hEART 2020 – 9th Symposium of the European Association for Research in Transportation
September 1–3, 2020, École Nationale des Travaux Publics de l'État (ENTPE), Lyon, France

Enhancement of SPaT-messages with Machine Learning based Time-To-Green predictions

Alexander Genser[1*] Lukas Ambühl[1] Kaidi Yang[2] Monica Menendez[3]
Anastasios Kouvelas[1]

[1] Traffic Engineering Group, Institute for Transport Planning and Systems
Swiss Federal Institute of Technology (ETH) Zurich, Switzerland
[2] Autonomous Systems Lab
Stanford University, United States of America.
[3] Division of Engineering
New York University (NYU) Abu Dhabi, United Arab Emirates.
[*]Corresponding author

Keywords: SPaT message enhancement; supervised machine learning; Time-to-Green prediction; actuated signal control.

Abstract

Introduction and background

The involvement of digital technology has changed the transportation domain significantly in the last decade. The availability of several new data sources (i.e., sensor technology or vehicle technology) postulates for data-driven methodologies that can be incorporated into well-established traffic management systems on a macro- and micro-scopic level. Furthermore, the upcoming developments, such as Vehicle-to-Infrastructure (V2I), open the door for new approaches that allow considering communication between vehicles and infrastructure. Recent evolution in traffic signal control of urban intersections (e.g., actuated signal control, self-control algorithms, etc.) influence the signal phases and result in variable green, red and cycle times. Hence, speed advisory systems would benefit from the information about when the next green phase starts so that vehicles do not have to stop when crossing an intersection. Nevertheless, predictions for residual times of these quantities are not trivial and require a sophisticated modeling approach.

Recently, efforts have been made to standardize Signal Phase and Timing (SPaT) messages. Such messages contain the current phase with a prediction for the corresponding residual time for all approaches of a signalized intersection. Hence, the information can be utilized for the motion planning of human-driven and/or autonomously operated individual or public transport vehicles. This would lead to more homogeneous traffic flow and a smoother speed profile (i.e., the absence of speeding and heavy breaking between traffic lights). Most existing speed advisory systems rely on SPaT information (Yu et al., 2018, Asadi and Vahidi, 2011), which includes elements such as the start time of a signal phase, phase duration, or the next time this signal phase starts. Although it is widely accepted that broadcasting SPaT information is potentially beneficial for traffic systems (Misener et al., 2010, Amelink, 2015), SPaT information is rarely provided in reality. No intersection in the U.S. today broadcasts SPaT messages (Ibrahim et al., 2019). One challenge lies in the accurate prediction of SPaT information. Modern traffic control systems (e.g., in the city of Zurich) are typically adaptive to

vehicle demand, resulting in non-repetitive control sequences. Although adaptive control systems can efficiently prioritize public transport and serve car demand, it is difficult to predict SPaT information for such systems accurately.

Most existing works develop methods to obtain SPaT information for pre-timed traffic signals based on aggregated trajectory data. In such works, signal timings are unknown and can be either fixed or change slowly with time using traffic models (Fayazi et al., 2015, Fayazi and Vahidi, 2016, Wang and Jiang, 2012, Ban et al., 2009), or machine learning approaches (Yu and Lu, 2016, Protschky et al., 2015). For example, Fayazi et al., 2015 and Fayazi and Vahidi, 2016 employed a queue discharging model to estimate the start of green signals based on aggregated low-frequency bus and probe data. Yu and Lu, 2016 formulated the SPaT estimation problem into a general approximate greatest common divisor problem, aiming to obtain the cycle lengths, green times, and the phase schemes based on historical sparse taxi trajectories. Protschky et al., 2015 used a Bayesian learning approach to reconstruct the cycle length from historical trajectory data for traffic signals where cycle length is fixed within a certain period. These methods typically rely on the underlying assumption that cycle length is fixed, although some of them (e.g., Fayazi and Vahidi, 2016 and Protschky et al., 2015) are able to identify the occasional changes in the traffic signal timing plan. Moreover, these works are based on the aggregation of historical vehicle trajectories, assuming that the historical signal timings are unknown.

Some other studies in the literature propose probabilistic methods to predict SPaT information such as Protschky et al., 2014, by employing a Kalman filter to estimate the probability of phase switches. Ibrahim et al., 2019 estimates the conditional distribution of each signal phase given real-time measurements to predict the phase duration as a conditional expectation and a confidence interval. Nevertheless, this approach requires separate aggregations for different cycle lengths.

All the mentioned works are based on (a) vehicle trajectories data, or (b) historical signal data, and do not incorporate vehicle detection into the modeling. However, in a complex signalized intersection with multiple approaches and movements, the signal timing can be determined jointly by many Loop Detectors (LD). Also, it is important to consider the temporal relation between the traffic signals and the LDs because a delay between the LD activation and the change of signal state could be apparent. Therefore, the tuning of a model that can identify the nonlinear relationship and incorporate both data sources to estimate the residual time is promising and still open to research.

In this paper we propose a machine learning approach to predict the residual time of each phase of an intersection. To capture the nonlinear relationship between the signal information and the multiple detectors of an intersection, we construct the problem as a time series forecast and apply a Random Survival Forest (RSF) model to predict the time series of the Time-to-Green (T2G). To prove the concept, historical LD and signal timing data from an intersection in the city of Zurich is utilized. The area under investigation includes signal priority for public transportation (i.e., signal priorities change the control behavior of the intersection irregularly). We compare the novel time series forecast by an RSF model to the baselines (a) Auto-Regressive Integrated Moving Average (ARIMA) and (b) Linear Regression (LR). A conclusion and proposal for future work is stated at the end of this paper.

Methodology

We introduce T2G prediction as a time-series problem. Historical data is utilized to determine the T2G for every phase when a traffic light is red. During green, the T2G remains zero. Consequently, a linear time series is constructed, as shown in Figure 1 for a short time period and a sample traffic signal (named with the id 1) in a morning period. The signal peaks in Figure 1 indicate that the maximum of the T2G per cycle (i.e., length of red phase) is not constant and varies between 31 and 66 sec. Note that the latter value might result from approaching public transport that is detected and gets signal priority. Hence, the green time of the corresponding approaches gets extended, which results in higher waiting times for other streams.

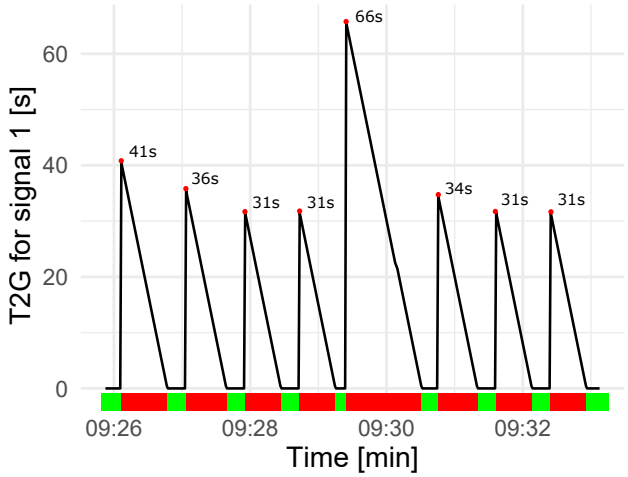


Figure 1: T2G time series for signal device 1; the highlighted signal peaks in red correspond to the maximal waiting time per signal phase.

In the following, we introduce a T2G prediction framework that allows a generic application to any intersection. The architecture of the proposed framework is depicted in Figure 2. The denoted blocks (1) – (3) follow the general working steps of a supervised machine learning procedure. The input

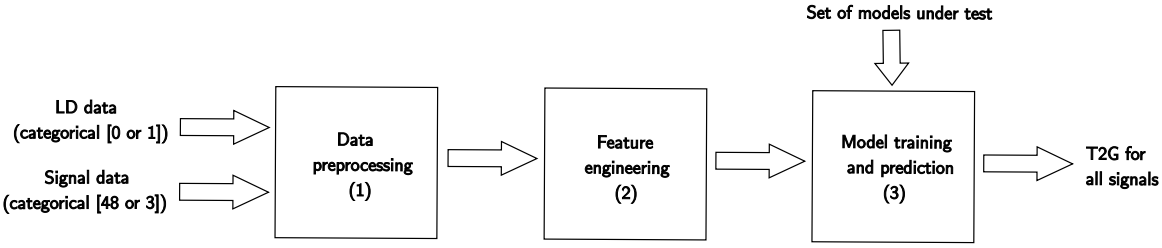


Figure 2: T2G prediction framework. The input data is represented by categorical variables. When an LD is occupied or not occupied, the corresponding is variable is 1 or 0, respectively. A red and green signal is denoted by the identifieres 3 and 48, respectively.

data (i.e., LD and signal data) functions as an input to the data pre-processing (Block (1)). Within this step, the data cleaning, data aggregation and transformation take place. In Block (2) the feature engineering is performed. Here the most relevant input variables are determined, variable correlations are eliminated and - if necessary - variable transformations are computed. The result of Block (2) is an input to Block (3), where the actual training and testing of a model is implemented. Consequently, a set of models $\mathcal{M}$ is considered, which is specified by the user. After training and testing all elements of $\mathcal{M}$ the best model is selected and predictions for the set of traffic signals are generated. Note that the process of tuning a proper prediction model requires a number of iterations between Block (2) and (3).

In this paper, we define an LR, an ARIMA and an RSF model as the members of $\mathcal{M}$. The LR and RSF model is trained with LD and signal data (categorical variables, respectively). Furthermore, a moving time window is implemented with size w (sec). This procedure improves the results significantly as the training process provides inputs at time step t for multiple time instances. First, we introduce the LR which can be defined as follows:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_p x_{i,p} + E_i, \ \forall i = 1, .., T, \tag{1}$$

where $\hat{y}_i$ is the T2G (response variable) for observation i, β_0 represents the intercept term and β_1 to β_p are the regression coefficients for the p predictors $x_{i,1}$ to $x_{i,p}$ (i.e., the possible data inputs described above), respectively. The error term is denoted by E_i and follows a Gaussian distribution (i.e., $E_i \sim \mathcal{N}(0, \sigma_{E_i})$); T denotes the prediction horizon. The solution for $\hat{y}_i$ is found by applying the Ordinary Least Square (OLS) method. The fitted model can be used to determine a prediction of the T2G for all given traffic signals by obtaining the conditional expected value of the response. To obtain the LR model, the build-in R-package was utilized. Research that similarly introduces LR models within this context can be found in, e.g., Branston and van Zuylen, 1978.

As the second model incorporated into the T2G framework, the ARIMA model is introduced below. ARIMA is a widely used tool for forecasting time series. The univariate model can be applied to a stationary time series. The model is a combination of differencing (by order of d) with auto-regression (by order of p) and the moving average method (by order of q). Consequently, the model is introduced by the following equation:

$$\hat{y}'_t = c + \phi_1 y'_{t-1} + ... + \phi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + ... + \theta_p \epsilon_{t-q} + \epsilon_t. \tag{2}$$

$\hat{y}'_t$ represents the differenced time series denoted as the response variable. The right-hand side of the equation contains the lagged values of y'_t by the degree of first-order differencing p. In addition, the lagged errors ϵ_t with the order of the moving average q are included. $\phi_{1...p}$ and $\theta_{1...p}$ are parameters. c denotes the intercept. Finally, a suitable model ARIMA(p, d, q) must be determined for forecasting (Sowell, 1992).

As a third model, an RSF is utilized. The RSF method is based on the concept of survival analysis, i.e., the analysis of time duration until a certain event occurs. Our constructed time series can be analyzed in such a way as it is expected that after the T2G the traffic light switches to green, meaning the time series is 0 for a certain time horizon. Therefore, we utilize the R-package *randomForestSRC*, which implements the following pseudo-code (Algorithm 1), proposed by Ishwaran et al., 2011. For

Algorithm 1 RSF pseudo code

procedure DORSF
  $\mathcal{B} \leftarrow$ bootstrap samples from the original training data set
  $\mathcal{T} \leftarrow$ grow a survival tree $\forall s \in \mathcal{B}$, where s is a sample from $\mathcal{B}$
  $\mathcal{O} \leftarrow$ Exclude out-of-bag data from all $s \in \mathcal{B}$
  Grow tree to full size with the constraint that a leaf node $l \in \mathcal{L}$ has no less than $d_0 > 0$ unique deaths; $\mathcal{L}$ is a set of leaf nodes
  $\mathcal{C} \leftarrow$ Calculate the Cumulative Hazard Function (CHF) $\forall t \in \mathcal{T}$, where t is a survival tree for one bootstrap sample
  $\mathcal{P} \leftarrow$ Calculate prediction error using the set $\mathcal{O}$

the detailed mathematical background of RSF, the interested reader is referred to Nasejje, 2017.

After training the model, the testing is performed with the test data set. For computation of the error value, the Mean-Average-Error (MAE) is utilized, defined as

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left| \hat{y}_i - y_i \right|. \tag{3}$$

$\hat{y}_i$ again represents the estimated T2G and y_i is the T2G from the test data set. As a prediction would only be applied when the T2G is decreasing (i.e., when the signal is red), only those parts of the signals $\hat{y}_i$ and y_i are considered for computation of MAE. We obtain these parts automatically by considering the derivative of T2G signals.

In the following section, we show the application of the T2G framework to the test intersection in the city of Zurich.

Case study and results

The analysis is based on a data set from an intersection in the city center of Zurich. From north to south and vice versa, there are tram lines that are prioritized by the signal control. Figure 3 depicts the intersection with the 10 associated LDs (indicated as rectangles at the intersection approaches) and 12 traffic signals (indicated by the circled numbers). The tram lines are indicated by the red dashed lines. Note that traffic signal 3 is for bicycles who are allowed to go straight ahead. No separate detector data is available for this signal. Hence, its phase corresponds to the one of signal 2. Signal 11 is installed for the tram to indicate potentially leaving the stop, but Signal 6 remains crucial at this approach. 15

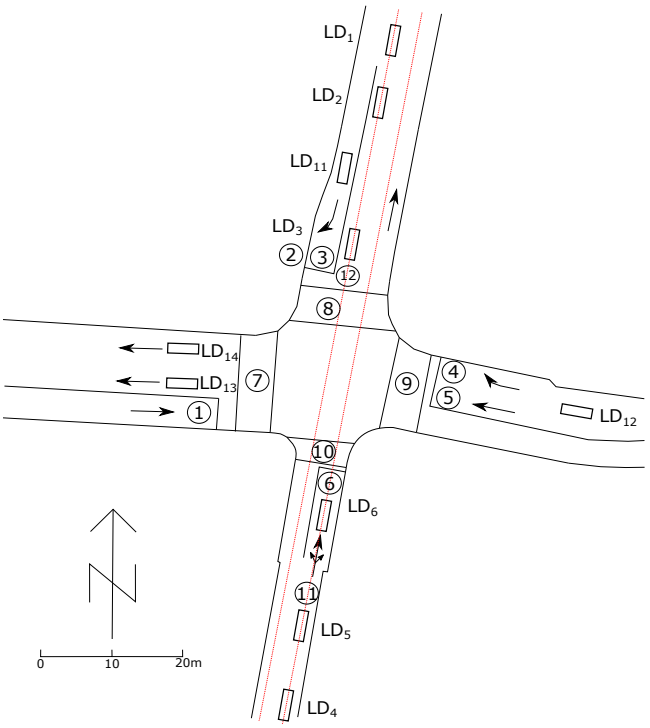
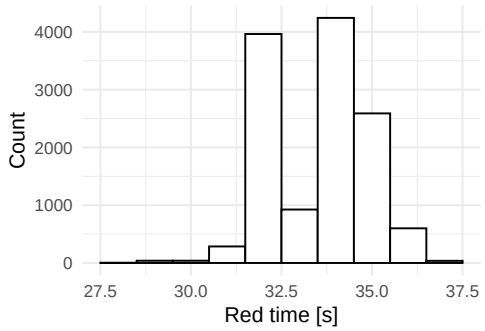


Figure 3: Test intersection in the city center (Kreis 1) of Zurich, Switzerland.

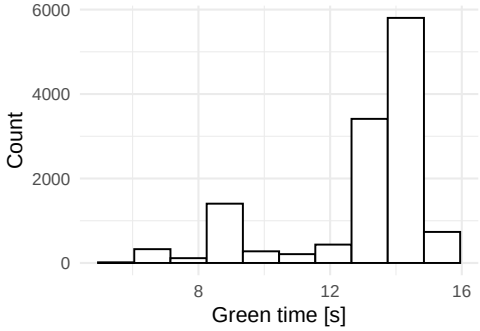
days of data from January 2019 are available. The intersection is operated with an actuated control with no fixed cycle times. The data set contains events from every device installed at the test site (i.e., LD and signal data). With a resolution of 0.1 seconds, a device sends a telegram to a centralized control unit. The telegram contains a timestamp, the device id and the state. In other words, we know when an LD is activated or a signal changes from red to green. Besides, LD or signal failures can be detected. For all further analysis, the data is aggregated to a resolution of 1 second. This leads to a reduction of data structure size and is also sufficient for the required prediction accuracy.

For simplicity, we prove our concept by obtaining a subset of 1 day out of the complete data set. The descriptive statistics of measurements from 7:00am to 10:00pm of one day are computed. During night, traffic signals are switched to a program where lights are flashing yellow and hence no analysis can be performed. As a first step, the red, green and cycle times are calculated. This step allows for determining if an intersection is operated by a fixed or actuated control. The data set of signal 1 shows

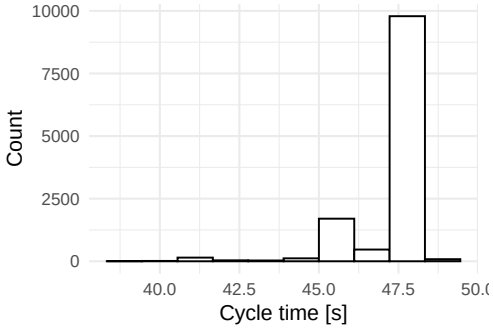
variations in red, green and cycle time, respectively. The distributions of these quantities are depicted in Figure 4.



(a) Distribution of red times.



(b) Distribution of green times.



(c) Distribution of cycle times.

Figure 4: Distributions of processed signal data (sg1), i.e., red, green and cycle time, respectively. The data indicates an actuated control which justifies a prediction approach for T2G.

The mean red time is determined to 33.51 sec and a corresponding variance of 1.84 sec. For the green and cycle times, the mean values are given by 12.82 sec and 47.54 sec and a variance of 4.03 sec and 1.84 sec, respectively. Therefore, it is evident that (a) a prediction of the T2G can not be given with simple reverse engineering, and (b) the variances give an indication that simple prediction approaches might fail to make a prediction that satisfies accuracy requirements.

A complete compilation of the computed T2G descriptive statistics for all traffic signals is given in Table 1. Note that the high values for signals 11 and 12 occur because these traffic signals are specifically for public transport. If no vehicle is detected, the traffic light remains red.

As the potential correlation between variables is degrading the model quality, we investigate all LDs and signal devices (Figure 5). The data shows almost no correlation between LD devices and between specific LD and signal devices. A more detailed analysis of the signal devices depict some significant positive correlations (e.g., signals 1 and 5 or signals 2 and 6). Nevertheless, this is expected as these signals control compatible traffic streams (see Figure 3). The same argument holds for the signals regulating the pedestrian streams (signals pairs (7,9) and (8, 10)). The correlation between signals 2 and 3 might occur because the two streams get green at the same time, but the green time for bicycles is smaller as a longer clearing time is required. Consequently, we decide to remove one of the variables that are members of correlated pairs. Note that for this analysis we decide to keep all variables with a correlation coefficient smaller than 0.8.

Finally, we apply our set of models $\mathcal{M}$ to the training and testing procedure. We asses the model

Table 1: Descriptive statistics T2G for all traffic signals.

T2G for signal	Min	1st Q.	Median	Mean	3rd Q.	Max
1	0.00	0.00	11.45	14.11	24.45	150.45
2	0.00	4.45	17.45	18.68	30.45	148.45
3	0.00	1.45	14.45	16.21	27.45	152.45
4	0.00	0.45	13.45	15.67	26.45	170.45
5	0.00	0.45	13.25	15.13	25.45	146.35
6	0.00	5.45	18.45	19.56	31.45	141.45
7	0.00	0.00	7.45	11.33	20.45	146.45
8	0.00	0.00	3.55	9.54	16.45	153.45
9	0.00	0.00	3.45	8.82	16.45	141.45
10	0.00	0.00	0.00	7.36	12.45	195.45
11	0.00	39.45	114.45	134.90	213.45	1086.45
12	0.00	39.45	115.45	142.94	222.45	1056.45

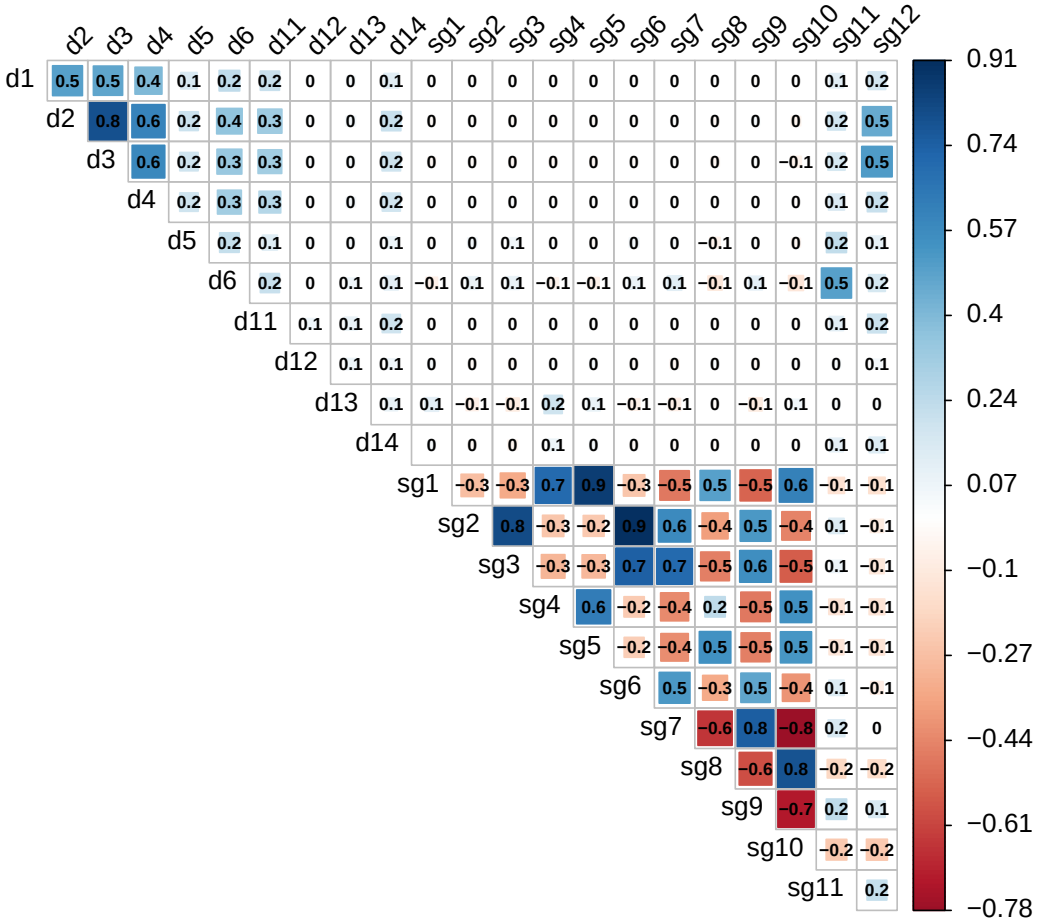


Figure 5: Correlations of all LD and signal devices.

quality by utilizing the a subset of the data (7:00am to 10:00am of one day). We first apply the ARIMA model to our test data set. Again note that an ARIMA model is univariate and only depends on the computed historical T2G time series. The parameters for ARIMA are chosen as $d = 2$, $p = 0$, and $q = 3$. The LR model is applied with standard settings and no variable transformations. Nevertheless, the general assumptions required to justify the application of OLS are considered in the analysis. The RSF is applied with a number of trees $n = 100$; n is chosen concerning the trade-off between gain in improvement versus the increase in computational time. Furthermore, the data is centered and scaled to obtain better model performance. Both models depend on LD and signal data time windows, for which $w = 10$ is chosen.

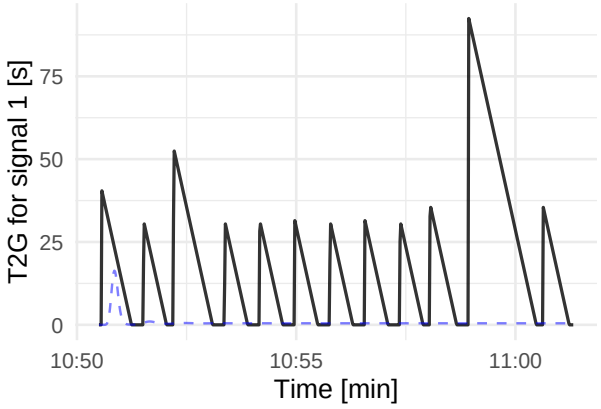
Consequently, we can state the $\overline{\text{MAE}}$ for all calculated errors and the corresponding standard deviation σ_{MAE} (Table 2).

Table 2: Model performance comparison models with the $\overline{\text{MAE}}$ and corresponding σ_{MAE}, respectively for ARIMA, LR and RSF.

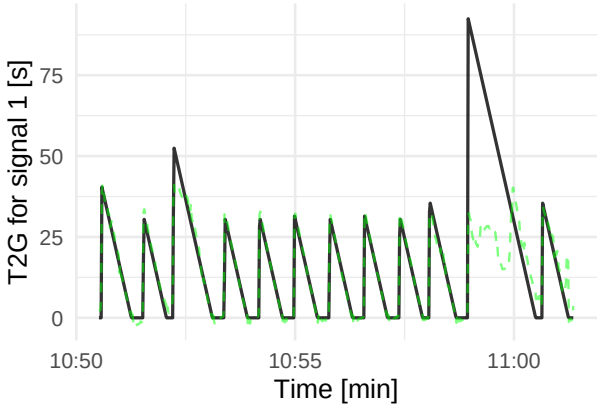
	ARIMA		LR		RSF	
Signal device	$\overline{\text{MAE}}$	σ_{MAE}	$\overline{\text{MAE}}$	σ_{MAE}	$\overline{\text{MAE}}$	σ_{MAE}
1	12.66	1.48	5.04	0.46	1.00	0.24
2	16.31	0.62	4.88	0.41	1.01	0.21
3	15.06	1.11	4.83	0.43	0.97	0.27
4	13.77	1.40	4.85	0.65	1.00	0.49
5	14.21	1.54	4.54	0.51	0.99	0.25
6	16.45	0.57	5.07	0.43	0.99	0.12
7	10.59	0.59	3.28	0.20	0.57	0.08
8	8.71	1.26	4.20	0.71	1.31	0.29
9	8.96	0.79	3.10	0.04	0.44	0.03
10	6.71	2.14	4.11	1.07	1.13	0.36
11	112.14	364.22	74.45	142.45	69.06	147.64
12	123.15	2731.73	90.16	927.00	81.53	927.46

The results show that a simple ARIMA model is not a promising approach for T2G predictions. $\overline{\text{MAE}}$ values with the corresponding variances σ_{MAE} are exceeding by far the accuracy requirements for application (values between 6.71 and 123.15 sec, with a corresponding high σ_{MAE}). The result underlines the hypotheses that simple forecasting models that only consider the T2G time series are not suitable for such prediction problems. Furthermore, Figure 6a shows a prediction sample from 10:50am to 11:00am. The ARIMA model is not capturing the pattern of a T2G time series, which results in high under- or over-estimation of the residual times. The model provides an estimate for the next phase, but the signal remains constant afterwards. Hence, ARIMA can only be used to predict the next phase and gives predictions with unsatisfactory accuracy. The LR model provides $\overline{\text{MAE}}$ values between 3.10 and 5.07 sec for traffic lights that regulate vehicle traffic streams (signals 1 to 10). For public transport signals, the prediction accuracy shows a high error. This is because the model is not able to capture the signal peaks that occur when there is no demand on these streams. An example of this behavior is depicted in Figure 6b (signal peak at 11:00am). Besides, the linear model suffers from negative values during the period that T2G is zero and is also likely to miss such signal patterns, which results in a constant T2G greater than zero (Figure 6b 11:02am). A zero-inflated regression model could be utilized to improve the results.

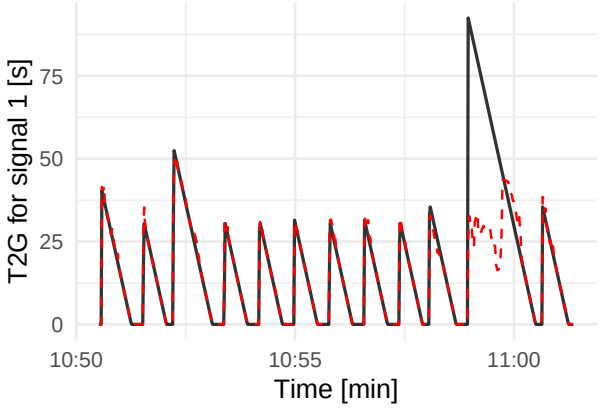
RSF model demonstrates the best performance best with $\overline{\text{MAE}}$ values ranging from 0.57 to 1.31 sec



(a) Result comparison with ARIMA.



(b) Result comparison with LM



(c) Result comparison with RSF.

Figure 6: Prediction of the T2G for a 10 minutes time frame with a) the ARIMA model b) the LR model and c) the RSF model.

for signals 1 – 10, with a corresponding low variance σ_{MAE}. For signal 11 and signal 12, the model also fails to capture high signal peaks accurately. Nevertheless, the RSF prediction never misses a green phase (when T2G is zero). A sample prediction is shown in Figure 6c with the described behavior.

Conclusion and future work

The paper proposed a framework for T2G predictions at an urban intersection to enhance the quality of SPaT messages. The problem was constructed as a time series forecast. The framework implementation is generic and can be applied to any intersection that provides LD and signal data. In the case study, the work was tested on an intersection in the city of Zurich with actuated signal control and public transport priority. Results show that the RSF is a promising tool for the prediction of residual phase times and outperforms the baseline models (i.e., LR as well as the ARIMA model). Nevertheless, the performance for predictions of public transport traffic lights needs to be further investigated. Future work will extend the present research with the possibility of predicting the T2G for all signals simultaneously. Consequently, we will also look at the parameter tuning of the models under test concerning computational time. This opens up an approach to justify a potential real-time application. Interesting would also be to include a Long-Short-Term-Memory (LSTM) neural network in the set of models. Literature shows that this model is promising for time series forecasts and should, therefore, be considered for this problem.

References

Amelink, M. (2015). Signal phase and time (spat) and map data (map).
URL: *https://amsterdamgroup.mett.nl/downloads/handlerdownloadfiles.ashx?idnv=500795.*

Asadi, B. and Vahidi, A. (2011). Predictive cruise control: Utilizing upcoming traffic signal information for improving fuel economy and reducing trip time, *IEEE Transactions on Control Systems Technology* **19**(3): 707–714.

Ban, X. J., Herring, R., Hao, P. and Bayen, A. M. (2009). Delay pattern estimation for signalized intersections using sampled travel times, *Transportation Research Record* **2130**(1): 109–119.

Branston, D. and van Zuylen, H. (1978). The estimation of saturation flow, effective green time and passenger car equivalents at traffic signals by multiple linear regression, *Transportation Research* **12**(1): 47 – 53.

Fayazi, S. A. and Vahidi, A. (2016). Crowdsourcing phase and timing of pre-timed traffic signals in the presence of queues: Algorithms and back-end system architecture, *IEEE Transactions on Intelligent Transportation Systems* **17**(3): 870–881.

Fayazi, S. A., Vahidi, A., Mahler, G. and Winckler, A. (2015). Traffic signal phase and timing estimation from low-frequency transit bus data, *IEEE Transactions on Intelligent Transportation Systems* **16**(1): 19–28.

Ibrahim, S., Kalathil, D., Sanchez, R. O. and Varaiya, P. (2019). Estimating phase duration for spat messages, *IEEE Transactions on Intelligent Transportation Systems* **20**(7): 2668–2676.

Ishwaran, H., Kogalur, U. B., Chen, X. and Minn, A. J. (2011). Random survival forests for high-dimensional data, *Statistical Analysis and Data Mining: The ASA Data Science Journal* **4**(1): 115–132.

Misener, J., Shladover, S. and Dickey, S. (2010). Investigating the potential benefits of broadcasted signal phase and timing (spat) data under intellidrive sm, *ITS America Annual Meeting* .

Nasejje, J.B., M. H. D. K. e. a. (2017). A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data, *BMC Med Res Methodol* **17**(115).

Protschky, V., Ruhhammer, C. and Feit, S. (2015). Learning traffic light parameters with floating car data, *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 2438–2443.

Protschky, V., Wiesner, K. and Feit, S. (2014). Adaptive traffic light prediction via kalman filtering, *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pp. 151–157.

Sowell, F. (1992). Modeling long-run behavior with the fractional arima model, *Journal of Monetary Economics* **29**(2): 277 – 302.

Wang, C. and Jiang, S. (2012). Traffic signal phases' estimation by floating car data, *2012 12th International Conference on ITS Telecommunications*, pp. 568–573.

Yu, C., Feng, Y., Liu, H. X., Ma, W. and Yang, X. (2018). Integrated optimization of traffic signals and vehicle trajectories at isolated urban intersections, *Transportation Research Part B: Methodological* **112**: 89 – 112.

Yu, J. and Lu, P. (2016). Learning traffic signal phase and timing information from low-sampling rate taxi gps trajectories, *Knowl.-Based Syst.* **110**: 275–292.