A TRAIN DATA COLLECTION AND ARRIVAL TIME PREDICTION SYSTEM FOR HIGHWAY-RAIL GRADE CROSSINGS

Yifeng Chen, Ph.D.
(corresponding author)
Traffic Engineer, AECOM
27777 Franklin Road, Suite 2000
Southfield, Michigan, 48034
Tel: 402-570-8771
Email: chaserchen98@hotmail.com

Laurence R. Rilett, Ph.D., P.E.
Keith W. Klaasmeyer Chair in Engineering and Technology
Director, Nebraska Transportation Center
262D Prem S. Paul Research Center at Whittier School
2200 Vine Street
University of Nebraska–Lincoln
Lincoln, NE 68583-0815, P.O. Box 830851
Telephone: 402-472-1992
Email: lrilett2@unl.edu

Word Count: 5,728 + 4 Figures + 3 Tables =7,478

Revision date: November 15, 2016

ABSTRACT

A key component to the safe and efficient operation of traffic signals near highway-rail grade crossings (HRGCs) is an accurate estimate of a given train's arrival time at each crossing. An improvement in the accuracy and timing of the predicted train arrival would allow for improvements in traffic signal preemption algorithms which would, in turn, lead to increased driver and pedestrian safety and/or reduced network delay. This paper introduces a system for train data collection and arrival time prediction at an HRGC using real-time data from a combined video-Doppler radar vehicle detection system. A dual-track test bed system in Lincoln, Nebraska was chosen for train speed data collection. Both kinematic equation-based and multiple linear regression models were developed and used to predict train arrival time at the HRGC. The best models, based on average absolute errors (AAE), were identified. It was found that on average, regression models were more accurate than kinematic models by approximately 13%. As detection time increases, AAEs of both the kinematic and regression models decrease. In addition, the confidence intervals (CI) about mean prediction errors were obtained using a bootstrap method. The average prediction errors of the regression models were approximately 50% smaller than the kinematic models, all else being equal.

INTRODUCTION

Highway-rail grade crossings (HRGCs) are locations where a roadway and a railroad intersect *(1)*. When a highway-rail grade crossing (HRGC) is located close, e.g., within 61 meters (200 feet), to a signalized intersection, the standard practice is to interconnect the railroad signal control equipment and the highway signal control equipment and use a preemption signal control strategy when a train is approaching *(1)*. According to the *Manual on Uniform Traffic Control Devices* (MUTCD) *(1)* and the *AREMA/AAR Signal Manual (3)*, a minimum of 20 seconds of warning time should be provided to the traffic control system before a train arrives at a crossing. This allows the railroad crossing warning devices to be activated prior to the arrival of a train and to clear any vehicles that may be stopped on the track(s) as part of the signal preemption strategy. To accomplish this task, a train detection system with train arrival forecasting functionality is required to determine when the railroad warning devices and the highway signal preemption will be activated.

The current preemption system uses railroad track circuits to detect arriving trains and to activate railroad warning devices and signal preemption at the intersection near HRGCs (IHRGC). This is known as first generation train detection technology *(4, 5)*. There are two types of first-generation prediction technologies in practice: fixed distance and constant warning time *(5, 6)*. The fixed distance system activates the warning devices whenever a train's presence is detected at a fixed point in the track. The location is calibrated so that the warning devices and signal controller will be notified at least 20 seconds before the fastest possible train arrives at the HRGC. However, the warning time would double for a train operating at half of the maximum speed. The constant warning time (CWT) system measures the train speed upon detection, and uses a simple distance-speed relationship to forecast train arrival time, assuming that the train will travel at this constant speed *(4, 5, 6, 7)*. The resulting warning time algorithm ensures a minimum warning time of 20 seconds, and is relatively consistent compared to the fixed distance system. However, if the trains are accelerating (decelerating) then the predicted arrival time may be in error and traffic signal preemption may start earlier (later) than the ideal preemption warning time (*5*). For example, a study by L-P Tardif & Associates, Parviainen & Associates, and CANAC International, Inc. showed that the CWT system warning time varied from 20 seconds to 38 seconds for a sample of 68 trains traveling at speeds greater than 48 km/h (30 mph), with 93% of times falling between 20 and 33 seconds (*5*).

Second generation detection technologies, including acoustic sensors, infrared sensors, Autoscope video detection, and Doppler radar detectors, are now available for train-detection applications. These detectors are non-intrusive devices mounted off the railroad right-of-way and can detect train movements continuously upstream from the railroad circuits. With this setup, trains will be detected considerably earlier than 20 seconds (e.g., 60 seconds). Larger train warning times mean that new advanced preemption strategies, such as transition preemption strategies (TPS1, TPS2 and TPS3) *(2)*, can be used, which allows signal timings to be optimized at an earlier stage of train arrival, and consequently improves safety and efficiency at IHRGCs *(2, 8)*. It should be noted that the second train detection technologies are intended to supplement, not replace, the current train preemption system *(5)*.

Third generation technology is based on on-board computers, Geographic Positioning Systems (GPS), and two-way communication between the train and the signal controller/warning devices. These systems have the ability to provide continuous and more accurate information on the location and speed of the approaching train than either the first or second generation technologies. To date, third-generation technology is mostly in the development or testing phase

and has not been widely implemented. Closer cooperation between the railroad and roadway agencies needs to be established before the technology will have widespread deployment.

This paper focuses on using the second-generation train detection technologies (e.g., video and radar detectors) to measure train speeds, and subsequently uses this information to predict train arrival times and their associated prediction intervals at HRGCs. The organization of this paper is as follows: the next section introduces the test bed and the data collection site for the train prediction system, followed by the architecture of the train data collection and prediction system. Then, the prediction models used in the system are described mathematically. Evaluation and comparison of the models are performed. Next, error bounds of the prediction models are estimated using a bootstrap method. Lastly, concluding remarks complete the paper.

TEST BED

The BNSF railroad in Lincoln, Nebraska was chosen as the test bed for the study and is shown in **FIGURE 1**. It is a dual-track mainline railroad with the tracks running in a northeast-southwest direction. The railroad intersects the urban roadway network, at grade, at 33rd Street, Adams Street, and 44th Street. The HRGC at Adams Street was chosen as the target HRGC because: 1) it has considerable train traffic, on the order of 50-70 trains per day; 2) it has been identified as one of the most hazardous HRGCs in Lincoln, NE by the FRA, based on the accident history dating back to 1975*(9)*; and 3) it is near a signalized intersection (35th Street and Cornhusker Highway), which has a unique geometry and high potential for traffic to back up on to the railroad tracks *(8, 9, 10)*. As **FIGURE 1** shows, the test bed system also includes an upstream location located at Salt Creek Roadway and Antelope Valley Parkway, which is approximately 2.91 km (1.81 miles) southwest from the Adams Street HRGC. This upstream location is used for eastbound (EB) train detection and speed measurement, using two second-generation detectors: a Doppler radar detector and an Autoscope video detector.

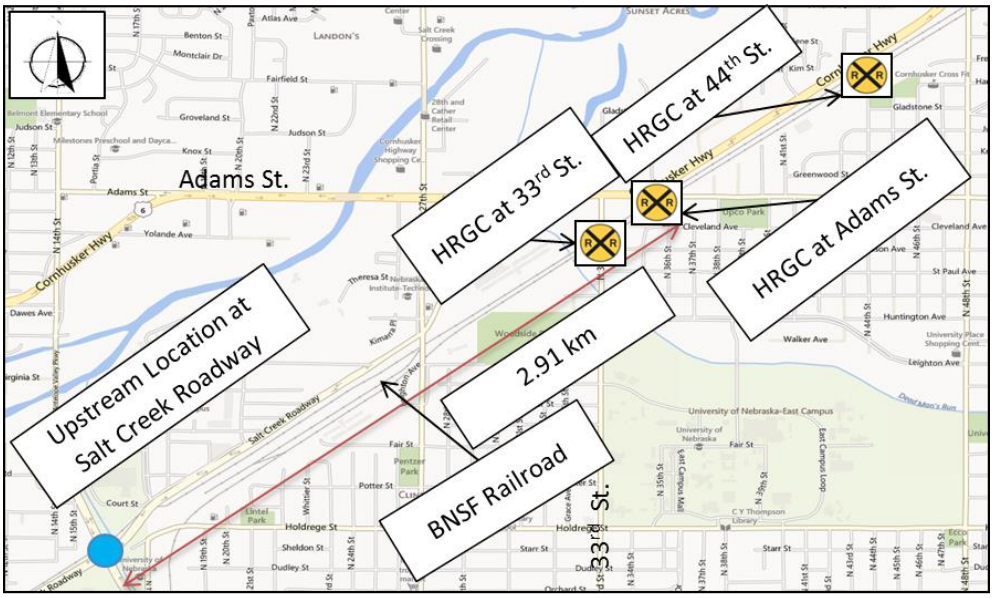


FIGURE 1 Test bed in Lincoln, Nebraska.

FIGURE 2 shows the layout of the upstream location at Salt Creek Roadway. The railway and roadway are grade-separated, and the intersection of Salt Creek Roadway and

Antelope Valley Parkway is comprised of a bridge that stands approximately 12.19 meters (40 feet) above the railroad tracks. A Doppler radar and a day/night video camera with an Autoscope video image processor *(10)* are installed on a traffic signal pole at the northeast corner of the intersection, as shown in **FIGURE 2**. The Doppler radar and the video camera are aimed at the same detection zone on the railroad, which is located approximately 213 meters (698 feet) east of the bridge. This detection zone is represented in **FIGURE 2** as a green rectangle.

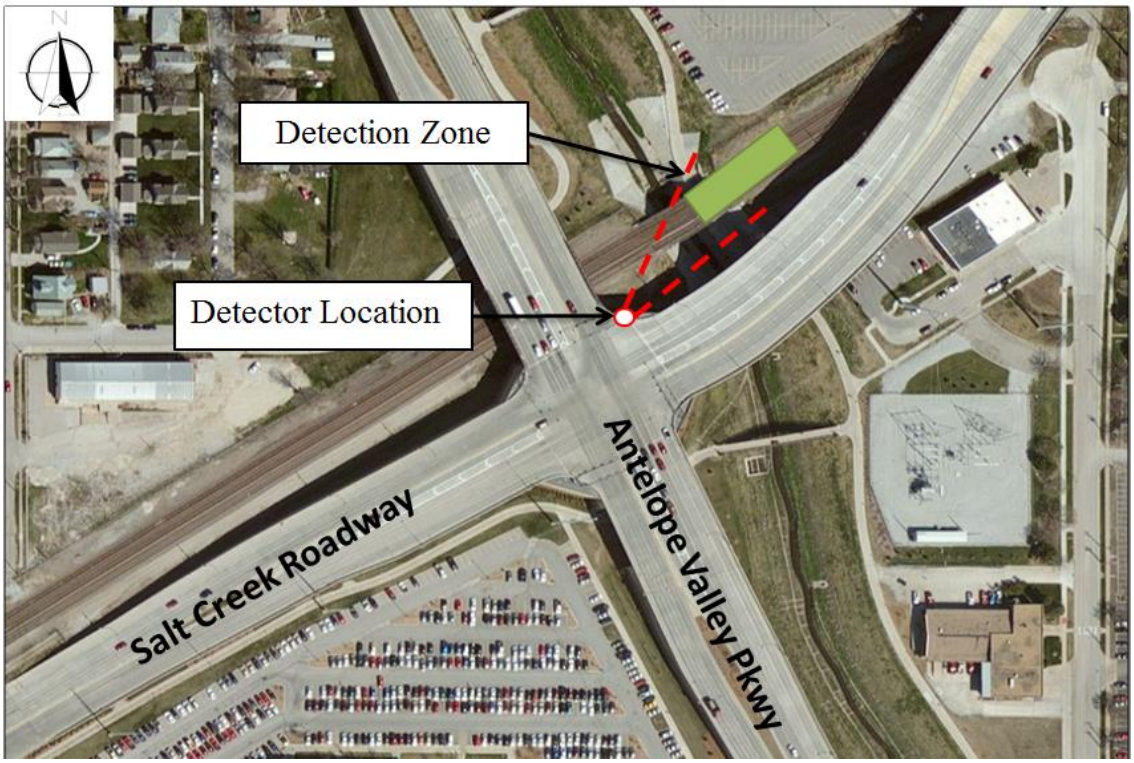


FIGURE 2 Upstream location at Salt Creek roadway.

ARCHETECTURE OF THE PROPOSED SYSTEM

The data collection and arrival time prediction system consists of three parts: (1) online data collection, (2) offline data processing, and (3) offline detector calibration. **FIGURE 3** illustrates the architecture of the speed measurement and arrival time prediction system.

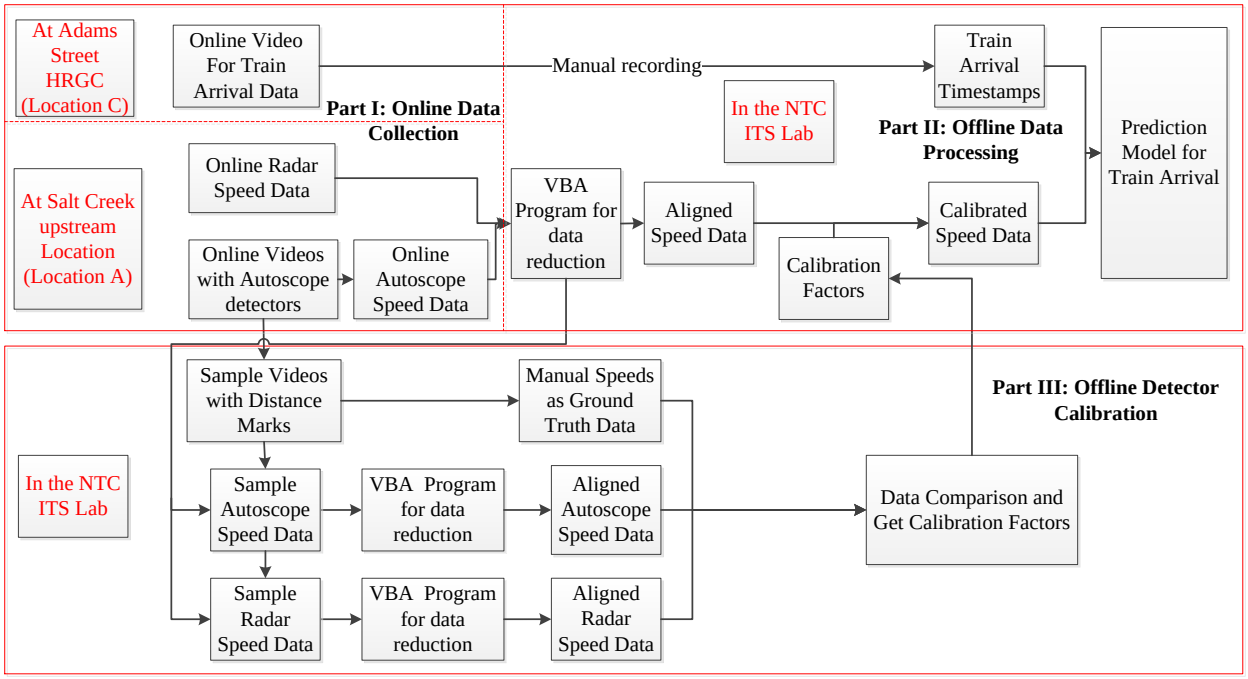


FIGURE 3 Architecture of the data collection and arrival time prediction system.

At the upstream Salt Creek location and the Adams Street HRGC, real-time train videos are recorded on a digital video recorder (DVR) and sent back to the Nebraska Transportation Center (NTC) ITS lab through the City of Lincoln's fiber optic network. Simultaneously, train speed data from radar and Autoscope video detectors are collected at the Salt Creek location and transmitted back to the ITS lab through the same network. This constitutes the online data collection listed as Part I in **FIGURE 3**.

In the offline detector calibration (Part III), train videos are randomly sampled from the videos transmitted back to the NTC ITS lab. Subsequently, the corresponding online speed data from radar and Autoscope speed detectors are identified. The sampled train videos are used to obtain ground truth train speed data. The aligned radar and Autoscope video speed data are subsequently used, along with the ground truth data, to calibrate the detectors. A detailed description of the detector calibration methodology can be found elsewhere *(8, 10, 11)*. The calibration factors obtained offline are then applied to the online data collection.

With regard to offline data processing (Part II), a program for data reduction was developed in Visual Basic for Applications (VBA), which eliminates outliers and false calls in the speed data. The other function of the program is to align speed measurements of the two speed sensors to one-second increments, as these two detectors measure train speed at different resolutions *(10)*.

The aligned speed data were then adjusted using the calibration factors obtained from the offline detector calibration process. The train arrival time stamps at the Adams Street HRGC were recorded manually by observing the videos transmitted from the field. In Part II, the calibrated train speed data and train arrival timestamps were used to develop the train arrival prediction models. It was found that the data from the radar detector were more comprehensive and more robust than those from the Autoscope detector, and consequently, radar data were used

in the prediction model development. The videos and Autoscope speed data were used to verify and validate the radar speed data.

TRAIN ARRIVAL PREDICTION MODELS

During the data collection process, a total of 138 EB trains were randomly sampled from May 2013 through September 2013. Of these, 100 were randomly selected and used for developing the train arrival prediction models. The remaining 38 trains were used to validate the prediction models.

Intuitively, in order to provide the most accurate train arrival estimates, the predicted train arrival time should be continuously updated as the train approaches the HRGC *(6)*. Consequently, the model for predicting train arrival time should also be a continuous model that is updated at specific time intervals as the train approaches the HRGC *(6)*. Previous studies *(4, 6)* have found that it is advantageous to predict train arrival time using models that are updated every 10 seconds. Therefore, it was decided to use the speed reading from every 10 seconds to develop the prediction models.

The average absolute error (AAE) at the end of every 10-second interval is used as the performance index of prediction models. It is calculated as shown in Equation 1.

$$\boldsymbol{AAE}_t = \frac{\sum_{x=1}^{N_t} |T_t^x - P_t^x|}{N_t} \quad \forall\, t = 10, 20, \ldots, 300 \tag{1}$$

where:

AAE_t average absolute error for prediction time t

T_t^x actual train arrival time for train x at prediction time t

P_t^x predicted train arrival time for train x at prediction time t

N_t number of trains in detection at prediction time t

There are two categories of models used in the analysis. The first category contains kinematic models that are based on first or second order motion equations and can be used directly. The second category consists of regression models that attempt to identify a statistical relationship between train speed measurements and the train's arrival time at the HRGC. This category of models requires extensive datasets so that the models can be calibrated and validated. In the next section, both categories of models are presented.

Kinematic Models (Model 1- Model 6)

Model 1 Basic train prediction method

This model uses the instantaneous speed of a train to calculate the arrival time at the target downstream crossing. Equation 2 and Equation 3 show the calculation of the remaining

estimated distance $\boldsymbol{d_t^x}$ and the predicted arrival time $\boldsymbol{p_t^x}$ for train x for every 10-second interval. This approach assumes that the train will travel at the current speed until it arrives at the crossing. This model is the most straightforward and is based on the CWT forecasting procedure *(2)*. However, if the train changes speed over time, this model will result in fairly large errors *(4, 5)*.

$$\boldsymbol{d_t^x} = \boldsymbol{D} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{t}} v_i^x \quad \forall\, t = 10, 20, \ldots, 300 \tag{2}$$

$$\boldsymbol{p_t^x} = \frac{\boldsymbol{d_t^x}}{v_t^x} \quad \forall\, t = 10, 20, \ldots, 300 \tag{3}$$

where:

$\boldsymbol{d_t^x}$	estimated distance from the head of train x to the downstream grade crossing t seconds after detection (m)
D	distance between the upstream detection location and the HRGC. For this testbed D is equal to 2910 m
$\boldsymbol{p_t^x}$	predicted arrival time for train x, t seconds after detection (s)
v_t^x	train speed t seconds after detection (m/s)
v_i^x	train speed $\boldsymbol{i}$ seconds after detection (m/s), i =1, 2, 3… $\boldsymbol{t}$

Model 2 Basic prediction method with average train speeds

Model 2 is a similar model to Model 1 and uses a distance-speed equation to predict train arrival times. The difference is that in Model 2, average train speed of train x at t seconds in detection (the end of the current 10-second interval) is used instead of the instantaneous speed at t seconds in detection, as shown in Equation 5.

$$\boldsymbol{p_t^x} = \frac{\boldsymbol{d_t^x}}{\overline{\boldsymbol{v_t^x}}} \quad \forall\, t = 10, 20, \ldots, 300 \tag{4}$$

$$\overline{\boldsymbol{v_t^x}} = \frac{\sum_{i=1}^{t} v_i^x}{t} \quad \forall\, t = 10, 20, \ldots, 300 \tag{5}$$

where:

$\overline{v_t^x}$ average speed of train x, t seconds after detection (m/s)

Model 3 Basic prediction method with moving average train speeds

In Model 3, the moving average speed over the current 10-second interval is used, as shown in Equation 6 and Equation 7.

$$p_t^x = \frac{d_t^x}{\widehat{v_t^x}} \quad \forall\, t = 10\ s, 20\ s, \ldots 300\ s \tag{6}$$

$$\widehat{v_t^x} = \frac{\sum_{j=1}^{9} v_{t-j}^x}{10} \quad \forall\, t = 10\ s, 20\ s, \ldots 300\ s \tag{7}$$

where:

$\widehat{v_t^x}$ average train speed for train x, t seconds after detection (m/s)

v_{t-j}^x train speed j seconds ahead of v_t^x, j =1,2,…9

Model 4 Train prediction method using second order kinematic equation and instant speed

This model calculates the arrival time by using the kinematic equation with the moving average acceleration rate, instantaneous speed of a train, and remaining distance to the crossing, as shown in Equation 8 and Equation 10. The moving average acceleration is calculated as the speed difference of the two speeds by the end of the current and last 10-second intervals divided by 10 seconds, as shown in Equation 9. The principle behind this model is that by including information on whether the train is increasing or decreasing its speed, the resulting model would be more accurate.

$$d_t^x = \frac{\widehat{a_t^x} * (p_t^x)^2}{2} + v_t^x * p_t^x \quad \forall\, t = 10, 20, \ldots, 300 \tag{8}$$

$$\widehat{a_t^x} = \frac{v_t^x - v_{t-10}^x}{10} \quad \forall\, t = 10, 20, \ldots, 300 \tag{9}$$

$$\boldsymbol{p_t^x} = \frac{-\boldsymbol{v_t^x} + \sqrt{\boldsymbol{v_t^{x2}} + \mathbf{2} * \widehat{a_t^x} * \boldsymbol{d_t^x}}}{\widehat{a_t^x}} \qquad when \;\; \widehat{a_t^x} > 0$$

$$\forall\, t = 10, 20, \ldots, 300 \qquad (10)$$

$$\boldsymbol{p_t^x} = \frac{\boldsymbol{d_t^x}}{\boldsymbol{v_t^x}} \qquad when \;\; \widehat{a_t^x} \le 0$$

where:

$\widehat{a_t^x}$ moving average acceleration/deceleration rate of train x, t seconds after detection (m/s^2)

The limitation of this model is that the one part in the square root (i.e.$\boldsymbol{v_t^{x2}} + \mathbf{2} * \widehat{a_t^x} * \boldsymbol{d_t^x}$) in Equation 10 may yield a negative value when $\widehat{a_t^x}$ is negative. Therefore, Equation 10 will work only when the acceleration rate is positive. As an alternative, $\widehat{a_t^x}$ is set to 0 when it is negative in Equation 9. In this case, Equation 10 reduces to Equation 3, and consequently Model 4 has the same results as Model 1 when the estimated acceleration is zero or negative.

Model 5 Train prediction method using second order kinematic equation and average speed

Model 5 uses average speed in place of instantaneous speed, as shown in Equation 11. Similar to Model 4, the value of $\widehat{a_t^x}$ is set to 0 if $\widehat{a_t^x} \le 0$. In this situation, Model 5 is equivalent to Model 2.

$$\boldsymbol{p_t^x} = \frac{-\overline{\boldsymbol{v_t^x}} + \sqrt{\overline{\boldsymbol{v_t^x}}^{\mathbf{2}} + \mathbf{2} * \widehat{a_t^x} * \boldsymbol{d_t^x}}}{\widehat{a_t^x}} \qquad when \;\; \widehat{a_t^x} > 0$$

$$\forall\, t = 10, 20, \ldots 300 \qquad (11)$$

$$\boldsymbol{p_t^x} = \frac{\boldsymbol{d_t^x}}{\overline{\boldsymbol{v_t^x}}} \qquad when \;\; \widehat{a_t^x} \le 0$$

Model 6 Train prediction method using second order kinematic equation and moving average speed

Model 6 uses moving average speeds in place of instantaneous speed, as Equation 12 shows. Similar to Model 4 and Model 5, the value of $\widehat{a_t^x}$ is set to 0 when $\widehat{a_t^x} \le 0$. In this situation, Model 6 is equivalent to Model 3.

$$\boldsymbol{p_t^x} = \frac{-\widehat{\boldsymbol{v_t^x}} + \sqrt{\widehat{\boldsymbol{v_t^x}}^{\mathbf{2}} + \mathbf{2} * \widehat{a_t^x} * \boldsymbol{d_t^x}}}{\widehat{a_t^x}} \qquad when \;\; \widehat{\boldsymbol{a_t}} > 0$$

$$\forall\, t = 10, 20, \ldots 300 \qquad (12)$$

$$\boldsymbol{p_t^x} = \frac{\boldsymbol{d_t^x}}{\widehat{\boldsymbol{v_t^x}}} \qquad when \;\; \widehat{\boldsymbol{a_t}} \le 0$$

The speed data from 100 trains were used to test the above six model types. **TABLE 1** lists the AAEs of these models at each time in detection (time during which a train dwells in the detection zone) from 10 seconds to 300 seconds. The number of trains present in each interval is also included. It may be seen that the number of trains decreases as the time in detection increases because some trains will have arrived at the crossing. Obviously, once this occurs it is not necessary to include them in the calibration.

In **TABLE 1**, the best prediction models (e.g., the ones with the lowest AAE) for each prediction interval are marked with an asterisk. Generally, the AAEs for each model type decreased as time in detection increased. It can be seen that the second order models (Model 4, Model 5, and Model 6) have lower AAEs than the first order models (Model 1, Model 2, and Model 3) for all the detection intervals except for the first one. Thus, it was concluded that the second order models are better than the first order models for predicting train arrival times. It can also be seen from **TABLE 1** that Model 4 using instantaneous speed, and Model 6 using moving average speed, yield the lowest AAEs in 27 out of the 30 prediction intervals. Model 6 yields the lowest AAEs during most of the detection intervals between 20 seconds and 150 seconds. Although Model 4 yields most of the lowest AAEs at detection times between 160 seconds to 300 seconds, the differences in AAEs between Model 4 and Model 6 are less than 1 second. If only one model were to be used, for consistency, then Model 6 is recommended. Note that Model 6 is 30% better, on average, than Model 1, which is based on current practice.

TABLE 1 Kinematic Model AAE Results

Time in Detection k (s)	Model1	Model2	Model3	Model4	Model5	Model6	# of Trains
10	138.7	138.7	138.7*	172.4	172.4	172.4	100
20	111.7	114.8	114.8	76.6	75.8	75.8*	100
30	107.4	111.7	108.9	73.5	71.6*	71.6*	100
40	103.3	109.2	105.1	69.0	68.0*	68.2	100
50	99.8	106.9	101.3	70.7	69.6*	69.6*	100
60	94.7	104.4	96.7	65.7	66.0	64.7*	100
70	90.6	101.8	92.5	55.4	56.7	55.3*	99
80	86.8	99.2	88.2	55.4	56.5	55.0*	99
90	83.2	96.6	84.6	54.9	57.8	54.3*	99
100	80.7	95.0	81.5	54.6	57.4	53.9*	97
110	77.5	92.6	78.1	49.2	54.1	48.9*	97
120	73.7	90.5	75.4	43.3	46.8	43.3*	95
130	71.1	88.2	72.1	49.8	55.1	49.6*	95
140	68.1	85.7	69.2	45.3	51.6	45.1*	94
150	65.7	83.7	67.1	46.0	52.0	45.9*	93
160	64.4	81.5	65.5	43.1*	50.8	43.2	90
170	62.3	79.6	63.4	40.3*	47.4	40.5	89
180	60.7	78.2	61.7	46.0*	54.5	46.1	87
190	59.1	77.4	60.5	43.3*	51.7	43.6	84
200	57.0	75.6	58.5	41.4*	50.9	41.7	83
210	59.3	78.5	60.0	42.6*	55.2	42.9	74

220	59.3	79.7	60.4	41.6*	52.5	41.9	68
230	63.7	86.7	64.2	47.3	61.7	47.2*	56
240	64.4	88.9	65.4	44.5*	59.4	44.5*	51
250	66.2	93.0	67.7	43.6*	57.6	44.3	45
260	69.9	97.3	70.9	52.8*	70.7	53.2	39
270	71.9	102.4	73.1	45.5*	62.5	46.0	34
280	73.3	108.8	73.9	51.6	69.2	51.2*	30
290	75.9	115.6	78.1	42.2*	62.6	43.1	25
300	75.0	118.7	77.5	47.2*	73.4	48.3	23
Average	**77.8**	**96.9**	**79.1**	**55.0**	**63.5**	**54.9**	

* Best prediction model for the given interval

Regression Models (Model 7-Model 21)

Models 7 through 21 are regression models. As in the kinematic model section, the same 100 trains were used as training data for model calibration, and the remaining 38 trains were used for model validation.

Model 7, Model 8, and Model 9 use speed as the only independent variable, as shown in Equations 13, 14, and 15. Similar to Model 1 through Model 3, instant speed, average speed, and moving average speed were used to represent speed in Model 7, Model 8, and Model 9, respectively.

Model 7

$$\boldsymbol{p_t^x} = c_1 + \boldsymbol{v_t^x} * c_2 \quad \forall\, t = 10\,,20\,,\ldots,300 \tag{13}$$

Model 8

$$\boldsymbol{p_t^x} = c_1 + \overline{\boldsymbol{v_t^x}} * c_2 \quad \forall\, t = 10\,,20\,,\ldots,300 \tag{14}$$

Model 9

$$\boldsymbol{p_t^x} = c_1 + \widehat{\boldsymbol{v_t^x}} * c_2 \quad \forall\, t = 10\,,20\,,\ldots,300 \tag{15}$$

Model 10 through Model 12 include speed and squared speed as independent variables. The squared speed variable was chosen in order to examine if there were any second order relationships between speed and arrival time.

Model 10

$$\boldsymbol{p_t^x} = c_1 + \boldsymbol{v_t^x} * c_2 + (\boldsymbol{v_t^x})^2 * c_3 \quad \forall\, t = 10\,,20\,,\ldots,300 \tag{16}$$

Model 11

$$\boldsymbol{p_t^x} = c_1 + \overline{\boldsymbol{v_t^x}} * c_2 + \overline{\boldsymbol{v_t^x}}^2 * c_3 \quad \forall\, t = 10\,,20\,,\ldots,300 \tag{17}$$

Model 12

$$\boldsymbol{p_t^x} = c_1 + \widehat{\boldsymbol{v_t^x}} * c_2 + \widehat{\boldsymbol{v_t^x}}^2 * c_3 \;\; \forall\, t = 10\,,20\,,\ldots,300 \tag{18}$$

Models 13 through 15 use speed, squared speed, and estimated remaining distance to the crossing as independent variables, as shown in Equations 19 to 21. As the train approaches the crossing, arrival time should decrease given that train speed does not change dramatically. Therefore, it is hypothesized that the estimated remaining distance is negatively correlated to arrival time. As before, the three different speed definitions were used in these models.

Model 13

$$\boldsymbol{p_t^x} = c_1 + \boldsymbol{v_t^x} * c_2 + \boldsymbol{v_t^x}^2 * c_3 + \boldsymbol{d_t^x} * c_4 \;\; \forall\, t = 10\,,20\,,\ldots,300 \tag{19}$$

Model 14

$$\boldsymbol{p_t^x} = c_1 + \overline{\boldsymbol{v_t^x}} * c_2 + \overline{\boldsymbol{v_t^x}}^2 * c_3 + \boldsymbol{d_t^x} * c_4 \;\forall\, t = 10\,,20\,,\ldots,300 \tag{20}$$

Model 15

$$\boldsymbol{p_t^x} = c_1 + \widehat{\boldsymbol{v_t^x}} * c_2 + \widehat{\boldsymbol{v_t^x}}^2 * c_3 + \boldsymbol{d_t^x} * c_4 \;\forall\, t = 10\,,20\,,\ldots,300 \tag{21}$$

Model 16 through Model 21 are natural logarithm-based models. Specifically, Model 16, Model 17, and Model 18 use the natural logarithm of speed as independent variables, while Models 19 through 21 also include the logarithm of the estimated remaining distance to the HRGC. The natural logarithm was chosen to see if there was a log-linear relationship between the independent variables and the arrival time.

Model 16

$$\boldsymbol{Ln}\,(\boldsymbol{p_t^x}) = c_1 + Ln\,(\boldsymbol{v_t^x}) * c_2 \;\forall\, t = 10\,,20\,,\ldots,300 \tag{22}$$

Model 17

$$\boldsymbol{Ln}\,(\boldsymbol{p_t^x}) = c_1 + Ln\left(\overline{\boldsymbol{v_t^x}}\right) * c_2 \;\; \forall\, t = 10\,,20\,,\ldots,300 \tag{23}$$

Model 18

$$\boldsymbol{Ln}\,(\boldsymbol{p_t^x}) = c_1 + Ln\left(\widehat{\boldsymbol{v_t^x}}\right) * c_2 \;\; \forall\, t = 10\,,20\,,\ldots,300 \tag{24}$$

Model 19

$$\boldsymbol{Ln}\,(\boldsymbol{p_t^x}) = c_1 + Ln\,(\boldsymbol{v_t^x}) * c_2 + Ln(\boldsymbol{d_t^x}) * c_3 \;\; \forall\, t = 10\,,20\,,\ldots,300 \tag{25}$$

Model 20

$$\boldsymbol{Ln}\left(\boldsymbol{p_t^x}\right) = c_1 + Ln\left(\overline{\boldsymbol{v_t^x}}\right) * c_2 + Ln(\boldsymbol{d_t^x}) * c_3 \;\; \forall\, t = 10\,,20\,,\ldots,300 \qquad (26)$$

Model 21

$$\boldsymbol{Ln}\left(\boldsymbol{p_t^x}\right) = c_1 + Ln\left(\widehat{\boldsymbol{v_t^x}}\right) * c_2 + Ln(\boldsymbol{d_t^x}) * c_3 \;\; \forall\, t = 10\,,20\,,\ldots,300 \qquad (27)$$

where:

c_1 intercept coefficient

c_2, c_3 coefficients for the independent variables

The overall results for each 10-second time period are shown in **TABLE 2**. This table shows the "best" models, which are defined as the models with the lowest AAE for the given ten second detection interval, for each of the 30 detection intervals. The equations, the AAEs, and the coefficient of determination (i.e. R^2) value of these models are shown in **TABLE 2** as well.

It can be seen that the best models in each interval are either natural logarithm-based models or second order models. This indicates that the relationship between measured speed and arrival time is non-linear. This would be expected because, in general, as the trains leave Lincoln, they are allowed to travel at a faster speed and this section of track will experience trains with fairly high acceleration rates. It was found that as the time in detection increases, the coefficient of determination of the best model increases, and the AAE decreases slightly. In other words, as more information is gathered about the train, the model accuracy improves. This makes intuitive sense because as the train travels closer to the HRGC, it would be expected that model accuracy would increase.

TABLE 2 Best Regression Model AAE Results

Time in detection (s)	Best Regression Model	Equations of the Regression Models *	AAE (s)	R^2
10	Model 19	Ln(p)=104.78-1.02*Ln(v)-12.21Ln(d)	33.6	0.64
20	Model 19	Ln(p)=82.22-1.18*Ln(v)-9.36*Ln(d)	32.6	0.65
30	Model 19	Ln(p)=68.32-1.31*Ln(v)-7.61*Ln(d)	31.7	0.66
40	Model 21	Ln(p)=55.51-1.37*Ln(ṽ)-5.995*Ln(d)	32.0	0.67
50	Model 19	Ln(p)=41.63-1.336*Ln(v)-4.26*Ln(d)	30.3	0.70
60	Model 19	Ln(p)=34.33-1.35*Ln(v)-3.34*Ln(d)	30.0	0.71
70	Model 21	Ln(p)=26.948-1.31*Ln(ṽ)-2.42*Ln(d)	30.0	0.71
80	Model 15	p=1864.6-121.52*ṽ+3.77*ṽ^2-0.39*d	31.2	0.72
90	Model 16	Ln(p)=7.24-0.88*Ln(v)	30.9	0.72
100	Model 16	Ln(p)=7.28-0.92*Ln(v)	30.4	0.73
110	Model 16	Ln(p)=7.35-0.97*Ln(v)	29.8	0.73
120	Model 16	Ln(p)=7.39-1.01*Ln(v)	29.7	0.74

130	Model 18	Ln(p)=7.46-1.06*Ln(ṽ)	29.2	0.75
140	Model 18	Ln(p)=7.51-1.10*Ln(ṽ)	28.7	0.75
150	Model 14	p=5259.86-324.38*ū+3.18*ū^2-1.58*d	27.6	0.70
160	Model 14	p=3845.27-265.26*ū+3.44*ū^2-1.10*d	28.1	0.70
170	Model 14	p=4039.1-286.78*ū+3.34*ū^2-1.17*d	28.1	0.71
180	Model 14	p=6333.22-434.97*ū+2.73*ū^2-1.96*d	26.3	0.72
190	Model 16	Ln(p)=8.01-1.46*Ln(v)	28.7	0.78
200	Model 12	p=652.76-88.87*ṽ+3.32*ṽ^2	28.0	0.75
210	Model 10	p=657.203-90.94*v+3.42*v^2	27.6	0.76
220	Model 10	p=657.19-90.34*v+3.46*v^2	29.0	0.74
230	Model 10	p=670.546-98.64*v+3.96*v^2	28.8	0.74
240	Model 10	p=656.23-95.90*v+3.79*v^2	29.8	0.76
250	Model 10	p=643.91-94.03*v+3.67*v^2	31.5	0.75
260	Model 10	p=639.29-93.45*v+3.60*v^2	32.1	0.76
270	Model 15	p=1150.6-147.13*ṽ+5.20*ṽ^2-0.17*d	30.1	0.84
280	Model 15	p=1062-138.97*ṽ+5.0*ṽ^2-0.15*d	32.7	0.82
290	Model 7	p=491.46-47.71*v	33.2	0.75
300	Model 16	Ln(p)=8.0-1.60*Ln(v)	29.6	0.81

***** v: instantaneous speed, ū: average speed, ṽ: moving average speed

Model Comparison and Validation

TABLE 3 shows the validation results of the regression models and the kinematic models using the data set of 38 trains. In 22 of the 30 intervals, the regression models have lower AAEs than the kinematic models, indicating that overall the best regression models are more accurate than the best kinematic models in train arrival time prediction. On average, the AAE values for the regression models are 13% less than the AAE values for the kinematic models. The regression models in the 27^{th} through 30^{th} intervals have higher AAEs than the kinematic models when using the validation data. It is hypothesized that this is because the decrease in the number of observations in these latter intervals decreases the ability of the regression models to capture the wide variety of train behavior.

TABLE 3 Model Validation AAEs

Time in detection (s)	AAE (s)		Difference (%)	# of Trains
	Best Kinematic Model	Best Regression Model	$\frac{\text{Kinematic} - \text{Regression}}{\text{Regression}}$	
10	101.6	47.3	53.5%	38
20	58.6	48.3	17.5%	38
30	62.2	50.4	18.9%	38
40	54.1	47.6	11.9%	38
50	58.1	49.7	14.5%	38
60	52.9	50.3	4.9%	38
70	49.7	50.1	-0.9%	38

80	53.1	46.3	12.7%	38
90	52.4	46.1	11.9%	38
100	48.7	46.8	3.8%	38
110	46.7	45.7	2.0%	38
120	39.2	42.7	-8.9%	37
130	47.8	41.8	12.5%	36
140	45.8	42.0	8.3%	36
150	39.3	34.1	13.3%	36
160	43.6	37.3	14.4%	35
170	45.4	35.4	21.9%	34
180	45.8	30.6	33.2%	33
190	42.7	43.1	-1.0%	31
200	44.6	47.6	-6.6%	31
210	46.4	45.1	3.0%	26
220	48.5	34.6	28.5%	23
230	48.6	39.9	17.8%	23
240	50.2	36.5	27.3%	20
250	51.6	29.5	42.9%	19
260	50.6	30.3	40.0%	18
270	45.7	44.9	1.6%	13
280	41.2	52.9	-28.5%	10
290	40.1	51.1	-27.5%	8
300	43.5	54.0	-24.3%	7
Average	49.9	43.4	13.1%	

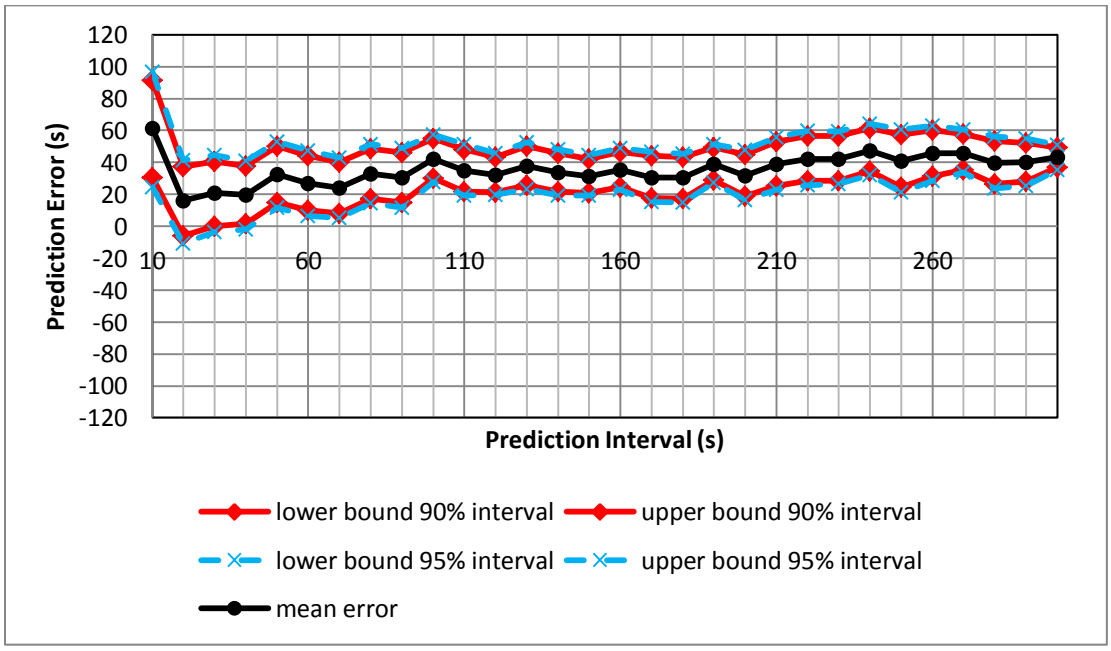
ESTIMATION OF PREDICTION ERROR INTERVALS

In addition to the earlier predicted train arrival time, the confidence interval (CI) about the predicted train arrival is also very useful in the advanced traffic signal preemption algorithms to adjust the predicted arrival times and mitigate the prediction errors *(2).* Both standard statistical methods *(12)* and bootstrap methods *(13)* can be used to estimate confidence intervals. The former are used for multiple linear regression (MLR) modeling because the theory and equations are well known. For non-linear models, such as the kinematic models and the local polynomial models *(14)*, standard statistical methods do not apply because they rely on prior information related to the models and parameter distributions *(12)*. The bootstrap method is a computer-based numerical method for statistical inference, especially for statistics such as sample median and sample correlation, which are difficult to infer using an analytical approach *(12, 13)*. Therefore, the bootstrap method was used to calculate the confidence intervals of mean prediction errors for both kinematic models and regression models.

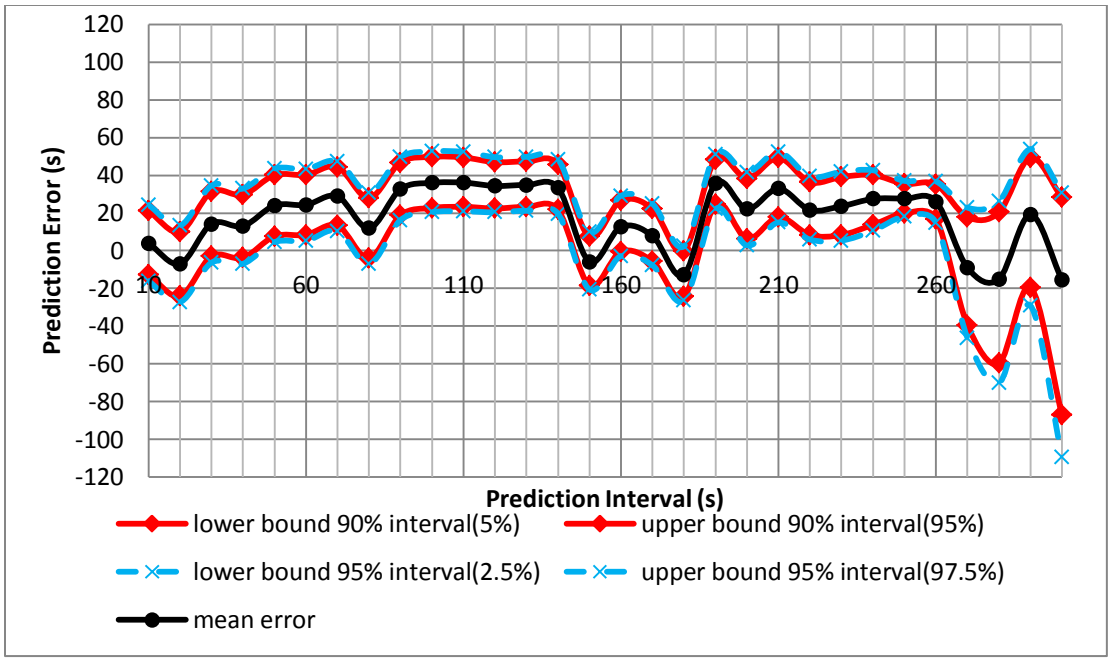
FIGURE 4 a) illustrates the mean prediction error (actual arrival time minus predicted arrival time) and the bootstrap CIs at 90 and 95 percentiles from the prediction using the best kinematic models, while **FIGURE 4 b)** illustrates those from the prediction using the best regression models. It may be seen that both the kinematic models and regression models can predict the train arrival time to within approximately $\pm$ 60 s of the true arrival time for most

prediction intervals. This accuracy is adequate for applying advanced preemption algorithm for an average train arrival time of 200 seconds and above (*6*).On average, the regression models had a prediction error of 17.3 seconds, which is 51.4% lower than that of the kinematic models (35.6 seconds).

However, the performance of the regression models degrades after 260 seconds in detection, while the performance of the kinematic models remains stable. It is hypothesized that this is due to the small size of the training data sample in the regression models. It is suggested that the best regression models be used prior to 260 seconds in detection. After 260 seconds of detection, either the best kinematic models can be used or the last predicted arrival time can be used as the final estimate. As the train data set increases, this range may vary. For the application, the analysis can be redone at set intervals (e.g., every few months), so that the fittest models can be applied.



a) Kinematic models



b) Regression models

FIGURE 4 Mean errors and the bootstrap CIs.

CONCLUDING REMARKS

This paper developed a system for train speed measurements and train arrival time prediction at HRGCs using second-generation train detection technologies and competitive kinematic and regression models. On average, the regression models were proven to be 13% more accurate than the kinematic models based on the validation data. The bootstrap method was used to estimate the error bounds of the mean estimates of the models, which, along with the predicted arrival times, are useful to new advanced traffic signal preemption algorithms *(2, 8).* It was found that the average prediction error for the regression models was approximately 50% smaller than the kinematic models, all else being equal.

With respect to the practicality of the models, the regression approach needs sufficient data to calibrate and validate the models at regular intervals. The advantage of the kinematic models is that they do not require calibration and are ready to implement in the field. Intuitively, they would be especially useful at locations with low train volumes and/or locations where train speeds do not vary much. The trade-off between the accuracy and ease-of-use of the different model categories would be application dependent, and the transportation agencies would have to decide which best fits their needs. In situations where there is not enough sample data, a combined use of kinematic and regression models would be a solution. It should also be pointed out that although the prediction models in the paper are developed using only a freight train sample, the models can be extended to passenger and/or light rail systems.

Future work will focus on streamlining the components of this system and integrating the system in new advanced traffic signal preemption algorithms *(2, 8).* In this paper, the methodology was used to predict train arrival times for EB trains at a single HRGC. This methodology can be extended to predict train arrival times for both EB and WB trains at multiple HRGCs (e.g., the 33rd Street, Adams Street, and 44thStreet HRGC along the test bed).

ACKNOWLEDGEMENTS

The authors would like to thank the U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology and the University Transportation Center for Railway Safety for their sponsorship and support of this research. The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein, and are not necessarily representative of the sponsoring agency.

REFERENCES

1. FHWA. *Manual on Uniform Traffic Control Devices: 2009 Edition*, U.S. Department of Transportation, 2009.
2. Cho, H. Preemption Strategy for Intersections near Highway-Railroad Crossings. Ph.D. Dissertation, Texas A&M University, College Station, TX, 2003.
3. Association of American Railroads. *Communication and Signal Section, Signal Manual of Recommended Practice*, Vol. 1, Section 3, Washington, D.C., 1995.
4. Cho, H. and Rilett, L.R. Forecasting train travel times at At-Grade Crossings, In: *Transportation Research Record, No.1844*, TRB, Washington, D.C., 2003, pp. 94-102.
5. Korve, H.W. *Traffic Signal Operations Near Highway-Railroad Grade Crossings*, NCHRP Synthesis 271, TRB, National Research Council, Washington, D.C., 1999.
6. Esters, R.M. and Rilett, L.R. Advanced Prediction of Train Arrival and Crossing Times at Highway-railroad Grade Crossings, In *Transportation Research Record,No.1844*, Transportation Research Board, Washington, D.C. 2003, pp. 94-102.
7. Richards, S.H., K.W. Heathington, and D.B. Fambro. Evaluation of Constant Warning Times Using Train Predictors at a Grade Crossing with Flashing Light Signals. In *Transportation Research Record 1254*, TRB, National Research Council, Washington, D.C., 1990, pp. 60-71.
8. Chen, Y. An Adaptive Corridor-wide Signal Timing Optimization Methodology for Traffic Networks with Multiple Highway-Rail Grade Crossings. PhD Dissertation. Department of Civil Engineering, University of Nebraska-Lincoln, 2015.
9. Jones, Elizabeth G., Khattak, Aemal H., and Rilett, L.R. The University Nebraska-Lincoln's Highway Rail Grade Crossing Test Bed System. In *Transportation Research Record: Journal of the Transportation Research Board*. CD-ROM. TRB of the National Academics, Washington, D.C., 2009.
10. Franca, D. Estimation Train Arrival Times at Highway-Railroad Grade Crossings Using Multiple Sensors, Master Thesis, University of Nebraska, Lincoln, NE, 2009.
11. Chen, Y.F. and Rilett, L. R. A train speed measurement and arrival time prediction system for highway-rail grade crossings (HRGCs). *Proceedings of 21ST World Congress on Intelligent Transport System*, Detroit, Michigan, September 8, 2014.
12. Spiegelman, Clifford H., Park, Eun Sug, and Rilett, L.R.*: Transportation Statistics and Microsimulation.* CRC Press, 2011.
13. Efron, B. and Tibshirani, R.J. *An Introduction to Bootstrap*. Chapman and Hall, New York, 1993.
14. Hang Yue, E. Jones, and Peter Revesz. Local Polynomial Regression Models for Average Traffic Speed Estimation and Forecasting in Linear Constraint Databases.*17th International Symposium on Temporal Representation and Reasoning (TIME'10)*, IEEE, Paris, France, September, 2010.