Spatiotemporal Bus Route Profiling using Odometer Data

Xiqi Fei
George Mason University
xfei@gmu.edu

Olga Gkountouna
George Mason University
ogkounto@gmu.edu

Dieter Pfoser
George Mason University
dpfoser@gmu.edu

Andreas Züfle
George Mason University
azufle@gmu.edu

ABSTRACT

Fixed-route bus systems are an important part of the urban transportation mix. 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 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 in relation to bus stops throughout the day. This gives us an insight on whether the routes are affected by traffic congestion or more random events such as traffic lights. We first employ a macro-level qualitative analysis to identify patterns across different trips. A micro-level quantitative analysis further refines this approach by analyzing the speed patterns around bus stops. Our analysis is based on bus odometer data, which is a one-dimensional representation of trips that has considerable accuracy when looking at speed patterns. Exploiting route metadata in relation to stops, we use Dynamic Time Warping to cluster different stops based on their speed profiles throughout the day. The clustering can be used to generate a spatiotemporal route profile and we show how such a profile provides actionable intelligence for route planning purposes.

CCS CONCEPTS

• **Information systems** → Data analytics; • **Applied computing** → **Transportation**.

KEYWORDS

Bus Data, Odometer, Clustering, Outlier Detection, Traffic

ACM Reference Format:
Xiqi Fei, Olga Gkountouna, Dieter Pfoser, and Andreas Züfle. 2019. Spatiotemporal Bus Route Profiling using Odometer Data. In *27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), November 5–8, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3347146.3359350

SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6909-1/19/11…$15.00
https://doi.org/10.1145/3347146.3359350

1 INTRODUCTION

The National Transit Database which is maintained by the United States Department of Transportation provides a yearly data report on "National Transit Summary and Trends" [14]. This report shows that more than 16 billion passenger miles were travelled in the US in 2017, and more than 4.5 billion unique trips were taken. It also shows that more than 150 million bus hours were in service in 2017, making fixed-route bus transportation the "most common form of public transportation service provided in the United States" [14].

While the research community has thoroughly researched public train transportation [12, 21], and road traffic [17, 23, 26, 27], contributions towards a better understanding of traffic in relation to bus routes have been widely neglected. A considerable disadvantage of fixed-route buses is their slow speed due to frequent stops, but also due to the lack of segregation from other vehicles in traffic. At a single stop, we often observe buses stopping multiples times due to adjacent traffic lights, or due to being located on roads with heavy traffic. As such, assessing bus stops that are prone to such issues is an important aspect of route planning and scheduling.

A typical approach to analyzing traffic is the use of GPS tracking data, e.g., [17]. In urban areas, GPS signals are often affected by high rise buildings ("urban canyons") [5], and the related GPS error can be observed in Figure 1(a), which shows the GPS locations of buses as recorded in relation to the station locations. We can see that for some stations, the resulting positioning error is considerable, randomly distributing the location of a bus by hundreds of meters.

Although map-matching methods (cf. [3]) can be used to align GPS trajectories to road networks, the significant (two-dimensional) measurement error results in uneven travel time distributions along the projected route. In the case of our bus tracking, odometers measure progress of the vehicle along a fixed one-dimensional route. Figure 1(b) shows odometer readings of different buses mapped to their location. While we still observe a similar uncertainty of the true stop location as we observed using GPS, much of this uncertainty is removed by aligning the starting points of the odometer readings of different trips of the same route. In our work, we reduce this uncertainty even further by not considering the entire trajectory, but only segments of it around bus stops as delineated by geofences (GPS coordinates) around bus stops, which are part of the collected metadata (cf. Figure 2).

Specifically, this work focuses on odometer data recorded by an Automatic Vehicle Location (AVL) system for fixed-route buses in the Washington D.C. area, which was made available by the Washington Metropolitan Area Transit Authority (WMATA). We use these readings to find spatiotemporal speed patterns in relation

to bus stops. These speed patterns are used to categorize bus stops into different types, with the goal to possibly identify problematic stops, i.e., stops affected by traffic and/or their location in relation to road infrastructure such as traffic lights and stops.

The data comprises trips of the same route during different times of the day. Each trip consists of a time series of odometer readings from a specific bus, with sampling intervals ranging from 1 to 10s. An odometer reading consists of a timestamp and the distance covered from the beginning of the route up to that point. Since the route of a bus is immutable, this absolute location on the route allows us to uniquely infer the absolute location of a bus. To align and verify the results of our approach, we exploit additional metadata in relation to odometer readings, such as (i) entry and exit markers when a bus enters or leaves a prescribed geofence around a bus stop, (ii) GPS locations, and (iii) door open and close information.

As a first step towards the analysis of this enriched odometer data, we have to account for the varying sampling rate. Here we discretize the route into constant length intervals and calculate the average bus speed over those specific periods for each trip. Even so, the odometer readings between trips are not aligned spatially, e.g., two buses may be at different locations after 1,000 odometer-measured feet. The reasons for this are not properly calibrated devices, different devices and bus types, tire pressure, and buses moving in traffic by switching lanes. As a result, bus stops as recorded in the data are not always assigned to the same interval bucket. To address this, we propose an alignment algorithm not for the entire trajectories, but for segments around bus stops. Here we utilize the geofencing metadata that records when a bus enters end exits the area around a bus stop. Using this so-called bus stop speed signatures and aligning them using DTW, we can cluster them and identify different stop categories. Our goal is to discover if there are underused stops that may need to be moved to a more popular location, or stops that due to their location on the road network significantly contribute to delays. Using this unsupervised learning approach, public transit authorities are able to make informed decisions with respect to route re-design, or to propose infrastructure modifications such as dedicated bus lanes.

To summarize, the contributions of this work are as follows:

- Analyzing bus routes to identify delays.
- Odometer alignment based on cross-correlation of the bus trips.
- Unsupervised learning approach to cluster route segments, i.e., bus stop categorization.
- Spatiotemporal profiling of bus stops using speed segment clusters.

The remainder of the paper is structured as follows. An overview of the related work is presented in Section 2. Then, Section 3 describes the specifics of odometer data and provides a qualitative and visual analysis of bus trips using fixed-route bus odometer data. In Section 4 we focus our analysis on individual bus stops (rather then whole bus trips) and describe a micro-level unsupervised approach to find similar bus stop speed profiles over space and time. Our experimental results are presented in Section 5. Finally, in Section 6 we discuss our conclusions and directions of future work.

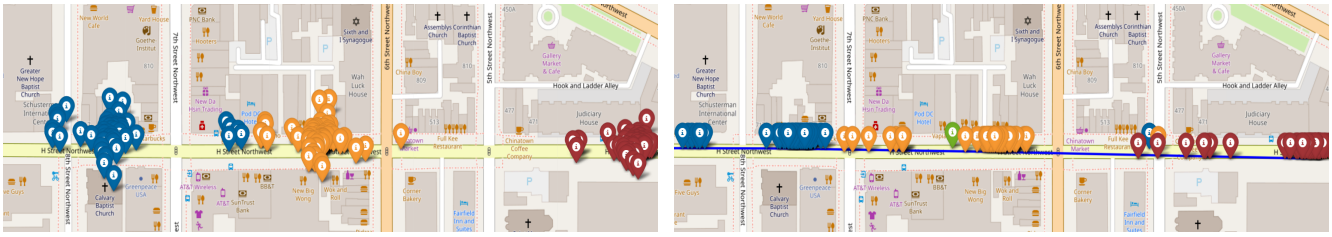
2 RELATED WORK

Traditional solutions for traffic estimation employ floating car data (FCD) and individual GPS data to analyze traffic conditions. FCD is data generated by cell-phones in vehicles used to determine the traffic speed and probe overall traffic conditions. A number of studies have explored FCD for estimating travel times [16, 19, 23], traffic conditions [1, 9], and traffic speed [15]. While FCD can be used for our task of analyzing traffic conditions at bus stops, odometer data is much more easily accessible, as it does not require to access to private phones of individuals.

GPS trajectories generated by taxis are being used in literature for traffic analysis [17, 23, 26, 27] and transportation networks improvement [28]. An overview of challenges and solutions of mining GPS trajectories is found in [29]. For the purpose of analyzing bus routes, buses can be easily equipped with GPS. However, this approach suffers from uncertainty due to signals blocked by buildings [5]. While it has been shown that this uncertainty can be overcome for the purpose of mapping signals to trajectories [13], it remains a challenging problem to accurately measure speed of individual vehicles using GPS [25]. In contrast, there are solutions to accurately measure the speed of traffic using GPS in urban areas [20] by averaging over many individual vehicles. However, such approach is inappropriate to estimate the speed of fixed-route buses, which must make a stop at bus stops. Using odometer data, we can exploit the fixed-route property of buses for highly accurate positioning, independent of buildings and signal strength.

As a traditional means of transportation, buses can collect a large amount of urban traffic data on a daily basis. The collected data have gained considerable attention for estimating traffic conditions. Bus-related studies are mainly focused on travel time prediction and traffic pattern analysis. By examining the relation between travel times of a transit vehicle and of an automobile, [4] shows that buses with AVL systems can be used as a probe to collect travel time data at regular intervals cost effectively. Kumar et al.[11] developed a bus arrival time prediction system which considers spatial-temporal variations, using a time-space discretization approach. Bai et al.[2] predicted bus travel time from both the offline prediction made by SVM using historical travel data and the dynamic adjustment made with Kalman Filter. Wang et al.[22] proposed a bus management system to analyses bus delay based on GPS and Automatic Fare Collection (AFC) system data. However, none of these works tackles the problem of finding spatio-temporal traffic patterns on bus routes for the purpose of planing better bus routes.

The exploration of bus stops has also been the focus of several works. Koshy and Arasan[10] studied the influence of bus stops on traffic flow under heterogeneous traffic conditions through a simulation technique. Yonezawa et al. [24] performed Random Forest classification of bus operation states by using bus sensor data, through which three states in services including "Stopping at bus stop" and "Arriving or departing at bus stop" are able to be distinguished. Yet, this approach does not consider traffic conditions and speed. By combining feature mining with SVM and Random Forest classification algorithms,[8] classified bus stops and non-bus-stop automatically from bus GPS trajectories based on speed-based features (speed and acceleration) and histogram-based features. While



(a) Measured GPS Locations (b) Measured Odometer-based Locations

Figure 1: Example of unaligned bus data. Each group (color) of points corresponds to different locations of the same bus stop, using GPS locations and odometer readings from different bus trips.

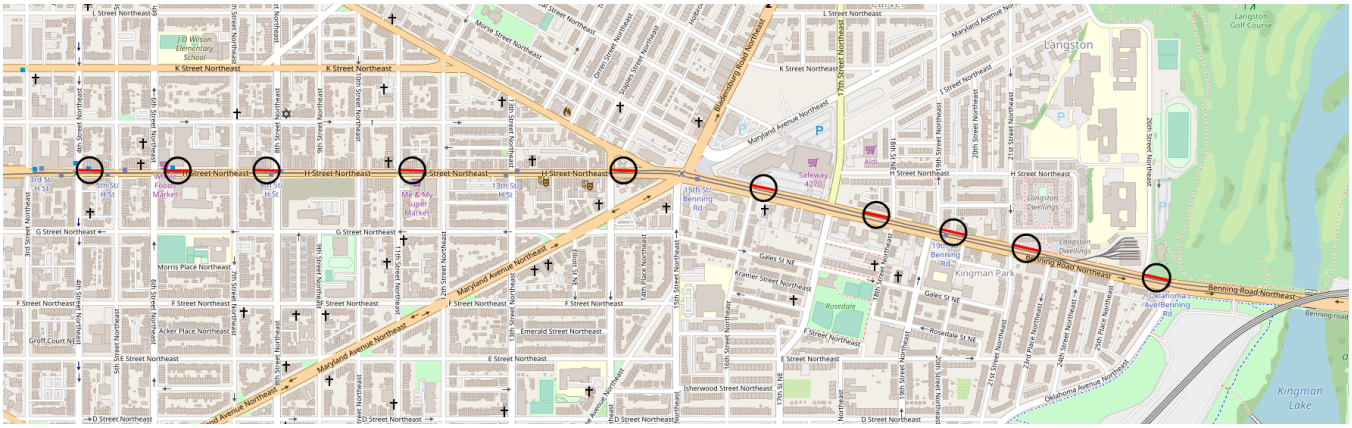
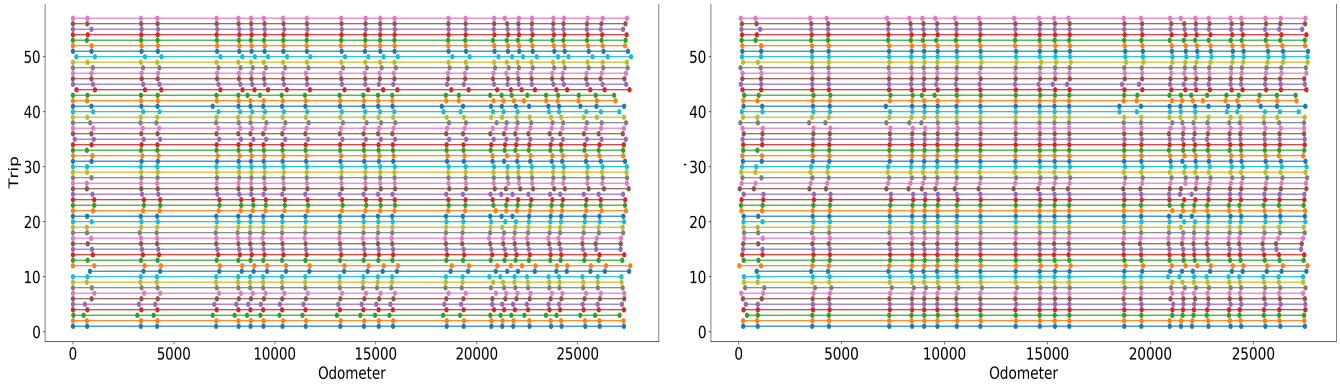


Figure 2: Geo-fenced bus stops and corresponding trajectory segments of part of the route.



(a) Original trips (b) Aligned trips

Figure 3: Odometer alignment based on bus stops.

this approach has a similar goal, they are using GPS trajectories only, whose uncertainty can be limited in urban areas.

Wang et al.[22] clustered bus stops according to the number of passengers who boarded, but did not consider traffic conditions. Fei et al.[7] used speed patterns of bus routes from AVL to identify different categories of route segments which include bus stop locations. This preliminary approach prescribes first ideas to analyze the traffic at bus stops, but does not have ground-truth data to evaluate their results. Delmelle et al. [6] integrated location coverage model within a GIS environment to identify bus stop redundancy

for transit planning optimization, which is an approach orthogonal to ours.

Finally, a lot of previous works have studied traffic conditions in metro and train networks Metro/Train traffic papers [12, 21]. However, such networks do not commonly exhibit traffic delays, nor can routes be redesigned to circumvent traffic.

3 MACRO-LEVEL QUALITATIVE ANALYSIS

The objective is to employ odometers instead of GPS devices to track location and speed of fixed-route buses. We argue that in this fixed-route case, odometer data has a smaller error, as it measures the location in a one-dimensional space (distance from origin) rather than a two-dimensional space. We know that there is a vast difference between the two data types (as illustrated in Figure 1) and we ran experiments to compare GPS (the buses we are studying also provide GPS data) and odometer data. We decided not to include these experiments, as is hard to experimentally assess the quality of the two without any authoritative ground truth. Thus, instead of measuring the average distance between odometer and GPS locations, we provided Figure 1, to give a qualitative intuition that odometer data is more accurate, by showing that GPS frequently puts the buses location off-road and inside buildings.

In our work, we use fixed-route odometer data of the X201 bus route of the Washington D.C. Metropolitan Area. Our dataset consists of 80 trips made on 10/04/2016. In the following, we describe the data that we use for our analysis, explain the preprocessing steps we applied to align different bus trips on the same route, and show a first qualitative analysis of spatial-temporal patterns in this data.

3.1 Fixed-Route Odometer Data

Table 1 presents a small sample of the data. Each line is a reading that includes the trip id, the GPS coordinates (Latitude, Longitude), the door status (O: open or C: closed), the bus status (M: moving or S: stopped), the odometer reading in feet from the origin of the trip, the timestamp in seconds from the beginning of the trip, and a geofence tag when entering (E) or exiting (X) a bus stop geofence. This specific sample contains readings from trip #24 around the area of a bus stop. The third line of the sample has the indication 'E', which signals that the bus entered the geofence of the 15^{th} bus stop at time = 1310s. At that point the doors are still closed (C) and the bus is moving (M). A few lines later at time = 1319s (the full record was omitted to save space), the bus stops (S) and the door opens (O). At time = 1366s, there is a corresponding indication 'X' for exiting the geofence of this bus stop.

3.2 Data Preprocessing

During data cleaning, we removed any bus trips that significantly deviated from the examined route. This resulted in a set of 58 trips travelling in the same direction.

Furthermore, we preprocessed the data by aligning the corresponding bus stops from different bus trips to each other. As we have seen in Figure 1(b), odometer readings of different trips may be significantly misaligned, due to buses starting their odometer early/late, longer travel due to frequent lane changes, or even due to varying tire pressure. To illustrate this problem in more detail, consider Figure 3(a), which shows the location of bus stops for each of the 58 trips of the X201 route. The stops are identified as metadata tags in the odometer data. Each trip is depicted with a different color and the trips are ordered by time starting with the earliest in the day. We clearly observe a misalignment between the stop locations. A first observation that we had was that this misalignment is often due to an initial offset.

To align bus stops between different trips, we used metadata that marks when a bus enters a geofence around a bus stop. We create a boolean vector with elements that are set to 1 whenever the bus enters the bus stop area, and set to 0 otherwise. This results in a time series of 0s and 1s for each trip. We use the cross-correlation between two series, t and t', to find their best alignment. The maximum of the cross-correlation function indicates the point where the series are best aligned, i.e., the delay between the two series is determined by the argmax of the Pearson correlation coefficient between the two times series:

$$\text{delay}(t, t') = \arg\max_d \left(\frac{\sum_{i=1}^{n} (t_i - \bar{t})(t'_{i+d} - \bar{t'})}{\sqrt{\sum_{i=1}^{n} (t_i - \bar{t})^2} \cdot \sqrt{\sum_{i=1}^{n} (t'_{i+d} - \bar{t'})^2}} \right), \quad (1)$$

where t_i corresponds to 1 if time series t had a geofence entering event during it's i^{th} $10ft$ odometer interval and n is the number of interval of time series t. The parameter d specifies an offset between the two time series, and the offset yielding the highest correlation between the two time series is returned by Equation 1. For this alignment, we chose $10ft$ segments and consider them neither too small to be noisy, nor large enough such that several bus stops and/or traffic lights would be contained within any single segment.

Since Equation 1 only aligns a pair of time series, we arbitrarily select one trip as a reference trip, then calculate its cross-correlation to any other trip, and the adjust the latter by moving it forward or backwards according to the calculated delay. The result of this alignment can be seen in Figure 3(b). We visually observe an improved alignment of bus stops for most trips. For some trips, we still observe a misalignment due to the whole trip being consistently shorter or longer. This may be due to an badly calibrated

Table 1: Sample from Fixed-Route Bus Odometer data.

Trip Id	Latitude	Longitude	Door	Status	Odom. (ft)	Time (sec)	E/X
24	38.900248	-77.012113	C	M	19298	1309	
24	38.900253	-77.012217	C	M	19327	1310	
24	38.900253	-77.012217	C	M	19336	1310	E
24	38.900258	-77.012318	C	M	19356	1311	
...							
24	38.900265	-77.012625	C	M	19439	1315	
24	38.900267	-77.012682	C	M	19453	1317	
24	38.900267	-77.012688	O	S	19454	1319	
24	38.900267	-77.012688	O	S	19454	1338	
24	38.900267	-77.012688	C	S	19454	1357	
24	38.900268	-77.012732	C	M	19467	1362	
...							
24	38.900265	-77.012898	C	M	19516	1365	
24	38.900265	-77.012898	C	M	19537	1366	X
24	38.900262	-77.013072	C	M	19566	1367	
24	38.900260	-77.013170	C	M	19595	1368	

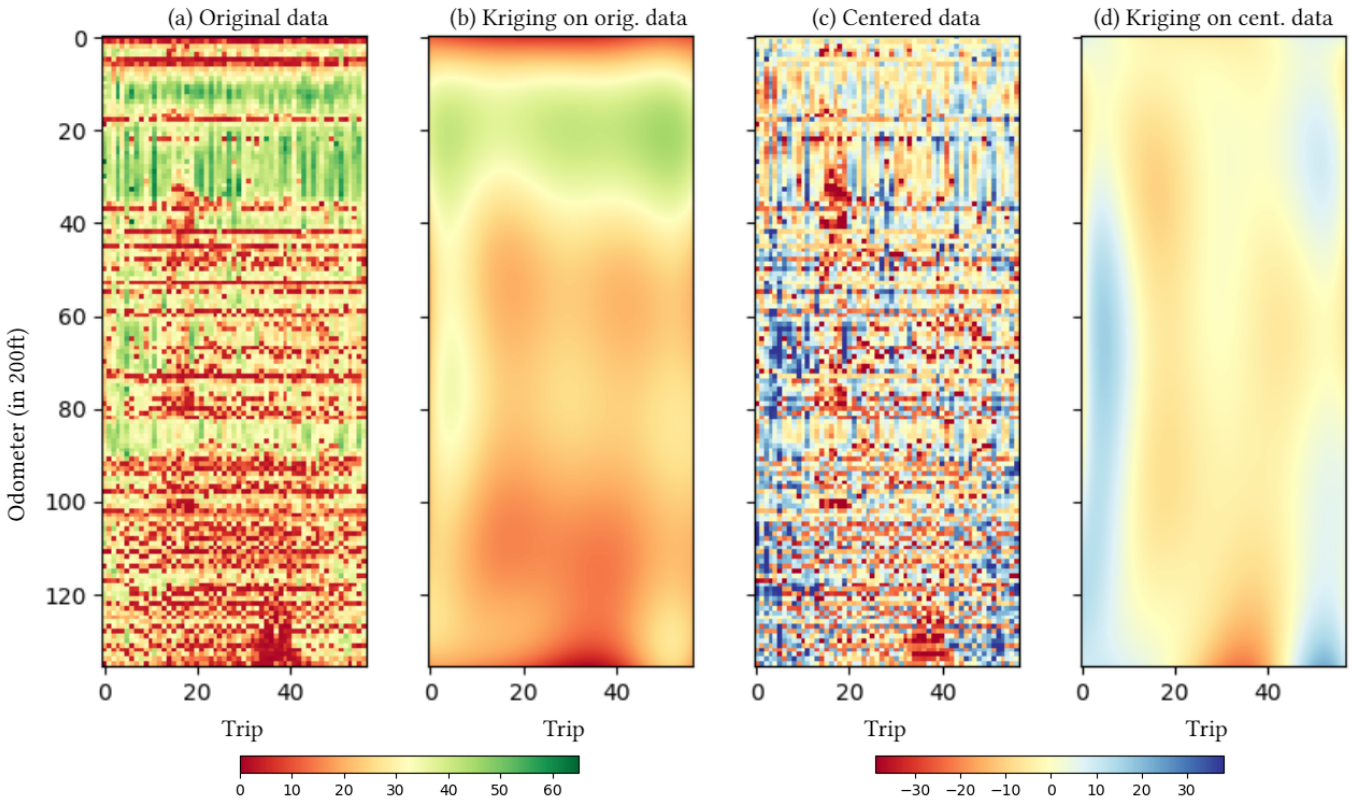


Figure 4: Modelling traffic patterns of the bus route over odometer space and time. High values of speed are depicted as green in the original data plots (a) and (b), while slower speed is depicted as red. The centered data plots (c) and (d) show high values of speed in blue color and lower ones with red respectively.

odometer, or due to different tire pressure, which causes more (or less) distance to be measured per trip. Finding a solution to account for such trends is beyond the scope of this paper and part of our future work.

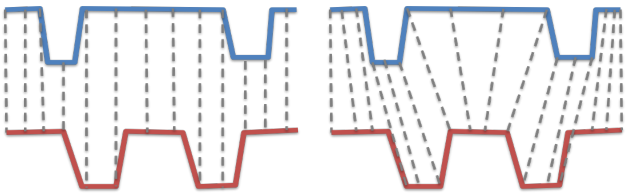
3.3 Visualizing Spatio-Temporal Patterns in Fixed-Route Bus Odometer Data

Following this preprocessing, our first approach is to provide a qualitative analysis of the bus routes using spatial interpolation to discover trends of low traffic speed in either space or time. For this purpose, we first estimate the average speed (in km/h) of a bus in each $200ft$ ($61m$) long road segment using the time stamps of a bus entering (t_{enter}) and exiting (t_{exit}) that segment:

$$v = \frac{200ft}{(t_{\text{exit}} - t_{\text{enter}})}$$

Figure 4(a) presents a visualization of our bus speed calculations. Every row corresponds to a $200ft$ route segment ordered by odometer distance from the origin. Every column represents a trip, with all trips being ordered by time. Trips take place approximately every $20mins$, so the x-axis is an indication of the time of the day. The speed values are depicted in color, ranging from red (slow) to green (high). The original data seems noisy, but we can observe parts of the route that are faster overall, and other parts being slower. Yet, it is hard to observe any temporal (vertical) patterns. These can be better observed in Figure 4(b) where we smoothed this two-dimensional speed map using Gaussian process regression (Kriging) having a Radial-Basis-Function kernel. We can see very clearly a region of free-flow towards the beginning of the bus route, and regions of lower speed towards the destination. This result is due to the direction of this bus route, which originates outside the DC city center and has its destination close to the center. Figure 4(b) helps us identify traffic patterns along the segments of the route, but it is hard to draw any conclusions about patterns based on the time of the day. For this reason, we centered every row of the speed matrix around its mean, by subtracting the row mean from each value. The result is shown in Figure 4(c) with relative faster times of the day depicted as blue, and relative slower ones as red. Kriging on the centered data is depicted in Figure 4(d), where two patterns of heavy congestion are evident around the morning and afternoon rush hours.

To summarize, the qualitative evaluation in Figure 4 presented in this section shows global spatial and temporal patterns of slow traffic. While this evaluation shows global trends over space in time, it does not allow to find local patterns and trends such as individual bus stations experience slow traffic at certain times of the day. We



(a) Euclidean distance pairs (b) DTW distance pairs

Figure 5: Example of the Euclidean vs. the DTW distance between two bus stop speed signatures, shown in blue and in red color respectively, with two slowdowns each. Each dashed line indicates the matching between a point in the first time series and its match in the second time series.

are more interested in finding spatiotemporal patterns, i.e., patterns of $(location, time)$ pairs of slow traffic. Such results would not only indicate problematic bus stops, but also show us problematic bus stops that are impacted by traffic and/or infrastructure problems. For this purpose, we split each bus trip into smaller sections that each correspond to a single bus stop. We obtain these sections by using geofences around bus stations to ensure alignment. In the following, we propose an unsupervised approach of finding unusual and problematic $(location, time)$ pairs by clustering the these speed time-series of buses inside the geofences.

4 MICRO-LEVEL QUANTITATIVE ANALYSIS

Our goal is to find spatiotemporal traffic patterns of the areas at and around each bus stop. Bus stops are of great importance for route planning purposes and while an adequate geographic coverage based on demand is of primary importance, the specific placement within limits can greatly affect the overall travel time and trip duration. By more closely analyzing bus speeds in relation to stops, we will be able to identify bottlenecks, such as many buses simultaneously clogging a bus stop, or the impact of congestion and traffic lights on stop duration. For this reason, we isolate those parts of each trip that are within a distance d around a bus stop. Again we utilize the geofence as recorded in the metadata to extract such segments from the overall trajectory. This results in $n \times m$ segments of length $2d$ in the odometer space, where n is the number of bus trips and m is the number of bus stops of the route. Each of these segments corresponds to a (bus stop, trip) pair. As these segments have significantly smaller lengths than the full length of the trip, We discretize them using a more refined granularity of 15 feet (4.6m). Within each 15ft segment, we compute the average speed v. We term the resulting sequence as the *traffic signature* of each (bus stop, trip) pair.

To find a clustering of speed signatures, we also use, alternatively to the original speed values v, the inverse of the average speed (v^{-1}), as well as a 3-level discretizing of the speed values in buckets of slow, moderate and average speed.

The reason for experimenting with all the above options is that we try to capture the local traffic behavior within bus stops, and distinguish those where bus drivers have a normal behavior (single slow down per stop) from those with irregular patterns where there are multiple slow downs per stop due to traffic lights or congestion. Such insight could result in bus stops being moved , e.g., to after the traffic light, or in creating dedicated bus lanes assigned to the corresponding road segments during rush hour. We are also able to identify groups of stops with no slowdowns, indicating unpopular locations where bus stops may not be necessary. Those bus stops may be moved to other more appropriate locations, potentially serving more passengers.

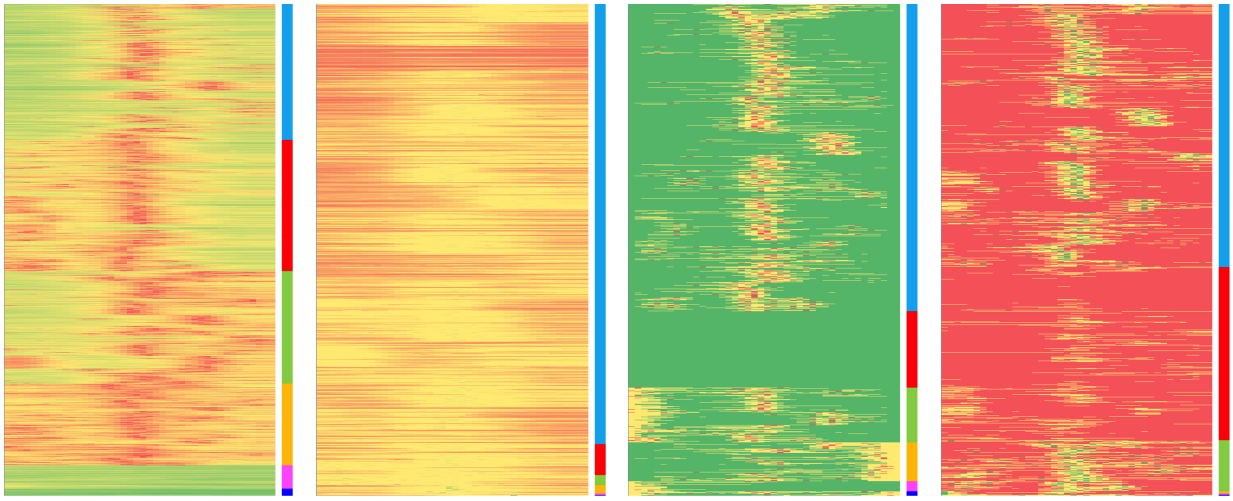
We use hierarchical agglomerative clustering to find categories of bus stops across time (different trips taking place at different times of the day) and space (different locations). As a distance metric, we DTW distance, which can capture the signal of one, none, or multiple drops of speed (slowdowns) in the speed signatures. The signatures of two bus stops with the same number of slowdowns or stops will have a smaller DTW distance than any two signatures having a different number of slowdowns. Further, the location of these slowdowns with respect to the bus stop area does not affect the distance significantly. Consider for example the two traffic signatures of Figure 5. Both have two major slowdowns at different locations with respect to the start and end of the bus stop area. The grey dashed lines of Figure 5(a) indicate which pairs of speed values (one from each signature) will be used to calculate the Euclidean distance. Since the slowdowns are not aligned, the Euclidean distance will be quite large. On the other hand, Figure 5(b) shows that DTW tries to match each slowdown of the first signature to the corresponding slowdown of the second signature. Thus, the distance will be much smaller.

The reason that we choose this distance metric is that we are not interested in finding groups of speed signatures that slow down and speed up exactly at the same relative location. Rather, we are interested in creating *groups of bus stops having identical number of slow downs in their speed pattern*, regardless of where exactly the slowdown occurred. Normally, buses are expected to slow down and stop once per bus stop. Multiple stops or multiple consecutive slowdown-speedup patterns at a bus stop area indicate a potential problem that the bus driver had to stop the vehicle multiple times. On the other hand, bus stop areas with no slowdown patterns (free flow) indicate that the bus driver did not have to stop since no passenger wanted to disembark and/or nobody was waiting to board on the bus.

5 EXPERIMENTAL EVALUATION

In the following experiments, the focus is on detecting groups of bus stops, based on similar speed profiles. We use the data of 58 bus trips for route X201 collected on 10/04/2016 in Washington, D.C. Metropolitan Area. The trips in terms of their speeds are visualized in Figure 4(a). Each column corresponds to an individual bus trip, and each row corresponds to a $200ft$ long segment of the fixed-route.

As stated previously, our goal is to find "similar" bus stops. In this section, we evaluate our spatiotemporal clustering approach of bus stops as described in Section 4 using geofence metadata to identify bus stop areas in the trip data.



(a) v (b) v^{-1} (c) 3-Level v (d) 3-Level v^{-1}

Figure 6: Speed signatures of each (bus stop, trip) pair and its corresponding cluster label for each of the four examined cases of speed-based features.

5.1 Evaluation of Different Speed Signatures

Figure 6 shows the resulting speed signatures using different speed-based features as described in Section 4. Each row corresponds to one (bus stop, trip) pair. All but the last column of each chart correspond to the feature values of the speed signature, while the color of the last (magnified) column indicates the cluster membership of the (bus stop, trip) pair. Each image uses a different speed feature transformation. To find similar speed patterns, we first use the raw speed values depicted in Figure 6(a). In this case, we observe four major clusters of comparable sizes emerge. However, they do not seem to distinguish well between multi-stop and single stop patterns (and corresponding drops in speed). Instead, it seems that the overall average speed of a signature is the most common characteristic of each cluster member. Also, it appears that the low/high speed locations within a bus stop area (beginning, middle, end) also affect the clustering result. Cluster 0 (blue label) has moderate speeds overall, Cluster 1 (red label) has lower speed at the beginning of the stop, Cluster 2 (green label) has lower speeds towards the end, Cluster 3 (yellow label) has lower average overall speeds, while the smaller Clusters 4 (magenta label) and 5 (indigo label) seem to capture the free-flow speed signatures, i.e, buses not stopping at all.

The main problem of this approach is the usage of raw speed values using DTW distance (Section 4), which ignores the magnitude of speed values. For example, the difference between 40 and $50km/h$ is the same as between 0 and $10km/h$, thus discriminating clusters by their free-flow speeds rather than by the, more interesting, lower speed sections.

In an attempt to address this issue, Figure 6(b) presents the result using the inverse speed (v^{-1}) as signatures of the (bus stop, trip) pairs. As can be observed, this approach actually confuses the clustering algorithm, as 89.53% of (bus stop, trip)-pairs (1001 out of 1118) belong to one big heterogeneous cluster. The problem here is that only very small speed values can be discriminated by the clustering algorithm. A speed change from $0km/h$ to $10km/h$ yields a reduction of inverse speed from $\frac{1}{1} = 1$ to $\frac{1}{10} = 0.1$, while any further speed increase cannot contribute more than another 0.1 reduction of inverse speed. Thus, this approach discriminates between different cases of consistently low speed, but considers higher speeds as too similar.

Table 2: Representative values per speed category

category	v (km/h)	v^{-1}	3-Lvl v	3-Lvl v^{-1}
slow	< 1.75	> 0.57	0.87	1.15
moderate	$1.75 - 15$	$0.067 - 0.57$	8.5	0.118
fast	> 15	< 0.067	37.5	0.027

These first two experiments pointed us as a problem: It is difficult to define a function that maps speed values to perceived traffic speeds tailored to the Washington D.C. area. In our next approach, we help the clustering algorithm by manually defining traffic speed categories. Specifically, we discretize the speed values into three categories: slow (below $1.75km/h$), moderate (1.75-$15km/h$), and fast (above $15km/h$). The reason for this choice is that, in this way, we distinguish between segments of a bus stop where the vehicle actually stopped, segments where buses move very slowly possibly due to congestion, but do not actually make a stop, and free flow segments. We replace any speed value within a range –slow, medium, or fast– by the median value of the corresponding range, i.e., 0.87, 8.5, and $37.5km/h$, respectively. Note that we experimented with different numbers of speed categories and concluded that the 3-level discretization yields more discriminative clustering results. The cut-off and the representative values of our speed categories are summarised in Table 2.

Table 3: Cluster size distributions for each clustering result, based on the speed signature of each (bus stop, trip)-pair, for each case of speed signature features.

	Cluster size distributions using			
	original values of		3-Level discrete values of	
cluster	speed	speed inverse	speed	speed inverse
0	312	1001	700	600
1	297	70	173	392
2	255	23	124	116
3	185	19	88	5
4	53	4	23	3
5	16	1	10	2

Figure 6(c) includes the clustering results using the 3-Level discretized speed values. This approach proves much better capability in distinguishing between signatures of free flow (Cluster 1 depicted with a red label), and clusters with at least one stop (Cluster 0 - blue label). We also observe Clusters 3 (green) and 4 (yellow), which seem to have more than one stop and differing only when the major slowdown occurred, i.e., beginning vs. end of the stop area. However, there are several signatures with high to moderate speed values (i.e., the bus did not stop there), and they were grouped into Cluster 0 together with the single-stop signatures. To distinguish better between these two categories, we inverse the 3-level speed. As such, the difference between moderate and high speed values is smaller, compared to the distance between the moderate and slow speed values. The result of this experiment is shown in Figure 6(d). The majority of the speed signatures in first large cluster (blue) have 1 main stop, which indicates the expected behavior. Whereas, the signatures of the second cluster (red) show either free flow or moderate slowdowns, but no stops. This could indicate that the bus slowed down due to moderate congestion, or to check if a passenger was waiting at the bus stop. These signatures correspond to the (bus stop, trip) pairs when no passenger boarded or disembarked the vehicle. We can also observe that, in the third cluster (green label), there are mainly 2-stop traffic signatures, whereas the three smaller clusters capture outliers of multiple (more than two) stops per bus stop area.

5.2 Spatiotemporal Bus Stop Profiling

Having used speed signatures to derive bus stop categories, we can now map them back to actual bus stop locations and see whether any trends emerge. Here, Figure 7 shows the clustering results for the case of the inverse 3-Level speed as a two dimensional array. Each column corresponds to a bus stop on our route. Each row of this matrix corresponds to a bus trip ordered by time, with the earliest trip at 0:12am on top and the latest at 11:57pm at the bottom. The cell color corresponds to one of the clusters shown in Figure 6(d). The blue arrows below each column point to the actual location of the corresponding bus stop. The travel direction is from east to west, thus the numbering of stops is ordered from right to left (1-20). For comparison, we also present the mean bus speed values per (bus stop, trip) pair in Figure 8.

In Figure 7, we can clearly distinguish the underused cases of bus stops 3 and 12, with no stops during most of the day. Stops 4 and 14 are not used half of the time. Using only the average speeds shown in Figure 8 these stops look different, because the average speed is affected by congestion, which is more common towards the end of the route closer to the city center. Even though Stop 14 has lower average speeds than Stop 4, they both have a similar stop rate (about 50%). In fact, Stop 14 has a similar traffic pattern as Stop 8, which is a perfect example of a normal stop where most buses stop exactly once at that bus stop area, for most trips of the day. The traffic speed pattern of Stop 8 in Figure 8 would indicate that it may be a candidate for multiple slowdowns around the middle of the day, this is not the case however, as we can see from the corresponding column of Figure 7.

We can also see that at the beginning and end of the route, Stops 1 and 20 frequently experience multiple stops throughout the day (orange and red cells). This is not easy to infer from the speed values of Figure 8, as their speed patterns are similar to other bus stops such as 18, where we observe overall much fewer multi-stops.

The bus stops that experience at least some multi-stops belong to two major categories: stops where the problems occur during certain hours of the day (traffic), and stops which experience such problems at random times throughout the day (traffic lights). Examples of the former are Stops 1, 13, 18, and 20. Stop 13 seems to be particularly affected by traffic in the morning. On the other hand, Stops such as 7, 11, 15, and 17 have multiple stops that seem to happen randomly throughout the day, possibly due to traffic lights and/or other buses clogging the stop.

To summarize, our approach of clustering bus stops based on speed patterns throughout the day allows us to identify related operational challenges that need to be addressed. For stops that are affected by traffic, the bus operator could request dedicated bus lanes. Further, a disadvantageous bus stop location and configuration (next to a light, small stop area, multiple vehicles clogging stop) could be changed, and underutilized stops could be relocated.

6 CONCLUSIONS

This paper focuses on identifying and profiling bus stops to identify operational problems along a route. In an initial phase, we discretize the bus trajectory and use cross-correlation to align different trips of the same route. Using mean speed values, we model and visualize the speed patterns of the trips along the route and over time. Using various modeling techniques, we are able to identify parts of slow speed along the route throughout the day. To complement this macro-level analysis and to gain further insight into speed patterns that characterize bus stops, we create traffic signatures for each (bus stop, trip) pair and use them to cluster the stops over time. This results in three bus stop profiles: (i) expected behavior - the bus slows down, stops once and continues, (ii) underused bus stops - the bus slows down but does not stop, and (iii) cases which exhibit a bus stopping multiple times at the same stop. We show that our multi-dimensional clustering approach can provide better insight into this profiling problem rather than a baseline approach that simply looks at average speeds.

We see this work as a first step towards a comprehensive data-driven spatiotemporal pattern mining framework for fixed-route

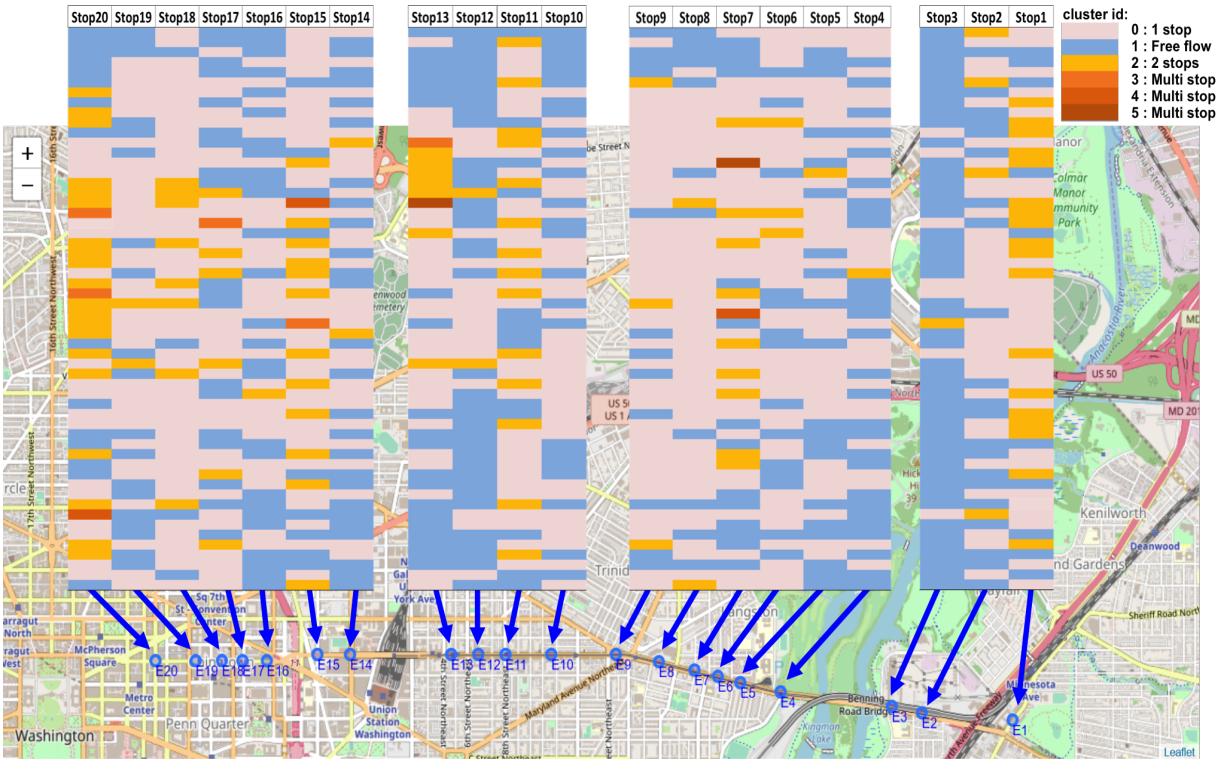


Figure 7: Spatiotemporal clustering of (bus stop, time) pairs.

public transport systems. In future work, we will look at ground-truth data such as driver feedback in relation to problems identified in this approach, e.g., do specific bus stop configurations lead to multiple stops. This will allow us to reformulate the problem of bus stop profiling as a supervised classification problem. Another direction of future work is to work with the bus operators and recommend changes of bus routes, bus stop configuration, and the creation of bus lanes to mitigate the effects of traffic. While this direction would require ubiquitous traffic data, as well as the power to adjust bus routes and traffic lanes, we can employ traffic simulation frameworks, such as the SMARTS Traffic Simulator [18] to perform this research in a sandbox.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] O. Altintasi, H. Tuydes-Yaman, and K. Tuncay. Detection of urban traffic patterns from floating car data (fcd). *Transportation research procedia*, 22:382–391, 2017.
[2] C. Bai, Z.-R. Peng, Q.-C. Lu, and J. Sun. Dynamic bus travel time prediction models on road with multiple bus routes. *Computational intelligence and neuroscience*, 2015:63, 2015.
[3] S. Brakatsoulas, D. Pfoser, R. Salas, and C. Wenk. On map-matching vehicle tracking data. In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pages 853–864. VLDB Endowment, 2005.
[4] 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.
[5] Y. Cui and S. S. Ge. Autonomous vehicle positioning with gps in urban canyon environments. *IEEE transactions on robotics and automation*, 19(1):15–25, 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] X. Fei and O. Gkountouna. Spatiotemporal clustering in urban transportation: a bus route case study in washington dc. *SIGSPATIAL Special*, 10(2):26–33, 2018.
[8] N. Garg, G. Ramadurai, and S. Ranu. Mining bus stops from raw gps data of bus trajectories. In *2018 10th International Conference on Communication Systems & Networks (COMSNETS)*, pages 583–588. IEEE, 2018.
[9] A. Gühnemann, R.-P. Schäfer, K.-U. Thiessenhusen, and P. Wagner. Monitoring traffic and emissions by floating car data. *ITLS Working Paper*, 2004.
[10] 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.
[11] 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.
[12] E. Lin, J. D. Park, and A. Züfle. Real-time bayesian micro-analysis for metro traffic prediction. In *Proceedings of the 3rd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, page 12. ACM, 2017.
[13] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang. Map-matching for low-sampling-rate gps trajectories. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 352–361. ACM, 2009.
[14] Office of Budget and Policy. 2017 national transit summary and trends. https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/ntd/130636/2017-national-transit-summaries-and-trends.pdf, October 2018.
[15] Q. Ou, R. L. Bertini, J. Van Lint, and S. P. Hoogendoorn. A theoretical framework for traffic speed estimation by fusing low-resolution probe vehicle data. *IEEE Transactions on Intelligent Transportation Systems*, 12(3):747–756, 2011.
[16] 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.
[17] D. Pfoser, N. Tryfona, and A. Voisard. Dynamic travel time maps - enabling efficient navigation. In *SSDBM '06: Proceedings of the 18th International Conference on Scientific and Statistical Database Management*, pages 369–378. IEEE Computer

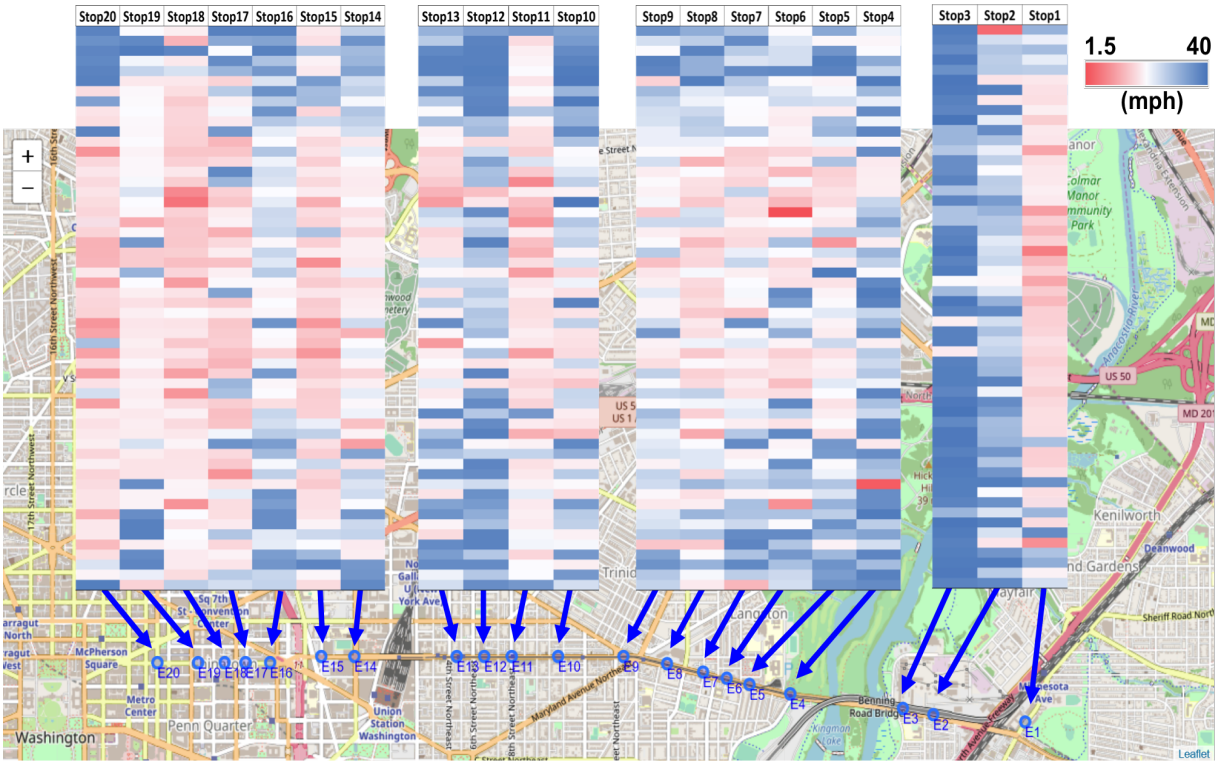


Figure 8: Average speed of each (bus stop, time) pair.

Society, 2006.

[18] K. Ramamohanarao, H. Xie, L. Kulik, S. Karunasekera, E. Tanin, R. Zhang, and E. B. Khunayn. Smarts: Scalable microscopic adaptive road traffic simulator. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2):26, 2017.

[19] R. Sevlian and R. Rajagopal. Travel time estimation using floating car data. *arXiv preprint arXiv:1012.4249*, 2010.

[20] S. Tao, V. Manolopoulos, S. Rodriguez Duenas, and A. Rusu. Real-time urban traffic state estimation with a-gps mobile phones as probes. *Journal of Transportation Technologies*, 2(1):22–31, 2012.

[21] V. Van Breusegem, G. Campion, and G. Bastin. Traffic modeling and state feedback control for metro lines. *IEEE Transactions on automatic control*, 36(7):770–784, 1991.

[22] Y. Wang, S. Ram, F. Currim, E. Dantas, and L. A. Sabóia. A big data approach for smart transportation management on bus network. In *2016 IEEE International Smart Cities Conference (ISC2)*, pages 1–6. IEEE, 2016.

[23] Y. Wang, Y. Zheng, and Y. Xue. Travel time estimation of a path using sparse trajectories. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 25–34. ACM, 2014.

[24] T. Yonezawa, I. Arai, T. Akiyama, and K. Fujikawa. Random forest based bus operation states classification using vehicle sensor data. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 747–752. IEEE, 2018.

[25] J. Yu, H. Zhu, H. Han, Y. J. Chen, J. Yang, Y. Zhu, Z. Chen, G. Xue, and M. Li. Senspeed: Sensing driving conditions to estimate vehicle speed in urban environments. *IEEE Transactions on Mobile Computing*, 15(1):202–216, 2016.

[26] J. Yuan, Y. Zheng, X. Xie, and G. Sun. T-drive: Enhancing driving directions with taxi drivers' intelligence. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):220–232, 2013.

[27] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun. Where to find my next passenger. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 109–118. ACM, 2011.

[28] 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.

[29] Y. Zheng. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):29, 2015.