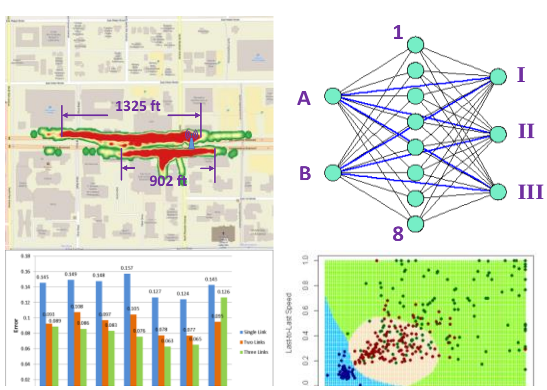
Title
Multi-Modal Arterial Performance Measurement Using Multi-Source ITS Data

Author
Wu, Yao-Jan

Publication Date
2016-10-01

Multi-Modal Arterial Performance Measurement Using Multi-Source ITS Data



Submission Date: October 2016

Prepared by

Yao-Jan Wu, Assistant Professor, PhD, P.E.

Yi-Chang Chiu, Associate Professor, PhD

Shu Yang, Graduate Research Assistant

The University of Arizona

UCONNECT

Final Report

Project Number TR201407

Multi-modal Arterial Performance Measurement Using Multi-source ITS Data

Prepared for

The University of California Center on Economic Competitiveness in Transportation

By

Yao-Jan Wu
Assistant Professor, PhD, P.E.
Department of Civil Engineering and Engineering Mechanics
University of Arizona

Yi-Chang Chiu
Associate Professor, PhD
Department of Civil Engineering and Engineering Mechanics
University of Arizona

Shu Yang
Graduate Research Assistant
Department of Civil Engineering and Engineering Mechanics
University of Arizona

October 2016

TECHNICAL REPORT STANDARD TITLE PAGE

1. REPORT NO.	2. GOVERNMENT ACCESSION NO.	3. RECIPIENT'S CATALOG NO.
4. TITLE AND SUBTITLE Multi-modal Arterial Performance Measurement Using Multi-source ITS Data		5. REPORT DATE **October** 2016 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Yao-Jan Wu, Yi-Chang Chiu, Shu Yang		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Civil Engineering and Engineering Mechanics University of Arizona, 1209 2nd Street Tucson, AZ, 85705		10. WORK UNIT NO. 11. CONTRACT OR GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS The University of California Center on Economic Competitiveness in Transportation (UCCONNECT)		13. TYPE OF REPORT AND PERIOD COVERED Final Report 14. SPONSORING AGENCY CODE
15. SUPPLEMENTARY NOTES		

16. ABSTRACT

With the rapid development of Intelligent Transportation Systems (ITS) technologies, surface-transportation data can now be collected by a wide variety of ITS traffic sensors, including Bluetooth sensors, automatic vehicle location (AVL) devices. Bluetooth technology has been widely used in transportation studies to collect traffic data (e.g. travel time and partial Origin-destination data). Bluetooth media access control (MAC) readers can be installed along roadways to collect Bluetooth-based data. This data is commonly used to measure traffic performance. One of the advantages of using Bluetooth technology to measure traffic performance is that travel time, one of the most important traffic performance measures, can be measured directly instead of being estimated. However, travel time outliers can commonly be observed due to different travel mode on arterials. Since travel mode information cannot be directly obtained from the raw Bluetooth-based data, a mathematical methodology is in need to identify travel mode. In this report, a genetic algorithm and neural network (GANN)-based model was developed. GPS-enabled devices were used to collect ground truth travel time. In order to additionally compare the model performance, K-NN was also implemented. Since the model performances depend on the model inputs, seven collections of model inputs were designed in order to achieve the best performance. An arterial corridor with four consecutive links and three intersections was selected to be the study corridor. The results suggested that correctly identifying the three travel modes successfully every time was not possible, although the GANN based model had low misidentification rates. In our study, 94% of autos were identified as bikes and 90% of bikes were identified as autos using three links.

17. KEY WORDS		18. DISTRIBUTION STATEMENT	
Multi-modal transportation, performance measurement, traffic data collection, and intelligent transportation systems		No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22616	
19. SECURITY CLASSIF. (of this report) None	20. SECURITY CLASSIF. (of this page) None	21. NO. OF PAGES	22. PRICE

Table of Contents

Contents

Acknowledgements

The authors would like to thank the University of California Center on Economic Competitiveness in Transportation (UCCONNECT), City of Tucson and the Regional Transportation Authority (RTA) for funding support. Also, the authors would like to thank PAG and ADOT for the Bluetooth MAC reader installation. The authors are grateful for the technical support from Mr. Mike Hicks in the City of Tucson and Mr. Simon Ramos of ADOT. Paul Casertano from PAG also provides useful advice on the development of the proposed system. Special thanks go to Mr. Payton Cooke for editing assistance.

Disclaimer

The content 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 Missouri Department of Transportation, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Executive Summary

Under the Moving Ahead for Progress in the 21st Century Act of 2012 (MAP-21), transportation agencies now face greater requirements with respect to the collection and analysis of surface-transportation data. Data-driven performance measurement is expected to play an important role in assisting transportation agencies with their transportation operations and planning decisions. The transportation world is experiencing a major shift from a “data desert” to a “data ocean”. With the emerging development of Intelligent Transportation Systems (ITS) technologies, surface-transportation data can now be collected by a wide variety of ITS traffic sensors, including Bluetooth sensors, automatic vehicle location (AVL) devices, inductive loop sensors, and radar-based detectors. In practice, ITS data are collected from multiple sources but individually analyzed or processed. It has been challenging to take full advantage of the ITS data from multiple sources by enabling them to exchange information with each other to compensate for their various disadvantages.

Bluetooth technology has been widely used in transportation studies to collect traffic data (e.g. travel time and partial Origin-destination data). Bluetooth media access control (MAC) readers can be installed along roadways to collect Bluetooth-based data. This data is commonly used to measure traffic performance. One of the advantages of using Bluetooth technology to measure traffic performance is that travel time, one of the most important traffic performance measures, can be measured directly instead of by estimation. Bluetooth-based travel time accurately represents ground truth on long freeway corridors in most circumstances. However, in urban environments, more effort is required to use Bluetooth-based data to measure arterial travel time. This is because traffic on arterials is controlled by traffic signals, and therefore, traffic conditions are more complex than on freeways. Additionally, heterogeneous traffic is observed, and multiple travel modes travel simultaneously, including transit, bicyclists, and pedestrians. Knowing the travel mode of Bluetooth-based data would help practitioners and researchers accurately estimate mode-specific travel time. The primary goal of this report is to identify the travel mode using the Bluetooth-based data.

In order to accomplish this goal, the first step is to fully understand the Bluetooth technology. Five aspects have been summarized:

1. Privacy concerns. A MAC address represents a unique identify of a Bluetooth-based device. A mechanism should be designed to protect the privacy.
2. Data types. Three major data types can be collected, including MAC, timestamp and location identifier.
3. Detection range. Few previous studies physically examine the detection range of Bluetooth-based MAC readers, especially for different travel modes (e.g. pedestrian and bike).
4. Multiple detections. The application of using multiple detections remains largely undefined.
5. Limitations on Bluetooth-based data applications. The Bluetooth-base data are primarily used to measure travel time and estimate O-D information. Traffic volume, lane-by-lane information and etc. are nearly impossible with Bluetooth technology.

The second step is to collect Bluetooth-based data, and then a custom Client-Server architecture based Bluetooth-based data collection system was designed:

1. Over 40 Bluetooth-based MAC readers have been installed at major intersections in Tucson, Arizona.
2. The MAC readers have been successfully collect the Bluetooth-based signals and transmit to a computer server at the University of Arizona (UA) .

3. A database was designed and implemented in the computer server to receive and store the transmitted Bluetooth-based data. The successful use of Bluetooth-based data highly depends on the efficiency of the database.

At the beginning of the project, multiple data sources were identified. Travel time estimation was based on the Bluetooth data collected from those installed Bluetooth-based MAC readers. GPS-enabled devices with high updating frequency were used to help collect ground truth trajectories in order to measure the accuracy of estimated Bluetooth-based travel time. Three travel modes were also identified, including pedestrians, bicyclists and autos. Since few studies physically examine the Bluetooth-based MAC reader detection ranges, especially with the consideration of travel modes, electronic devices with both GPS and Bluetooth modules enabled were used to examine the detection ranges. Two findings were noted: 1) most of these detection ranges were less than 300 m (985 ft.) in our study; 2) the detection ranges varied depending on the intersection and travel mode.

A genetic algorithm and neural network (GANN)-based travel mode identification methodology was developed using the collected Bluetooth-based data. A traditional K-NN algorithm was also implemented to compare the results with those produced from the GANN-based models. Seven collections of model inputs were tested using the Bluetooth-based travel time measures in the previous studies. Four consecutive links with three intersections on the Speedway Bulverde was selected to be the study corridor. The results of the GANN-based model suggested that:

1. Using both First-to-First (FF) and Last-to-Last (LL) speed as inputs performed better than using FF or LL speed alone.
2. The detection ranges of the travel modes had little impact on travel mode identification.
3. The travel mode misidentification rate can be decreased by considering higher numbers of arterial links.

4. Multiple detection data may not improve the rate of successful travel mode identification.
5. The GANN based model outperformed the KNN. Using the KNN, even pedestrians were sometimes misidentified as other modes.

Section 1 BACKGROUND

Under the Moving Ahead for Progress in the 21st Century Act of 2012 (MAP-21), transportation agencies now face greater requirements with respect to the collection and analysis of surface-transportation data such as freeway data, arterial data, and transit data. Data-driven performance measurement is expected to play an important role in assisting transportation agencies with their transportation operations and planning decisions. Once the data is further processed and disseminated, travelers should also benefit from more accurate and timely information about current traffic conditions to help them determine optimal travel routes.

It is generally accepted that the transportation world is experiencing a major shift from a “data desert” to a “data ocean”. Due to lack of mobile and fixed traffic sensors on the roads, it has been difficult to collect multiple sources of data; however, with the emerging development of Intelligent Transportation Systems (ITS) technologies, surface-transportation data can now be collected by a wide variety of ITS traffic sensors, including Bluetooth sensors, automatic vehicle location (AVL) devices, inductive loop sensors, and radar-based detectors and the information gathered applied for various purposes. Meanwhile, issues have begun to appear due to the transformation in the data environment. Traffic data collection and utilization are generally conducted by a body such as a city transportation department for a single purpose (e.g. travel time). In practice, ITS data are collected from multiple sources but individually analyzed or processed. It has been challenging to take full advantage of the ITS data from multiple sources by enabling them to exchange information with each other to compensate for their various disadvantages.

In order to measure arterial performance measures, the data used in the report used multi-source ITS data, primarily including Bluetooth-based data and Global Positioning Systems-based (GPS-based) data. Bluetooth technology has been widely used in transportation studies to collect traffic data (e.g. travel time and origin-destination data). Bluetooth media access control (MAC) readers can be installed along roadways to collect Bluetooth-based data. This data is commonly used to measure traffic performance (e.g. Araghi et al., 2013; Barceló et al., 2013; Khoei et al., 2013; Qiao et al., 2013; Aliari and Haghani, 2012; Barceló et al., 2010; Quayle et al., 2010; Haghani et al., 2009; and Wasson et al., 2008). One of the advantages of using Bluetooth technology to measure traffic performance is that travel time, one of the most important traffic performance measures, can be measured directly instead of being estimated. In recent years, the number of personal Bluetooth devices (e.g. laptops, smart phones and smart watches) has grown significantly, enlarging the size of Bluetooth-based data samples (increased penetration rate). Therefore, travel time can be more accurately measured since the penetration rate has increased.

Bluetooth-based travel time accurately represents ground truth on long freeway corridors in most circumstances (Haghani et al., 2009). However, in urban environments, more effort is required to use Bluetooth-based data to measure arterial travel time. This is because traffic on arterials is controlled by traffic signals, and therefore, traffic conditions are more complex than on freeways. Additionally, heterogeneous traffic is observed, and multiple travel modes travel simultaneously, including transit, bicyclists, and pedestrians. Several travel time outlier detection algorithms have been developed to clean Bluetooth-based data before use (e.g. Moghaddam et al., 2013(a); and Van Boxel et al., 2011). In the work as stated by Moghaddam et al. (2013(a)), the authors tested several outlier detection algorithms based on autos and buses. Based on this approach, if travel modes can be identified, then outliers can be eliminated and Bluetooth-based travel time can be more precisely estimated. Therefore, knowing the travel mode of Bluetooth-based data would help practitioners and researchers develop mode-specific travel time outlier detection algorithms and accurately estimate mode-specific travel time.

The rest of the report is organized as follows: first, relevant Bluetooth technology is overviewed. Next, a study corridor in Tucson, AZ, and its corresponding dataset is presented. A genetic algorithm neural network (GANN) based model is introduced to identify travel modes using the Bluetooth-based data. The model performance is presented before drawing final conclusions.

Section 2 STUDY SITE AND DATA

2.1 Multi-source data in Tucson, Arizona

Three real-time data sources can be primarily collected in Tucson, Arizona, including Bluetooth-based data, General Transit Feed Specification (GTFS) transit data, and video-based data. Figure 2-1 shows the overview of the three data sources. Sun Tran, which manages the regional public transportation system in Tucson, AZ, has installed automatic vehicle location (AVL) devices in approximately 180 of its transit buses. Real-time data are saved in the GTFS format, a generalized data format for transit information exchange. Both static and real-time GTFS data can be collected through a free online public portal hosted by Sun Tran. In addition to the GTFS data, traffic detection (Autoscope®) and Pan-Tilt-Zoom (PTZ) surveillance cameras have already been installed at major intersections. The video stream facilitates travel time ground truth data collection by manually matching vehicles, and counting vehicles and pedestrians.

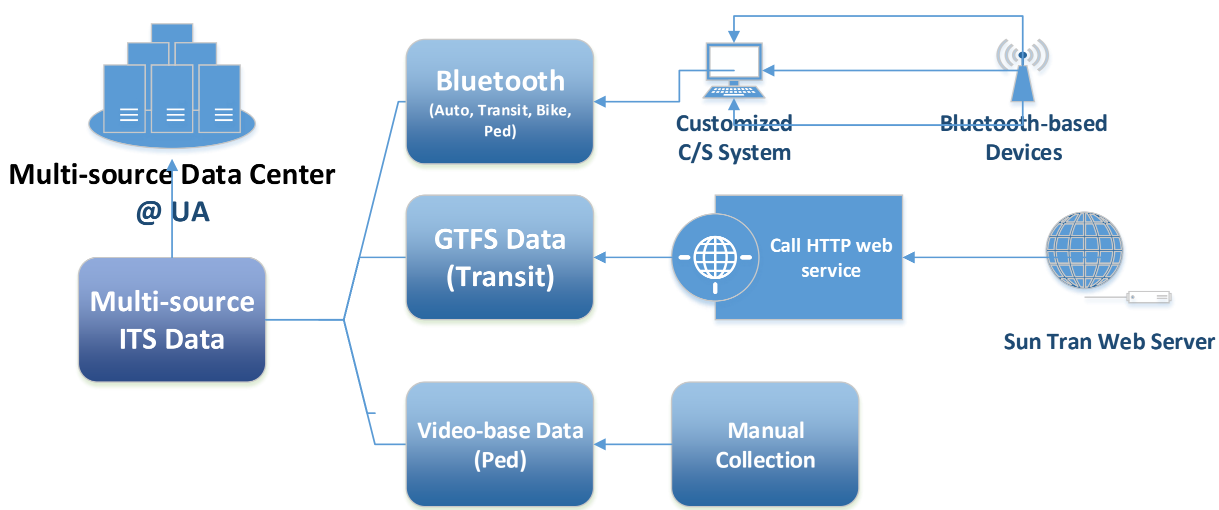


Figure 2-1. Three Major Data Sources in Tucson, AZ, U.S.

The department of transportation in the City of Tucson has been using the traffic detection and PTZ surveillance cameras to manage traffic. However, only one traffic detection camera was configured properly for experimental testing at the intersection of Cherry Avenue and Speedway Blvd during the project. Due to limited installation, our project team was unable to collect the data from traffic detection cameras. Therefore, the Bluetooth-based data served the major data source for measuring the multi-modal arterial performances.

2.2 Bluetooth Technology Overview

2.2.1 Privacy Concerns

Bluetooth technology has been widely used for short-range wireless communication. For example, data can be shared between two Bluetooth-enabled devices, and certain devices (e.g. smart phones) can be remotely controlled by other devices (e.g. smart watches) via Bluetooth connections. These wireless operations require a share or control agreement between the two devices. These agreements are typically unnecessary in transportation studies, because: 1) only the Bluetooth signals broadcasted from devices that have Bluetooth turned on are detected. The detected signals are usually encrypted before analysis. 2) No communication between the detecting and detected devices can be established since no agreement is initialized. 3) The MAC addresses of the detected devices are used anonymously and are not connected to specific individuals. Therefore, the privacy concerns regarding detected Bluetooth-enabled devices are not an issue.

2.2.2 Bluetooth-Based Data

Three primary types of Bluetooth-based data are used in transportation studies:

- Media access control (MAC) addresses: every electronic device with a Bluetooth module built in has a global unique identifier. Travel times are usually estimated by matching identical MAC addresses detected at upstream and downstream MAC readers.
- Timestamp: the time of detection for the Bluetooth-enabled device.
- Location identifier: the location where the Bluetooth-enabled device is detected.

The Bluetooth received signal strength indication (RSSI) can also be used. The RSSI could be recorded depending on the functionality and configuration of the Bluetooth MAC readers. A few studies have used the RSSI in traffic studies to estimate or predict travel time (see, for example, Araghi et al., 2013; and Saeedi et al., 2013). These authors concluded that RSSI-based travel time may be a better representation of ground truth travel time.

2.2.3 Detection Range

Bluetooth MAC readers detect Bluetooth-enabled devices within a certain range. Several factors influence detection range, including the types and power gains of Bluetooth antennas and the antenna installation position. Additionally, previous studies have also shown that the Bluetooth signal strength inside cars may be half of the normal range due to the vehicle's metal body (Quayle et al., 2010), resulting in a smaller detection range. Therefore, detection range depends on both hardware and travel mode. This characteristic could be helpful in travel mode determination.

2.2.4 Multiple Detections

Bluetooth-enabled devices may be detected multiple times within a particular detection range. However, the specific locations of the devices remain unknown and the location identifier is the only known spatial information. For example, consider a Bluetooth MAC reader installed at an intersection. The information collected by the Bluetooth MAC reader includes the MAC addresses of the detected devices, multiple timestamps for each device due to repeat detections, and the location identifier of the reader. The timespan of repeat detections is also called duration. A few studies have tried to explore the value of duration at intersections and establish the relationship between the duration and traffic congestion (Tsubota et al., 2011). However, the relationship remains largely undefined.

2.2.5 Limitations on Bluetooth-Based Data Applications

Most existing Bluetooth-based data applications are focused on two areas: 1) travel time estimation and prediction (e.g. Araghi et al., 2013; Khoei et al., 2013; Qiao et al., 2013; Aliari and Haghani, 2012; Quayle et al., 2010; Haghani et al., 2009; and Wasson et al., 2008). 2) Origin-destination matrix estimation (e.g. Barceló et al., 2013; and Barceló et al., 2010). Several researchers have conducted work zone analysis (e.g. Haseman et al., 2010) or route choice analysis (e.g. Hainen et al., 2011) based on either estimated or predicted Bluetooth travel time. Few studies have used Bluetooth-based data to study bike travel time (Mei et al., 2012). Unlike other traffic data sources (e.g. loop sensors or GPS), which have been applied to various subjects, Bluetooth-based data applications have been limited. Therefore, Bluetooth-based data has been considered as complementary transport data (Bhaskar and Chung, 2013). Because of the limited applications of Bluetooth data, Bluetooth-based data types are few in number, and errors, mainly caused by detection range, multiple detections, and various travel modes, are common.

Three major types of traffic data are difficult to collect with Bluetooth technology:

Traffic volume information and turning movements: previous studies have shown that only 2.0% to 3.4% of the total traffic volume is detected by the average Bluetooth system (Aliari and Haghani, 2012); therefore accurate traffic volume cannot be estimated.

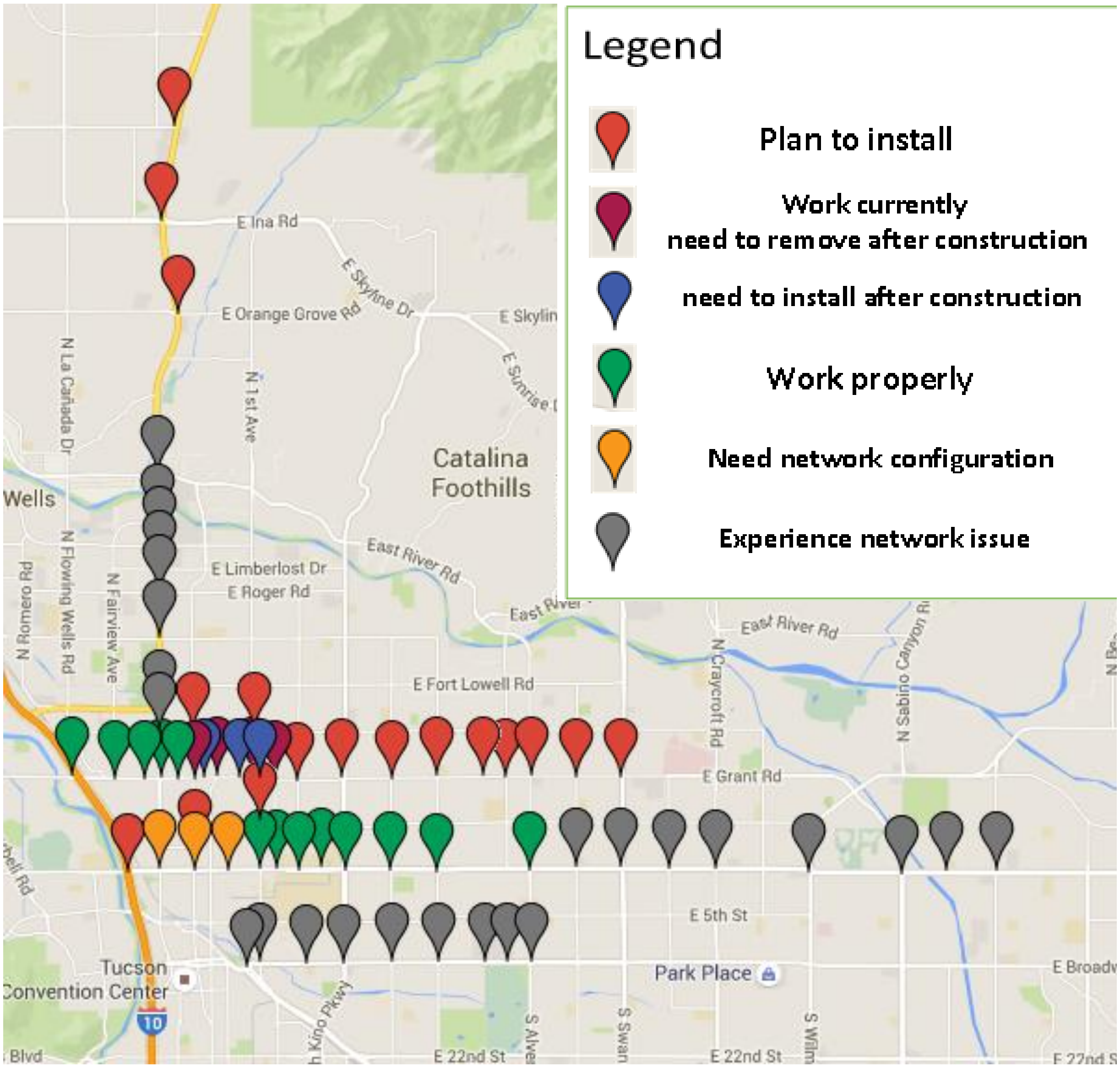
Lane-by-lane information is nearly impossible with Bluetooth technology. Detailed lane-by-lane information is valuable for studying driving behavior, such as lane changing maneuvers. Lane-by-lane information is critical for locating vehicles in a connected vehicle environment. Additionally, Haghani et al. (2010) concluded that Bluetooth technology was not suitable for facilities with managed lanes.

Travel mode information. On freeways, autos are the primary travel mode, including passenger cars, motorcycles, trucks, etc. However, in an urban context, multiple travel modes, including autos, bikes, and pedestrians, are mixed and share the roadway. Since Bluetooth-enabled devices are independent of travel mode and device privacy is protected, travel mode information is unavailable. Our study seeks to enable mode identification using modeling methods.

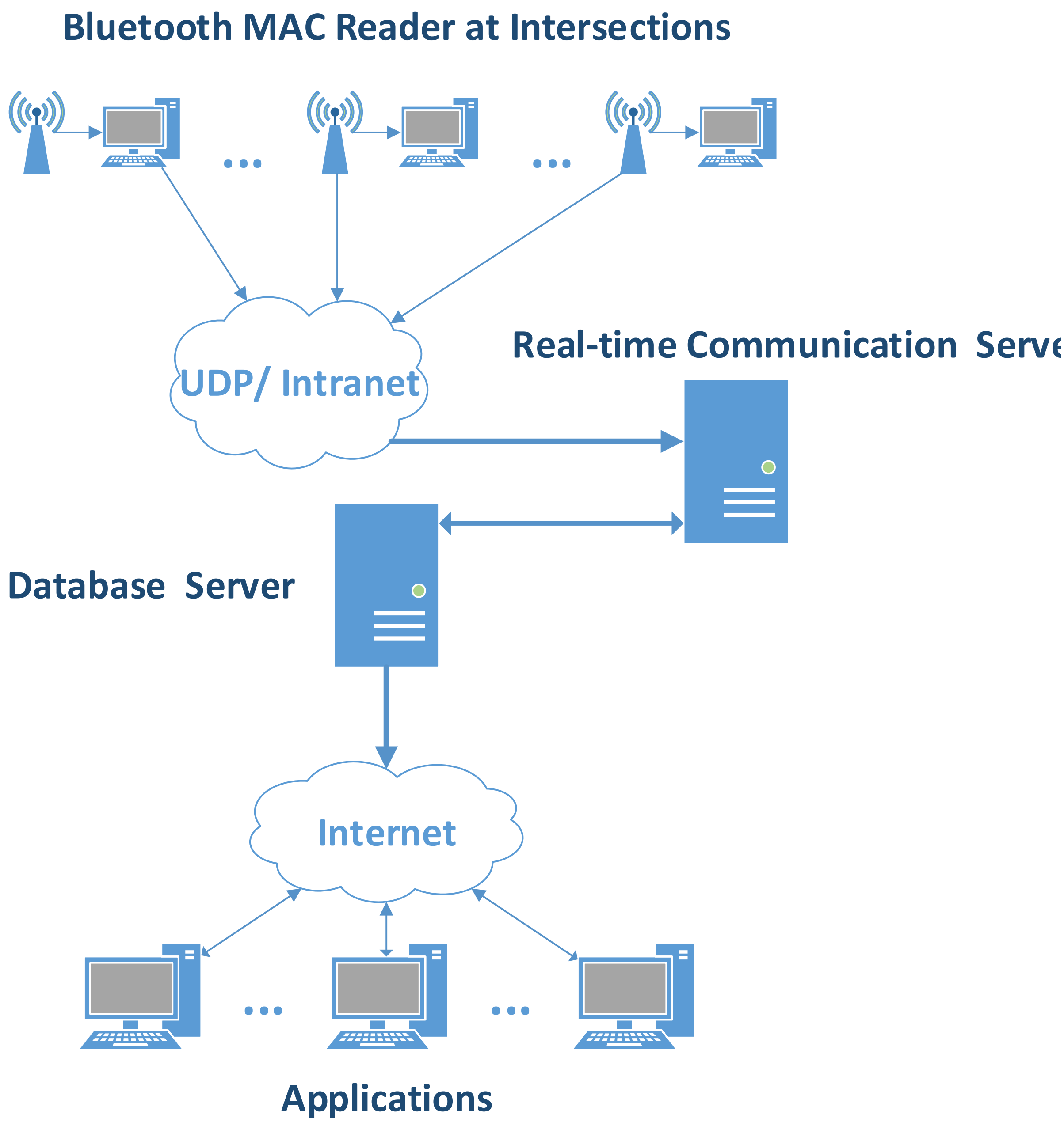
Errors created by detection range and multiple detections are difficult to correct. Moghaddam and Hellinga (2013) categorized the characteristics of Bluetooth measurement errors into three types: sampling error, sampling bias, and measurement error. These errors were mainly caused by detection range, multiple detections, and different travel modes. For example, without knowing the travel mode of detected devices, estimated travel time could be biased. Slower modes, such as pedestrians, could result in a longer estimated travel time. A recent study quantified the relationship between estimated speed errors and arterial corridor length (Haghani et al., 2010). The authors concluded that estimated speed errors increased with decreasing arterial corridor length. However, few studies have developed models to mathematically correct these errors.

2.3 Bluetooth Data Collection

With the help of the City of Tucson, the Pima Association of Governments (PAG), and the Arizona Department of Transportation (ADOT), a Bluetooth-based data collection system has been developed and maintained in the Tucson, Arizona, area since 2013. Figure 2-2(a) shows the locations of Bluetooth MAC readers in the Tucson area. These MAC readers were installed inside traffic control cabinets located at major intersections. Instead of using commercial MAC readers, custom MAC readers with 9 dbi antennas, Bluetooth external adapters, and mini PCs were created. The Bluetooth channel scan time interval was programed to be 3.84 seconds instead of the default value of 10.24 seconds in order to more precisely record entering and exiting times for each detector (Saeedi et al., 2013, p.92). After completing a Bluetooth scan time interval, the MAC readers sent the detected MAC addresses back to a central computer server located at the University of Arizona (UA) using user datagram protocol (UDP). Figure 2-2(b) shows the data collection system architecture.



(a) Bluetooth MAC reader locations



(b) Data collection system architecture

Figure 2-2. Bluetooth-based Data Collection System in Tucson, AZ, U.S.

2.2 Study Corridor and Ground Truth Data Collection

Speedway Boulevard is one of the busiest roadways in Tucson. Since UA is located next to Speedway and Tucson is a bicycle-friendly city, sidewalks and exclusive bike lanes had been built along the corridor. Because of the high volume of multiple transportation modes along the route, Speedway between Park Avenue and Campbell Avenue was chosen as the study corridor.

This corridor included four intersections, three westbound links, and three eastbound links. Each intersection was configured with a custom MAC reader. To identify travel modes using Bluetooth-based data, ground truth data was collected. Several components of the data collection plan are shown below.

1) Three travel mode categories were classified, including autos (passenger cars, motorcycles, and trucks), bicyclists, and pedestrians. A preliminary study showed that Bluetooth signals from devices in transit buses were weak and could not be reliably detected by the MAC readers, possibly due to the metal body of the buses (Quayle et al., 2010). Therefore, transit was not included in our study.

2) Table 2-1 shows the ground truth data collection plan. A trip was defined as either eastbound Speedway from Park to Campbell or westbound Speedway from Campbell to Park. The data was split into two parts: data used for travel mode identification model calibration (training data) and data used for verification (testing data).

3) The Bluetooth devices used for ground truth collection included two Blackberry cellphones, a Samsung cell phone, an iPhone, and an iPad. The GPS module for each device was also enabled to track the device's location every second. The GPS data was used only to estimate the MAC reader detection ranges.

4) As noted in Table 2-1, to account for real vehicle behavior, some trips ended in right, left, or U-turns at Campbell and Park, rather than continuing straight on Speedway beyond the study corridor.

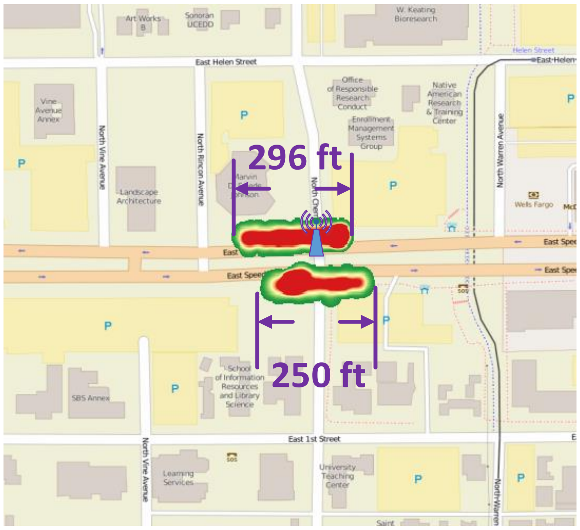
Table 2-1. Ground Truth Data Collection

Mode / Date	Autos		Bike		Pedestrian	
	Collection	Number of trips	Collection	Number of trips	Collection	Number of trips
Data used for model calibration (training data) (GPS enabled during data collection)						
2015-08-31					✓	4
2015-09-01			✓	12	✓	8
2015-09-04	✓	18			✓	8
2015-09-11	✓	16			✓	8
2015-09-16			✓	12	✓	8
2015-09-17	✓	20*	✓	12		
2015-09-18			✓	12		
Data used for model verification (testing data)						
2015-09-28			✓	14		
2015-09-29	✓	20				
2015-10-01	✓	20				
2015-10-05	✓	20	✓	14		
2015-10-06			✓	14	✓	8
2015-10-10					✓	8
2015-10-12					✓	8

* Five trips ended in non-through movements at the intersection with Campbell.

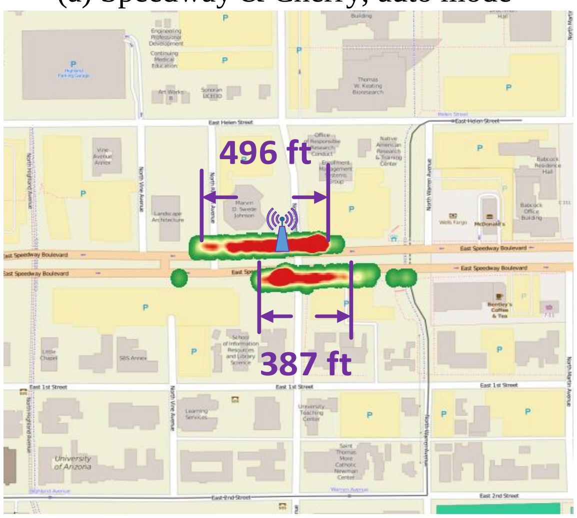
2.3 Detection Range Examination

Although previous studies have determined theoretical MAC reader detection ranges (e.g. 100 – 300 m (Araghi et al., 2013) or 300 ft. (Haghani et al., 2010)), few studies have physically examined actual MAC reader detection ranges. Our study used both GPS and Bluetooth-based data to examine the MAC reader detection ranges by matching the timestamps collected from both data. Figure 2-3 shows the detection ranges for each of the three travel modes at two intersections. The red areas in Figure 2-3 represent the regions where most of the tested devices were detected. Two findings were noted: 1) most of these detection ranges were less than 300 m (985 ft.) in our study; 2) the detection ranges varied depending on the intersection and travel mode.

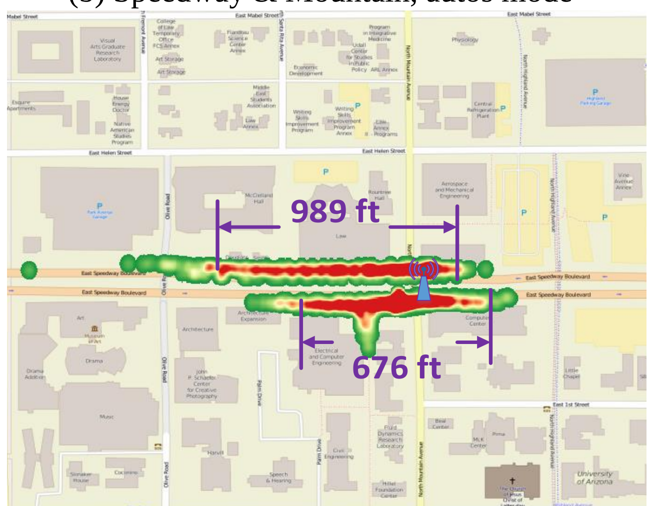


(a) Speedway & Cherry, auto mode

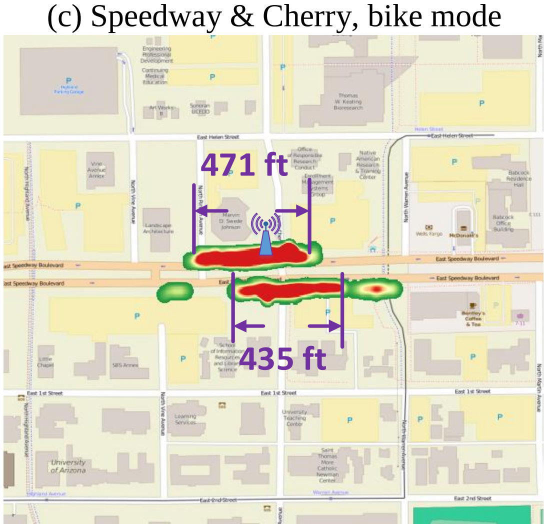
(b) Speedway & Mountain, autos mode



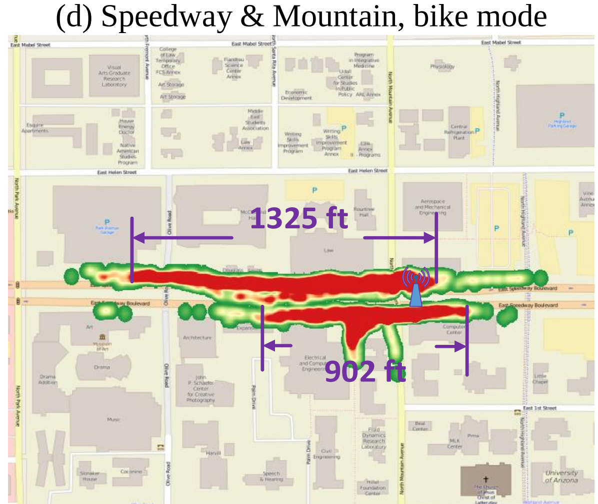
(c) Speedway & Cherry, bike mode



(d) Speedway & Mountain, bike mode



(e) Speedway & Cherry, pedestrian mode



(f) Speedway & Mountain, pedestrian mode

Figure 2-3. Detection Range by Three Travel Modes

(Background images from OpenStreetMap)

Section 3 GENETIC ALGORITHM AND NEURAL NETWORK-BASED MODE IDENTIFICATION

3.1 Justification for Using a Genetic Algorithm to Train a Neural Network

Neural networks and the K-nearest neighborhood (KNN) are two primary approaches to training nonparametric models (Qiao et al., 2013, p.166). Neural networks are widely used for data classification and prediction because of their high accuracy. Three common types of neural networks include feed-forward, recurrent, and high-order. One of the most popular neural network structures is the single hidden layer feed-forward neural network (SHLFFNN). Many commercial and open source software implementations of the SHLFFNN can be found, such as the neural network toolbox in MATLAB and the “RANN” package in R language. SCHLFFNNs are composed of three layers in the following order: input, hidden, and output layers. Each layer contains one or more neurons. The neuron connections strictly follow these rules: 1) connections are only made between two consecutive layers, such as the input layer to the hidden layer; or the hidden layer to the output layer. 2) Neurons in a layer must fully connect to every single neuron in the consecutive layer. 3) During the neural network training procedure, only the connection weights can be updated.

With regards to the SHLFFNN, recent research has found that: 1) several factors may affect the accuracy and efficiency of training SHLFFNNs, including learning rate, number of iterations, and initial connection weights (Michalewicz, 1996; and Koehn, 1994). 2) The back-propagation algorithm (BP) (Rojas, 1996) is commonly used to train the SHLFFNNs. However, the connection weight “often gets trapped in a local minimum of the error function and is incapable of finding a global minimum” (Yao, 1996, p. 1425). 3) Optimal combinations of the three abovementioned training factors are typically found by trial-and-error, making this a time consuming experiment. 4) Not only the connection weights but also the topology of neural networks can be updated during the training procedure. Updating neural network topology can improve accuracy and find near-optimal solutions (Yao, 1996).

The genetic algorithm is a near-optimal search algorithm commonly used for solving problems that are difficult to solve using mathematical equations. Therefore, the Genetic Algorithm Neural Network (GANN) was used in our study to obtain more accurate results through changing connection weights, topology, and to avoid the time consuming trial-and-error approach, since 21 neural networks were trained. Details regarding these networks are provided in the following section.

3.2 Genetic Algorithm and Neural Network (GANN)

3.2.1 Topology in the Genetic Algorithm and Neural Network

Essentially, the neural network used in our study was a SHLFFNN. The implementation details can be found on the authors' website (Yang, 2015). However, the topology of our neural network did not follow the strict rules listed in Section 4.1 and was organized in a more flexible manner: 1) the three layers were connected to each other and 2) connections between two neurons (connectivity) were not required to be established. Figure 3-1 shows an example of GANN topology. The black links represent the connections in a typical SHLFFNN; and the blue links represent the connections from the input to the output layers. The neurons in the three layers are labeled as A, B; 1 – 8; and I –III, respectively.

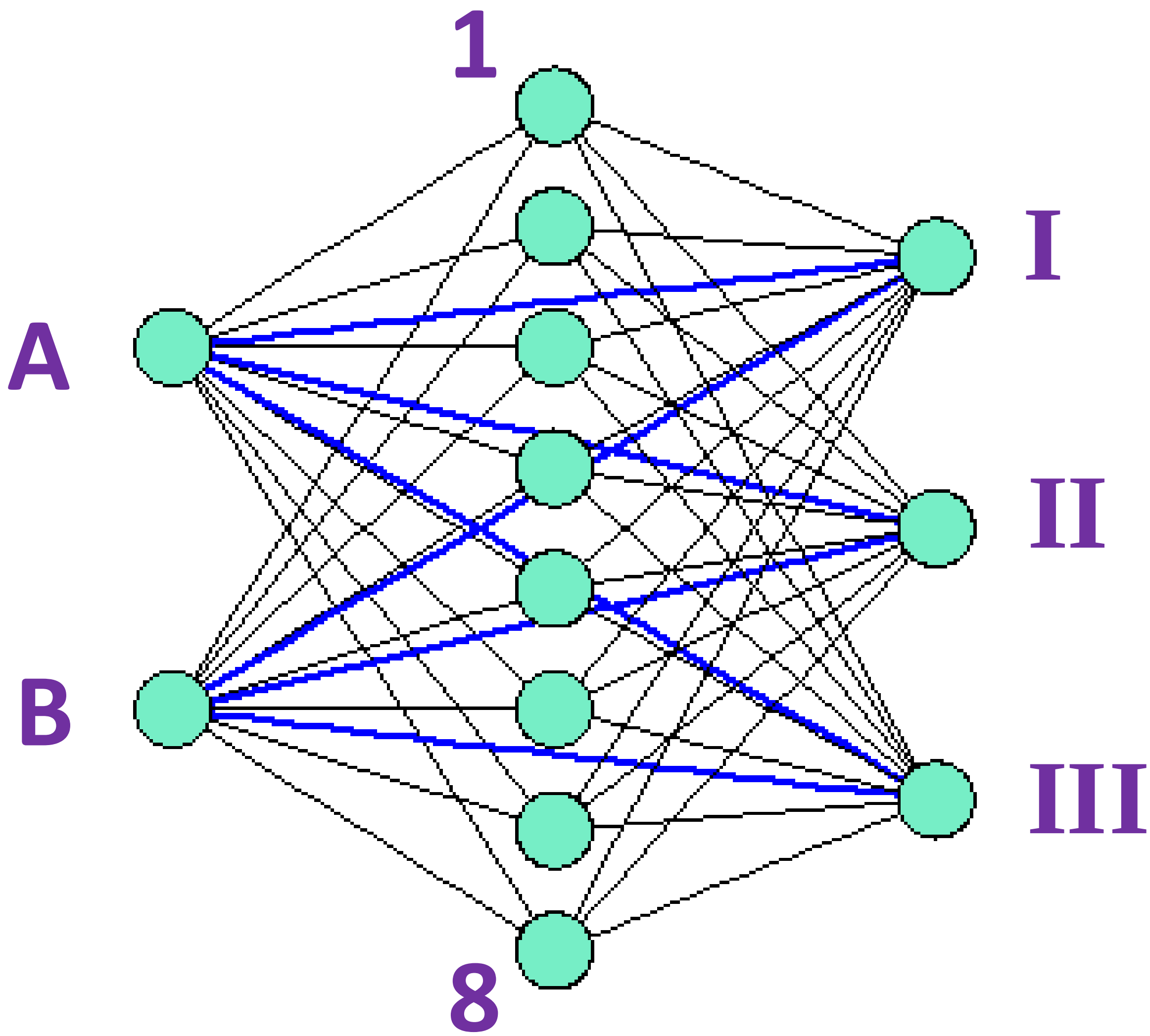


Figure 3-1. Feed-forward GANN Model with Single Hidden Layer and Full Connectivity

3.2.2 Connection Weights and Connectivity Encoding Scheme in the Genetic Algorithm

Solutions to the connection weights and connectivity are usually coded as string-based schema in genetic algorithms. The string-based schema is designed to easily operate the crossover and mutate solutions. In order to update the topology, a sparse-matrix-based code scheme is proposed. Figure 3-2 shows the basic encoding scheme and the crossover operation. Each cell represents either a real-value connection weight or binary value connectivity. For example, the cell A-1 represents the connection weight (or connectivity) from the A neuron in the input layer to the 1 neuron in the hidden layer. According to the proposed solution encoding scheme, the uniform crossover technique was used to complete the crossover operation.

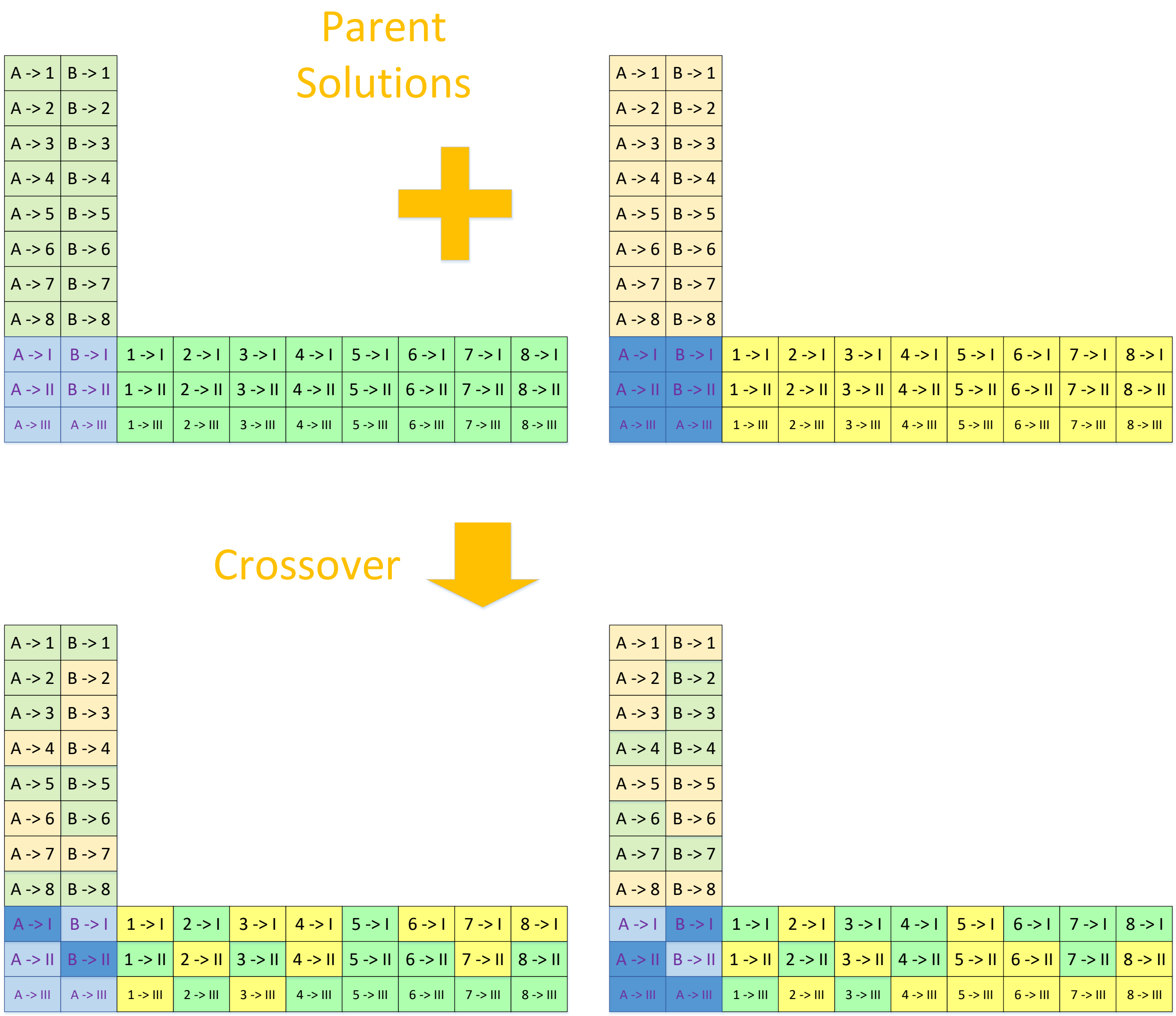


Figure 3-2. Solution Encode Scheme and Crossover Operation Using the Uniform Crossover

3.2.3 Error Calculation in GANN

The traditional SHLFFNN uses the BP algorithm to minimize output error. Due to the requirements of the BP algorithm, the output error is calculated based on a single input and is reduced back to the connection weights. Equation 1 shows the error calculation if the activation function is based on a binary function, while Equation 2 shows the error calculation if the activation function is based on the Sigmoid function. Equation 3, unlike Equations 1 and 2 which only consider the error of a single input, calculates the error in GANN-based models by summing up the errors of all inputs (Yao, 1996, p. 1425). Equation 3 also serves as the fitness function in the genetic algorithm. The objective of the genetic algorithm is to minimize the error calculated from the inputs.

$$Error_{O_i} = Calculated_{O_i} - Target_{O_i} \tag{1}$$

$$Error_{O_i} = \left(Calculated_{O_i}\right)\left(1 - Calculated_{O_i}\right)\left(Calculated_{O_i} - Target_{O_i}\right) \tag{2}$$

$$Error = \frac{\sum_{j=1}^{N}\sum_{i=1}^{M}\left(Calculated_{O_i} - Target_{O_i}\right)^2}{N} \tag{3}$$

Where: $Error_{O_i}$ represents the error of the i_{th} neuron in the output layer;
$Calculated_{O_i}$ is the calculated value of the i_{th} neuron in the output layer;
$Target_{O_i}$ is the ground truth value of the i_{th} neuron in the output layer;
N is the number of training data samples; and
M is the number of classifications. In our study $M = 3$.

Figure 3-3 shows the genetic algorithm flow chart used to train the sparse-matrix encoding schema neural network. Parallel computing was used to accelerate the model training procedure. With the data from our study corridor, each GANN-based model was well trained within five minutes using standard computers.

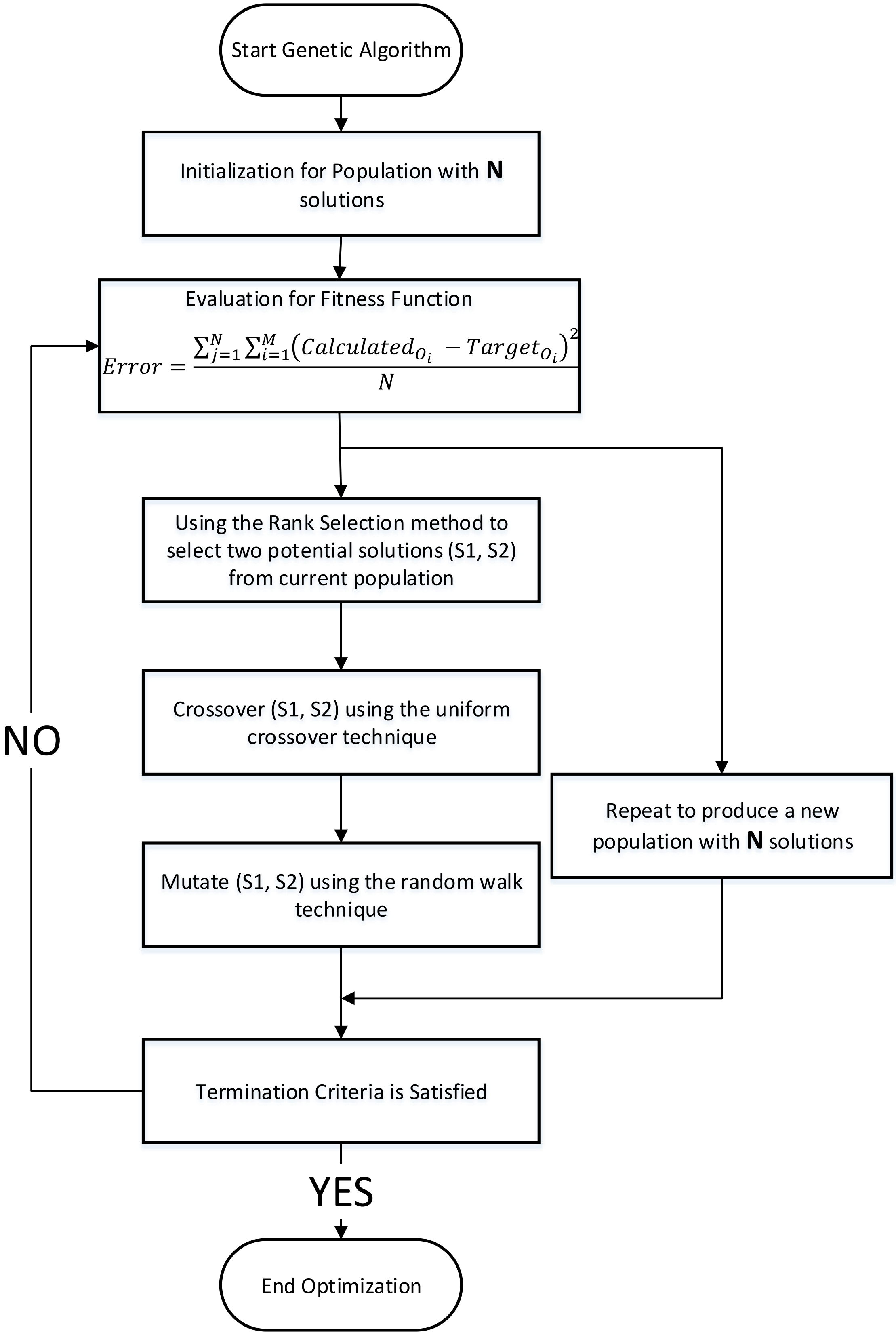


Figure 3-3. Genetic Algorithm Flow Chart

3.3 Input Selection

Traffic performance measures, such as travel time and speed, can help identify travel modes. One of the measures obtained from Bluetooth-based data is travel time, which can then be used as the primary performance measure to identify travel modes. For example, autos usually travel faster than both bikes and pedestrians on arterials. However, travel time is dependent on the specific link, and therefore, a single model is required to identify travel mode on that link. Speed (normalized travel time divided by link length) is an alternative measure that can be used to develop link-independent mode identification models.

First-to-first (FF) and last-to-last (LL) travel time were favored in previous studies (for example, Araghi et al., 2013 and Saeedi et al., 2013). Since the RSSI is unavailable, peak-to-peak travel time (Araghi et al., 2013) was not calculated in our study. Based on the detection range determined in Section 3.3, the FF and LL distances for the three travel modes are listed in Table 3-1. Then, the FF and LL speeds were calculated. In order to examine whether the detection ranges could significantly affect travel mode identification, a baseline speed was also calculated using the intersection-to-intersection distances.

Table 3-1. Measured Distance by Mode

From	To*	# of Links	Autos Mode (ft)		Bike Mode (ft)		Pedestrian Mode (ft)		Intersection-Intersection (ft)
			FF	LL	FF	LL	FF	LL	
1	2	One	1526.0	1639.1	1346.3	1456.5	1289.1	1575.1	1602.7
1	3	Two	3234.3	3110.2	3307.3	3037.3	3256.2	3015.3	3181.8
1	4	Three	4665.6	4804.3	4767.7	4753.0	4694.5	4536.0	4665.6
2	1	One	1474.9	1398.3	1467.6	941.9	1598.9	798.0	1602.7
2	3	One	1781.5	1522.2	1985.9	1584.3	2026.1	1596.0	1579.1
2	4	Two	3205.6	3169.1	3431.9	3285.8	3548.4	3168.8	3062.9
3	1	Two	3095.6	3132.1	2956.9	3095.6	2869.3	3183.2	3181.8
3	2	One	1606.2	1850.8	1482.2	2161.1	1361.6	2445.9	1579.1
3	4	One	1457.0	1599.3	1467.8	1752.5	1372.7	1566.3	1483.8
4	1	Three	4607.0	4650.8	4468.2	4409.8	4431.8	4329.5	1483.8
4	2	Two	3110.5	3380.7	3051.9	3548.3	3052.1	3687.0	3062.9
4	3	One	1533.3	1486.0	1551.1	1387.4	1586.0	1241.2	1483.8

* 1: Speedway Blvd. & Park Ave.;
2: Speedway Blvd. & Mountain Ave.;
3: Speedway Blvd. & Cherry Ave.;
4: Speedway Blvd. & Campbell Ave.

The GANN-based model was sensitive to its inputs and could have produced greatly different results depending on the inputs. Seven scenarios designed to identify the best collection of inputs are presented in Table 3-2. Scenarios 1a and 1b used FF speeds. Scenario 1a was based on measured distances between detection ranges, while Scenario 1b was based on the intersection-to-intersection length. Scenario 1b could be used in situations where measured distances were unknown or not available. Similarly, Scenarios 2a and 2b used LL speeds, and Scenarios 3a and 3b used both FF and LL speeds. Scenario 4 was designed based on the assumption that the travel mode identification accuracy is improved by adding duration data, which is determined as a result of multiple detections. The duration data at upstream and downstream locations was included in Scenario 4, while the other input parameters depended on the best-performing scenario among Scenarios 1-6.

To investigate the effect of the number of links on travel mode identification, data from a single link, two links, and three links was collected and used in each scenario. Therefore, a total 21 GANN-based models were developed. Note that real-value inputs in the GANN-based models were scaled between zero and one for computing convenience.

Table 3-2. Input Selection Candidates Scenarios

Scenario Name \ Input	Input Parameters	Number of Inputs
Scenario 1a	FF speeds using measured distances for the three travel modes	3
Scenario 1b	FF speed using intersection-to-intersection length	1
Scenario 2a	LL speeds by the three travel modes	3

Scenario 2b	LL speed using intersection-to-intersection length	1
Scenario 3a	Using both FF and LL speeds by the three travel modes	6
Scenario 3b	Using both FF and LL speeds based on intersection-to-intersection length	2
Scenario 4	Adding duration data to the best-performing scenario among Scenarios 1-6	Determined based on Scenario 1-6 results

Section 4 MODEL PERFORMANCE AND COMPARISONS

4.1 Best Input and GANN-based Model Selection

Low error calculated using Equation 3 indicated good performance of a particular GANN-based model. Figure 4-1 shows the errors of the 21 GANN-based models. Several findings are summarized below:

- The training errors decreased with an increasing number of links.
- Given a fixed corridor (regardless of the number of links), using both FF and LL speeds was better than using either speed type alone.
- Detection ranges by travel mode had limited impact on travel mode identification. For example, the errors for the three modes in Scenario 3a were 0.127, 0.078, and 0.063, while the errors in Scenario 3b were 0.124, 0.077, and 0.065. These differences were minor.
- The models in Scenario 3a and Scenario 3b outperformed the models in Scenarios 1a-2b. Considering the minor performance differences between Scenario 3a and Scenario 3b and overall model complexity, Scenario 3b was identified as the best input because the number of input parameters (FF and LL speeds) was less than that in Scenario 3a (FF and LL speeds by the three travel modes).
- Adding the duration information in Scenario 4 did not improve the accuracy of travel mode identification. The errors in Scenario 3b were 0.124, 0.077, and 0.065 for single link, two links, and three links, while the errors in Scenario 4 were 15.06%, 22.91%, and 93.19% worse, respectively.

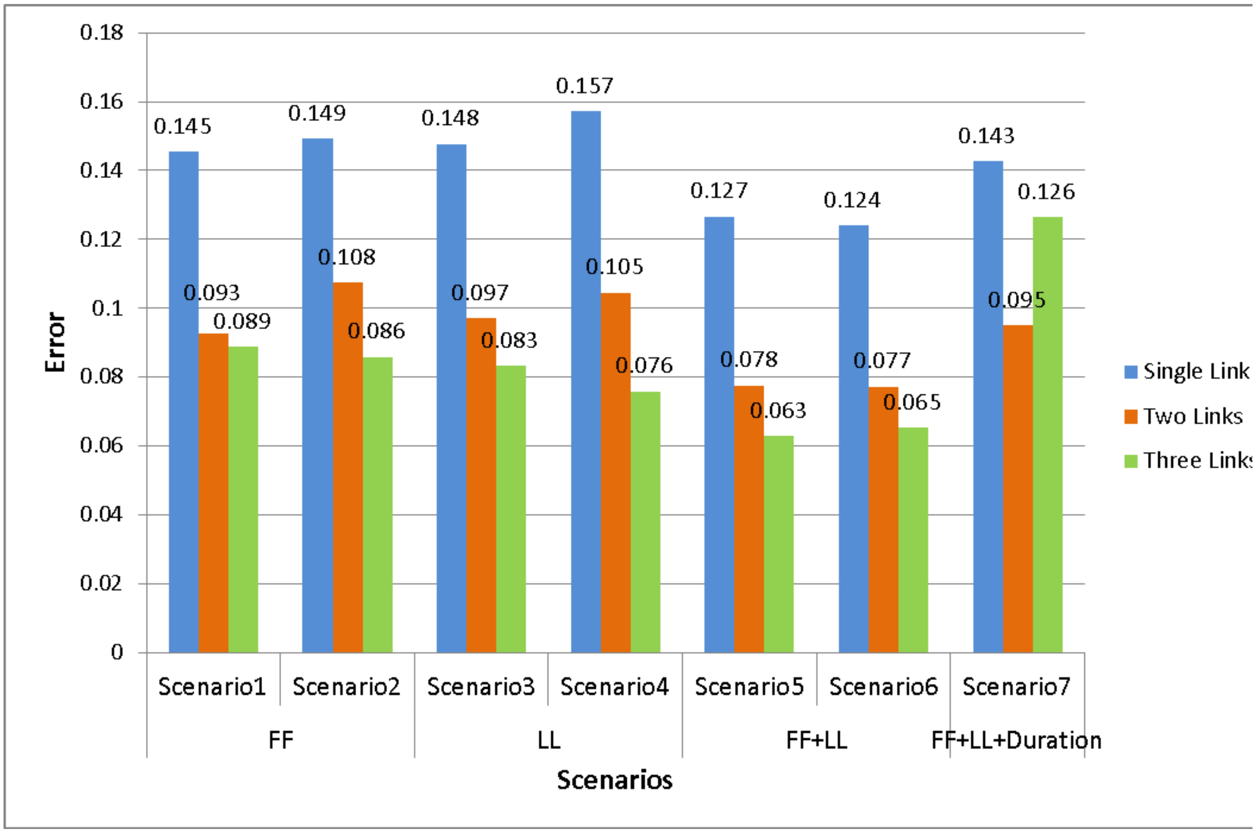
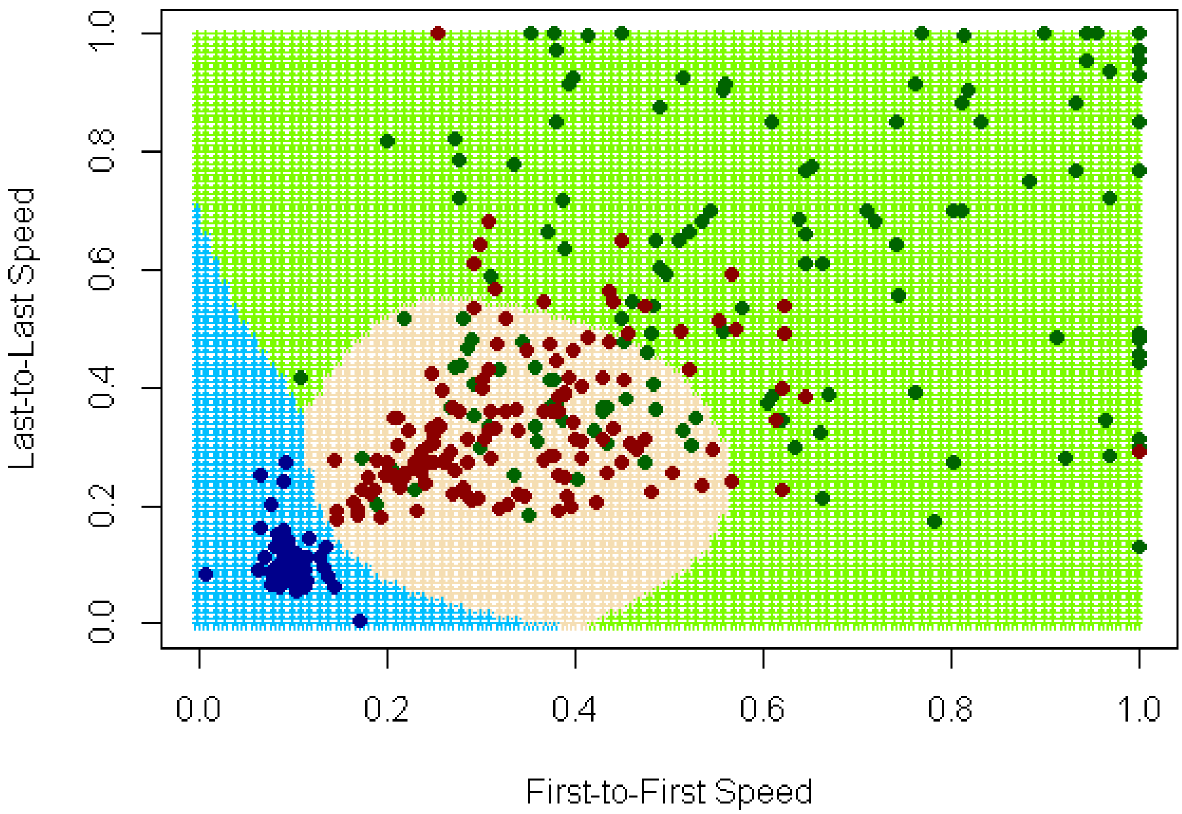


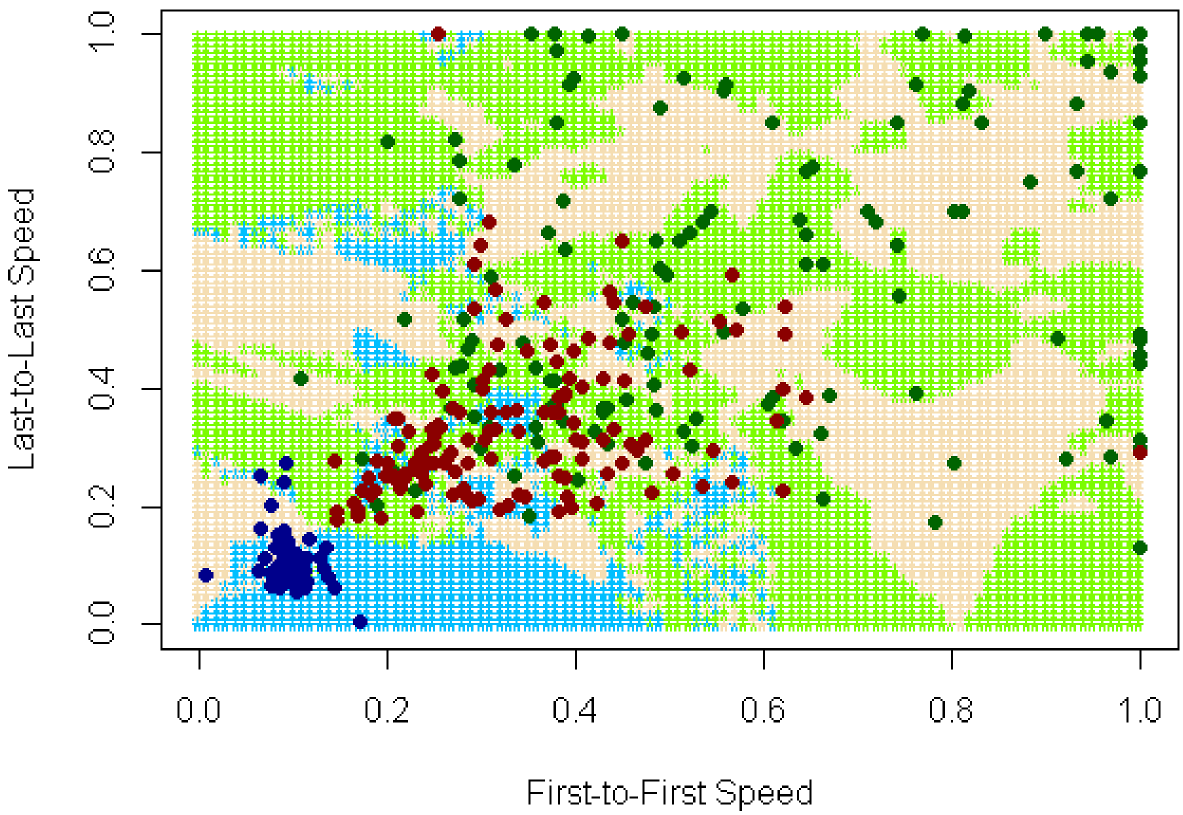
Figure 4-1. Peer Comparisons among 21 GANN-based Models

4.2 GANN-based Models vs. KNN

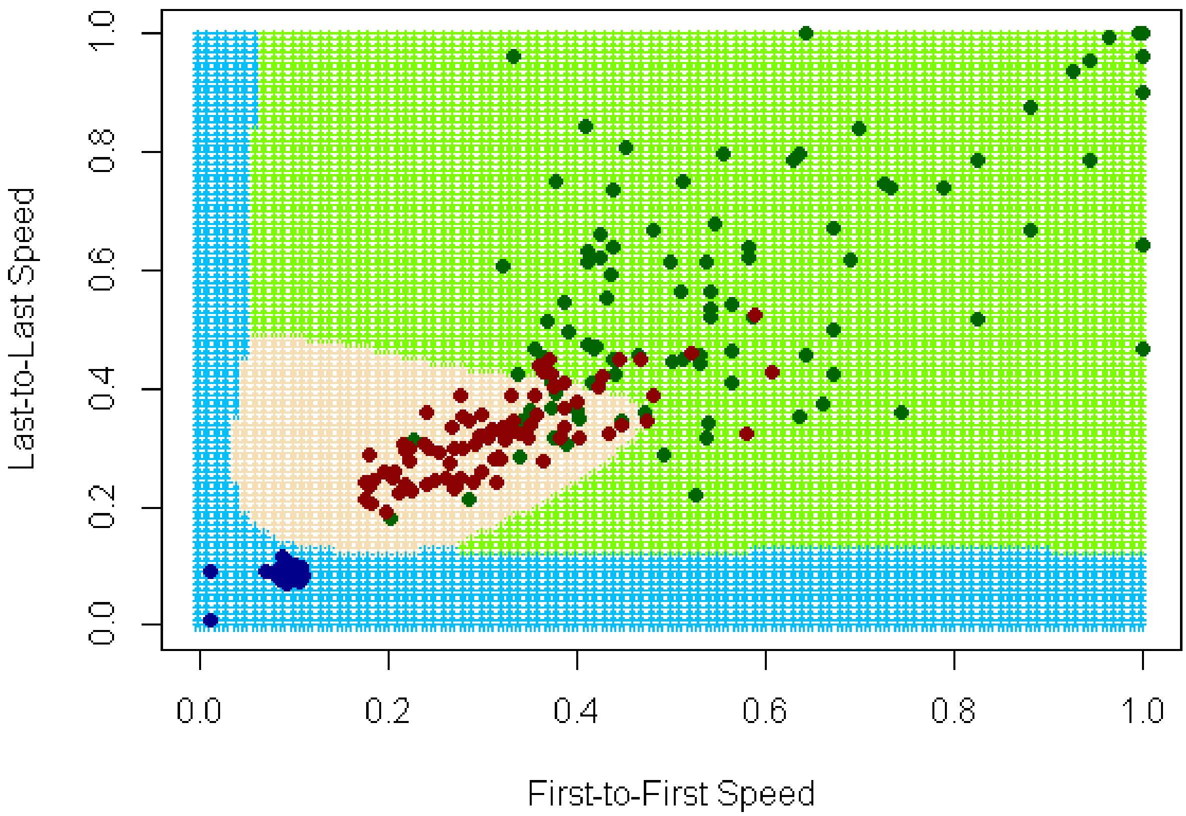
The test data in Table 2-1 was used to examine the GANN-based models' performance. As a comparison, the K-nearest neighborhood (KNN) (Altman, 1992) was also applied on the same dataset. Figure 4-2 compares the two models using graphs. The axes on each graph represent the FF and LL speed scaled from 0 to 1 based on the lowest and highest recorded speeds. Each graph is split into three colored regions representing the different travel modes. For example, if the normalized FF and LL speeds of a particular Bluetooth-enabled device were 0.8 and 0.2, respectively, then the device's speed would be located in the green region of Figure 4-2(a), indicating that the device was most likely an auto. However, the same device would be located in the tan area in Figure 4-2(b), indicating the travel mode was a bike. The test data is also separated by number of links.



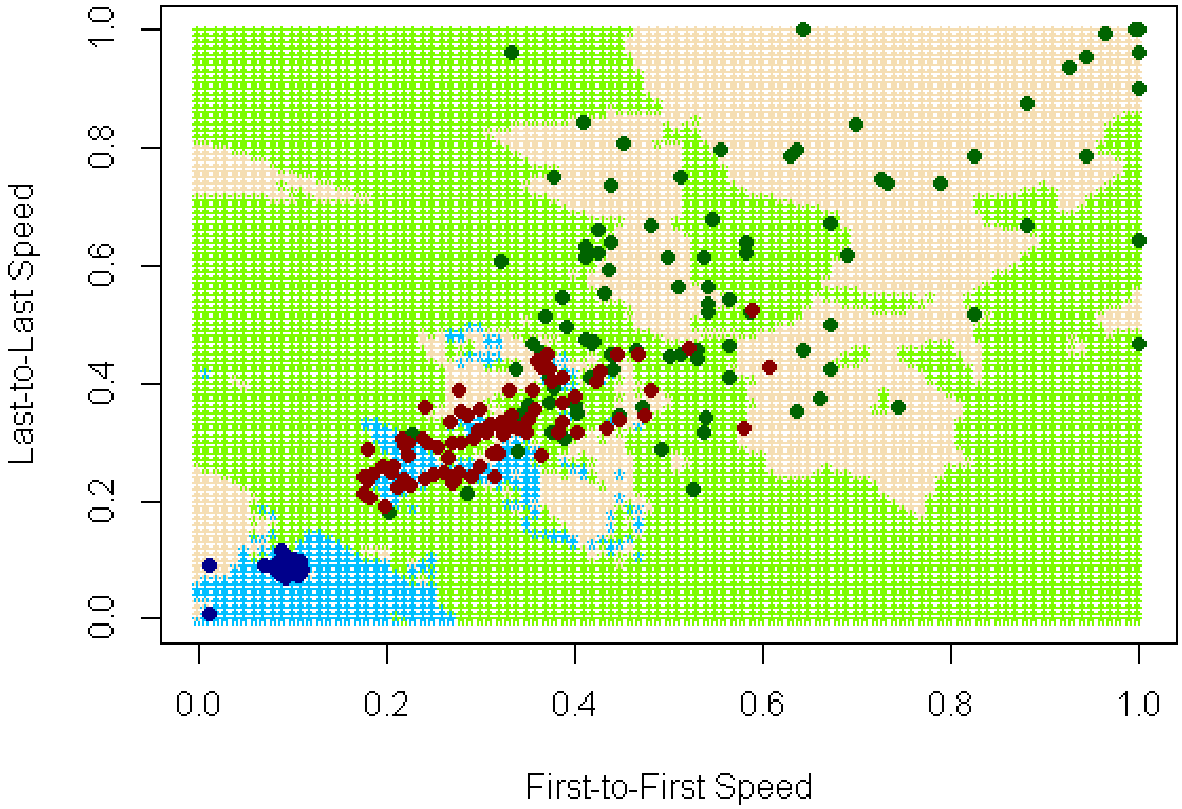
(a) GANN-based model, single link



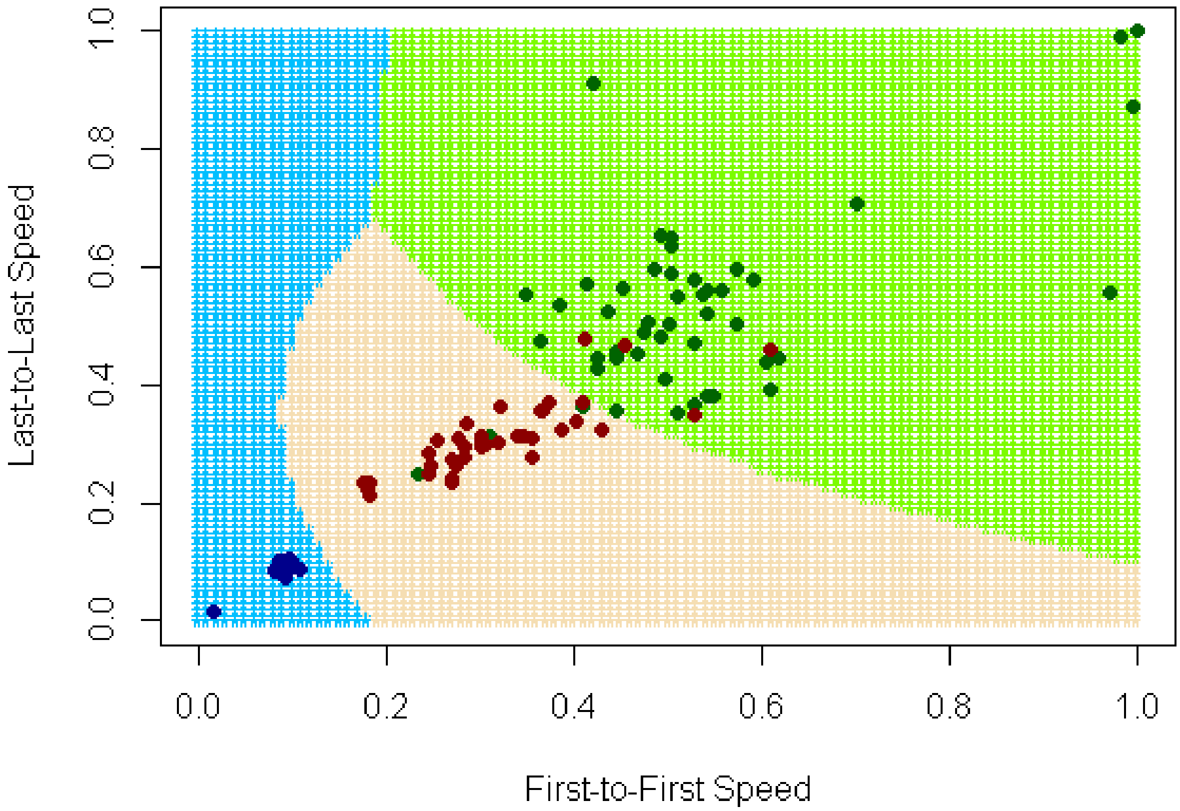
(b) KNN, single link



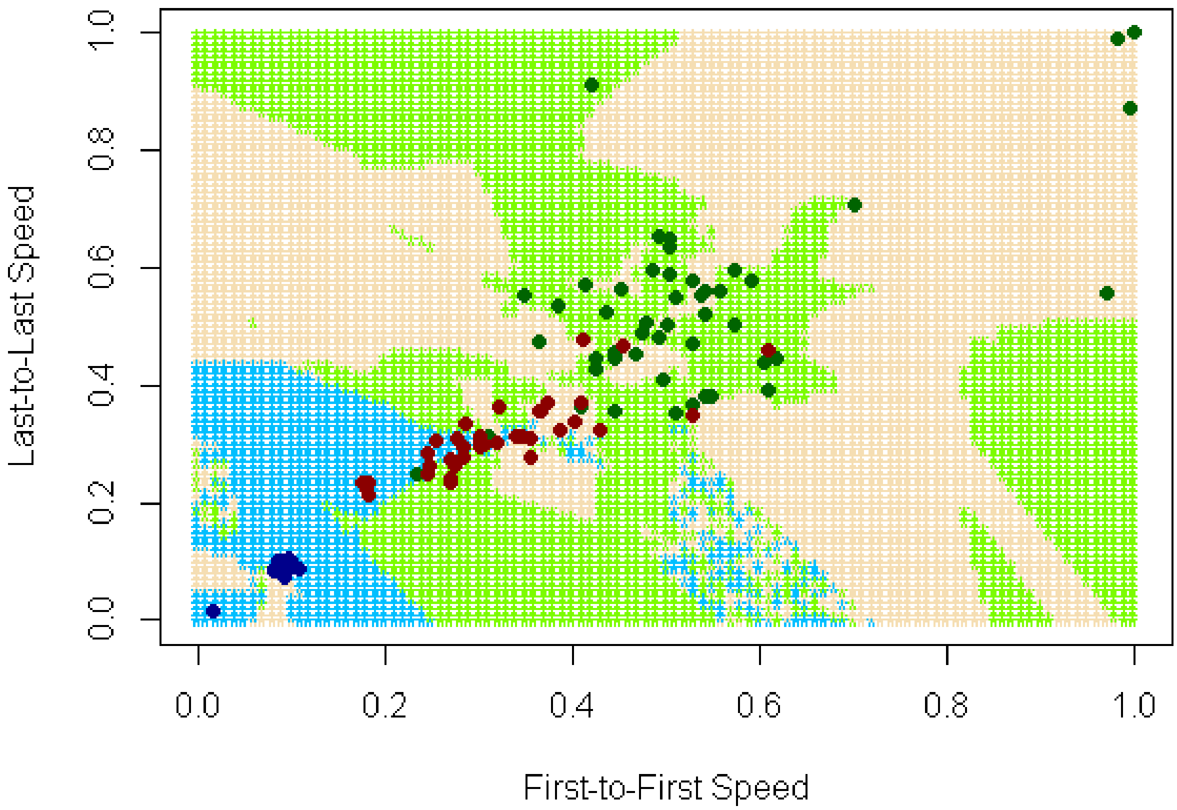
(c) GANN-based model, two links



(d) KNN, two links



(e) GANN-based model, three links



(f) KNN, three links

Figure 4-2. Ground Truth Data vs. Classification Areas by Modes

Note that the input parameters in the GANN-based model were scaled for calculation convenience between zero and one.

Figure 4-3 quantifies the performance differences between the GANN-based model based on Scenario 3b and the KNN. Several findings are summarized below: Overall, the GANN-based model outperformed the KNN. For example, in the three links case, 6.12% of autos were misidentified as bikes and 10.53% of bikes were misidentified as autos using the GANN-based model, while the corresponding misidentification rates were 22.45% and 34.21% using the KNN.

The GANN-based model separated the entire graph plane into three continuous regions. Test data was clearly separated between the regions.

The KNN could not adequately distinguish the pedestrian mode from other modes. Intuitively, pedestrians should be easily distinguishable due to their low speed (approximately 1.5 – 3 mph). The GANN-based model successfully identified pedestrians with a 0% misidentification rate. However, the KNN failed to identify pedestrians in many cases. The failure of the KNN may account for the extensive overlap between the three modes (see Figure 4-4), especially autos and bicyclists. The KNN searched for k nearest points with the same or similar attributes, and then classified the k points as a group. Therefore, instead of separating the entire plane into three areas, the KNN organized data into separated groups. This KNN behavior resulted in test data located outside travel mode groups being assigned to the closest group, causing misidentification. Figure 4-4 (b), (d) and (f) depict the identified travel mode groups using the KNN.

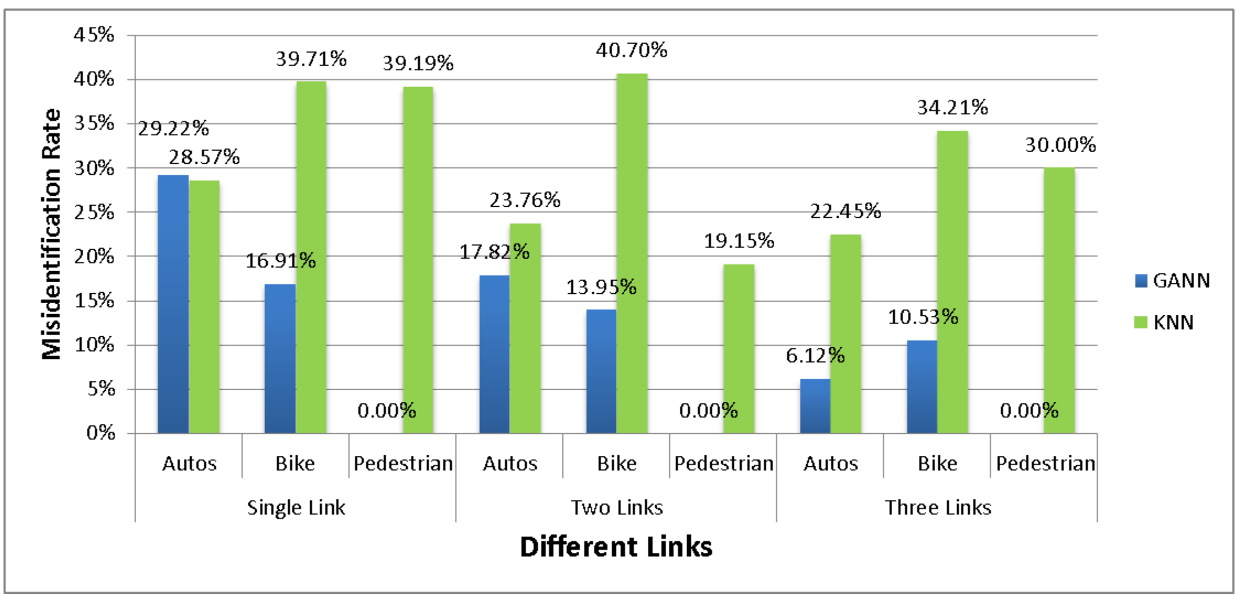


Figure 4-3. Performance Differences between the GANN model and KNN

4.3 Discussion

4.3.1 Similar Traffic Performance Measures on Short Corridors

Figure 4-4 (a) shows that bike speeds overlapped with auto speeds. The inherent difficulties of travel mode identification may explain this effect. The speeds of autos, bikes, and pedestrians on Speedway were approximately 35 mph, 12 mph, and 2.5 mph, respectively. Intuitively, they should be distinguishable. However, several factors can significantly affect speed estimation using Bluetooth-based data:

1) Short corridor length: Haghani et al. (2010) showed that estimated speed errors were approximately 4.5 mph for an arterial link length of 0.5 miles and speed limit of 30 mph. However, the link length in our study site was 0.32 miles, indicating that the estimated speed errors may have been greater than 4.5 mph. These speed errors could have resulted in speeds of different travel modes overlapping. However, our study also proved that the misidentification rates in the three-link corridor were lower than those in the single-link corridor, suggesting that speeds were more accurately estimated in the three-link corridor.

2) Poorly coordinated traffic signals: if two consecutive traffic signals are not well coordinated, an auto stopped at the first signal may have to stop again at the second signal. Low speed travel modes, such as bikes, would have enough time to catch up to the auto as it waited at the second signal. In this case, both the travel time and average speed of the auto and the low speed mode would be similar.

3) Traffic congestion: since bikes and pedestrians often travel on bike lanes and sidewalks, they are much less affected by vehicular traffic congestion. However, auto speeds would be lower due to the delay caused by traffic congestion. Therefore, average bike speeds were sometimes faster than auto speeds in our study corridor.

4.3.2 Potential Applications

Travel mode identification on arterials can assist estimation of mode-specific traffic performance measures (e.g. travel time and speed). In addition, a Bluetooth-based travel time outlier detection algorithm could be developed based on mode identification. If a Bluetooth-enabled device is identified as an auto during a relatively short time period, the data from this device could be used for further traffic measure estimation.

Section 5 CONCLUSION

Under the Moving Ahead for Progress in the 21st Century Act of 2012 (MAP-21), transportation agencies now face greater requirements with respect to the collection and analysis of surface-transportation data. Data-driven performance measurement is expected to play an important role in assisting transportation agencies with their transportation operations and planning decisions. The transportation world is experiencing a major shift from a "data desert" to a "data ocean". With the emerging development of Intelligent Transportation Systems (ITS) technologies, surface-transportation data can now be collected by a wide variety of ITS traffic sensors, including Bluetooth sensors, automatic vehicle location (AVL) devices, inductive loop sensors, and radar-based detectors. In practice, ITS data are collected from multiple sources but individually analyzed or processed. It has been challenging to take full advantage of the ITS data from multiple sources by enabling them to exchange information with each other to compensate for their various disadvantages.

Many previous studies have utilized Bluetooth-based data to measure traffic performance. However, travel time on arterials may be inaccurate because of mixed travel modes traveling at different speeds. Therefore, travel mode identification becomes necessary before further data processing. Our report proposed a genetic algorithm neural network (GANN) based model to identify travel modes on the study corridor in Tucson, Arizona. Twenty-one groups of input candidates were tested. To calibrate and verify these GANN-based models, Bluetooth-based data with known travel modes were collected. The Bluetooth-based infrastructure on the study corridor, which had been developed and maintained since 2013, facilitated data collection.

Several important findings from our studies are summarized below:

- Using both First-to-First (FF) and Last-to-Last (LL) speed as inputs performed better than using FF or LL speed alone.
- The detection ranges of the travel modes had little impact on travel mode identification.
- The travel mode misidentification rate can be decreased by considering higher numbers of arterial links.
- Duration data may not improve the rate of successful travel mode identification.
- The GANN based model outperformed the KNN. Using the KNN, even pedestrians were sometimes misidentified as other modes.

It was found challenging to identify the three travel modes, pedestrians, bicyclists and autos successfully. However, the GANN-based model developed in our study had low misidentification rates, i.e. only 6.12% of autos were misidentified as bikes and 10.53% of bikes were misidentified as autos. The GANN-based travel mode identification model showed its potential to detect travel time outliers and further clean Bluetooth-based data. Future studies will focus on two areas: 1) development of the outlier detection algorithm based on the GANN model. Obtaining the percentages of bike and autos in reality could help further improve the GANN model. 2) Mode-specific travel time could be reported after travel modes were identified.

References

Altman, N. (1992). An introduction to Kernel and Nearest-neighbor Nonparametric Regression. *The American Statistician*, *46*(3), 175–185. doi:10.1080/00031305.1992.10475879

Aliari, Y., & Haghani, A. (2012). Bluetooth Sensor Data and Ground Truth Testing of Reported Travel Times. *Transportation Research Record: Journal of the Transportation Research Board*, *2308*, 167–172. doi:10.3141/2308-18

Araghi, B. N., Christensen, L. T., Krishnan, R., Olesen, J. H., & Lahrmann, H. (2013). Use of Low-Level Sensor Data to Improve the Accuracy of Bluetooth-Based Travel Time Estimation. *Transportation Research Record: Journal of the Transportation Research Board*, *2338*, 29–34. doi:10.3141/2338-04

Barceló, J., Montero, L., Bullejos, M., Serch, O., & Carmona, C. (2013). A Kalman Filter Approach for Exploiting Bluetooth Traffic Data When Estimating Time-Dependent OD Matrices. *Journal of Intelligent Transportation Systems*, *17*(2), 123–141. doi:10.1080/15472450.2013.764793

Barceló, J., Montero, L., Marqués, L., & Carmona, C. (2010). Travel Time Forecasting and Dynamic Origin-Destination Estimation for Freeways Based on Bluetooth Traffic Monitoring. *Transportation Research Record: Journal of the Transportation Research Board*, *2175*, 19–27. doi:10.3141/2175-03

Bhaskar, A., & Chung, E. (2013). Fundamental Understanding on the Use of Bluetooth Scanner as a Complementary Transport Data. *Transportation Research Part C: Emerging Technologies*, *37*, 42–72.

Haghani, A., Hamedi, M., Sadabadi, K. F., Young, S., & Tarnoff, P. (2009). Data Collection of Freeway Travel Time Ground Truth with Bluetooth Sensors. *Transportation Research Record: Journal of the Transportation Research Board*, *2160*, 60–68. doi:10.3141/2160-07

Hainen, A., Wasson, J., Hubbard, S., Remias, S., Farnsworth, G., & Bullock, D. (2011). Estimating Route Choice and Travel Time Reliability with Field Observations of Bluetooth Probe Vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, *2256*, 43–50. doi:10.3141/2256-06

Haseman, R., Wasson, J., & Bullock, D. (2010). Real-Time Measurement of Travel Time Delay in Work Zones and Evaluation Metrics Using Bluetooth Probe Tracking. *Transportation Research Record: Journal of the Transportation Research Board*, *2169*, 40–53. doi:10.3141/2169-05

Khoei, A. M., Bhaskar, A., & Chung, E. (2013). Travel Time Prediction on Signalised Urban Arterials by Applying SARIMA Modelling on Bluetooth Data. In *In 36th Australasian*

Transport Research Forum (ATRF) 2013. Queensland University of Technology, Brisbane, QLD.

Koehn, P. (1994). *Combining Genetic Algorithms and Neural Networks : The Encoding Problem*. The University of Tennessee. Retrieved from http://homepages.inf.ed.ac.uk/pkoehn/publications/gann94.pdf

Mei, Z., Wang, D., & Chen, J. (2012). Investigation with Bluetooth Sensors of Bicycle Travel Time Estimation on a Short Corridor. *International Journal of Distributed Sensor Networks*, *2012*(1), 1–7. doi:10.1155/2012/303521

Michalewicz, Z. (1996). *Genetic Algorithms + Data Structures = Evolution Programs* (3rd, rev.). Springer.

Moghaddam, S. S., & Hellinga, B. (2013a). Evaluating the Performance of Algorithms for the Detection of Travel Time Outliers. *Transportation Research Record: Journal of the Transportation Research Board*, *2338*, 67–77. doi:10.3141/2338-08

Moghaddam, S. S., & Hellinga, B. R. (2013b). Quantifying Measurement Error in Arterial Travel Times Measured by Bluetooth Detectors. *Transportation Research Record: Journal of the Transportation Research Board*, *2395*, 111–122.

Qiao, W., Haghani, A., & Hamedi, M. (2013). A Nonparametric Model for Short-Term Travel Time Prediction Using Bluetooth Data. *Journal of Intelligent Transportation Systems*, *17*(2), 165–175. doi:10.1080/15472450.2012.748555

Quayle, S. M., Koonce, P., DePencier, D., & Bullock, D. M. (2010). Arterial Performance Measures with Media Access Control Readers. *Transportation Research Record: Journal of the Transportation Research Board*, *2192*(-1), 185–193. doi:10.3141/2192-18

Rojas, R. (1996b). The Backpropagation Algorithm. In *Neural Networks - A Systematic Introduction* (pp. 151–184). Springer. doi:10.1109/78.127967

Saeedi, A., Park, S., Kim, D. S., & Porter, J. D. (2013). Improving Accuracy and Precision of Travel Time Samples Collected at Signalized Arterial Roads with Bluetooth Sensors. *Transportation Research Record: Journal of the Transportation Research Board*, *2380*(-1), 90–98. doi:10.3141/2380-10

Tsubota, T., Bhaskar, A., Chung, E., & Billot, R. (2011). Arterial Traffic Congestion Analysis Using Bluetooth Duration Data. *Australasian Transport Research Forum*, (September), 28–30.

Van Boxel, D., Schneider, W. H., & Bakula, C. (2011). Innovative Real-Time Methodology for Detecting Travel Time Outliers on Interstate Highways and Urban Arterials. *Transportation Research Record: Journal of the Transportation Research Board*, *2256*, 60–67. doi:10.3141/2256-08

Wasson, J. S., Sturdevant, J. R., & Bullock, D. M. (2008). Real-Time Travel Time Estimates Using Media Access Control Address Matching. *ITE Journal*, 20–23.

Yao, X. (1999). Evolving Artificial Neural Networks. In *Proceedings of the IEEE* (Vol. 87, pp. 1423–1447). doi:10.1109/5.784219

Yang, S. (2015). Teach Traffic Smart. Retrieved Oct. 12, 2015, from https://onthejeep.wordpress.com/2015/10/12/feedforward-neural-network-development-single-layer/