small.

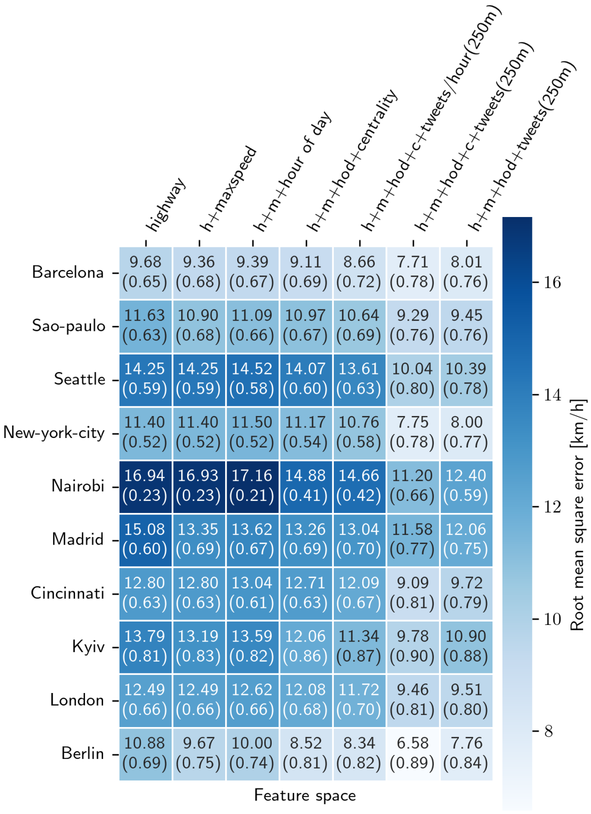

Figure 2. RMSE and R^2 (in brackets) of models predicting the 85th percentile of traffic speed using different features.

Several Twitter indicators based on different spatio-temporal aggregation methods were evaluated (Figure 3). Again, the same pattern could be observed across all cities. Regarding spatial aggregation, model performance considerably improved with increasing distance from streets. For example, in Seattle, considering Tweets up to 500 meters away from the street reduces the RMSE by 5.13 km/h, while considering Tweets within a distance of 100 meters only reduces RMSE by 0.31 km/h. As for temporal aggregation, aggregating Tweets by the hour of the day instead of taking the total Tweet count within the vicinity of the street does not significantly improve the models. Due to memory issues, it was not possible to aggregate Twitter messages using a 500-meter buffer in cities with very large Twitter data (New York City and London). These limitations can be overcome by improving the implementations's efficiency, but it also highlights the high computational effort required to compute this indicator for all street segments.

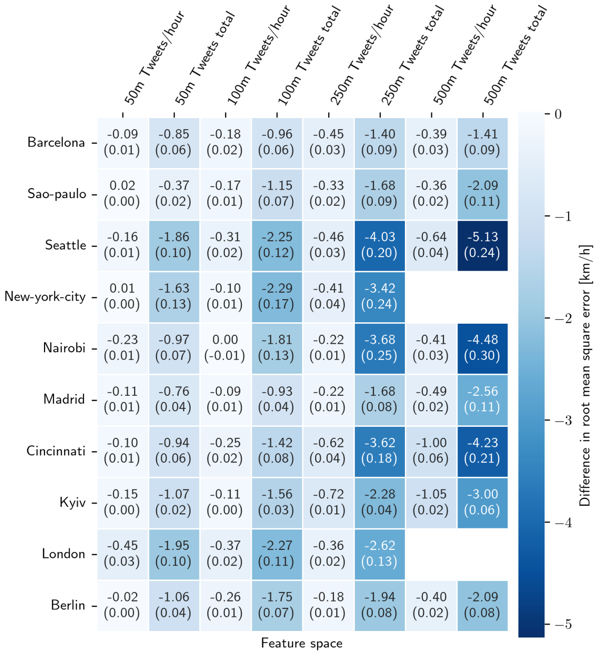

Figure 3. Difference in RMSE and R^2 (in brackets) of models using different Twitter indicators compared to the model h+m+hod+centrality in Figure 2. Twitter indicators differ depending on the buffer sizes and temporal aggregation.

The prediction of traffic speed improved to varying degrees depending on the road type. For traffic-calmed sectors (highway=living_street), construction sites (highway=construction), as well as motorways and links, model performance improved when considering the total Tweet count within the vicinity of the street. In Berlin, for example, the RMSE for construction sites decreased from 7.78 km/h to 3.79 km/h (Figure 4). However, for small to medium-sized roads such as residential streets, no improvement was observed by incorporating Twitter or centrality indicators (Figure A.1 in the appendix).

Regarding feature importance, the influence of the features on the model prediction is generally well-balanced (Figure 5). However, for roads with high traffic speeds such as motorways, the OSM tag max_speed=* has a significantly higher influence on the model prediction than the other indicators.

6.2 Evaluation of travel time estimation

To assess the impact of different traffic speed data sets on travel time estimation in openrouteservice, routes were generated using the Google Routing API for each city: Berlin (3097 routes), Nairobi (2744 routes), and Cincinnati (2796 routes). The variation in the number of routes is due to the different numbers of alternative routes provided for each request. Only a subset of

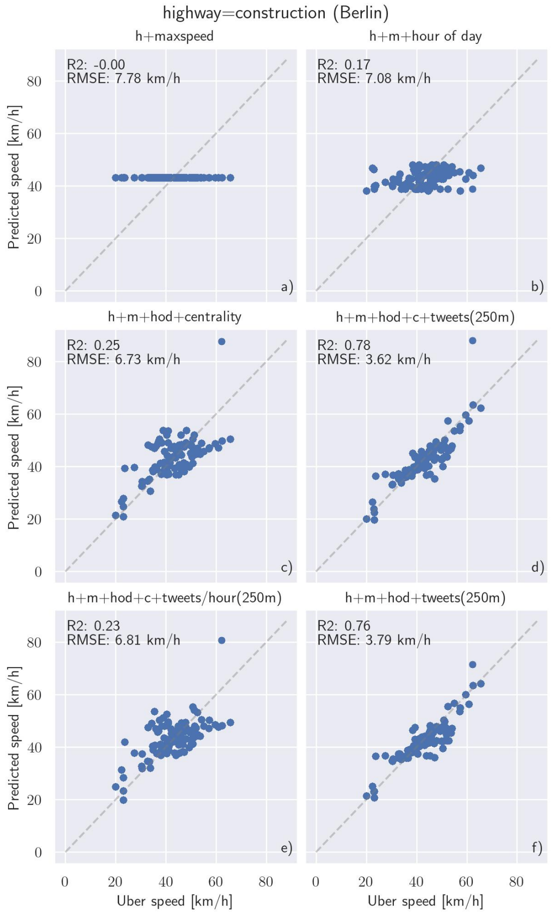

Figure 4. Residuals of OSM road features with tag highway=construction in Berlin for different models. Twitter and centrality indicators improve the model considerably.

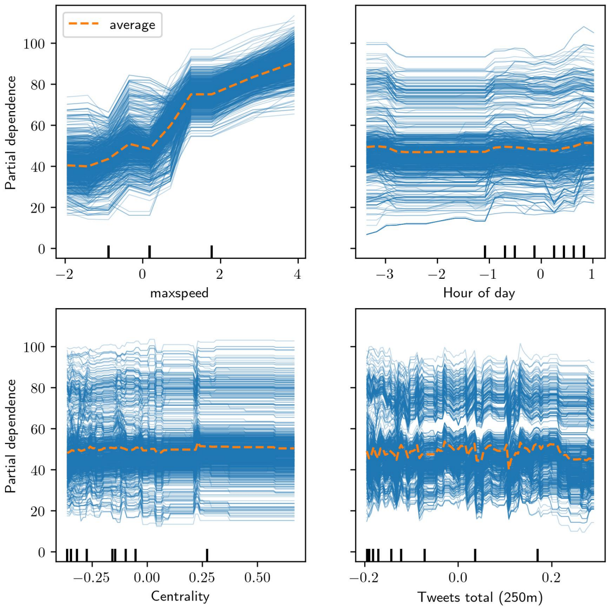

Figure 5. Feature importance represented using a individual conditional expectation (ICE) plot for the model predicting the 85th percentile of traffic speed in Barcelona.

these routes could be accurately replicated using all five openrouteservice instances: 210 in Berlin, 489 in Cincinnati, and 43 in Nairobi.

In Berlin, incorporating traffic speed data in openrouteservice did not significantly improve the accuracy of estimated travel times (Figure 6). Without traffic data, travel times were overestimated on average by 7.77%, whereas with modelled traffic speed data, travel times were generally underestimated, particularly when using the 85th percentile (-35.99%). The best results were achieved when using the modelled median traffic speed (-8.89%). The utilization of raw traffic speed data provided by Uber, available for approximately 40% of OSM street segments, did not considerably enhance travel time estimation.

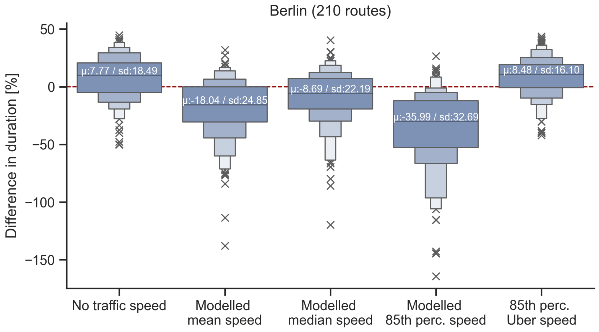

Figure 6. Differences in estimated travel time between Google Routing API and openrouteservice with different traffic speed data sets for Berlin. Positive values mean that estimated travel time of openrouteservice is longer than Google.

In Cincinnati, openrouteservice without traffic speed data overestimated travel times on average by 24.74% (Figure 7). These discrepancies were slightly reduced to 22.49% when incorporating the modelled 85th percentile traffic speed data. The utilization of raw Uber traffic speed data, which is accessible for 17% of street segments, further narrows the gap in travel times to 19.23%.

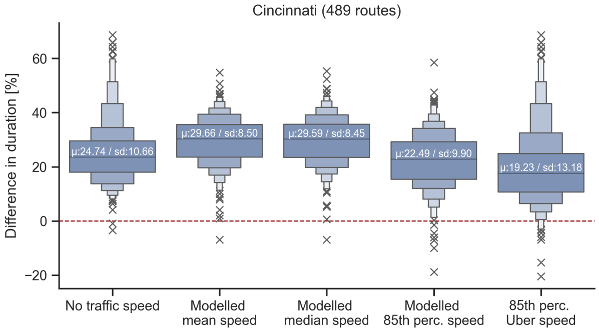

Figure 7. Differences in estimated travel time between Google Routing API and openrouteservice with different traffic speed data sets for Cincinnati. Positive values mean that estimated travel time of openrouteservice is longer than Google.

In contrast to Berlin and Cincinnati, openrouteservice without traffic speed data underestimated travel times in Nairobi on average by 41.53% (Figure 8). However, incorporating the modelled or raw traffic speed data significantly improved the estimation of travel times. The most accurate results were obtained when using the modelled 85th percentile traffic speed data, reducing the deviation in travel times to just 1.42%.

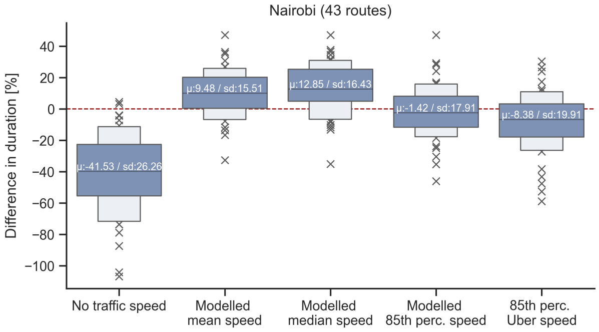

Figure 8. Differences in estimated travel time between Google Routing API and openrouteservice with different traffic speed data sets for Nairobi. Positive values mean that estimated travel time of openrouteservice is longer than Google.

7. DISCUSSION

The influence of Twitter indicators on the accuracy of the traffic speed model was similar in all ten cities. Using the total count of Twitter messages in the vicinity of a street segment led to higher model accuracy than using hourly aggregation. This could be due to the fact that the density of Twitter messages is not high enough to extract both a spatial and a temporal signal. It might also be the reason why considering Tweets up to 500 meters away from the streets yielded the best model results. This means that aggregating the data at a street segment level is not necessary, and that it can also be done on a smaller neighborhood scale, for which a less computationally intensive method can be found. In this way, Twitter indicators with buffers of 500 meters and above could also be calculated for larger cities such as New York City and London, which previously failed due to memory issues. The geographically adapted betweenness centrality indicator also improved the models, especially in cities with a low density of Twitter data. This shows that centrality indicators can complement the Twitter indicators quite well. However, further analysis regarding the adaptability of these indicators to the geospatial domain should be conducted to better integrate them into traffic speed modeling.

In this study, Twitter data was used as an indicator to predict traffic speed, but technically, other datasets containing geolocation and timestamp information, such as other social media platforms or mobile phone data, could be integrated as well with minimal effort. Therefore, future studies should investigate the potential to replace or enrich Twitter data to further improve model performance. The main focus of this study was to investigate the potential of different open datasets for traffic modeling rather than optimizing the model itself. In future studies, the potential of the investigated indicators in deep learning models should be explored. Additionally, the transferability of the models to cities without Uber data is another important aspect that needs to be addressed.

The comparison of travel times between Google and openrouteservice yielded different patterns in all cities. In Berlin and Cincinnati, the original openrouteservice without traffic data overestimated travel times, while in Nairobi, it was strongly underestimated. This could be attributed to a lower level of attribute completeness of OSM highway features in Nairobi or may indicate the need for more country-specific adaptations in the heuristics used for travel time estimation in openrouteservice.

Using the raw or modeled traffic speed data in openrouteservice corrected this bias in Nairobi, demonstrating that traffic speed information could be useful in alleviating regional differences in accuracy. In some cases, the raw Uber traffic speed data improved travel time estimation more than the modeled traffic speed. This observation might indicate that the traffic speed models are not accurately predicting traffic speed in exceptionally busy streets. Further investigation is needed to better understand and improve the models.

Generally, the approach to compare travel times between two different routing engines needs improvement, as it was only possible to reproduce a fraction of Google routes with the required accuracy. Therefore, the analysis should be repeated using a better method and a larger number of routes. Nevertheless, the results demonstrate that evaluating the performance of a traffic speed model in combination with its usage in a routing engine is worthwhile, since strong improvements in the traffic models might not necessarily lead to significant improvements in travel time estimation. For example, in Berlin, the accuracy of the traffic models was the highest, but it did not result in a significant improvement in travel time estimation.

Technically, integrating the Uber traffic speed data in openrouteservice was quite easy since no map matching was necessary. However, this data structure resulted in very large CSV files containing the traffic speed data, requiring a significant amount of RAM during graph building. Therefore, the structure should be adapted to store the traffic speed information in a more efficient way. One option to consider is reducing the temporal resolution from one-hour intervals to three-hour intervals, which may help address this issue.

8. CONCLUSION AND OUTLOOK

The results of the traffic speed models have shown that the integration of Twitter data can considerably improve traffic speed models. This is especially true for traffic-calmed sectors and construction sites but not as much for medium-sized roads, such as residential streets. Aggregating the Twitter data on a neighborhood level was more effective than on a street level, and the temporal information contained in the tweets was not relevant for improving the model. The comparison between travel times in Google and openrouteservice showed regional differences in the accuracy of estimated travel times. These differences could be partly alleviated by incorporating raw or modeled traffic speed information. Generally, the method used to evaluate the travel time difference between two routing engines needs improvement to yield more reliable results. Investigating traffic speed models in combination with their usage in routing engines is worthwhile since the results showed that improvements in the models do not necessarily lead to improvements in travel time estimation.

9. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1503-1 FUGG. Authors employed at HeiGIT gGmbH acknowledge support by the Klaus Tschira Stiftung, Germany.

References

Boeing, G., 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65, 126–139. doi.org/10.1016/j.compenvurbsys.2017.05.004.

Camargo, C. Q., Bright, J., McNeill, G., Raman, S., Hale, S. A., 2020. Estimating Traffic Disruption Patterns with Volunteered Geographic Information. *Scientific Reports*, 10(1), 1271. doi.org/10.1038/s41598-020-57882-2. Number: 1 Publisher: Nature Publishing Group.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. doi.org/10.1145/2939672.2939785.

Cui, Z., Ke, R., Pu, Z., Ma, X., Wang, Y., 2020. Learning traffic as a graph: A gated graph wavelet recurrent neural network for network-scale traffic prediction. *Transportation Research Part C: Emerging Technologies*, 115, 102620. doi.org/10.1016/j.trc.2020.102620.

Das, R. D., Purves, R., 2019. Exploring the Potential of Twitter to Understand Traffic Events and Their Locations in Greater Mumbai, India. *IEEE Transactions on Intelligent Transportation Systems*, PP, 1–10. doi.org/10.1109/TITS.2019.2950782.

Derrow-Pinion, A., She, J., Wong, D., Lange, O., Hester, T., Perez, L., Nunkesser, M., Lee, S., Guo, X., Wiltshire, B., Battaglia, P. W., Gupta, V., Li, A., Xu, Z., Sanchez-Gonzalez, A., Li, Y., Velickovic, P., 2021. ETA Prediction with Graph Neural Networks in Google Maps. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, Association for Computing Machinery, New York, NY, USA, 3767–3776. doi.org/10.1145/3459637.3481916.

Fan, J., Stewart, K., 2021. Understanding collective human movement dynamics during large-scale events using big geosocial data analytics. *Computers, Environment and Urban Systems*, 87, 101605. doi.org/10.1016/j.compenvurbsys.2021.101605.

Fang, S., Zhang, Q., Meng, G., Xiang, S., Pan, C., 2019. GST-Net: Global Spatial-Temporal Network for Traffic Flow Prediction. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, Macao, China, 2286–2293. doi.org/10.24963/ijcai.2019/317.

Fang, X., Huang, J., Wang, F., Zeng, L., Liang, H., Wang, H., 2020. ConSTGAT: Contextual Spatial-Temporal Graph Attention Network for Travel Time Estimation at Baidu Maps. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, Association for Computing Machinery, New York, NY, USA, 2697–2705. doi.org/10.1145/3394486.3403320.

Gao, S., Wang, Y., Gao, Y., Liu, Y., 2013. Understanding Urban Traffic-Flow Characteristics: A Rethinking of Betweenness Centrality. *Environment and Planning B: Planning and Design*, 40(1), 135–153. doi.org/10.1068/b38141.

Ge, L., Li, S., Wang, Y., Chang, F., Wu, K., 2020. Global Spatial-Temporal Graph Convolutional Network for Urban Traffic Speed Prediction. *Applied Sciences*, 10(4), 1509. doi.org/10.3390/app10041509. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Geofabrik, 2023. Geofabrik Download Server. https://download.geofabrik.de/.

Graphhopper, 2019. GraphHopper Traffic Data Integration. https://github.com/karussell/graphhopper-traffic-data-integration.

GraphHopper Open Traffic Collection, 2023. https://github.com/graphhopper/open-traffic-collection.

Hu, W., Jin, P. J., 2017. An adaptive hawkes process formulation for estimating time-of-day zonal trip arrivals with location-based social networking check-in data. *Transportation Research Part C: Emerging Technologies*, 79, 136–155. doi.org/10.1016/j.trc.2017.02.002.

Integrating Household Travel Survey and Social Media Data to Improve the Quality of OD Matrix: A Comparative Case Study, n.d. Integrating Household Travel Survey and Social Media Data to Improve the Quality of OD Matrix: A Comparative Case Study. 21. https://doi.org/10.1109/TITS.2019.2958673.

Jayasinghe, A., Sano, K., 2017. Estimation of Annual Average Daily Traffic on Road Segments: Network Centrality-Based Method for Metropolitan Areas. *Transportation Research Record Journal of the Transportation Research Board*, 1–18.

Liao, Y., Yeh, S., Gil, J., 2022. Feasibility of estimating travel demand using geolocations of social media data. *Transportation*, 49(1), 137–161. doi.org/10.1007/s11116-021-10171-x.

Open Source Routing Machine, 2023. Open Source Routing Machine Documentation. https://github.com/Project-OSRM/osrm-backend/wiki/Traffic.

openrouteservice, 2023. Openrouteservice. https://github.com/GIScience/openrouteservice.

Pandhare, K. R., Shah, M. A., 2017. Real time road traffic event detection using Twitter and spark. *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 445–449. doi.org/10.1109/ICICCT.2017.7975237.

Pedreira Junior, J. U., Assirati, L., Pitombo, C. S., 2021. Improving travel pattern analysis with urban morphology features: A panel data study case in a Brazilian university campus. *Case Studies on Transport Policy*, 9(4), 1715–1726. doi.org/10.1016/j.cstp.2021.07.019.

Pereira, A. S., Braga Silva, T. R. M., Silva, F. A., Correia, L. H. A., Loureiro, A. A., 2021. A Workflow to Detect Traffic Events Using Multiple Algorithms and Data Sources. *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 164–170. doi.org/10.1109/DCOSS52077.2021.00038.

Ren, Y., Zhao, D., Luo, D., Ma, H., Duan, P., 2022. Global-Local Temporal Convolutional Network for Traffic Flow Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 23(2), 1578–1584. doi.org/10.1109/TITS.2020.3025076.

Schiavina, M., Freire, S., MacManus, K., 2022. GHS-POP R2022A—GHS population grid multitemporal (1975–2030). *European Commission, Joint Research Centre*. doi.org/10.2905/2FF68A52-5B5B-4A22-8F40-C41DA8332CFE.

Uber Technologies Inc., 2023. Uber Movement: Speeds. https://movement.uber.com.

Valhalla, 2023. Valhalla: Traffic Influenced Routing - Proof of Concept. https://valhalla.github.io/valhalla/thor/simple_traffic/.

Wang, S., Yu, D., Ma, X., Xing, X., 2018. Analyzing urban traffic demand distribution and the correlation between traffic flow and the built environment based on detector data and POIs. *European Transport Research Review*, 10(2), 50. doi.org/10.1186/s12544-018-0325-5.

World Population Review, 2023. 2023 World Population by Country (Live). https://worldpopulationreview.com/.

Yang, C., Gidófalvi, G., 2018. Fast map matching, an algorithm integrating hidden Markov model with precomputation. *International Journal of Geographical Information Science*, 32(3), 547–570. doi.org/10.1080/13658816.2017.1400548. Publisher: Taylor & Francis.

Zhao, S., Zhao, P., Cui, Y., 2017. A network centrality measure framework for analyzing urban traffic flow: A case study of Wuhan, China. *Physica A: Statistical Mechanics and its Applications*, 478, 143–157. doi.org/10.1016/j.physa.2017.02.069.

A. APPENDIX

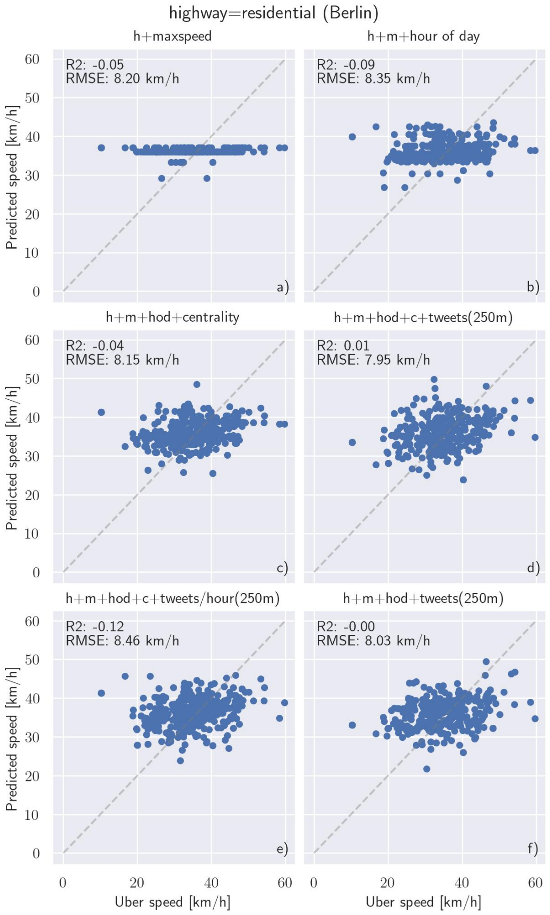

Figure A.1. Residuals of OSM road features with tag highway=residential in Berlin for different models. Twitter and centrality indicators do not improve the model.