Spatiotemporal Clustering in Urban Transportation - a Bus Route Case Study in Washington D.C.

Xiqi Fei, Olga Gkountouna
George Mason University, USA
{xfei, ogkounto}@gmu.edu

Abstract

Public buses are an important part of the urban transportation mix. However, a considerable disadvantage of buses is their slow speed, which is in part due to frequent stops, but also due to the lack of segregation from other vehicles in traffic. As such, assessing bus routes and the respective sections that are prone to congestion is an important aspect of route planning, scheduling, and the creation of dedicated bus lanes. In this work we use bus tracking data from the Washington Metropolitan Area Transit Authority, to discover speed patterns of specific bus routes in relation to the road network throughout the day. Specifically, we focus on using these patterns to identify free flow segments, bus stop locations, traffic light locations and road segments prone to congestion.

1 Introduction

Buses, like other forms of public transportation, provide an essential service to users that depend on this service to commute to and from work, and to other places. Such services are especially important in large cities, where increasing vehicular traffic flows continues to a be a major challenge for urban planners, who must content with associated road congestion in cities. However, buses face several challenges, one of which is having notoriously slow speeds (as low as 17mph for some bus routes in DC), thus resulting in longer commute times for passengers. Besides frequent stops, which are prescribed for this means of transportation, the speed is also impacted by the lack of segregation from other vehicular traffic. As such, the assessment of traffic conditions along bus routes forms an integral part of route planning, scheduling, and the creation of dedicated bus lanes in cities.

In our study, we discretize a bus route and calculate the average speed using odometer and time stamp values to discover patterns of slowdowns throughout a 24-hour period. This slowdown pattern may be persistent throughout the day, random, or may appear at specific times of the day. Each of these cases are due to different causes. Random slowdowns throughout the day are indicative of traffic lights. More persistent slowdowns are indicative of bus stops, while time-dependent slowdowns are more likely related to traffic congestion. From a public transportation planning perspective, route segments prone to traffic congestion would be prime candidates for dedicated bus lanes.

We performed our analysis on real (Metrobus) data from WMATA [20], the public transport authority of the Washington DC area. We cluster all road segments along a bus route, using features derived from the bus speeds at each segment, and sampled at hourly intervals. Our results reveal different categories of road segments which can be associated with free-flow of traffic, and different types of slow-downs.

The remainder of this paper is structured as follows. Section 2 presents a brief survey of the related work on bus data. Section 3 includes an overview of the main challenges we faced in working with this type of data. In Section 4, we present our method, experimental setup and an evaluation of our Washington D.C. Metrobus case study. Finally, Section 5 concludes the paper.

2 Related Work

Bus data has received considerable attention for the estimation of traffic conditions. Floating car data (FCD) or probe vehicle data (PVD) refers to the use of data generated by one vehicle as a sample to assess to overall traffic conditions (cork swimming in the river). PVD from automobiles have previously been studied to estimate travel times and traffic conditions [15, 10, 2, 18, 22, 21] and traffic speed [9]. Specifically, as it relates to bus data, the focus of our work, a number of works [12, 16, 17, 3, 1] have used this data to study travel times and related traffic flows in urban areas. It was shown in [3] that the difference between travel times of a bus and that of an car was relatively stable, and that buses with automated vehicle locators (AVL) can be used as a probes to collect travel time data at regular intervals with minimum cost. AVL bus data is used for characterizing the performance of arterial roads in Oregon [1]. [12] examine real-time sensitivity between buses and cars to study the feasibility of a real bus probe application in an urban traffic environment. [17] predicted travel times under heterogeneous traffic conditions by applying a Kalman filtering technique to GPS data collected from buses.Further, [16] use bus probe data to evaluate the travel time variability and the level of service of roads. Kumar et al. [8] developed a bus arrival time prediction system, considering both spatial and temporal variations of travel times. In [7] a simulation technique was used to study the influence of these stops on traffic flow under heterogeneous traffic conditions.

The location optimization of bus stops has also been the focus of several works. Saka [14] developed a model for determining optimum bus-stop spacing in urban areas, with the aim of decreasing travel time, headway, and the fleet size. Chien et al. [5] focus on optimizing bus routes in areas with a commuter (many-to-one) travel pattern. [4] address the problem of optimizing the placement of bus stop locations, with the goal to improve the accessibility of a bus service. [13] used a GIS-based methodology to identify hazardous bus stop locations prone to auto-pedestrian collisions. [6] developed a spatial interaction coverage model for identifying bus stop redundancy in order to optimize transit planning. A work more relevant to ours [11] proposes a methodology to de-noise GPS AVL data, identify bus stops, and detect time schedule information.

While [11] clusters all the bus recordings along a route to form groups, with each group corresponding to one stop, our approach aims to discover the different categories of segments within the bus route. Ideally, all stops should appear in one cluster, the traffic lights in another, etc.

3 Challenges

While most existing approaches are based on GPS data, in our study we use odometer readings recorded using a bus AVL system. The data comprises of bus trips for different routes collected during a 24-hour period. Each trip consists of a time series of odometer readings from a specific bus, with sampling interval varying from 1 to 10 seconds.

3.1 Rate of recordings

The rate at which the location and time stamp information were recorded is not constant. It varies for different buses, as well as for the same bus over time. The time delay between any two consecutive measurements varies from one to several seconds. To overcome this inconsistency, we discretized time into constant time intervals, and calculated the average bus speed over those specific periods.

3.2 Odometer alignment

Even though they are more reliable than GPS, still the odometer readings between any two different bus trips are not perfectly aligned. Specifically, two buses that follow the same path may be at different locations after 1,000 odometer-measured feet. One reason for this misalignment is the choice of lane, especially when turning

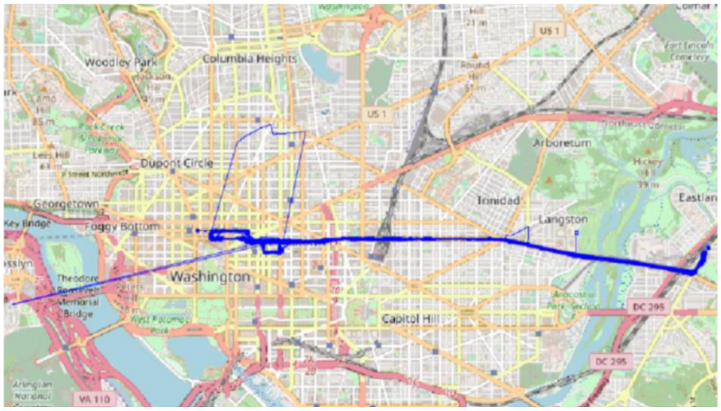


Figure 1: Bus trips of X2 route.

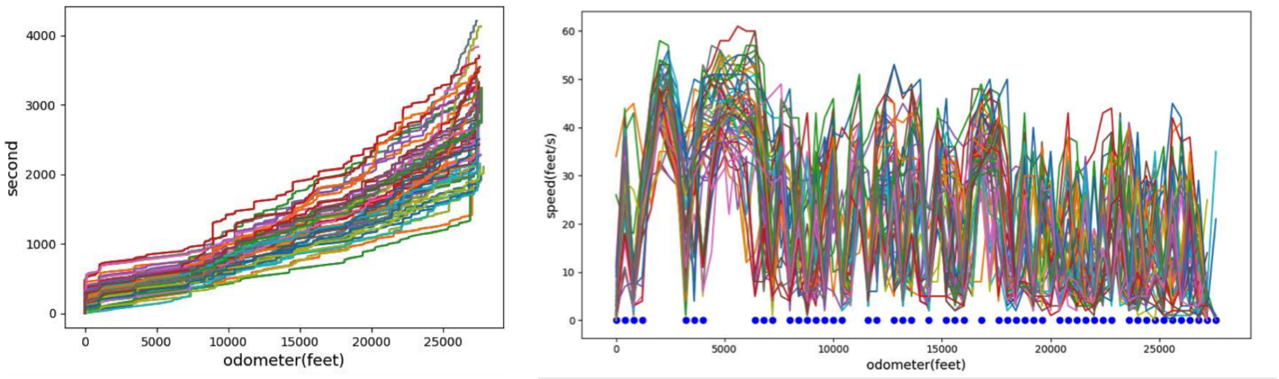
the bus. A bus that goes around the outer lane to turn will record a longer odometer distance compared to a bus that has used the inner lane in the same direction of traffic flow. Another reason is that the pressure of the tires may contribute to recording more feet over the same distance travelled. This misalignment creates a problem wherein when we divide the odometer space into segments of 200ft, these segments are not the same for every bus trip. Consequently, this makes the locations of bus stops appear differently for every bus. To address these issues, we plan to align the bus routes as part of our future work, using a special indication of bus stop locations. Whenever a bus enters or exits a geo-fence around a bus stop, it is recorded in the data. These recordings should include all bus stops, regardless of whether the bus actually opened its doors or not.

3.3 Bus stop location identification

We want to know the bus stop locations in the odometer space, i.e., how far each stop is from the beginning of the route. This forms an important part of this research; we aim to identify what segments of the route have traffic patterns indicating the existence of a bus stop, as opposed to slow traffic due to congestion, or traffic lights. The dataset contains indications of an ”Open door” whenever passengers board or disembark the vehicle, and which happens almost exclusively at bus stop locations. Combining these indications from all the bus trips is not a trivial task as the odometers of different bus routes are not aligned. Using the geo-fence based indication mentioned in challenge 3.2 can solve this misalignment. However, this method includes all bus stop locations, even if no buses ever stop there (for example a very unpopular bus stop, or an old bus stop that is no longer in use). To identify valid bus stops as our ground truth, we consider a road segment as containing a bus stop only if more than a minimum number of buses per day open their doors at that location.

4 Case Study: Washington D.C. Metrobus

The main bus service in the Washington D.C. metropolitan area, Metrobus, provides more than 400,000 trips each weekday, and serves 11,500 bus stops in the District of Columbia, Maryland, and Virginia respectively. Metrobus has more than 1,500 buses operating on 325 routes, and is the sixth busiest bus agency in the United States [20]. Over the last two decades, there has been gradual decrease in the ridership of Metrobus, particular due to increased congestion, and which has resulted in significant lose of revenue [19]. This makes the specific study area and bus service of particular interest in the study of traffic flow dynamics. The sections that follow provide an overview of our findings from an analysis of real data acquired from the Metrobus service.



(a) Time travelled vs. Distance covered.

(b) Average Speed vs. Distance covered.

Figure 2: Relationship between travel time, average speed and odometer readings of all the buses.

4.1 Data

We used bus trips from the X2 route of the Washington D.C. metropolitan area shown in Figure 1. Our dataset contained 489 trips from 10/04/2016. During data cleaning, we removed any bus trips that significantly deviated from the examined route. This resulted in a dataset containing 58 trips travelling on one direction. We understand the limitation of performing analysis on such a small dataset, and for only a 24 hour period. Future work will undertake a more in-depth analysis, using a larger collection of trips that are acquired for a much longer time period.

4.2 Preprocessing

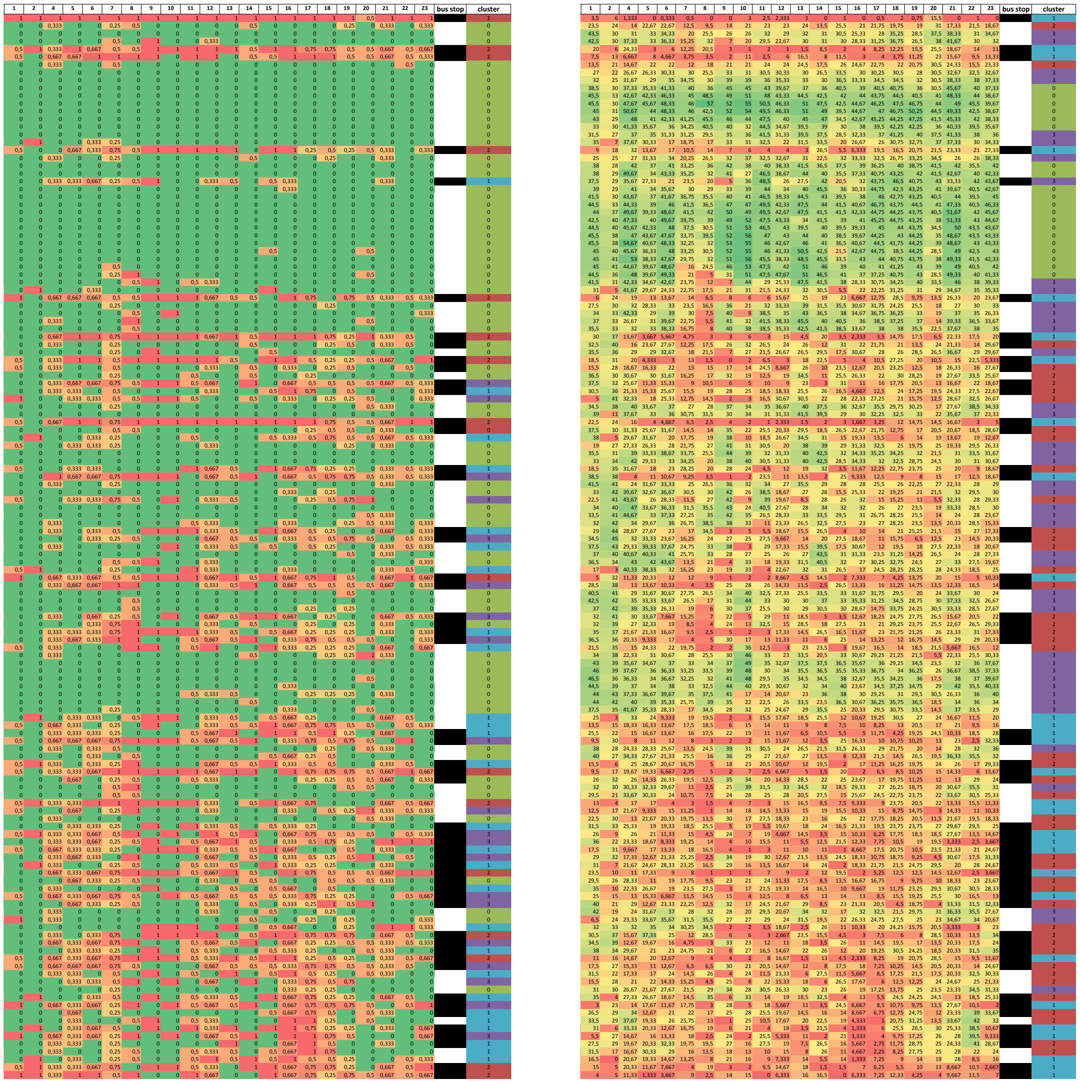
We discretized the bus route into 200ft segments. This segment length was considered neither too small to be noisy, nor large enough so that several bus stops and/or traffic lights would be contained within any single segment. The odometer readings (i.e., distance covered) versus the time from the beginning of the trip are shown in Figure 2(a), for all the buses in our data collection.

Using the above routes, we estimated the average speed of every bus, for each 200ft road segment. These results are visualized in Figure 2(b). The blue dots at the bottom of this figure correspond to indications of a bus door opening. These indications are used in our evaluation as ground truth for where bus stops are located. To avoid false indications (e.g. when the bus driver opens a door for an emergency passenger request), we consider a road segment having a bus stop only if at least five buses have opened their doors at that segment.

4.3 Clustering Segments

We use the aforementioned bus speeds to derive a set of features for clustering road segments along the bus route. Our first approach uses the percentage of slow buses calculated over hourly time buckets. Given a speed threshold τ, we consider a bus as "slow" if it moves at a speed $v < \tau$. For our experiments, we set the value of τ to be 10km/h. Our second approach uses the average speed of all buses calculated for each road segment for each hourly time bucket. Using these approaches, we generate a set of 24 features per road segment. Principal component analysis was then used to reduce our feature space to 4 principal components, following which these latent features were fed to a k-means clustering algorithm. A k value of 4 was used since we expect to discover 4 types of road segments.

Intuitively, we expect *bus stop* locations to be those where almost all buses would stop. On the other hand, *traffic lights* should cause some buses to stop at a red light, while others would pass with a green light, regardless



(a) Fraction of slow buses

(b) Average bus speed

Figure 3: Clustering results.

of time of the day. *Congestion* should affect specific road segments at specific times of the day (e.g. rush hours). This would be a periodic phenomenon, and it would be very interesting to examine this aspect as part of future research. Finally, in *free flow* segments, buses should have relatively high values of speed throughout the day. These segments are identified as those having between 0 and 20 slow buses for at least 20 out of the 24 hour buckets of the day. The remaining segments are labelled as other congestion or traffic lights. The results of our two approaches are presented in the following section.

4.4 Experimental Results

A visualization of our results is presented in Figure 3. Each row corresponds to a 200ft road segment along the X2 bus route. Columns labels 1 through 24 are the hourly buckets, and correspond to the 24 derived features. For Figure 3(a), these features are the fraction of slow buses, colored in a scale from green to red, with red depicting a high number of slow buses, while green depicts few or no slow buses. In the case of Figure 3(b), the

Table 1: Summary of Results for the Fraction of Slow Buses Approach.

Cluster Label	0	1	2	3	Total
Bus stop	5	16	15	16	52
Other congestion or traffic light	14	11	1	4	30
Free flow	55	0	0	0	55
Total	74	27	16	20	137

Table 2: Summary of Results for the Average Bus Speed Approach.

Cluster Label	0	1	2	3	Total
Bus stop	0	24	27	1	52
Other congestion or traffic light	0	5	14	11	30
Free flow	20	0	4	31	55
Total	20	29	45	43	137

features are the average hourly bus speeds for each segment. In this Figure, green shows high speeds, yellow and orange show moderate speeds, while red depicts low bus speeds. The buses start from an origin outside the city, where there is typically less traffic, and travel towards a destination in the center, which is more prone to congestion. This explains why there are many more green cells towards the top (origin) and more red cells at the bottom (destination). The 25th column contains the locations of bus stops depicted as black cells. White cells correspond to road segments that do not include bus stops. The final column shows the derived cluster label of each segment. For clarity, in both Figures 3(a) and 3(b) cluster '0' is colored green, '1' is blue, '2' is red, and '3' is purple. Note however that the numbering of the cluster labels is random. Clusters of the same number in the two figures do not necessarily correspond to each other. In the following we evaluate the quality of our findings.

4.4.1 Using the Fraction of Slow Buses.

The results of our cluster analysis are shown in Table 1. Cluster label 0 includes all the road segments of free-flow. Cluster 1 appears to be more related to congestion as in several of its segments there contain slow buses at certain hours, where there are no bus stops. We note that most road segments of cluster 1 are located towards the city center, compared to most free-flow segments (label 0) that are located outside the city center, and thus less prone to congestion. Most members of cluster 2 correspond to bus stops. However, all the bus stops are almost evenly distributed among labels 1, 2 and 3. In particular, 15 out of 16 road segments of cluster 2 contain bus stops (Precision=0.94), while 15 out of the 52 bus stops in total were labeled as cluster 2 (Recall=0.29). What remains to be tested is which road segments correspond to traffic lights, for which no data was available in this study. This data could be aligned with collected odometer readings to verify if one or more clusters correspond to these locations.

4.4.2 Using the Average Bus Speed.

Table 2 shows the results of using the average bus speed in our analysis. The members of cluster 0 correspond only to road segments of free-flow, where the average bus speed is high throughout the day. These segments are mainly located outside the city center, with no bus stops. Compared to the use of the percentage of slow buses, several free-flow segments are now placed in cluster 3. Those segments are closer to the city center, but are not bus stop or traffic light locations. Thus they are able to maintain high average speeds throughout the day.

Most bus stops are distributed between only two clusters, 1 and 2. While the majority of them (28 of 53) are in cluster 2, the purity of bus stop locations in cluster 1 is larger, with a precision of 0.83, compared to 0.60 for bust stops in cluster 2. The members of cluster 1 have low average speeds throughout the day, implying frequent bus stops. On the other hand, the members of cluster 2 have lower speeds at certain hours of the day, which indicates that they are also prone to traffic, or that the corresponding bus stops are more popular at certain hours of the day. This explains why cluster 2 contains 31% of "other congestion" segments. Cluster 3 almost never coincides with a bus stop and contains road segments with moderate, but not low, average speeds.

5 Conclusions

This analysis is a small step in the direction of spatiotemporal analysis of latent traffic patterns. We explored a data-driven approach to assess the traffic characteristics of road segments using public transportation data. In particular, we use data mining techniques to identify latent speed patterns in bus traffic related flows: traffic-light related delays, bus stop related delays, or free-flow. The lack of sufficient data is an obvious limitation of our analysis. More data are needed in order to assess the validity of our approach, but we believe that our preliminary results show a very promising direction of research.

Acknowledgement

This research has been supported by National Science Foundation "AitF: Collaborative Research: Modeling movement on transportation networks using uncertain data" grant NSF-CCF 1637541.

We would like to thank the Washington Metropolitan Area Transit Authority (WMATA) for providing us with their data.

References

[1] R. Bertini and S. Tantiyanugulchai. Transit buses as traffic probes: Use of geolocation data for empirical evaluation. *Transportation Research Record: Journal of the Transportation Research Board*, 1870(1870):35–45, 2004.

[2] A. Bhaskar, E. Chung, and A. Dumont. Fusing loop detector and probe vehicle data to estimate travel time statistics on signalized urban networks. *Comp.-Aided Civil and Infrastruct. Engineering*, 26(6):433–450, 2011.

[3] P. Chakroborty and S. Kikuchi. Using bus travel time data to estimate travel times on urban corridors. *Transportation Research Record: Journal of the Transportation Research Board*, 1870(1870):18–25, 2004.

[4] S. I. Chien and Z. Qin. Optimization of bus stop locations for improving transit accessibility. *Transportation planning and Technology*, 27(3):211–227, 2004.

[5] S. I.-J. Chien, B. V. Dimitrijevic, and L. N. Spasovic. Optimization of bus route planning in urban commuter networks. *Journal of Public Transportation*, 6(1):4, 2003.

[6] E. M. Delmelle, S. Li, and A. T. Murray. Identifying bus stop redundancy: A gis-based spatial optimization approach. *Computers, Environment and Urban Systems*, 36(5):445–455, 2012.

[7] R. Z. Koshy and V. T. Arasan. Influence of bus stops on flow characteristics of mixed traffic. *Journal of transportation engineering*, 131(8):640–643, 2005.

[8] B. A. Kumar, L. Vanajakshi, and S. C. Subramanian. Bus travel time prediction using a time-space discretization approach. *Transportation Research Part C: Emerging Technologies*, 79:308–332, 2017.

[9] Q. Ou, R. L. Bertini, H. van Lint, and S. P. Hoogendoorn. A theoretical framework for traffic speed estimation by fusing low-resolution probe vehicle data. *IEEE Trans. Intelligent Transportation Systems*, 12(3):747–756, 2011.

[10] D. Pfoser, S. Brakatsoulas, P. Brosch, M. Umlauft, N. Tryfona, and G. Tsironis. Dynamic travel time provision for road networks. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '08, pages 68:1–68:4, 2008.

[11] F. Pinelli, F. Calabrese, and E. P. Bouillet. Robust bus-stop identification and denoising methodology. In *Intelligent Transportation Systems-(ITSC), 2013 16th International IEEE Conference on*, pages 2298–2303. IEEE, 2013.

[12] W. Pu, J. Lin, and L. Long. Real-time estimation of urban street segment travel time using buses as speed probes. *Transportation Research Record: Journal of the Transportation Research Board*, 2129(2129):81–89, 2009.

[13] S. S. Pulugurtha and V. K. Vanapalli. Hazardous bus stops identification: An illustration using gis. *Journal of Public Transportation*, 11(2):4, 2008.

[14] A. A. Saka. Model for determining optimum bus-stop spacingin urban areas. *Journal of Transportation Engineering*, 127(3):195–199, 2001.

[15] R.-P. Schäfer, K.-U. Thiessenhusen, and P. Wagner. A traffic information system by means of real-time floating-car data. In *ITS world congress*, volume 2, 2002.

[16] N. Uno, F. Kurauchi, H. Tamura, and Y. Iida. Using bus probe data for analysis of travel time variability. *Journal of Intelligent Transportation Systems*, 13(1):2–15, 2009.

[17] L. Vanajakshi, S. C. Subramanian, and R. Sivanandan. Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses. *IET intelligent transport systems*, 3(1):1–9, 2009.

[18] Y. Wang, Y. Zheng, and Y. Xue. Travel time estimation of a path using sparse trajectories. In *The 20th ACM SIGKDD*, pages 25–34, 2014.

[19] Washington Metropolitan Area Transit Authority. FY2017 Budget: Ridership and Revenue, October, 2015.

[20] WMATA. Washington Metropolitan Area Transit Authority. `https://www.wmata.com`.

[21] X. Zhan, S. V. Ukkusuri, and C. Yang. A bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data. *Automation in Construction*, 72:237–246, 2016.

[22] F. Zheng and H. Van Zuylen. Urban link travel time estimation based on sparse probe vehicle data. *Trans. Res. Part C: Emerging Technologies*, 31:145–157, 2013.