Final Report

Developing an Intelligent Connected Vehicle Based Traffic State Estimator

Mohammad A. Aljamal, Ph.D.
m7md92@vt.edu

Mohamed Farag, Ph.D.
mfarag@vtti.vt.edu

Hossam M. Abdelghaffar, Ph.D.
hossamvt@vt.edu

Ahmed Abdelrahman, M.Sc.
anhafz@vt.edu
Virginia Tech Transportation Institute
3500 Transportation Research Plaza
Blacksburg, VA 24061

Hesham A. Rakha, Ph.D., P.Eng., FIEEE
Charles E. Via, Jr. Department of Civil and Environmental Engineering
Virginia Polytechnic Institute and State University
3500 Transportation Research Plaza
Blacksburg, VA 24061
Phone: (540) 231-1505 Fax: (540) 231-1555
hrakha@vt.edu

Date
March 2022

Prepared for the Urban Mobility & Equity Center, Morgan State University, CBEIS 327, 1700 E. Coldspring Lane, Baltimore, MD 21251

ACKNOWLEDGMENT

This research was funded by the Urban Mobility & Equity Center at Morgan State University and the University Transportation Center(s) Program of the U.S. Department of Transportation.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

1. Report No. UMEC-038

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
Developing an Intelligent Connected Vehicle based Traffic State Estimator

5. Report Date
March 2022

6. Performing Organization Code

7. Author(s)
Mohammad A. Aljamal (ORCID # 0000-0003-4251-1899)
Mohamed Farag
Hossam M. Abdelghaffar (ORCID # 0000-0003-4396-5913)
Ahmed Abdelrahman
Hesham A. Rakha (ORCID # 0000-0002-5845-2929)

8. Performing Organization Report No.

9. Performing Organization Name and Address
Virginia Tech Transportation Institute
3500 Transportation Research Road
Blacksburg, VA 24061

10. Work Unit No.

11. Contract or Grant No.
69A43551747123

12. Sponsoring Agency Name and Address
US Department of Transportation
Office of the Secretary-Research
UTC Program, RDT-30
1200 New Jersey Ave., SE
Washington, DC 20590

13. Type of Report and Period Covered
Final January 2021 - March 2022

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Estimation of traffic state parameters is crucial in advanced traffic management systems. However, measuring these parameters in the field is not practical since they are categorized as spatiotemporal parameters. This report presents three estimation approaches to estimate the traffic volume existing on signalized links. The first approach includes three model-driven approaches (Kalman filter [KF], adaptive KF [AKF], and particle filter [PF]) using a single average level of market penetration (ρ) in the state-space equations based on connected vehicle (CV) data only. The second approach develops an artificial neural network (ANN) approach to estimate two ρ variables; ρ_{in} and ρ_{out}, to be used in the state-space equations. Fused CV and camera data are utilized to build the ANN approach. After that, the second approach integrates the ANN with the KF approach (KFNN approach) to estimate the traffic volume on signalized links. The third approach develops three data-driven approaches (ANN, k-nearest neighbor, and RF) to estimate the traffic volumes using only CV data to build the data-driven approaches. The three approaches were applied on a signalized intersection in downtown Blacksburg, VA. The results showed that the use of CV data only is sufficient to provide accurate traffic volume estimates. In addition, using two predicted variable values in the state-space equations is not recommended, as it may produce undesired large errors in the state equation. It was found that the ANN approach may over-estimate the first variable and under-estimate the second variable or vice versa for the same estimation step. Consequently, the second research approach is not recommended. Finally, the ANN is the most accurate estimation approach. However, taking into consideration the huge amount of data needed to train and build the ANN approach, the long computational time needed to build the ANN, and the constraints on keeping the traffic behavior the same as the behavior in the training data set, the use of the KF approach is highly recommended for the application of traffic state estimation due to its simplicity and applicability in the field.

17. Key Words: Real-time estimation; probe vehicle; traffic density; neural network; connected vehicles; level of market penetration rate.

18. Distribution Statement

19. Security Classif. (of this report) : Unclassified

20. Security Classif. (of this page) Unclassified

21. No. of Pages 17

22. Price

Contents

Table of Figures

List of Tables

Abstract

Estimation of traffic state parameters is crucial in advanced traffic management systems. However, measuring these parameters in the field is not practical since they are categorized as spatiotemporal parameters. This report presents three estimation approaches to estimate the traffic volume existing on signalized links. The first approach includes three model-driven approaches (Kalman filter [KF], adaptive KF [AKF], and particle filter [PF]) using a single average level of market penetration (ρ) in the state-space equations based on connected vehicle (CV) data only. The second approach develops an artificial neural network (ANN) approach to estimate two ρ variables; ρin and ρout, to be used in the state-space equations. Fused CV and camera data are utilized to build the ANN approach. After that, the second approach integrates the ANN with the KF approach (KFNN approach) to estimate the traffic volume on signalized links. The third approach develops three data-driven approaches (ANN, k-nearest neighbor, and RF) to estimate the traffic volumes using only CV data to build the data-driven approaches. The three approaches were applied on a signalized intersection in downtown Blacksburg, VA. The results showed that the use of CV data only is sufficient to provide accurate traffic volume estimates. In addition, using two predicted variable values in the state-space equations is not recommended, as it may produce undesired large errors in the state equation. It was found that the ANN approach may over-estimate the first variable and under-estimate the second variable or vice versa for the same estimation step. Consequently, the second research approach is not recommended. Finally, the ANN is the most accurate estimation approach. However, taking into consideration the huge amount of data needed to train and build the ANN approach, the long computational time needed to build the ANN, and the constraints on keeping the traffic behavior the same as the behavior in the training data set, the use of the KF approach is highly recommended for the application of traffic state estimation due to its simplicity and applicability in the field.

1. Introduction

In the U.S., people wasted around 166 billion hours in traffic congestion in 2017, which led to a waste of around 3.8 billion gallons of fuel. Traffic engineers and researchers are making efforts to provide solutions for the traffic congestion problem. One efficient solution is to deploy Intelligent Transportation System (ITS) applications with the aim of increasing the capacity of the existing traffic infrastructure (Wang 2010). One such ITS application is the use of connected vehicle (CV) technology, which can allow the exchange of information between two CVs (vehicle-to-vehicle communication) and also the exchange of information between any CV and the traffic infrastructure (vehicle-to-infrastructure communication). In the case of traffic congestion, traffic infrastructure, such as a traffic signal controller, can send early messages to the surrounding CVs to find alternative routes, leading to a reduction in travel time.

Traffic congestion can be represented by the macroscopic traffic stream density (the number of vehicles that traverse a specific traffic segment divided by the length of that segment). Traffic density is considered a spatial rather than a temporal measurement. Consequently, the temporal traffic occupancy measurements, obtained from loop detectors, cannot be used to estimate the traffic density for the entire link unless multiple loop detectors are installed. However, this results in high installation costs. A more efficient way to estimate the traffic density is to exploit the main advantage of CV technology, which is its ability to share real-time information, such as the vehicle's location and speed, anywhere inside the link.

To estimate the number of vehicles in a link, researchers have developed different estimation approaches, such as model-driven approaches (filtering techniques) and data-driven approaches (machine learning). In addition, different data sources have been used to implement the proposed estimation approaches, such as data from fixed sensors (e.g., loop detectors), data from two different detection sources (fusion data), and CV data.

1.1. Model-Driven Estimation Approaches

For the use of fixed sensors, the input-output approach has been widely used to develop model-driven approaches. One study developed a Kalman filter (KF) approach to estimate the vehicle counts in a signalized link using at least three loop detectors—two at the boundaries of the tested link and the third one in the middle of the link (Vigos, Papageorgiou, and Wang 2008). Another study (Ghosh and Knapp 1978) employed data from four loop detectors to estimate the number of vehicles, resulting in accurate estimates. Traffic flow and occupancy data, measured from six loop detectors, were utilized to provide accurate estimates for the vehicle counts in an on-ramp segment (Bhouri et al. 1989). However, these studies require a high implementation cost for installing multiple fixed sensors. Moreover, it was found that fixed sensors always produce some noise in their data, requiring the use of additional data sources to reduce that noise (Mimbela and Klein 2007).

Fusion data has gained more attention following the introduction of advanced technologies such as CV technology. Recently, researchers have started using fixed sensors together with CV data for finding better estimation accuracy. One such study attempted to provide accurate estimates of traffic density using mobile sensors and loop detector data (Herrera and Bayen 2007), showing that estimation accuracy using fusion data outperformed estimation using loop detector data. A recent study utilized CVs and cameras to estimate traffic density in a 500 m highway segment. The model's development was based on the assumption that the average CV speed is approximately equal to the average speed of traditional vehicles (Bekiaris-Liberis, Roncoli, and Papageorgiou 2016). In that study, a KF model was developed under the consideration of having a linear parameter-varying system with known parameters. The state equation was based on the traffic flow continuity equation, while the measurement equation was based on the average speed of CVs. Wright and Horowitz (2016) developed a particle filter (PF) using fusion loop and CV measurements to estimate the number of vehicles in a freeway section, demonstrating that the use of fusion data resulted in improved estimation accuracy. Another study (Di, Liu, and Davis 2010) developed a KF approach using fused loop and CV data to estimate the number of vehicles in a signalized link.

Recently, a few studies have attempted to estimate the number of vehicles in signalized links using CV data only. In those studies, KF, adaptive KF (AKF), and PF model-driven approaches were developed to provide accurate estimates (Aljamal, Abdelghaffar, and Rakha 2020b) (Aljamal, Abdelghaffar, and Rakha 2019a) (Aljamal, Abdelghaffar, and Rakha 2019b) (Aljamal, Abdelghaffar, and Rakha 2020a).

1.2. Data-Driven Approaches

Machine learning techniques have always required a large amount of data to build mathematical models that draw the relationship between the model's inputs and outputs, and as such, machine learning is considered a data-driven technique. Data-driven approaches have been employed to estimate traffic state variables such as traffic density and speed (Aljamal, Abdelghaffar, and Rakha 2019a) (Fulari, Vanajakshi, and Subramanian 2017) (Antoniou and Koutsopoulos 2006) (Wassantachat et al. 2009) (Jahangiri, Rakha, and Dingus 2015) (Sekuła et al. 2018) (Raj, Bahuleyan, and Vanajakshi 2016). In previous studies, proposed estimation approaches have relied on different detection techniques, such as fixed sensors and fusion data.

Artificial neural network (ANN) and k-nearest neighbor (k-NN) data-driven approaches were developed to produce reliable estimates for vehicle counts (Raj, Bahuleyan, and Vanajakshi 2016). In that study, authors relied on fixed sensors to obtain traffic speed and flow measurements to build and train the ANN and the k-NN approaches. Fulari, Vanajakshi, and Subramanian (2017) developed an ANN approach to estimate the number of vehicles using video and Bluetooth data. It was found that the ANN approach performed well if a good quantity of training data was accessible. Fused loop and CV data were used to develop support vector machine and k-NN approaches, with the aim of estimating the level of traffic congestion in a freeway segment (Khan, Dey, and Chowdhury 2017). Another study (Sekuła et al. 2018) deployed data from fixed sensors and CVs to build different data-driven estimation approaches such as ANN, k-NN, and random forest (RF) to estimate hourly traffic volumes. In that study, the ANN approach was found to outperform the other approaches. Aljamal, Abdelghaffar, and Rakha (2019a) developed an ANN approach to estimate the CV level of market penetration (LMP) rate. In that study, the ANN approach provided the AKF approach with real-time values of the LMPs, resulting in an improved vehicle count estimation accuracy. The LMP represents the percentage of the CVs in relation to the total number of vehicles.

In summary, previous studies have shown the benefits of using data-driven approaches in addressing different aspects of the traffic state estimation problem. Therefore, the research described in this study aims to develop data-driven approaches in the application of traffic stream density estimation (vehicle counts). One commonality among the related previous studies is that they all estimated vehicle counts using data from fixed sensors or using fused source data (e.g., loop with CV data).

The research described in this study aims to develop different data-driven estimation techniques to estimate the vehicle counts using only CV data. The proposed estimation approaches are applied to test a signalized link in downtown Blacksburg, VA. The proposed research extends the state-of-the-art in vehicle count estimation by making three major contributions:

1. Developing three data-driven estimation approaches to estimate the vehicle counts in signalized links. The three data-driven approaches are developed using only CV data.
2. Developing a data-driven approach to estimate the LMP for CVs at the entrance and exit of the link.
3. Comparing the three proposed data-driven approaches with state-of-the-art model-driven estimation approaches.

2. Development of Simulation Data

A congested link in downtown Blacksburg, VA was selected to evaluate the proposed estimation approaches. The link falls between two traffic signals, as shown in **Error! Reference source not found.**. The link length is 97 meters. INTEGRATION traffic simulation software was used to simulate the network in **Error! Reference source not found.**. The traffic Origin-Destination (O-D) values for the network were calibrated using real count data. The speed limit of the tested link is 40 km/h, the speed-at-capacity is 32 km/h, the jam density is 160 veh/km/ln, and the saturation flow rate is 1800 veh/h/ln.

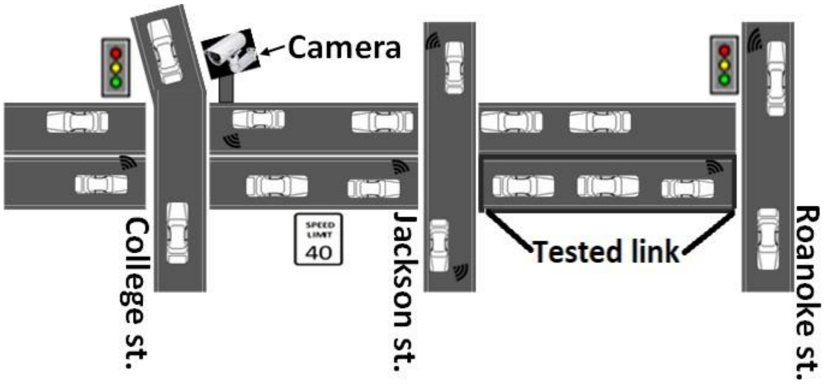

Figure 0.1: Tested link section in downtown Blacksburg, VA

2.1. Generation of the Training Dataset

Training data are needed to develop machine learning estimation approaches. INTEGRATION simulation software was used to facilitate the generation of the CV data, as the CV data are not easy to access. In the simulation input files, 400 scenarios, combining the O-D values and right turn traffic volumes that exit the Main street toward Jackson street, were considered. For the O-Ds and right turn traffic volumes, 20 different demand scaling factors generated from a normal distribution, ranging from 0.8 to 1.2, were used—for instance, a scenario of an 0.82 O-D demand scaling factor with 1.05 right turn volume demand scaling factor. In addition, 25 scenarios with different random seeds were considered for each LMP, resulting in a total of 1,000 scenarios (20 × 20 × 25) to build the training data. The INTEGRATION simulation software generates an output time-space file that includes real-time information about the CVs, such as each vehicle's location and speed. In section 1.3, more details are provided about the inputs and outputs that were considered in the training data set.

3. Methodology

In this section, three research approaches are presented: (1) model-driven approaches, (2) integrated data-driven and model-driven approaches, and (3) data-driven approaches. In the first research approach, linear and nonlinear filtering approaches were used to estimate the vehicle counts. The second approach first developed a data-driven approach to estimate the ratio of the number of CVs (NCV) to the total number of vehicles (NT), and then combined the data-driven approach with the most accurate model-driven approach to finally estimate the vehicle counts. The third approach developed data-driven approaches to directly estimate the vehicle counts.

3.1. First Approach: Model-Driven Approaches

Linear and nonlinear filtering approaches are presented in this section: (1) KF, (2) AKF, and (3) PF. These filtering techniques are always used to solve state-space models. A state-space model is represented by: (1) a state, and (2) a measurement system. The filtering techniques are mainly used to provide posterior estimates given some measurements, with the aim of minimizing the errors in the priori estimates.

In this chapter, the state-space model presented in (Aljamal, Abdelghaffar, and Rakha 2020b) is used to estimate the vehicle counts. The state and measurement equations are presented in Equations (1.1) and (1.3), respectively.

$$N(t) = N(t - \Delta t) + u(t) \tag{1.1}$$

$$u(t) = \frac{\Delta t \, [q^{in}(t) - q^{out}(t)]}{\max(\rho_{actual}, \rho_{min})} \tag{1.2}$$

$$TT(t) = H(t) \times N(t) \tag{1.3}$$

$$H(t) = \frac{1}{\bar{q}(t)} = \frac{2 \times \rho_{actual}}{q^{in}(t) + q^{out}(t)} \tag{1.4}$$

where N (t) is the number of vehicles crossing the link at time t, N (t − Δt) is the number of vehicles crossing the link in the preceding time interval, ρ is the CVs’ LMP, defined as the ratio of the CV counts to the total vehicle counts. In this research approach, the ρ is computed from historical data and assumed to remain constant for the entire simulation. For instance, if a scenario of 10% LMP is evaluated, the ρ value is assumed to be 10%. q_{in} and q_{out} represent the flow of CVs entering and exiting the link, respectively, during Δt. The Δt is updated when five CVs have traversed the tested link (Aljamal, Abdelghaffar, and Rakha 2020b). TT is the average travel time for CVs. The following subsections present three filtering techniques to solve the described state-space model.

The KF Approach

The KF is a linear filtering technique and can be implemented using the following equations:

$$\widehat{N}^{-}(t) = \widehat{N}^{+}(t - \Delta t) + \mathrm{u(t)} \tag{1.5}$$

$$\hat{T}T(t) = \ H\ (t)\ \times\ \widehat{N}^{-}(t) \tag{1.6}$$

$$\hat{P}^{-}(t) = \ \hat{P}^{+}(t - \Delta t) \tag{1.7}$$

$$G(t) = \ \hat{P}^{-}(t)H(t)^{T}\ [H(t)\hat{P}^{-}(t)\ \ H(t)^{T} + R]^{-1} \tag{1.8}$$

$$\widehat{N}^{+}(t) = \widehat{N}^{-}(t) + G(t)\ \ [TT(t) -\ \ \hat{T}T\ \ (t)] \tag{1.9}$$

$$\hat{P}^{+}(t) = \hat{P}^{-}(t) \times [1 - H\ \ (t)\ \ G\ \ (t)] \tag{1.10}$$

where $N^{\text{ˆ}-}$ and $N^{\text{ˆ}+}$ are the priori and the posterior vehicle count estimates, $TT^{\wedge}$ is the estimated average travel time, $P^{\text{ˆ}-}$ and $P^{\text{ˆ}+}$ are the priori and posterior covariance estimate for the state system, G is the Kalman gain, and R is the error covariance in the measurement system. For more details, readers can refer to (Aljamal, Abdelghaffar, and Rakha 2020b).

The AKF Approach

The linear AKF dynamically estimates the noise error values for the state and measurement systems every estimation step. The AKF approach can be solved using the following equations:

$$\widehat{N}^{-}(t) = \widehat{N}^{+}(t - \Delta t) + u(t) + m(t - \Delta t) \tag{1.11}$$

$$\hat{T}T(t) = \ H\ \ (t)\ \ \times\ \ \widehat{N}^{-}(t) \tag{1.12}$$

$$\widehat{P}^{-}(t) = \ \widehat{P}^{+}(t - \Delta t) + M(t - \Delta t) \tag{1.13}$$

$$r = \frac{1}{n}\sum_{t=1}^{n} r(t), \quad \text{where}\ \ r(t) = TT(t) - H(t)\,\widehat{N}^{-}(t) \tag{1.14}$$

$$R = \frac{1}{n-1}\sum_{t=1}^{n}\ [(r(t) - r).(r(t) - r)^{T} - (\frac{n-1}{n})H(t)\hat{P}^{-}(t)H^{T}(t)] \tag{1.15}$$

$$G(t) = \ \hat{P}^{-}(t)H(t)^{T}\ [H(t)\hat{P}^{-}(t)\ \ H(t)^{T} + R(t)]^{-1} \tag{1.16}$$

$$\widehat{N}^{+}(t) = \widehat{N}^{-}(t) + G(t)\ \ [TT(t) - H(t)\widehat{N}^{-}(t) - r(t)] \tag{1.17}$$

$$\widehat{P}^{+}(t) = \widehat{P}^{-}(t) \times [1 - H\ \ (t)\ \ G\ \ (t)] \tag{1.18}$$

$$m = \frac{1}{n}\sum_{t=1}^{n}\ \ m(t), \quad \text{where}\ \ m(t) = \widehat{N}^{+}(t) - \widehat{N}^{+}(t - \Delta t) - u(t) \tag{1.19}$$

$$M = \frac{1}{n-1}\sum_{t=1}^{n} \left[(m(t)-m).(m(t)-m)^T - (\frac{n-1}{n})\hat{P}^+(t-\Delta t) - \hat{P}^+(t)\right] \quad (1.20)$$

where r and R are the mean and covariance of the measurement noise, n is the number of state noise samples, and m and M are the mean and covariance of the state noise.

The PF Approach

The PF is a nonlinear filtering technique. First, the PF generates different particles with unique relative weights. In every estimation step, the system removes the particles with low relative weights and replaces them with new particles (resampling) and thus the system preserves only the important particles. The PF approach can be implemented using the following steps:

1) Initialization: t = 0
   N^+(0), R, V, and l.
   Generate particles:

$$N^l(0) \sim P(N_0) \quad (1.21)$$

2) For t = 1: T

$$N^l(t) = N^l(t-\Delta t) + u(t) \quad (1.22)$$

$$TT^l(t) = \ H(t) \times N^l(t) \quad (1.23)$$

$$w^l(t) = \frac{1}{\sqrt{2\pi R}}\ e^{-\left(TT-TT^l(t)\right)^2/2R} \quad (1.24)$$

$$\widehat{w}^l(t) = w^l(t)/\sum_{l=1}^{L} w^l(t) \quad (1.25)$$

After normalizing the weights using Equation (1.25), the low-weighted particles are replaced with new particles (resampling (Liu and Chen 1998)). After a few iterations in the PF process, the weight will focus on a few particles only and most particles will have insignificant weights, resulting in sample degeneracy (Li, Sattar, and Sun 2012). The resampling process is therefore used to tackle the degeneracy problem.

$$\widehat{N}^+(t) = \frac{1}{L}\sum_{l=1}^{L} N^l(t) \quad (1.26)$$

where V is the variance of the initial vehicle count estimate, N^l is the particles' locations from 1 to L, and TT is the observed measurement from the CVs. More details can be found in (Aljamal, Abdelghaffar, and Rakha 2020a)

3.2. Second Approach: Integrating Data-Driven and Model-Driven Approaches

In our state-space equations, the ρ variable is found to be the main source of noise in the state-space model (Aljamal, Abdelghaffar, and Rakha 2020b). Unlike the first research approach described in section 3.1, two ρ variables, instead of one ρ variable, are used in the state-space equations: (1) the ρ_{in}, and (2) the ρ_{out}. The ρ_{in} and ρ_{out} are observed at the entrance and exit of the link, respectively. The ρ_{in} and the ρ_{out} are displayed in Equations (1.27) and (1.28), respectively. The Acv, AT , Dcv, and DT are the number of CV arrivals, total number of arrivals, number of CV departures, and total number of departures, respectively. Equations (1.29) and (1.30) present the new formulation of the u(t) and H(t) using the two ρ variables.

$$\rho_{in}(t) = A_{cv}/AT \tag{1.27}$$

$$\rho_{out}(t) = D_{cv}/DT \tag{1.28}$$

$$u(t) = \Delta t \, [\frac{q_{in}(t)}{\rho_{in}(t)} - \frac{q_{out}(t)}{\rho_{out}(t)}] \tag{1.29}$$

$$H(t) = \frac{2}{\frac{q_{in}(t)}{\rho_{in}(t)} + \frac{q_{out}(t)}{\rho_{out}(t)}} \tag{1.30}$$

It should be noted that the two variables can be measured if two fixed sensors (e.g., cameras) are installed at the entry and exit of the tested link; however, the installation cost is high and thus makes this approach undesirable. A more efficient approach is to employ estimation techniques such as machine learning without the need to add any changes to the existing infrastructure. Hence, in this research approach, an ANN is developed to estimate the ρ_{in} and ρ_{out} variables.

ANN Approach

The ANN data-driven model is a combination of simple units (nodes) that are connected by links. The ANN aims to recognize relationships between an enormous amount of data by adding a certain number of neurous in the assigned hidden layers. The ANN contains three layers: the input layer, the hidden layer, and the output layer (Kubat 1999). The mechanism behind the ANN is that every node receives/sends signals from incoming/outgoing links by performing computations. The links that connect the nodes in the network have certain weight values, and these weights determine the strength of the connection between the nodes.

ANN Inputs and Outputs

In this section, the aim was to use the nearest existing fixed sensor with the CV data to build the ANN model. As seen in Figure 1.1, an existing camera was located upstream of the tested link (at the intersection of College Street). The camera in the field measures the total traffic counts at the intersection. Consequently, the total traffic count variable is used as an input for the ANN model. CVs are used to generate the inputs of the ANN model due to their their ability to provide measurements at any location inside the network.

Seven inputs are used to build the ANN approach, as follows:

1. The total traffic counts obtained from the camera (C_T),
2. The number of CVs on the tested link (N_{cv}),
3. The number of CVs at the entrance of the link (A_{cv}),
4. The space-mean speed of CVs (u_s),
5. The average speed for CVs at link entrance (S_1),
6. The average speed for CVs at link exit (S_2), and
7. The estimation interval time (Δt).

Figure 1.2 displays the ANN inputs and outputs. To build a strong ANN approach, the inputs must relate to the outputs, which allows the ANN to define the relationship between the two. For instance, a high value of traffic volumes (C_T , N_{cv}, and A_{cv}) means that we have more vehicles in the link, which results in large values in the denominator in Equations (1.27) and (1.28). The speed factor (us, S1, and S2) is also an important indicator of the level of congestion. A congested link can also result in having large values in the denominator in the two equations. It should be noted that the Δt variable strongly relates to the output variables. Remember that the Δt is not a constant value and is updated when five new CVs are observed at the end of the link. A high Δt value means that the number of CVs is low, which results in low output values. The ANN output variables are the ρ_{in} and ρ_{out}. In reality, the ρ output values vary between 0 and 1; a 0 means that no CVs are observed, while a 1 means that the number of CVs is equal to the total number of vehicles.

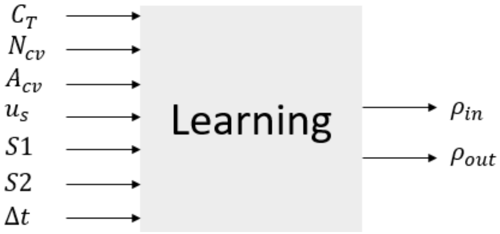

Figure 0.2: Estimate the ρ_{in} and ρ_{out} variables

The developed ANN approach consists of single hidden layer with 10 neurons using a transfer function of hyperbolic tansgent sigmoid and the Levenberg-Marquardt (LM) optimization method.

After developing and training the ANN approach, the estimated values for the ρ_{in} and ρ_{out} are used in our most accurate model-driven approach to estimate the main research goal, which is the vehicle count.

3.3. Third Approach: Data-Driven Approaches

The third research approach aims to directly estimate the vehicle counts by developing different data-driven estimation approaches: (1) ANN approach, (2) k- NN, and (3) RF approach. The data-driven approaches are developed using CV data only; data from the camera is not necessary. Six inputs are considered to train and build the data-driven approaches, as shown in Figure 1.3.

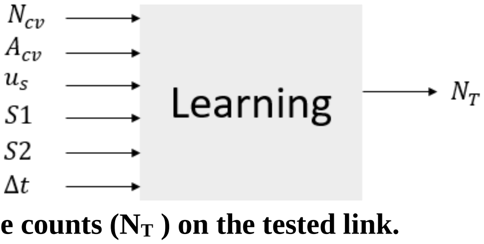

Figure 0.3: Estimate the vehicle counts (N_T) on the tested link.

ANN Approach

To estimate the vehicle counts using the ANN approach, the structure of the ANN consists of a single hidden layer with 10 neurons. A transfer function of hyperbolic tansgent sigmoid and the Levenberg-Marquardt (LM) optimization method are used.

k-NN Approach

The k-NN approach (Cover and Hart 1967) is used for classification and regression applications. The k-NN approach does not build a model but requires storing the entire data set. To estimate a new value using k-NN, the following information is required: (1) having access to the training records, (2) defining the distance metric to compute the distance between the records, and (3) identifying the value of the number of nearest neighbors (k). The results section will test different k values to find the optimal k value for the k-NN approach. The new estimated value is computed by taking the average value of the nearest neighbors.

RF Approach

The RF approach (Breima 2010) is a supervised learning technique and can be used in classification and regression. The RF is a set of decision trees. Each decision tree is constructed using a subset of inputs. The desired estimation values are given based on the majority votes from all trees. The advantage of using RF is its ability to handle large data sets without the need to create dummy variables. For the purpose of this study, 100 trees were used to develop the RF.

4. Results and Discussion

This section tests the accuracy of the three research approaches on a signalized link in downtown Blacksburg, VA. The Relative Root Mean Square Error (RRMSE) and the Root Mean Square Error (RMSE) are used to evaluate and compare the proposed estimation approaches. The RRMSE and RMSE can be computed using Equations (1.31) and (1.32), respectively.

$$RRMSE(\%) = 100\sqrt{S\sum_{s=1}^{S}[\widehat{N}^{+}(s)-N(s)]^{2}}/\sum_{s=1}^{S}N(S) \tag{1.31}$$

$$RMSE(veh) = \sqrt{\sum_{s=1}^{S}[\widehat{N}^{+}(s)-N(s)]^{2}/S} \tag{1.32}$$

where N (s) is the actual vehicle count, Nˆ+(s) is the estimated vehicle count value, and S is the total number of estimations.

4.1. First Research Approach

This section evaluates the three estimation model-driven approaches, (1) KF, (2) AKF, and (3) PF, using data from CVs only. The three approaches are used to estimate the number of vehicles crossing the tested link. Table 1.1 presents the RRMSE and RMSE values at different LMPs: 1, 3, 5, 8, 10, 15, 20, 30, 40, 50, 60, 70, 80, and 90%. For most of the LMP scenarios, the KF approach produces the lowest error values, while the PF approach outperforms the KF and the AKF for a few scenarios (1, 70, 80, and 90%). However, the nonlinear PF requires more computational time. Consequently, the use of the linear KF approach is highly recommended due to its simplicity and high-performance accuracy.

Table 0.1: RRMSE and RMSE values of KF, AKF, and PF approaches for different LMPs

LMPs %	RRMSE (%), RMSE (veh)		
	KF	AKF	PF
1	44, 3.2	58, 4.3	37, 2.8
3	39, 3.0	48, 3.6	39, 3.0
5	37, 2.8	44, 3.2	38, 2.9
8	36, 2.8	40, 3.0	37, 2.9
10	36, 2.8	38, 2.9	37, 2.9
15	36, 2.8	40, 3.0	39, 3.0
20	37, 2.8	39, 3.0	39, 3.0
30	38, 2.9	38, 2.9	42, 3.2
40	37, 2.9	38, 2.9	39, 3.0

4.2. Second Research Approach

First, the ANN approach is developed to estimate the percentage of the CVs to the total number of vehicles at the entry and the exit of the tested link; ρ_{in} and ρ_{out}, respectively. Table 1.2 presents the RRMSE values for estimating the two variables. The results demonstrate that the ANN produces reasonable error values; the errors for estimating the ρ_{in} vary between 14 and 25% while the error values are between 10 and 23% for the ρ_{out}.

After that, the estimated ρ values are used as inputs to the KF approach to estimate the vehicle counts on the tested link. A new approach, named KFNN, is developed based on integrating the KF and the ANN approaches. Remember that the KF approach uses an average one value of ρ in its equations, while the KFNN approach uses real-time values for the ρ_{in} and ρ_{out} in the KF equations at every estimation step. Table 1.3 shows the RRMSE values for estimating the vehicle counts using the KF and KFNN approaches. The table demonstrates that the KF approach outperforms the KFNN approach. Investigations were undertaken to find the reason for this. Findings suggested that the ANN may over-estimate the ρ_{in} and under-estimate the ρ_{out} or vice versa for the same estimation step, resulting in large errors in the state equation compared to the errors from using the average ρ. These large errors make the error correction from the KF more difficult. In conclusion, the use of one single rho value in the state-space equations is sufficient to produce accurate estimates.

Table 0.2: RRMSE of ρ_{in} and ρ_{out} at different LMPs.

LMPs %	RRMSE (%)	
	ρ_{in}	ρ_{out}
10	14	19
20	18	23
30	22	22
40	25	21
50	25	19
60	24	16
70	24	14
80	25	12
90	24	10

Table 0.3: RRMSE of KF and KFNN approaches at different LMPs

LMPs %	RRMSE (%)	
	KF	**KFNN**
10	36	64
20	37	50
30	38	52
40	37	54
50	37	58
60	31	59
70	26	57
80	23	52
90	20	35

4.3. Third Research Approach

This section utilizes the three data-driven approaches to estimate the number of vehicles traversing the tested link. CV data alone are used to train and build the three approaches; camera data are not required.

First, different neighbors (k) are tested to calibrate and train the k-NN approach, as shown in Table 0.4 The optimal k was found to be 14, with an RRMSE of 18.47%.

After calibrating the data-driven estimation approaches, external data are used to test and evaluate the estimation approaches' performance. Table 0.5 presents the RRMSE and RMSE values using the three data-driven estimation approaches: ANN, k-NN, and RF. The results demonstrate that the ANN outperforms the k-NN and the RF for all LMP scenarios.

Next, the we compare the performance of the model-driven approaches (KF, AKF, and PF) and the data-driven approaches (ANN, k-NN, and RF) for the application of the traffic stream density. Table 0.6 summarizes the RRMSE and RMSE values using the six estimation approaches. The table demonstrates that the ANN approach produces the most accurate estimates compared to other approaches. However, it is worth mentioning the difficulties in applying this approach in the field due to the huge amount of data needed to train and build the ANN approach, especially for a large network (e.g., Los Angeles, CA). Moreover, sudden changes in traffic behaviors (e.g., incidents) would not always ensure accurate estimates and thus might lead to worsening traffic signal controller performance. Consequently, we recommend using the KF approach for the application of traffic density estimation due to its simplicity and applicability in the field.

Table 0.4: RRMSE of k-NN approach using different k values

k	RRMSE (%)
1	24.63
2	21.63
3	20.47
4	19.91
5	19.49
6	19.27
7	19.07
8	18.88
9	18.81
10	18.68
11	18.54
12	18.58
13	18.48
14	18.47
15	18.47

Table 0.5: RRMSE and RMSE values using data-driven approaches

LMPs %	RRMSE (%), RMSE (veh)		
	ANN	k-NN	RF
1	27, 2.1	40, 3.1	42, 3.2
3	29, 2.2	37, 2.8	42, 3.2
5	29, 2.2	37, 2.8	38, 2.8
8	28, 2.1	36, 2.7	38, 2.8
10	27, 2.1	35, 2.8	38, 2.8
15	25, 1.9	33, 2.4	36, 2.7
20	24, 1.8	33, 2.4	36, 2.7
30	22, 1.7	30, 2.3	34, 2.6
40	20, 1.5	27, 2.1	30, 2.3
50	17, 1.3	24, 1.8	26, 2.0
60	14, 1.1	22, 1.6	23, 1.7
70	11, 0.9	20, 1.5	21, 1.6
80	9, 0.7	18, 1.4	19, 1.5
90	8, 0.6	17, 1.3	17, 1.3

Table 0.6: RRMSE and RMSE values using model- and data-driven approaches

LMPs %	Model-Driven Approaches			Data-Driven Approaches		
	KF	AKF	PF	ANN	k-NN	RF
1	44, 3.2	58, 4.3	37, 2.8	27, 2.1	40, 3.1	42, 3.2
3	39, 3.0	48, 3.6	39, 3.0	29, 2.2	37, 2.8	42, 3.2
5	37, 2.8	44, 3.2	38, 2.9	29, 2.2	37, 2.8	38, 2.8
8	36, 2.8	40, 3.0	37, 2.9	28, 2.1	36, 2.7	38, 2.8
10	36, 2.8	38, 2.9	37, 2.9	27, 2.1	35, 2.8	38, 2.8
15	36, 2.8	40, 3.0	39, 3.0	25, 1.9	33, 2.4	36, 2.7
20	37, 2.8	39, 3.0	39, 3.0	24, 1.8	33, 2.4	36, 2.7
30	38, 2.9	38, 2.9	42, 3.2	22, 1.7	30, 2.3	34, 2.6
40	37, 2.9	38, 2.9	39, 3.0	20, 1.5	27, 2.1	30, 2.3
50	37, 2.8	38, 2.9	37, 2.8	17, 1.3	24, 1.8	26, 2.0
60	31, 2.4	38, 2.9	31, 2.4	14, 1.1	22, 1.6	23, 1.7
70	26, 2.0	32, 2.5	25, 1.9	11, 0.9	20, 1.5	21, 1.6
80	23, 1.8	31, 2.4	20, 1.5	9, 0.7	18, 1.4	19, 1.5
90	20, 1.5	30, 2.2	14, 1.1	8, 0.6	17, 1.3	17, 1.3

5. Summary and Conclusions

This work presents three research approaches to estimate the number of vehicles along signalized links. The first research approach presents three model-driven approaches (KF, AKF, and PF) using a single average ρ in the state-space equations. This first approach relies solely on CV data. The second research approach develops the ANN approach to estimate two ρ variables, ρ_{in} and ρ_{out}, to be used in the state-space equations. Fused CV and camera data are utilized to build the ANN approach. After that, the second approach integrates the ANN with the KF approach (KFNN approach) to estimate the number of vehicles on signalized links. The third research approach develops three data-driven approaches (ANN, k-NN, and RF) to directly estimate the number of vehicles. This third approach uses only CV data to build the data-driven approaches. The three research approaches are applied on a signalized link in downtown Blacksburg, VA. The main findings and conclusions of the chapter are summarized as follows:

- The use of CV data is sufficient to provide accurate vehicle count estimates.
- Using two predicted variable values in the state-space equations is not recommended, as it may produce undesired large errors in the state equation. It was found that the ANN approach may over-estimate the first variable and under-estimate the second variable or vice versa for the same estimation step. Consequently, the second research approach is not recommended.
- The ANN is the most accurate estimation approach. However, it is also necessary to take into consideration the huge amount of data needed to train and build the ANN approach, the long computational time needed to build the ANN, and the constraints on keeping the traffic behavior the same as the behavior in the training data set. Based on these factors, the use of the KF approach is highly recommended for the application of traffic density due to its simplicity and applicability in the field.

6. References

Aljamal, Mohammad A, Hossam M Abdelghaffar, and Hesham A Rakha. 2019a. 'Developing a neural–Kalman filtering approach for estimating traffic stream density using probe vehicle data', *Sensors*, 19: 4325.

———. 2019b. "Kalman filter-based vehicle count estimation approach using probe data: A multi-lane road case study." In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 4374-79. IEEE.

———. 2020a. 'Estimation of traffic stream density using connected vehicle data: Linear and nonlinear filtering approaches', *Sensors*, 20: 4066.

———. 2020b. 'Real-time estimation of vehicle counts on signalized intersection approaches using probe vehicle data', *IEEE Transactions on Intelligent Transportation Systems*, 22: 2719-29.

Antoniou, Constantinos, and Haris N Koutsopoulos. 2006. 'Estimation of traffic dynamics models with machine-learning methods', *Transportation research record*, 1965: 103-11.

Bekiaris-Liberis, Nikolaos, Claudio Roncoli, and Markos Papageorgiou. 2016. 'Highway traffic state estimation with mixed connected and conventional vehicles', *IEEE Transactions on Intelligent Transportation Systems*, 17: 3484-97.

Bhouri, Neila, Habib Haj Salem, Markos Papageorgiou, and Jean Marc Blosseville. 1989. "Estimation of traffic density on motorways." In *IFAC/IFIP/IFORS International Symposium (AIPAC'89)*, 579-83.

Breima, L. 2010. 'Random Forests. Machine Learning'.

Cover, Thomas, and Peter Hart. 1967. 'Nearest neighbor pattern classification', *IEEE transactions on information theory*, 13: 21-27.

Di, Xuan, Henry X Liu, and Gary A Davis. 2010. 'Hybrid extended Kalman filtering approach for traffic density estimation along signalized arterials: Use of global positioning system data', *Transportation research record*, 2188: 165-73.

Fulari, Shrikant, Lelitha Vanajakshi, and Shankar C Subramanian. 2017. 'Artificial neural network–based traffic state estimation using erroneous automated sensor data', *Journal of Transportation Engineering, Part A: Systems*, 143: 05017003.

Ghosh, Dipankar, and CH Knapp. 1978. 'Estimation of traffic variables using a linear model of traffic flow', *Transportation Research*, 12: 395-402.

Herrera, Juan C, and Alexandre M Bayen. 2007. 'Traffic flow reconstruction using mobile sensors and loop detector data'.

Jahangiri, Arash, Hesham A Rakha, and Thomas A Dingus. 2015. "Adopting machine learning methods to predict red-light running violations." In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, 650-55. IEEE.

Khan, Sakib Mahmud, Kakan C Dey, and Mashrur Chowdhury. 2017. 'Real-time traffic state estimation with connected vehicles', *IEEE Transactions on Intelligent Transportation Systems*, 18: 1687-99.

Kubat, Miroslav. 1999. 'Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7', *The Knowledge Engineering Review*, 13: 409-12.

Li, Tiancheng, Tariq Pervez Sattar, and Shudong Sun. 2012. 'Deterministic resampling: unbiased sampling to avoid sample impoverishment in particle filters', *Signal Processing*, 92: 1637-45.

Liu, Jun S, and Rong Chen. 1998. 'Sequential Monte Carlo methods for dynamic systems', *Journal of the American statistical association*, 93: 1032-44.

Mimbela, Luz-Elena Y, and Lawrence A Klein. 2007. 'Summary of vehicle detection and surveillance technologies used in intelligent transportation systems'.

Raj, Jithin, Hareesh Bahuleyan, and Lelitha Devi Vanajakshi. 2016. 'Application of data mining techniques for traffic density estimation and prediction', *Transportation Research Procedia*, 17: 321-30.

Sekuła, Przemysław, Nikola Marković, Zachary Vander Laan, and Kaveh Farokhi Sadabadi. 2018. 'Estimating historical hourly traffic volumes via machine learning and vehicle probe data: A Maryland case study', *Transportation Research Part C: Emerging Technologies*, 97: 147-58.

Vigos, Georgios, Markos Papageorgiou, and Yibing Wang. 2008. 'Real-time estimation of vehicle-count within signalized links', *Transportation Research Part C: Emerging Technologies*, 16: 18-35.

Wang, Fei-Yue. 2010. 'Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications', *IEEE Transactions on Intelligent Transportation Systems*, 11: 630-38.

Wassantachat, Thanes, Zhidong Li, Jing Chen, Yang Wang, and Evan Tan. 2009. "Traffic density estimation with on-line SVM classifier." In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, 13-18. IEEE.

Wright, Matthew, and Roberto Horowitz. 2016. 'Fusing loop and GPS probe measurements to estimate freeway density', *IEEE Transactions on Intelligent Transportation Systems*, 17: 3577-90.