On Strategies Improving Accuracy of Speed Prediction from Floating Car Data (FCD)

K. Kocamaz[1], H. Tuydeş-Yaman[2], K. Tuncay[3], B. Çınar[4]

[1]Department of Civil Engineering, METU, Ankara, Turkey, kkocamaz@metu.edu.tr
[2]Department of Civil Engineering, METU, Ankara, Turkey, htuydes@metu.edu.tr
[3]Department of Civil Engineering, METU, Ankara, Turkey, ktuncay@metu.edu.tr
[4]Parabol Software Inc., Ankara, Turkey, busra.cinar@paraboly.com

Abstract

For smart mobility, speed data extracted from Floating Car Data (FCD) plays an important role in speed prediction accuracy. However, there are reliability issues for commercial FCD due to processing of individual vehicle tracking data, and imposed temporal averaging to compress data size. Furthermore, spatial discretization significantly affects the accuracy of the prediction due to uneven segment lengths and highly variable data availability in the network. In this study, these issues are examined in detail, and several strategies to improve average speed prediction are proposed. An extensive FCD data from a 75-km long corridor is utilized in the calculations. Firstly, for data reliability, several filters are applied to clean data, then, a robust algorithm is applied to smoothen the speed data. Secondly, to investigate and reduce prediction errors due to spatial segmentation, a number of segmentation approaches are developed, and their effects on the average speed prediction are assessed. Finally, several autoregressive prediction models are implemented and a comprehensive comparison of results is presented.

Keywords: *Floating Car Data, data filtering, data smoothening, Autoregressive prediction models*

1 Introduction

Concept of Floating Car Data (FCD) can be simply described as use of spatio-temporally coded vehicle routes as probe vehicle trajectories in traffic flows, which enables estimation of some of the link flow characteristics (i.e. travel time, speed, etc.), when processed. FCD has been increasingly used for traffic state estimation and monitoring in urban networks mainly due to its high spatial coverage and low cost. Due to hydrodynamic nature of traffic flows, it is also shown that a rather low FCD penetration rate of 15%-20% can provide considerable reliable estimations (Talebpour and Mahmassani, 2016), if sampled randomly. However, current state and nature of FCD sources, unfortunately, does not support the latter, as a) current penetration rates are still around 5%-6% in Turkey, and b) the vehicles are not accessorized to monitor traffic per se, but, for security or fleet monitoring purposes, instead (thus they do not provide data equally in space and time). This brings questions regarding the estimation power of FCD. Secondly, the road network topology is expected to have effect on reliability: smaller segments (50 m-200 m) due to increased number of road crossings, and thus, have very short FCD paths in urban/suburban corridors, whereas comparatively longer ones (more than 500m or even 1 km) segments in rural/intercity corridors may have more fluctuations within segment. But, combined with the traffic flow levels and probability of having the probed vehicles in traffic, the effect of network segmentation may be more emphasized spatially or temporally, thus, needs to be studied in relation to traffic state estimation, which is the aim of this study.

This paper focused on evaluation of speed estimation power of FCD with real vehicle trajectories, more specifically its quality over different corridor characteristics (i.e. urban corridors versus intercity ones). Raw GPS data from various commercially tracked vehicles traveling along (fully or partially) the 75km-long corridor between Polatlı-Ankara (city center) were extracted for December, 2019. Using the Open Street Map (OSM) network topology,

the study corridor was discretized into 655 segments of various lengths. Processing of the raw GPS data resulted in calculation of average segment speeds for 30-min intervals (u_{30}). Second level of analysis included development of autoregressive models (also known as ARIMA models) for the estimation of speeds $(\hat{u}_{30})$ using 3 weeks data, which are later used to test estimation quality with the fourth week values. The ARIMA models are developed individually for each segment separately. The comparisons of ARIMA based speed estimation quality using FCD for different segment types are performed by calculating two measures of effectiveness (MoEs) for selected segments along three sub-corridors (SCs) in urban, urban transition and intercity regions. The results simply show the power of using FCD for determination of travel characteristics of various network and traffic regime characteristics.

2 Literature Review

2.1 Urban speed estimations using FCD

FCD based traffic state estimation studies have been increasingly getting attention of the researchers, as the availability of vehicle trajectory data has been increasing due to increase of GPS-equipped vehicles/fleets for various purposes (taxi fleets, bus fleets, rental cars, construction vehicles, etc.). The map-matching of the raw GPS data for vehicle trajectories simply associates the vehicular movement with the selected road network segments, for which speed, travel time or congestion can be estimated. Also, there are companies that process the raw vehicle trajectories (sometimes fused with various traffic sensors) and produce commercially available FCD, which simply provide speed/travel time for road networks published in 1-minute intervals in real time (i.e. Be-Mobile, TomTom, etc.). While the quality and the penetration rate of this commercially available FCD has to be addressed properly, it is still a big dataset for developing smart arterial management systems at relatively low costs. Alternatively, simulated vehicle trajectories were also produced based on real ones using a next-generation simulation program (NGSIM) as in, but, these studies focus on more detailed evaluation of errors in prediction in a "controlled" environment rather than real case studies.

Altintasi et al. (2017) focused on the detection of recurrent congestion or bottleneck locations and even the length of queues formed before the bottlenecks. In a follow up study, Altintasi et al. (2019a, 2019b) evaluated the commercial FCD quality in terms of derivation of i) the traffic fundamental diagram by fusing GT and commercial FCD speed, ii) speed transformation functions representing the mathematical relationship between GT and commercial FCD, iii) speed and level of service (LOS) estimation performance, iv.) dominant traffic characteristics of the urban arterial by taking the longer duration of FCD to identify the recurrent bottleneck and congested locations as well as the free-flowing conditions, v.) queue formations and dissipations around bottleneck locations. An in-depth discussion of the FCD studies can be found in Altintasi et al. (2019a, 2019b), but will not be discussed in detail due to limitations.

2.2 ARIMA Models in Transport Studies

Seasonal Autoregressive Integrated Moving Average models require simple training process and offer simple solution as compared to neural networks and nonparametric approaches, thus, for the prediction of traffic speed or travel time, seasonal ARIMA model is one of the most widely used method (Williams and Hoel, 2003). Apart from this, ARIMA models are also implemented within hybrid models for traffic speed prediction problem (Wang, 2015). In addition to hybrid models, ARIMA models are also used in multi-layered models such that the prediction is conducted in multi steps calculation for multiple ARIMA models (Zou et al. 2015). By employing real-time data acquisition devices and the internet, ARIMA models were used for the real-time prediction for traffic state (Xu et al., 2017). Box and Jenkins (1970) set principles for the determination of SARIMA model coefficients and training.

3 Methodology

3.1 Study Corridor and Locations

The study corridor was selected as a long stretch including three different regions: i) an urban arterial starting from the city center, Inonu Boulevard, which turns into ii) an urban transition corridor (Dumlupınar Boulevard section) until the beltway, and finally including iii) an intercity stretch until Polatlı, a nearby town of Ankara. The corridor is 75.1 km in total, when measured from Ankara to Polatlı, consisting of 655 segments with lengths varying in the range of [0.31m; 1595m]. These segments are mainly the ones taken from the OSM database, which are checked and cleaned up to provide a consecutive and continuous network topology. For further analysis on speed estimations, three sub-corridors are selected on the aforementioned three different regions.

3.2 FCD Structure and Preprocessing

FCD used in this study is a commercial dataset which was provided for academic purposes by Parabol Software Inc., as a part of an ongoing research project supported by TUBITAK. The raw data itself has a simple structure with a VehiclelD, longitude (Long), latitude (Lat) and time stamp information as shown Table 1. From the whole Turkish road network, a buffer zone of approximately 1 km wide is defined around the study corridor segments, which is used to extract the GPS track data relevant to the study corridor. Later, the extracted data is sorted by date and VehicleID to be preprocessed via a map matching (MM) algorithm using a readily available open source routing machine (OSRM) algorithm, which employs Multi-level Dijkstra algorithm for route matching (Luxen and Vetter, 2011). This MM algorithm simply associates each GPS track point with an existing segment of the study corridor, if it satisfies the proximity and continuity conditions embedded. The final product of the MM process adds a projected longitude and latitude value on the matched segment ID (a local value given in the study corridor definition), as a part of a route created based on the raw GPS track data. If, in a given raw GPS track data, there are long gaps or missing data not matched with the study corridor, the MM algorithm may produce multiple paths for that vehicle ID. From this point on, all the speed estimations are performed using the map-matched attributes of the FCD data (FCD-MM in short)

Table 1. Geocoded raw and map-matched GPS track data attributes of a vehicle from FCD set

	Raw (R) Attributes			Map-matched (MM) attributes			
Vehicle ID	**Long_R**	**Lat_R**	**Timestamp**	**Long_MM**	**Lat_MM**	**Route ID**	**Segment ID**
1e11q58	32.723686	39.906059	18-11-19 8:11	32.723687	39.906040	1	270
1e11q58	32.726357	39.906174	18-11-19 8:11	32.726359	39.906152	1	271
1e11q58	32.729034	39.906288	18-11-19 8:11	32.729035	39.906282	1	272
1e11q58	32.770481	39.907974	18-11-19 8:55	32.770482	39.907956	2	289
1e11q58	32.772736	39.908077	18-11-19 8:56	32.772738	39.908048	2	289
1e11q58	32.774834	39.908165	18-11-19 8:56	32.774836	39.908134	2	289

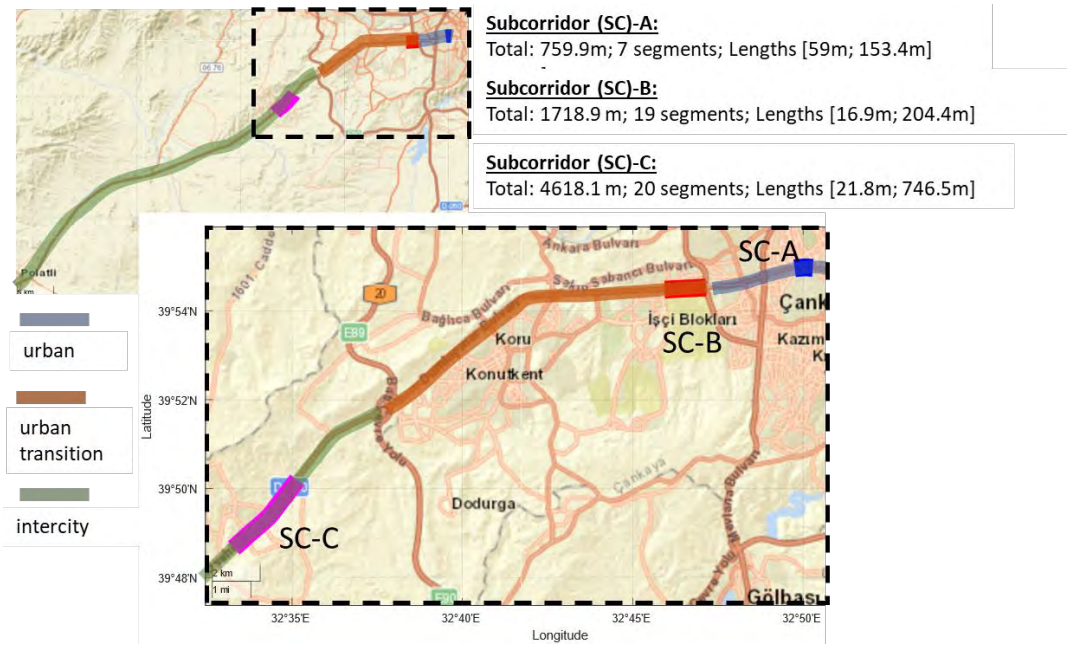

Figure 1. Study Corridor and Analysis Locations on the urban, urban transition and intercity

3.3 Average segment speed calculations

Once the vehicle-specific routes are produced, it is possible to aggregate the MM GPS points over each segment, to find the first and the last GPS point on it. However, as the GPS point locations along a route can be rather randomly distributed over the time and space (i.e. multiple points in one segment versus one or no points on some segments), and it is not necessary to calculate instantaneous speed between two consecutive GPS points, a simple constant speed approximation is made between the last GPS point in a segment and the next GPS point on the route to estimate the approximate time of the vehicle through the start point of the next segment (which is the end point of the current segment as well). The approximate time and space information of a vehicle at the beginning of each segment along its route, led to an augmented MM GPS dataset, in which the same location of each observed vehicle is timestamped accordingly. Time difference between these augmentations simply produces the average time of each vehicle observed a segment, which can be averaged to estimate "space mean speed" of the segment within a given time period.

A key issue is the selection of time windows for average speed calculations. For real-time estimations, it is necessary to create short enough time intervals (i.e. 5 min -15 min) to observe major changes in traffic conditions (i.e. start/end of peak hours, incidents, etc.). However, the main purpose of this study is to focus on long-term (weekly or monthly) or mid-term estimation (daily) of speed values for archival usage, which can be safely and efficiently evaluated in 30-min intervals. Further analysis of short-term speed predictions (15-min or shorter) more computational power as well as discussion of real-time traffic management aspects, which are not within the scope of this study, thus, left for future research.

3.4 Speed estimation using ARIMA Models

Based on the fact that urban traffic networks are mostly exposed to large number of repetitive commute trips that create ever continuing spatio-temporally distributed peak hours and bottlenecks, it is important to analyze the repeating nature of average speeds on road segments over a selected time series. For this purpose, a control data is created by preprocessing the GPS track based FCD data, from the month of December, 2019. The weekdays are selected as a similar dataset; daily average speed profiles for 30-min intervals (creating a total of 48 speed estimation for each road segment) are obtained for a total of 20 days; the scarcity of data for the low volume night period of 8 pm to 6 am are left out to focus on the speed profiles for daytime urban mobility conditions.

Combining a total of 580 average speed values (u_{30}) for a total of 20 weekdays created the time series fed into the ARIMA models run by "*statsmodels*" (Seabold et al., 2010) and Box-Jenkins procedure was followed. Using the 3-week daytime urban activity periods speed values, an ARIMA model is developed for each road segment. Use of this ARIMA model allows different options such as i) a long-term (LT) full week speed estimations $(\hat{u}_{30\_LT})$ can be made in advance (LT) and ii) a short-term (ST) daily speed estimations $(\hat{u}_{30\_ST})$ using the updated data (available until the day before the estimation)

3.3 Measures of effectiveness (MoEs) for speed estimation

Measures of effectiveness selected for the evaluation of quality of FCD speed included:
i) *MAPE* proposed in (Wang et al., 2014; Hu et al, 2016) and determined as Equation 1

$$\mathrm{MAPE} = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{\hat{u}_{30} - u_{30}}{u_{30}}\right| \tag{1}$$

ii) *RMSE* proposed in (Wang et al., 2014) and calculated as shown in Equation 2

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{u}_{30} - u_{30}\right)^2} \tag{2}$$

where *T* shows the analysis periods (number of 30-minute periods)

4 Numerical Results

While analyzing the results, first daily average speed profiles estimated from FCD will be provided as a base for the upcoming discussions on MoEs and corridor characteristics.

4.1 Daily Average Speed Profiles

For visual depiction of the daily speed profiles for different road segments, a segment from each sub-corridor (SC-A Segment 14, SC-B Segment 51 and SC-C Segment 189) are studied in detail showing the third week (end of training data) and fourth week (test data) average speeds, u_{30} values as shown in Figure 2. The x-axis shows the number of 30-min periods for the analysis period. As expected, there is a much stronger cyclic behavior in the speeds, where much slower speeds (<30 km/hr) are observed during the peak hours. Since there is a strong directionality during the peak hours, there is only one peak pattern in the Kızılay-Polatlı direction. Segments in the sub-corridors A and B have more repetitive pattern than the one in the intercity one in SC-C.

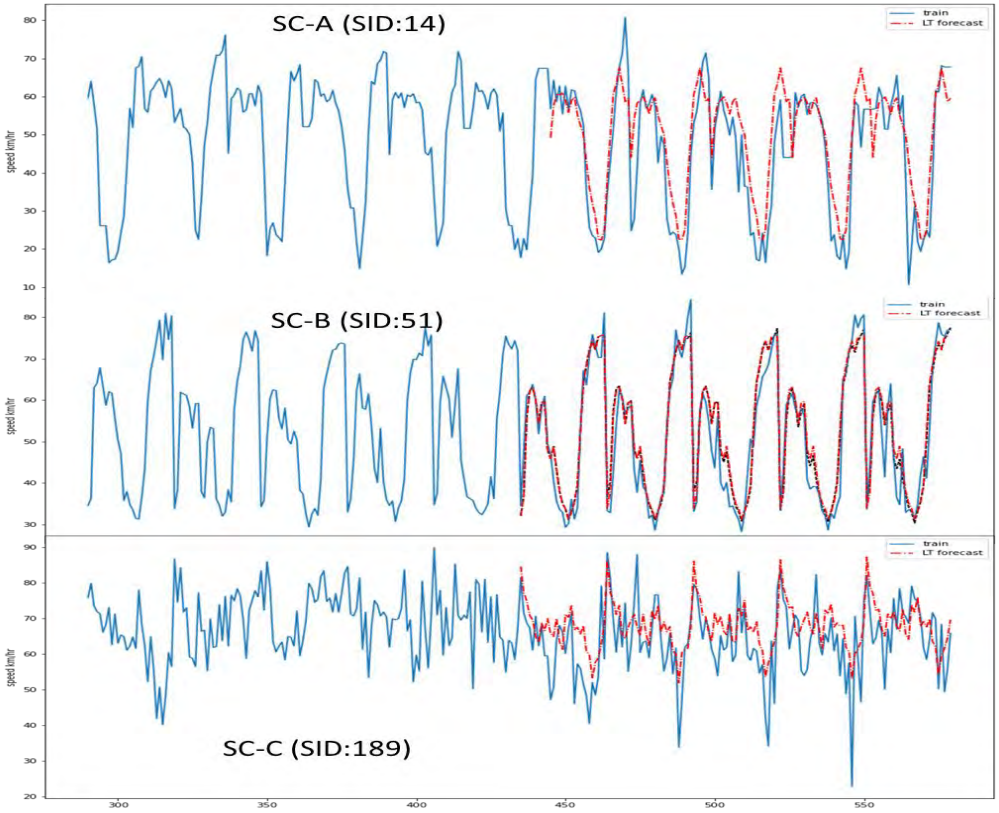

Figure 2. Daily average speed profiles for three selected road segments (SID= 14; SID=51 and SID=189)

4.2 ARIMA based Speed Estimations

The segment-specific ARIMA models are used to estimate the weekly expected speed profiles for the test week with no updated daily information (as LT estimations) as shown in the right-hand sides of the graphs in Figure 2. As can be seen, the cyclic behavior of the urban road segments in SC-A and SC-B provide much closer estimated speed, whereas the intercity segment on SC-C shows more variability in real data compared to the estimated ones. As an alternative, ST speed estimations for the next day produces similar results for the urban segments, compared to much larger variations in the intercity segment (See Figure 3)

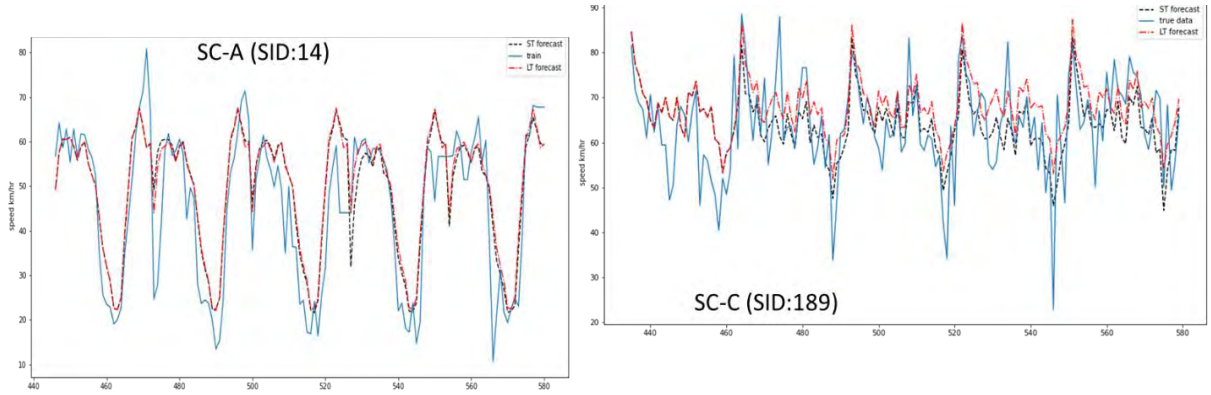

Figure 3. Comparisons of ST and LT speed predictions for study segments for urban and intercity regions

4.3 MoEs for Speed Estimations

As it is not possible to visually portray speed estimations, which may vary spatio-temporally, a more effective way of assessing the success of the estimations would be the monitoring of the MAPE and RMSE values for various conditions. Figure 4 portrays the distribution of these MoEs as a function of the segment length for all three sub-corridors for each estimation day. As ST speed estimations are expected to be better than the LT ones, the comparisons are made for the former case, which suggested that on the SC-C intercity region, MAPE values would be generally smaller than the 0.15, whereas they can reach to 0.30s for relatively shorter segments in the SC-A and SC-B locations. There are more significant differences among observation days than the length of the segments. But, analysis of RMSE values for the same locations suggested that this pattern is not as strong as in the case of MAPE relation to day of the estimation and length of the segment. One main reason behind this is the high sensitivity of MAPE measure in small observed values. For example, 5 km/hr difference for a real average speed of 25 km/hr would result in 20% error, where as it would be only 6.25% error for high intercity travel speed of 80 km/hr.

Furthermore, to locate the time intervals leading to more errors, the overall daily error values are further decomposed into peak (7 am-10 am and/pr 4 pm-7 pm) and off-peak periods as shown in Table 2. Peak period errors are sometimes smaller than the off-peak periods as congested traffic regimes minimize the variability in travel speeds spatio-temporarily, and thus make it easier to predict more correctly.

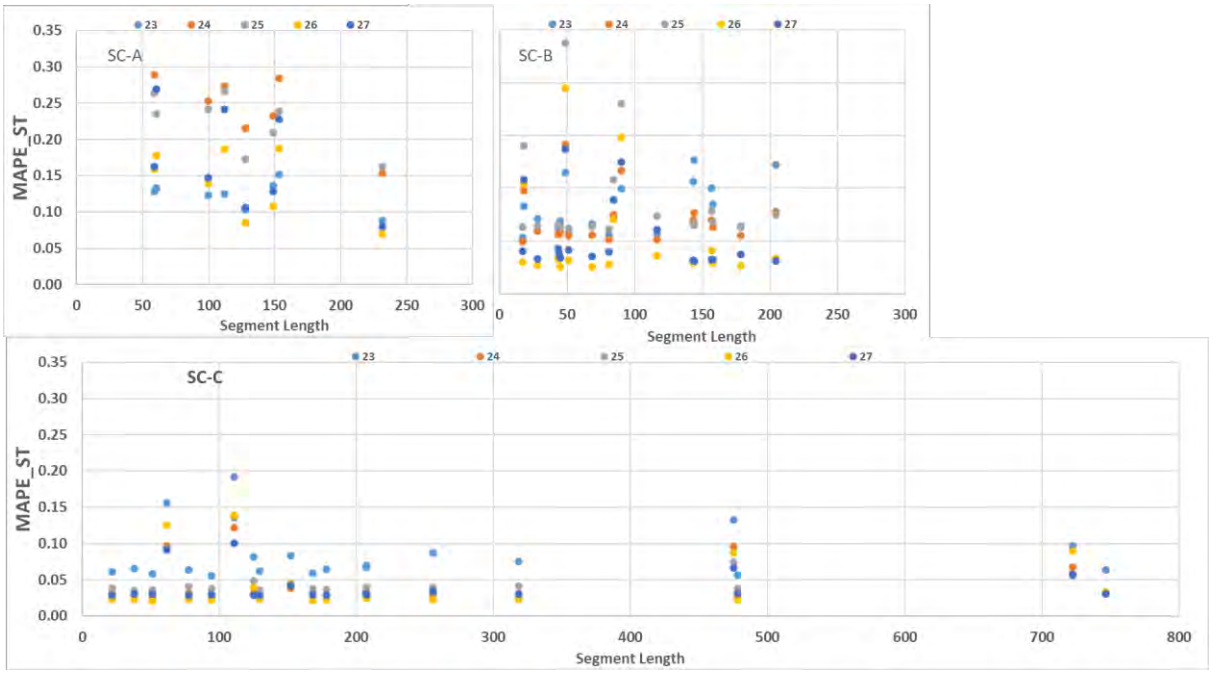

Figure 4a. MAPE values for different segment lengths in three study locations

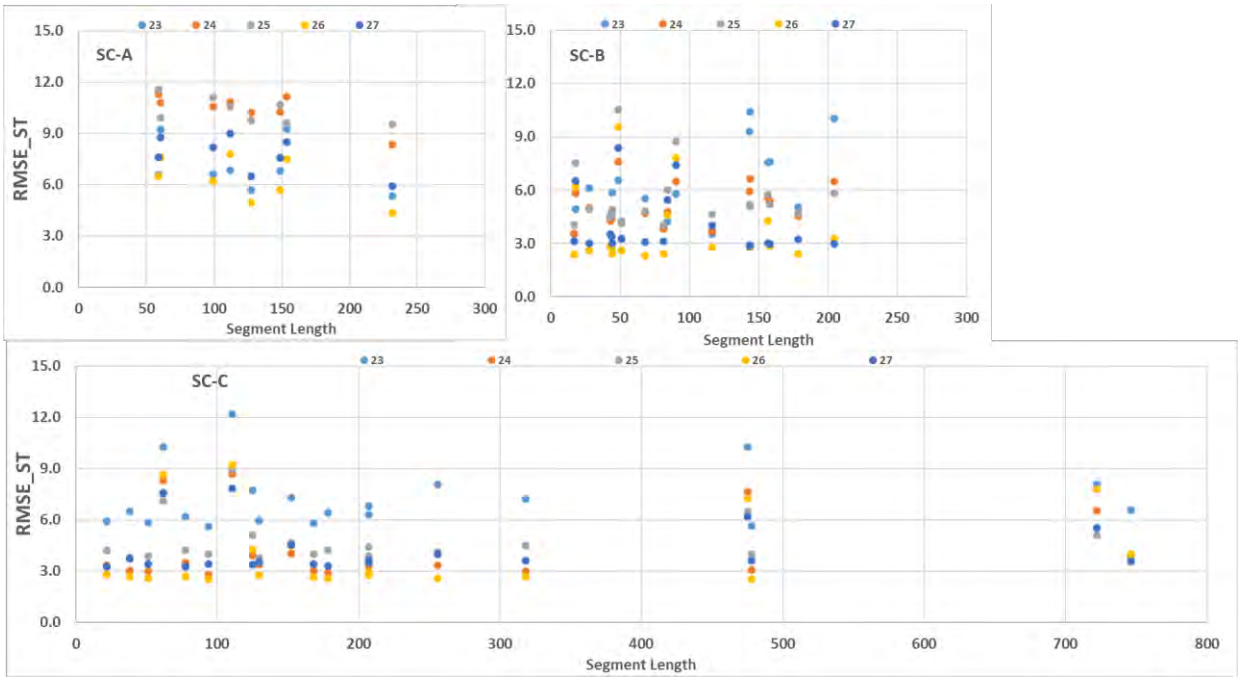

Figure 4b. RMSE values for different segment lengths in three study locations

Table 2. Distribution of MoEs values for different study locations and periods

	overall				peak hours				off peak hours			
Day	$MAPE_{ST}$	$MAPE_{LT}$	$RMSE_{ST}$	$RMSE_{LT}$	$MAPE_{ST}$	$MAPE_{LT}$	$RMSE_{ST}$	$RMSE_{LT}$	$MAPE_{ST}$	$MAPE_{LT}$	$RMSE_{ST}$	$RMSE_{LT}$
	SC-A											
day_23	0.08	0.08	7.23	7.23	0.09	0.09	7.40	7.40	0.08	0.08	7.05	7.05
day_24	0.04	0.07	4.17	6.23	0.05	0.09	4.84	7.44	0.03	0.05	3.35	4.80
day_25	0.05	0.08	4.69	7.48	0.06	0.10	5.13	8.80	0.04	0.06	4.23	5.98
day_26	0.04	0.09	4.01	8.30	0.06	0.11	4.75	9.40	0.03	0.08	3.12	7.09
day_27	0.04	0.07	4.21	7.08	0.05	0.09	4.76	7.80	0.03	0.06	3.59	6.31
	SC-B											
day_23	0.08	0.08	5.97	5.97	0.03	0.03	2.96	2.96	0.13	0.13	7.76	7.76
day_24	0.07	0.07	5.15	4.99	0.05	0.05	4.46	4.15	0.09	0.09	5.64	5.58
day_25	0.09	0.09	5.53	6.10	0.08	0.08	5.78	6.13	0.10	0.11	5.21	5.99
day_26	0.05	0.07	3.67	4.56	0.04	0.05	3.76	4.28	0.06	0.08	3.47	4.69
day_27	0.05	0.06	3.96	4.02	0.05	0.05	4.31	4.33	0.05	0.06	3.50	3.48
	SC-C											
day_23	0.12	0.12	7.04	7.04	0.09	0.09	7.88	7.88	0.16	0.16	5.90	5.90
day_24	0.25	0.24	10.44	10.08	0.09	0.09	6.51	6.57	0.41	0.41	13.43	12.81
day_25	0.22	0.23	10.34	10.38	0.21	0.21	11.94	11.72	0.24	0.25	8.24	8.64
day_26	0.14	0.14	6.33	5.95	0.06	0.06	4.66	4.29	0.22	0.22	7.65	7.20
day_27	0.17	0.18	7.75	8.01	0.09	0.09	6.18	6.32	0.26	0.28	8.99	9.33

5 Conclusions and Further Recommendations

As can been seen in the detailed evaluation of speed long-term and mid-term speed estimations based on FCD with vehicle trajectories, estimation errors may significantly be affected from the length of the segment, location of the estimated segment and general traffic regime characteristics. The longer segments in the intercity regions may have smaller MAPE values but not guarantee better speed estimation because of the length of the segment per se; instead it is possible to be governed with the traffic regime. Alternatively, smaller links in the city center does not necessarily guarantee better estimation solution, as they may have more vehicles to pass but create a bigger variability. It can be, however, suggested that combining smaller links to form longer urban segments may not

endanger the prediction quality of the road segments, but, it may decrease the computational and archiving needs for large urban areas. Furthermore, urban network segments show a more definite repeating pattern which can be predicted for long-term periods (such a week in advance) and used as archival data for road segments to aid real-time travel state prediction when data scarcity or connection problems occur. OSM network topology does not necessarily comply with the needs of the traffic state estimation, thus, should be processed further to optimize speed/travel time estimation potential of FCD network.

Acknowledgement

This paper is produced as a part of an TUBITAK 1501 project titled “Konum verisi ile ulaşım analiz, planlama ve modelleme için büyük veri platformu geliştirilmesi”.

References

Altintasi, O., Tuydes-Yaman, H. &Tuncay, K. (2019a). Quality of floating car data (FCD) as a surrogate measure for urban arterial speed. *Canadian Journal of Civil Engineering. 46(12)*, 1187-1198. https://doi.org/10.1139/cjce-2018-0422.

Altintasi, O., Tuydes-Yaman, H. &Tuncay, K. (2019b). *Monitoring Urban Traffic from Floating Car Data (FCD): Using Speed or a Los-Based State Measure*. Directions of Development of Transport Networks and Traffic Engineering. TSTP 2018. Lecture Notes in Networks and Systems, 51:163-173. Springer, Cham

Altintasi, O., Tuydes-Yaman, H. &Tuncay, K. (2017). Detection of urban traffic patterns from Floating Car Data (FCD). *Transportation Research Procedia*, *22*, 382-391.

Box, G. E. P., & Jenkins, G. M. (1976). Time series analysis: Forecasting and control. San Francisco: Holden-Day.

Hu, J., Fontaine, M.D., Ma, J. (2016). Quality of private sector travel-time data on arterials. *Journal of Transportation Engineering*, 142(4):04016010. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000815.

Luxen D. and Vetter, C. (2011). Real-time routing with OpenStreetMap data. *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Association for Computing Machinery, New York, NY, USA, 513–516. doi: https://doi.org/10.1145/2093973.2094062

Seabold, Skipper, and Josef Perktold. “statsmodels: Econometric and statistical modeling with python.” Proceedings of the 9th Python in Science Conference. 2010.

Talebpour, A., & Mahmassani, H. S. (2016). Influence of connected and autonomous vehicles on traffic flow stability and throughput. *Transportation Research Part C: Emerging Technologies, 71,* 143-163. https://doi.org/10.1016/j.trc.2016.07.007.

Wang X., Liu, H., Yu, R., Deng, B., Chen, X., &Wu, B. (2014) . Exploring operating speeds on urban arterials using Floating Car Data: Case study in Shanghai. *Journal of Transportation Engineering*, *140(9).* https://doi.org/10.1061/(ASCE)TE.1943-5436.0000685.

Wang, H., Liu, L., Dong, S., Qian, Z., & Wei, H. (2015). A novel work zone short-term vehicle-type specific traffic speed prediction model through the hybrid EMD–ARIMA framework. Transportmetrica B: Transport Dynamics, 4(3), 159–186. doi:10.1080/21680566.2015.1060582

Williams, B. M., & Hoel, L. A. (2003). Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results. Journal of Transportation Engineering, 129(6), 664–672. doi:10.1061/(asce)0733-947x(2003)129:6(664)

Xu, Dw., Wang, Yd., Jia, Lm. et al. Real-time road traffic state prediction based on ARIMA and Kalman filter. Frontiers Inf Technol Electronic Eng 18, 287–302 (2017). https://doi.org/10.1631/FITEE.1500381

Zou, Y., Hua X., Zhang Y., and Wang Y. Hybrid short-term freeway speed prediction methods based on periodic analysis. Canadian Journal of Civil Engineering. 42(8): 570-582. https://doi.org/10.1139/cjce-2014-0447