LARGE SCALE POINTS-TO-POINTS HIGHWAY LATENCY PREDICTION

Baiyue HE [a], Zehui LUO [b], Man Fung HO [c], Zhiquan LAO [d],
Tat Shing CHOI [e] and K. Y. Michael WONG [f]
Department of Physics, The Hong Kong University of Science and Technology, China
Email: baiyue.he@connect.ust.hk [a], *zluoap@connect.ust.hk* [b], *mfho@connect.ust.hk* [c],
zlaoaa@connect.ust.hk [d], *tschoiab@connect.ust.hk* [e], and *phkywong@ust.hk* [f]

ABSTRACT

We consider the simultaneous prediction of traffic latencies among various locations in metropolitan freeway systems, addressing the following new challenges. (1) We introduce flexibility in data engineering to tackle the non-uniform availability of latency data, which is especially significant for long-distance journeys, whose infrequent traffic leads to highly fluctuating and even absence of latency data in the monitoring period immediately before the instant of prediction. (2) We introduce an estimation technique supplementing actual latency measurements with historical data to deal with the segment-wise latencies in different freeway segments, such that the collected latency data balances the need for timely predictions and data sufficiency. (3) We compare the segment-added prediction with the single prediction. Numerical experiments showed that they are comparable for short distances, but the single predictor is more superior for long distances. This shows that a hybrid approach is suitable for system-wide latency predictions.

Keywords: points-to-points latency prediction, machine learning, single predictors, segment-added predictors, XGBoost

1 INTRODUCTION

Highway traffic latency prediction is an important issue to many stakeholders including commuters, logistics companies, navigation information providers, traffic controllers and designers of intelligent transportation systems (Zhang *et al*., 2011). With vast amount of historical data, there were numerous attempts to develop latency prediction techniques through machine learning (Vlahogianni, 2014). Typical studies focused on isolated highway segments of several kilometers (see, for example, Abbas *et al*. 2018), but in typical metropolitan areas, there are many applications demanding the simultaneous prediction of latencies among various combinations of origin-destination (OD) pairs spanning over 10 km. This poses the following new challenges to the development of latency prediction algorithms, and will be addressed in this paper.

First, since machine learning approaches rely on instantaneous measurements of latency to train the predictors and make predictions, the non-uniform availability of latency data is a problem. This is especially significant for long-distance OD pairs, for which infrequent traffic leads to highly fluctuating and even absence of latency data in the monitoring period immediately before the instant of prediction.

Second, as different highway segments have different segment-wise latencies, a uniform set of criteria is required to collect median latency data in a timely manner, such that the collected latency data are not too early in time for timely predictions, and not too late in time to avoid data insufficiency. When actual data during a monitoring period is not sufficient to derive the median latency, how will historical data supplement the estimation?

Third, for highway systems with multiple locations, it is tempting to divide the latency prediction task among the individual segments, and perform an addition of the individual latencies when the journey to be predicted consists of a few segments. For a highway system with N locations, this reduces the number of prediction tasks from $O(N^2)$ to $O(N)$, but the prediction accuracy may be compromised. A comparative study is necessary to assess the feasibility.

2 DATA ENGINEERING OF INFREQUENT OUTPUTS

The data used in this study is collected from the Electronic Toll Collection (ETC) system installed along the Sun Yat-Sen Freeway (Freeway Bureau, 2019). While the use of high-density sensors in short-term traffic prediction is increasingly popular in recent years (Hubert *et al.* 2020), the ETC data, typically spaced at distances of the order of 1 km, has no extra hardware requirements but is more challenging for predictions. The system records the detection time and vehicle ID as the vehicles pass the detectors, enabling us to estimate various descriptive variables of the traffic. We focus on a stretch of the freeway spanning from the east side of Taipei city to its suburb on the west, traversing the city center. This freeway stretch consists of 9 segments demarcated by 10 detectors. In this paper, the detectors are labeled by 3-digit numbers representing their east-to-west freeway positions measured in units of 0.1 km: 155, 182, 248, 293, 339, 376, 413, 467, 509. For example, the distance between the eastmost and westmost detectors (155 and 509, respectively) is 35.4 km.

Our data-driven prediction task is based on the westward traffic data collected from 6 am to 11 pm during weekdays for the entire year of 2017. We build XGBoost predictors (Chen & Guestrin, 2016) to predict the median latency for vehicles in the next 5 minutes for the OD pairs between the 10 detectors. For each OD pair, the input data used for prediction includes:

(1) influx at all detectors along each segment between the OD pairs,
(2) accumulation: total number of vehicles in each segment between the OD pairs,
(3) Gaussian process regression: the expected latency at a particular time of a weekday, as obtained by Gaussian process regression of the yearlong latency data (Wilson & Adam, 2013),
(4) Past history of median speed at all detectors along each segment between the OD pairs (to be explained in Section 3).

The above data is similar to the single-journey version of our previous work (Kwok *et al*. 2019, Li *et al*. 2021), but is further engineered for extension to large-scale predictions. The training of the XGBoost predictors is implemented by dividing the data of the 260 weekdays of 2017 into 5 sets, and each set becomes the test set of the predictor trained by minimizing the mean square error (MSE) of the predictions of the other 4 sets.

First, the raw output data of the examples used for training and testing the prediction algorithm has to be refined. Defined as the median latency of the vehicles in the next 5 minutes, the data is easy to collect for journeys with frequent traffic. However, for journeys with infrequent traffic, the median latency may be highly fluctuating. In the most extreme case, such data may even be absent if no vehicles are passing in the 5 minutes of the monitoring period. Such data inaccuracy or insufficiency often happens for a long segment with a long travel time required to reach the destination, especially during off-peak hours. Consequently, it leads to spurious spiky predictions as illustrated in Fig. 1(a).

Thus, to tackle this problem, we have broadened the data collection criterion to promote a more reliable estimation of median latency. Namely, we require the number of vehicles used for estimating the median latency to be not less than a threshold. That is, the period of estimation needs to be extended beyond 5 minutes until the number of vehicles reaches a threshold if necessary. As shown in Fig. 1(b), this seemingly simple criterion significantly eliminates the spurious predictions. Figure 2 shows that the predictor trained with the modified output performs much better than the one trained with raw output.

3 DATA ENGINEERING OF HISTORICAL LATENCY

Past history of median speed is an important input to the prediction task. To estimate the median latency (and hence the median speed) of vehicles through a detector at an instant from past history, a time interval previous to the instant needs to be defined so that the latencies of the vehicles originating from the upstream detector during the instant can be monitored, and their median can be obtained.

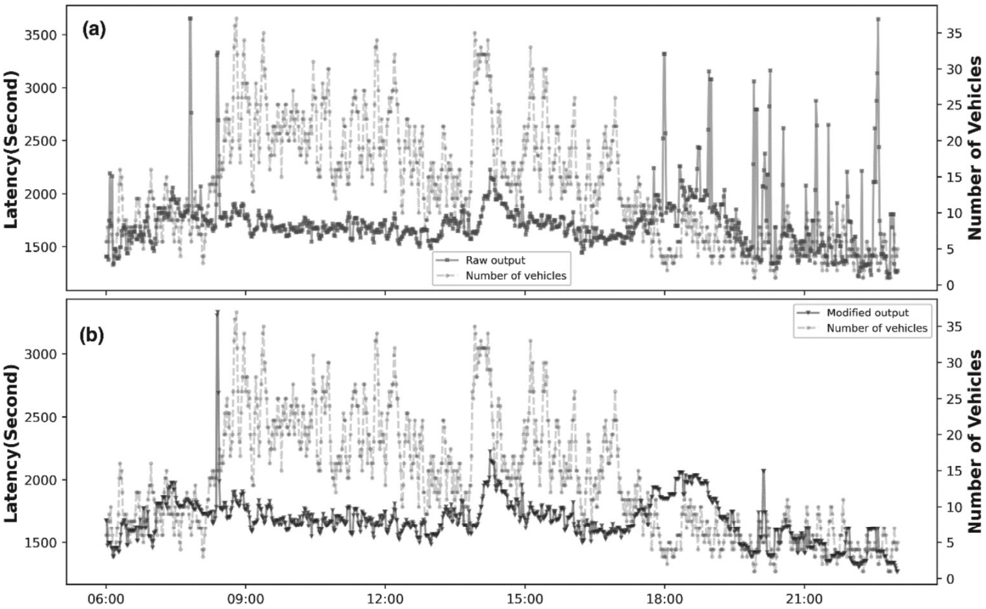

Figure 1 (a) Raw output of median latency from 155 to 509 during 06:00-23:00 on 2 March 2017 (black). (b) Modified output using a threshold of 15 vehicles during the same period (black). The grey data in both panels are the number of vehicles in each 5-minute interval, showing that the raw latencies are spurious when the estimation is based on less than 15 vehicles.

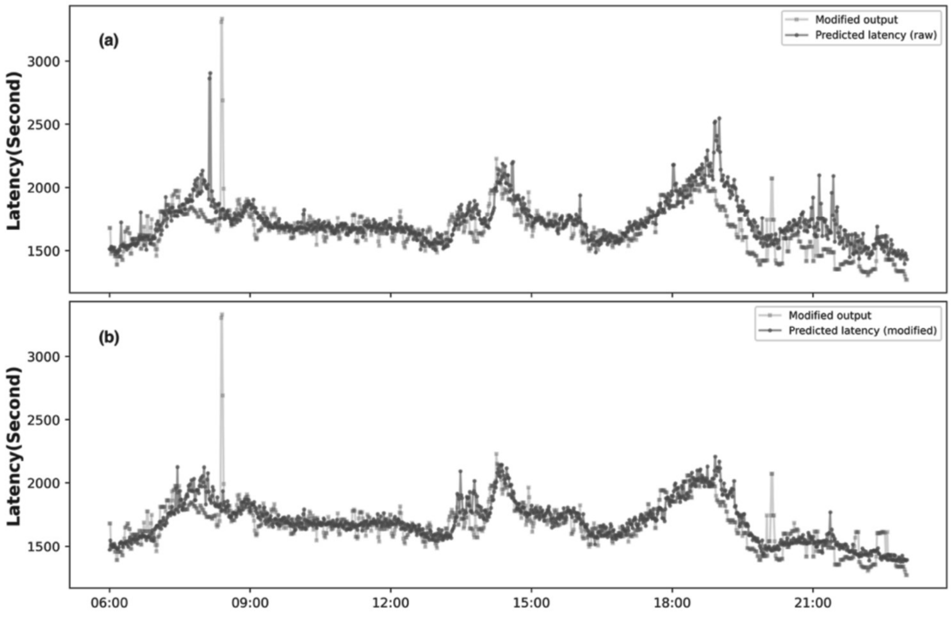

Figure 2 (a) Predicted median latency from detector 155 to 509 during 06:00-23:00 on 2 March 2017, trained with raw output (black). (b) Predicted latency during the same period, trained with modified output (black). The grey data in both panels are the modified output as shown in Fig. 1(b).

For convenience of discussion, we consider the estimation of the median latency at a detector at instant t based on the median latency of vehicles through the upstream detector during a one-minute time interval at the instant $t - p$. As different freeway segments have different segment-wise latencies, a uniform set of criteria is required to collect the median latency data in a timely manner, such that the collected latency data are not too late in time to avoid data insufficiency, and not too early in time for timely predictions. For example, with a speed limit of 100 km/h, choosing the initial time $p = 5$ min is too late because it is impossible for a vehicle to travel 17 km from detector 339 to 509 within 5 min, but the choice of $p = 30$ min may be too early to keep track of changes in traffic conditions.

The appropriate choice of the initial time p at a detector is made by collecting the yearlong latency data traversing the upstream segment and plotting the cumulative frequency curve of the latencies. The values of latencies corresponding to 50% and 84% of the cumulative frequency are used to construct two inputs of historical latency for prediction. This enables an adaptive determination of the initial time according to the length of the freeway segment, such that a sufficient percentage of vehicles is able to arrive at the considered detector within the designated time interval, thereby reflecting the timely freeway condition.

In real-time determination of the median latency, it is straightforward to read the value of the median latency from the data when more than 50% of vehicles have arrived at the designated detector before the instant t. However, we further need to deal with the case that the arrival percentage is less than 50%. As shown in Fig. 3, we expect that the arrival percentage is strongly correlated with the median latency. Hence, we plot the measured median latency versus the arrival percentage using the yearlong data, and construct a look-up table for each bin of size 1%, mapping the arrival percentage to the median of the data points in each bin. Compared with cruder estimates of median latencies of those yet-to-arrive vehicles, such as the fixed average of historical median latencies of the yet-to-arrivals (similar to Kwok *et al*., 2019), observations support that the look-up tables perform better in moderately heavy traffic, but further improvement is still required in extremely heavy traffic.

As another consideration of historical effects, we consider whether downstream or upstream traffic conditions facilitate the latency prediction in a segment. For example, if there is a congestion downstream, there will be a congestion wave propagating upstream, causing the drivers in a segment to slow down (Choi *et al.* 2019). To this end, we build single-segment predictors using the input data intended for the downstream or upstream segment in addition to the input data for the current segment. For example, to predict the latency of the segment 182 to 248, we also include the inputs intended for the segment 248 to 264. As shown in the mean absolute percentage error (MAPE) data in Fig. 4, it is effective to include traffic data of downstream and upstream segments for prediction. The MSE of the predictions also indicate that including downstream and upstream inputs improve the accuracy, except that the MSE data points have more outliers.

4 SINGLE PREDICTORS VERSUS SEGMENT-ADDED PREDICTORS

For freeway systems with *N* locations, it is tempting to divide the latency prediction task among the individual segments, and perform an addition of the individual latencies when the journey to be predicted consists of a few segments. This reduces the number of prediction tasks to O(N) and avoid the much more computationally intensive task of providing O(N^2) predictions. On the other hand, the multiple operations may introduce more errors. Hence, we perform a comparative study of the segment-added predictions with the single predictor approach.

As shown in the MAPE data in Fig. 5, the segment-added approach is observed to be comparable with the single predictors for short distances, but for long distance predictions, the MAPE of the segment-added approach is considerably larger than the single predictor approach.

The MSE data in Fig. 6 also confirms the advantage of the single predictor approach for long distance predictions, but shows that it has no advantage for short distance predictions. Given that the MSE may be more susceptible to outlying data, the single predictor approach is still preferred for accurate latency predictions, although its computational complexity is heavier.

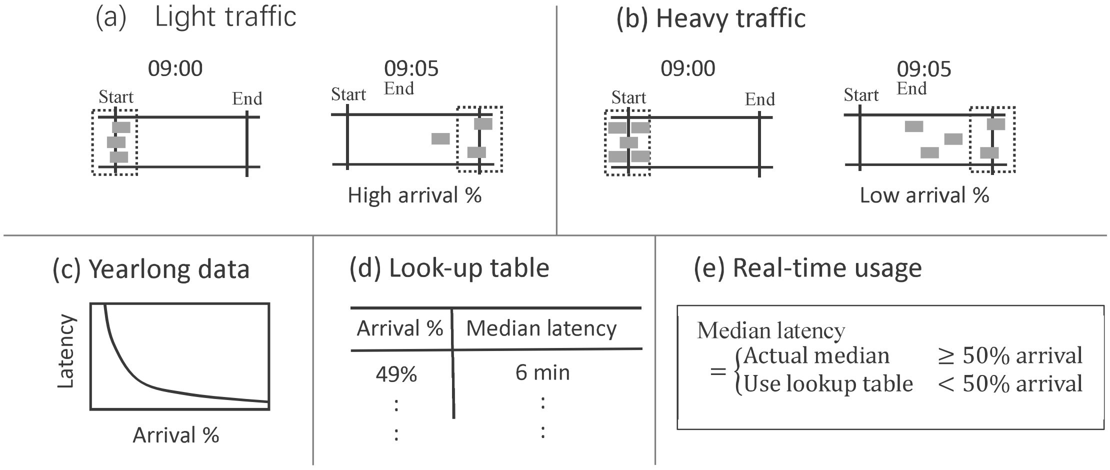

Figure 3 (a)-(b) Illustration of the case with the initial time $p = 5$ min when the traffic is light (a) and heavy (b), showing that the arrival percentage is expected to decrease with the traffic load, and hence the median latency. (c) A plot of measured median latency versus the arrival percentage is prepared using the yearlong data. (d) A look-up table is constructed. (e) The look-up table is used when the arrival percentage is less than 50%.

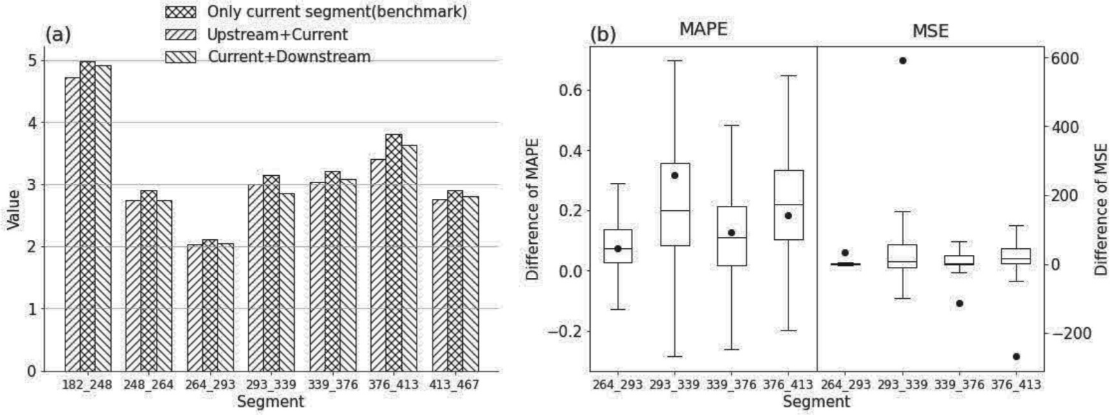

Figure 4 (a) The MAPE of latency prediction of the 9 segments. (b) The boxplot of the MAPE (left) and MSE (right) of the predictor using the current segment input only, subtracted by the MAPE (left) and MSE (right) of the predictor using the current and downstream inputs. The dots are the means of the distributions.

To get further insights into the comparative study, we plot in Fig. 7 the actual latencies used to train both predictors as well as the predicted latencies along the longest segment as a function of time of the day, averaged over the entire year. The actual latencies used to train the single predictor are collected exclusively from those vehicles that traverse from the first detector to the last one. For the segment-added predictor, the training data for each segment are collected from those vehicles traversing the considered journey, as well as those engaged in other shorter journeys. The sum of these actual latencies is displayed in Fig. 7 for comparison. We observe that the predicted results have an excellent agreement with the training data in both cases, showing that the XGBoost predictors are able to provide unbiased estimates of the training data, and the prediction errors mainly arise from the variances of the estimates. However, the sums of the latencies in the training data of the segment-added predictors are significantly lower than the training data for the single predictor.

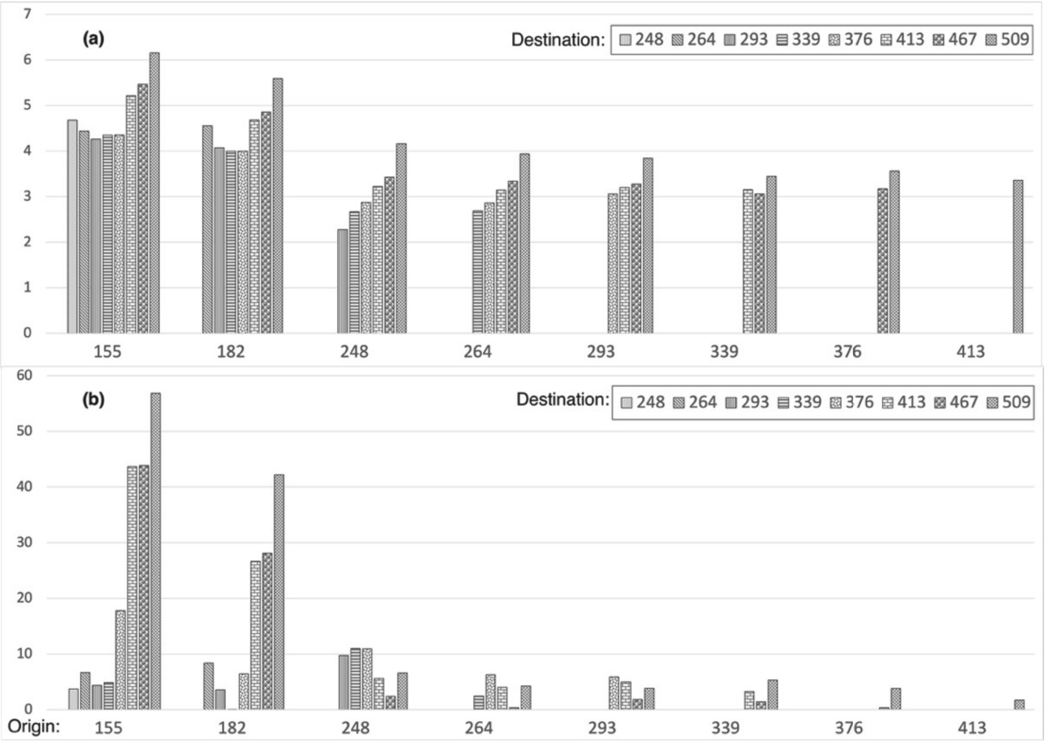

Figure 5 (a) MAPE of the single predictors for OD pairs in the freeway stretch of this study. (b) Percentage advantage of MAPE in using single predictors compared with segment-added predictors.

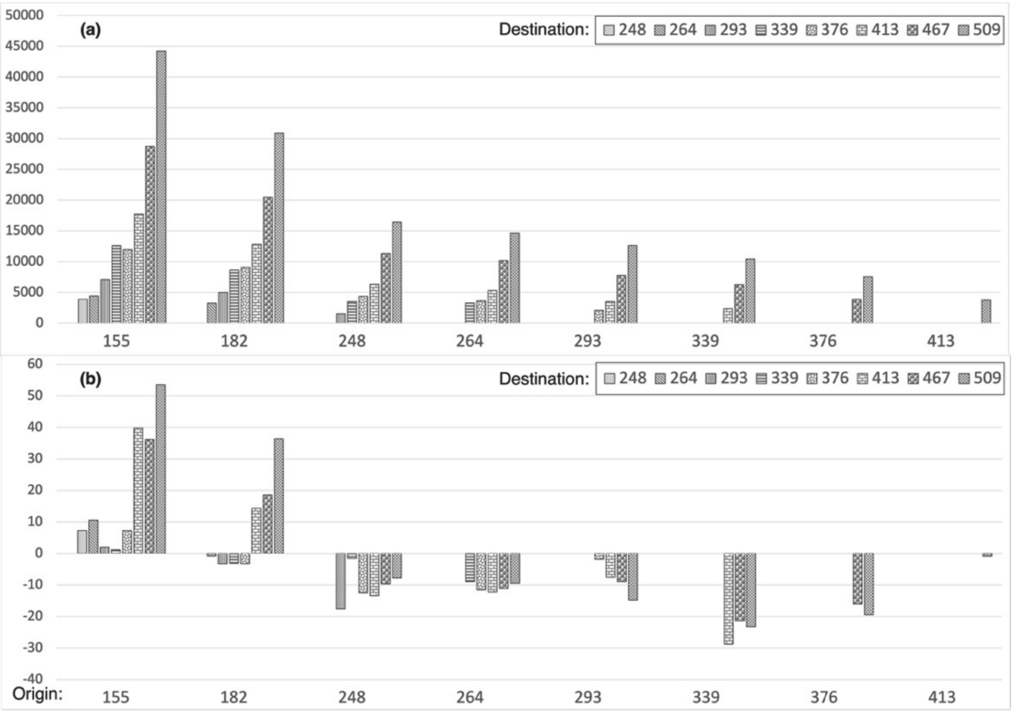

Figure 6 (a) MSE (in sec^2) of the single predictors for OD pairs in the freeway stretch of this study. (b) Percentage advantage of MSE in using single predictors compared with segment-added predictors.

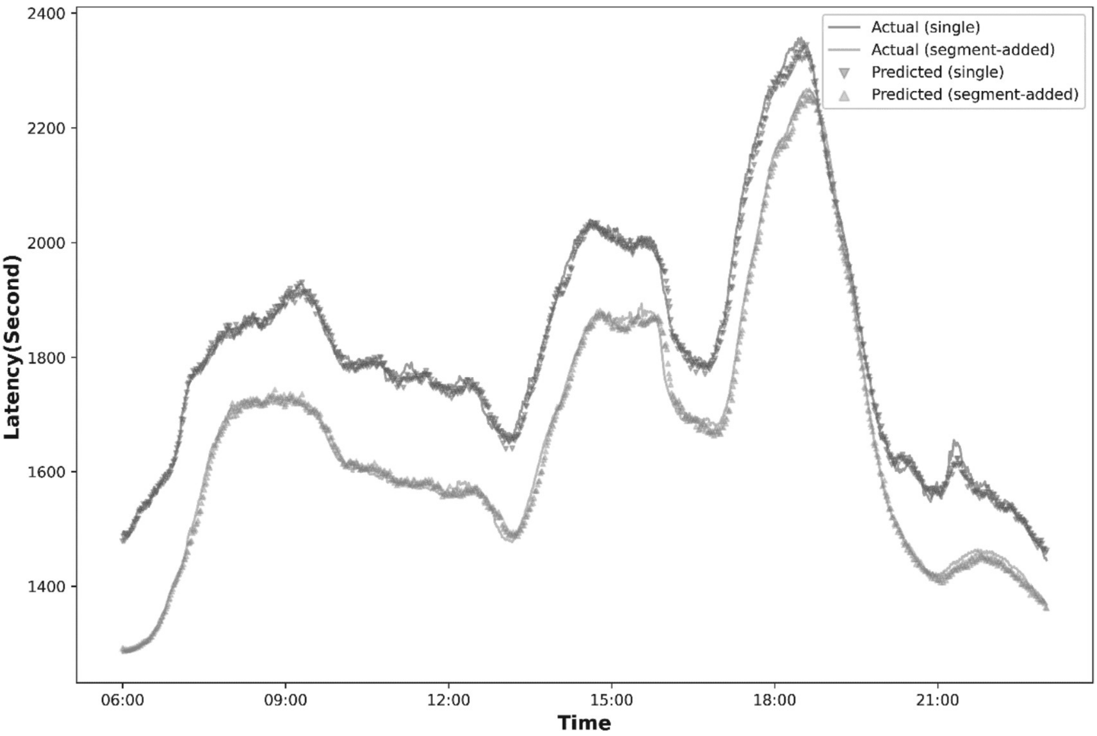

Figure 7 Actual latencies and predicted latencies for the single predictor and the segment-added predictor along the journey from detector 155 to 509 as a function of time of the day, averaged over the entire year.

Thus, this comparative study reveals a fundamental difference in the nature of the data between single journeys and segment-added journeys, even if perfect predictions are available. Specifically, single journeys consist of traveling records traversing from the origin to the destination exclusively, whereas segment-added journeys are the collections of those traveling records in each segment, including all possible journeys through that segment, local and system-wide. The different types of journeys may have their own statistical features, leading to discrepancies in the two approaches. The discrepancies may not be large if the journey whose latency is to be predicted shares the same features as the background traffic. As illustrated in Fig. 5(b) and Fig. 6(b), the MAPE and MSE differences of the journeys originated from detector 155 to the next three detectors are relatively minor, whereas there is a sharp rise for those journeys destinated at 413 and beyond. This may be due to the different traffic characteristics within the city (from detector 155 to 339) and its western suburb (from detector 413 to 559).

5 CONCLUSION AND OUTLOOK

We have considered the new challenges when single-journey latency predictions are to be extended to large scale, points-to-points latency predictions. First, latency data may not be uniformly available in different locations and OD pairs, especially for those with infrequent traffic or separated by long distances. It is difficult to apply a uniform criterion to all data collection steps. Flexible criteria need to be introduced to ensure that the collected data is statistically meaningful so that spurious predictions can be reduced. In our example, relaxing the criterion of measuring the median latency in a fixed period already results in a significant elimination of spiky predictions.

Second, for real-time measurements of quantities such as the median latency, it is important to determine the appropriate time interval for data collection so as to ensure that sufficient data is collected. To enhance the responsiveness of the prediction, the time interval cannot be too remote from the instant of prediction. To cope with the situation that the median latency is indeterminate due to an arrival percentage being less than 50%, we show that it is effective to use look-up tables based on historical data mapping from the arrival percentage to the expected median latency. We also found that adding the input data from downstream and upstream segments to the current segment can improve the predictions.

Third, we found that the performance of segment-added predictors is comparable to single predictors for short distances, but is inferior for long distance predictions. The performance difference is especially significant when the journeys to be predicted span geographical areas with different traffic characteristics. This implies that a hybrid approach combining single predictors for long-distance predictions with segment-added predictors for short-distance predictions may provide a trade-off between prediction accuracy and computational complexity. To implement this, the short-distance locations should be clustered with similar traffic characteristics. This will be an interesting topic for further investigation.

6 ACKNOWLEDGEMENT

This work is supported by grants from the Research Grant Council of Hong Kong (16302419 and 16302619).

7 REFERENCES

Abbas, Z., Al-Shishtawy, A., Girdzijauskas, S., & Vlassov, V. (2018). Short-Term Traffic Prediction Using Long Short-Term Memory Neural Networks. *2018 IEEE International Congress on Big Data (BigData Congress)*, 2018, pp. 57-65.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785-794.

Choi, T. S., To, K., & Wong, K. Y. M. (2018). Segment-wise description of the dynamics of traffic congestion. arXiv preprint arXiv:1808.01627.

Freeway Bureau (2019). *Traffic Data Collection System.* MOTC of the Taiwan ROC TDCS. Electronic database viewed 1 August 2019. http://tisvcloud.freeway.gov.tw/

Hubert, R., Koller, M., & Kaufmann, S. (2020). Data-Driven Traffic Engineering: Understanding of Traffic and Applications Based on Three-Phase Traffic Theory. Elsevier, Amsterdam.

Kwok, T. H., Xu, Y., Wong, T. C. T., Choi, T. S., & Wong, K. Y. M. (2019). Latency prediction based on real time data in the Taiwan highway system. *Proceedings of the 24th International Conference of Hong Kong Society for Transportation Studies*, pp. 491-498.

Li, Y., Chen, J., Wong, T. C. T., Kwok, T. H., Choi, T. S., & Wong, K. Y. M. (2021). Latency Prediction of Traffic Data Improved by Input-Output Dependent Pre-Clustering. *Proceedings of the 25th International Conference of Hong Kong Society for Transportation Studies,* pp. 340-348.

Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2014). Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43, pp. 3-19.

Wilson, A., & Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. *Proceedings of the 30th International Conference on Machine Learning* 28(3), pp. 1067-1075.

Zhang, J., Wang, F. Y., Wang, K., Lin, W. H., Xu, X., & Chen, C. (2011). Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems,* 12(4), pp. 1624-1639.