OTREC

FINAL REPORT

OREGON TRANSPORTATION RESEARCH AND EDUCATION CONSORTIUM

Exploratory Methods for Truck Re-Identification in a Statewide Network Based on Axle Weight and Axle Spacing Data to Enhance Freight Metrics

OTREC-RR-11-07
February 2011

A National University Transportation Center sponsored by the U.S. Department of Transportation's Research and Innovative Technology Administration

EXPLORATORY METHODS FOR TRUCK RE-IDENTIFICATION IN A STATEWIDE NETWORK BASED ON AXLE WEIGHT AND AXLE SPACING DATA TO ENHANCE FREIGHT METRICS

Research Report

OTREC-RR-11-07

by
Christopher M. Monsere
Portland State University, Portland, OR

Mecit Cetin
Old Dominion University, Norfolk, VA

Andrew P. Nichols
Marshall University, Huntington, WV

for
Oregon Transportation Research
and Education Consortium (OTREC)
P.O. Box 751
Portland, OR 97207

OTREC
OREGON TRANSPORTATION RESEARCH AND EDUCATION CONSORTIUM

February 2011

Technical Report Documentation Page

1. No.	2. Government Accession No.	3. Recipient's Catalog No.
OTREC-RR-11-07		

4. Title and Subtitle	5. Report Date
Exploratory Methods for Truck Re-Identification in a Statewide Network Based on Axle Weight and Axle Spacing Data to Enhance Freight Metrics	February 2011
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Christopher M. Monsere, Mecit Cetin, Andrew P. Nichols	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Portland State University and Old Dominion University	11. Contract or Grant No.
12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Oregon Transportation Research and Education Consortium (OTREC) P.O. Box 751 Portland, Oregon 97207	Final Report
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

The main objective of this project is to evaluate the feasibility of re-identifying commercial trucks based on vehicle-attribute data automatically collected by sensors installed at traffic data collection stations. To support this work, archived data from weigh-in-motion (WIM) stations in Oregon are used for developing, calibrating, and testing vehicle re-identification algorithms. The vehicle re-identification methods developed in this research consist of two main stages. In the first stage, each vehicle from the downstream station is matched to the most "similar" upstream vehicle by using a Bayesian model. In the second stage, several methods are introduced to screen out those vehicles that cross the downstream site but not the upstream site and to tradeoff accuracy versus the total number of vehicles being matched. These methods involve calculating both the highest and the second highest similarity measures for each vehicle being matched. It is demonstrated that the proposed screening approach improves the accuracy of the re-identification methods significantly. The models are applied to the truck data collected by WIM sensors at three stations in Oregon, which together create two different "links" that are 125 and 145 miles long, respectively. It is observed that the algorithms can match trucks with approximately 90% accuracy while the total number of trucks being matched at this accuracy level is about 95% of the actual common trucks that cross both upstream and downstream sites. These methods allow the user to trade-off the accuracy vs. total vehicles being matched by adjusting a threshold parameter. For example, trucks can be matched with 98% accuracy if one is willing to match about 40% of all common trucks. It is also found that when travel times of vehicles between the upstream and downstream sites exhibit larger variation, mismatch rate increases. Overall, for estimating travel times and origin-destination flows between two WIM sites, the methods developed in this project can be used to effectively match commercial vehicles crossing two data collection sites that are separated by long distances.

17. Key Words	18. Distribution Statement
vehicle re-identification, weigh-in-motion, freight performance measures	No restrictions. Copies available from OTREC: www.otrec.us

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	74	

ACKNOWLEDGEMENTS

The authors acknowledge the Oregon Transportation Research and Education Consortium (OTREC) for funding this research and the Oregon Department of Transportation Motor Carrier Division for providing the data. At ODOT, Dave Fifer and David McKane were particularly helpful. Christopher Higgins at Oregon State University contributed to the WIM archive by providing software tools and expertise. At Portland State University, Kristin Tufte and Michael Wolfe were instrumental in data management through the PORTAL umbrella. The National Science Foundation supported early development of PORTAL. The contents of this paper reflect the views and opinions of the authors, who are responsible for the facts and the accuracy of the data presented here. The authors also acknowledge the help of graduate students Faisal Mahmud and Ilyas Ustun from ODU who contributed to this work substantially by conducting data analyses.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

Most transportation agencies rely on point detectors (e.g., inductive loops, axle detectors) located at specific points on highways to collect data on traffic volumes, vehicle classes, and other relevant attributes of traffic. By utilizing the data collected from these point detectors, researchers have developed vehicle re-identification algorithms to match measurements at two sites that belong to the same vehicle. This enables tracking the movement of individual vehicles between different data collection sites, which in turn provides valuable information for the estimation of travel times, travel delays, and origin-destination flows.

The aim of this OTREC project is to investigate the feasibility of re-identifying trucks in a statewide network by developing and applying vehicle re-identification algorithms. Data from weigh-in-motion (WIM) stations provide a basis for the development and testing of these algorithms. The data supporting this research come from the WIM sites in Oregon, which are equipped with sensors that can measure axle weights, axle spacing, and gross vehicle weight estimates that are uniquely matched to each truck. Since some of the trucks (20-35%) are carrying Green Light transponders, these measured attributes are also uniquely matched to transponder-equipped trucks. These particular trucks provide the needed data for model development, calibration, and testing.

The vehicle re-identification method developed in this research consists of two main stages. In the first stage, each vehicle from the downstream station is matched to the most "similar" upstream vehicle, as is typically done in vehicle re-identification methods. Both a Euclidian distance method and a Bayesian method are utilized to solve the first-stage problem. In the second stage, several methods are introduced to screen out those vehicles that cross the downstream site but not the upstream site, and to trade off accuracy versus the total number of vehicles being matched. These methods involve calculating both the highest and the second highest similarity measures for each vehicle being matched. Several criteria are suggested and evaluated for screening mismatched vehicles of the first stage based on these similarity measures. As demonstrated in this report, the proposed screening approach improves the accuracy of the re-identification methods significantly.

The models are applied to the truck data collected by weigh-in-motion (WIM) and automatic vehicle classification (AVC) sensors at three stations in Oregon, which together create two different "links" that are 125 and 145 miles long, respectively. It is observed that the algorithms can match trucks with approximately 90% accuracy while the total number of trucks being matched at this accuracy level is about 95% of the actual common trucks that cross both upstream and downstream sites. These methods allow the user to trade off the accuracy versus total vehicles being matched by adjusting a threshold parameter. For example, trucks can be matched with 98% accuracy if one is willing to match about 40% of all common trucks. It is also observed that when travel times of vehicles between the upstream and downstream sites exhibit larger variation, the re-identification becomes more challenging. In other words, mismatch rate increases as travel-time variance increases. Overall, for estimating travel times and OD flows, the methods presented in this report can be used effectively to match commercial vehicles crossing two data collection sites that are separated by long distances.

1.0 INTRODUCTION

Most transportation agencies rely on point detectors (e.g., inductive loops, axle detectors) located at specific points on highways to collect data on traffic volumes, vehicle classes, and other relevant attributes of traffic. By utilizing the data collected from these point detectors, researchers have developed vehicle re-identification algorithms to match measurements at two sites that belong to the same vehicle. This enables tracking the movement of individual vehicles between different data collection sites, which in turn provides valuable information for the estimation of travel times, travel delays, and origin-destination flows.

Even though there are other technologies that can be utilized to track the vehicle movement over transportation networks, most of these technologies (e.g., automatic vehicle identification (AVI) tags, license plate recognition) require installation of additional in-car and/or roadside devices and may have related privacy concerns. However, vehicle re-identification methods that are based on the vehicle-attribute data collected by sensors already installed on roadways enable tracking vehicles anonymously and do not require substantial additional investment.

The aim of this OTREC project is to investigate the feasibility of re-identifying trucks in a statewide network by developing and applying vehicle re-identification algorithms. Data from weigh-in-motion (WIM) stations provide a basis for the development and testing of these algorithms. The data supporting this research come from the WIM sites in Oregon, which are equipped with sensors that can measure axle weights, axle spacing, and gross vehicle weight estimates that are uniquely matched to each truck. Since some of the trucks (20-35%) are carrying Green Light transponders, these measured attributes are also uniquely matched to transponder-equipped trucks. These particular trucks provide the needed data for model development, calibration, and testing.

This report describes the algorithms developed for matching trucks crossing both an upstream and a downstream site. Since these sites are separated by long distances (e.g., more than 100 miles), trucks traveling between these two points may stop for fuel or deliveries and thus may have different weights at two sites, which makes the re-identification a challenge. Furthermore, some trucks may have very similar attribute data (e.g., axle spacing) which makes distinguishing between individual trucks difficult. Despite these complications, as shown in the report, the methods perform reasonably well and can be potentially used in practice.

Overall, the methods developed in this research can be used to support programs and applications for monitoring freight over the highways. One of the key aspects of monitoring freight has to do with determining the flow patterns (and travel times) of trucks, which can be achieved by uniquely identifying trucks at specific points along the roads or by tracking individual trucks using technology such as GPS. The re-identification method, in some circumstances, can be more advantageous compared to other available options to track and re-identify trucks (e.g., GPS, AVI, license plate recognition) because of several reasons:

- Data from AVI transponders, such as Green Light, or from other types of electronic tracking systems might not be readily available to the public agencies involved in motor

freight planning (e.g., MPOs, DOTs) due to privacy, jurisdictional, and institutional issues;

- Not all trucks are equipped with AVI transponders. However, with the re-identification methods all trucks can be potentially tracked since they all cross the WIM stations; and

- The proposed approach does not require installation of any new sensors since the input data are already collected at existing WIM and automatic vehicle classification (AVC) stations, whereas alternative technologies like license plate recognition requires additional investment.

1.1 OBJECTIVES

By building upon past and ongoing research by the PIs and others in the areas of WIM data analysis, travel-time estimation for commercial trucks and vehicle re-identification methods, this research aims to contribute to the state-of-the-art and state-of-practice in freight movement by developing and testing novel vehicle re-identification methods to improve the ability to estimate truck movements in a transportation network. These methods capitalize on vehicle-attribute data, such as axle spacing and axle weights, which are already collected by numerous sensors installed on roadways.

The specific objectives of this project are:

- To evaluate the feasibility of re-identifying commercial trucks based on vehicle-attribute data automatically collected by sensors installed at traffic-data collection stations;

- To develop robust algorithms for truck re-identification based on these data; and

- To test and evaluate the level of accuracy of the matching algorithms under different scenarios (e.g., depending on the distance between stations, available vehicle data, truck volume, truck type).

1.2 ORGANIZATION OF THE REPORT

This report is organized as follows: The next chapter provides an overview of some relevant studies on vehicle re-identification methods and applications, and Chapter 3 describes the WIM data utilized for model development and testing in this project. Chapter 4 describes the problem of re-identification in detail and presents the algorithms developed in this project. Chapter 5 describes statistical finite mixture models and estimation of their parameters. These models are used in creating probability distributions needed for the re-identification algorithms described in Chapter 4. Chapter 6 presents the results of the application of the algorithms to the sample datasets. The study's conclusions are given in Chapter 7.

2.0 LITERATURE REVIEW

As explained in *A Concept for a National Freight Data Program: Special Report 276*, data on goods movements are needed to identify and evaluate options for mitigating congestion; improve regional and global economic competitiveness; inform investment and policy decisions about modal optimization; enhance transportation safety and security; identify transportation marketing opportunities; and reduce fuel consumption and improve air quality (TRB 2003). This project contributes to a better understanding of freight movement by developing re-identification algorithms to estimate truck O-D (origin-destination) flows and travel times for trucks. Even though determining truck counts at particular locations on a transportation network is relatively easy to do, obtaining O-D data is, in general, more difficult since it requires uniquely re-identifying trucks at multiple points.

Since the mid-1990's, many research efforts have focused on methods to anonymously track vehicular movements by re-identifying individual vehicles at multiple locations utilizing existing sensors. The predominant objective has been to estimate travel times in order to characterize link performance. For this reason, the re-identification has focused primarily on passenger cars and light trucks, which typically make up the majority of traffic in urban areas where the link performance varies the most. Various techniques and technologies have been employed for the re-identification of vehicles including video/imaging (Shuldiner and Upchurch 2001) and AVI (Dion and Rakha 2006; Hellinga 2001). A more detailed explanation of these technologies and the associated techniques can be found in the *Travel Time Data Collection Handbook* (Turner et al. 1998).

There have been several studies on re-identifying individual vehicles anonymously at multiple locations by utilizing data from existing inductive dual loop detectors (Sun et al. 1999; Coifman and Cassidy 2002; Coifman 2003). While most of the previous studies are based on data from dual loops, some researchers also extended the application of the re-identification algorithms to data from single loops (Coifman and Krishnamurthy 2007). Other than the traditional inductive loops that are embedded in the pavement, researchers have investigated new types of inductive loops, the so-called "blade sensors," to get more detailed characteristics of vehicles. These sensors are more sensitive than the typical inductive loops and are capable of capturing wheel locations (Oh et al. 2007). In general, magnetic vehicle signatures from loops provide the raw data, which is used to extract useful vehicle features or attributes to differentiate between different vehicles. The predominant application of vehicle re-identification has been to estimate travel times (Liu et al. 2002; Oh et al. 2005; Sun et al. 2003).

Less attention has been given to the techniques to re-identify commercial vehicles at multiple locations, even though such techniques can support numerous applications including estimating travel times for trucks, quantifying travel-time reliability, estimating truck-flow patterns (i.e., origins-destinations), estimating empty-truck movements, trip-length estimation, pavement management, WIM-sensor accuracy, and weigh-station enforcement.

Recently, the authors of this report explored the use of axle spacing and axle weight data to re-identify commercial trucks at two WIM stations in Indiana where commercial trucks cross both stations (Cetin and Nichols 2009). They developed matching algorithms based on statistical mixture models and tested the performance of the algorithms on the data from these two WIM stations that are separated by one mile. The results showed that trucks were matched with 99% accuracy when both axle spacing and weight were used; and with 97% accuracy when only axle spacing was used. However, the WIM stations in this study were only separated by one mile and all trucks in the sample crossed both the upstream and downstream stations. As explained in this report, the datasets used in this project come from WIM stations in Oregon that are separated by greater distances (more than 100 miles), which introduces additional complexities since travel times can vary significantly and trucks can leave and new ones enter the road in between the two stations (this was not the case in the Indiana dataset).

Other than the work by Cetin and Nichols in 2009, the only known previous application of WIM data for vehicle re-identification was conducted by the Norway Public Roads Administration for determining link travel times on the Oslo Toll Ring (Christiansen and Hauer 1996). A prototype of the system was tested at the Winter Olympic Games in Lillehammer in 1994 and was later refined with more advanced matching algorithms.

In general, vehicle re-identification methods rely on the variability within the vehicle population and the ability to accurately identify the pairs of measurements collected at upstream and downstream stations that are generated by the same vehicle. These measurements can either be the actual physical attributes of vehicles such as length (Coifman and Cassidy 2002) and axle spacing (Cetin and Nichols 2009) or some characteristics of the sensor waveform or inductive vehicle signature (Sun et al. 1999). Researchers have developed various methods, such as lexicographic optimization (Sun et al. 1999; Oh et al. 2007) and decision trees (Tawfik et al. 2004) to re-identify vehicles. In a typical implementation of these methods, a downstream vehicle is matched to the most "similar" upstream vehicle (or vice versa) based on some defined metric (e.g., Euclidian distance). The resulting accuracy of these methods depends on several factors, including the variation of the attribute data from vehicle to vehicle, number of attributes, the distance between data collection stations, variability of travel time, and type of re-identification algorithm used. Given a particular set of factors, this accuracy may or may not be satisfactory for a given application. It would be desirable to have a model to "adjust" the level of accuracy by perhaps being more judicious in matching vehicles. In other words, the model should match a (select) set of vehicles rather than all vehicles such that the accuracy is maintained at an acceptable level.

This research presents a new approach on how this can be done effectively. A recent paper summarizing some of the findings presented in this report recently was submitted for publication by the authors (Cetin et al. 2010).

3.0 WEIGH-IN-MOTION DATA

In this chapter, the assembly, processing, and storage of the weigh-in-motion (WIM) and automatic vehicle classification (AVC) data is described. Oregon's prescreening/preclearance program for commercial motor vehicles at fixed weigh and inspection stations is called Green Light. There are 22 equipped stations on the Oregon highway system. These locations are shown in Figure 3.1 with a corresponding list of stations shown in Table 3.1. At each of the Green Light stations, approaching trucks are directed into the appropriate lane on the mainline highway. At a location upstream from the static weigh station, transponder-equipped trucks are identified by the reader. Participation in the Green Light program is high; on average, about 40% of observed vehicles are equipped with transponders (though this varies from station to station). In addition to the transponder record, the vehicles are weighed in motion (by load cells). The observation consists of axle weights as well as axel spacing. These data also include speed, timestamp, the lane of observation (some stations are multilane), length (calculated), gross vehicle weight (calculated), and a count of the number of axles (calculated). As part of the proprietary control program by the equipment vendor (International Road Dynamics), a sieved-based classification algorithm uses the axle spacing information to classify vehicles. An example of the transponder reader, over-height detection, and load and axle sensors is shown in Figure 3.2. A more detailed description of the Oregon WIM system is provided by Elkins and Higgins (2008).

The unique aspect of Oregon's system is that this transponder and weight-related data are available together in one record. These transponder-equipped vehicles provide a large pool of data to develop, validate, and test the vehicle re-identification techniques described within.

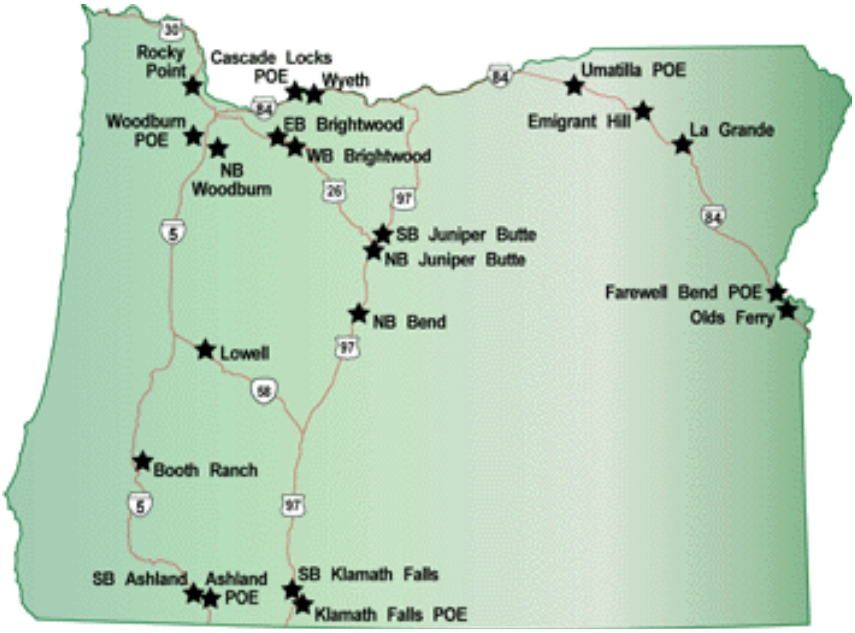


Figure 3.1: Oregon Green Light Locations

Figure 3.2: A Green Light Station Pre-clearance Arrangement

Table 3.1: List of stations

Number	Code	Name	Route	Direction	MP
1	FWB	Farewell Bend POE	I-84	WB	353.31
2	EMH	Emigrant Hill	I-84	WB	226.95
3	WYT	Wyeth	I-84	WB	54.3
4	CSL	Cascade Locks POE	I-84	EB	44.93
5	LGR	La Grande	I-84	EB	258.52
6	ODF	Olds Ferry	I-84	EB	354.38
7	ASP	Ashland POE	I-5	NB	18.08
8	BOR	Booth Ranch	I-5	NB	111.07
9	WDN	Woodburn, NB	I-5	NB	274.15
10	WDS	Woodburn, SB	I-5	SB	274.18
11	BRE	Brightwood, EB	US-26	EB	36.51
12	BRW	Brightwood, WB	US-26	WB	36.31
13	JBS	Juniper Butte	US-97	SB	108.2
14	LWL	Lowell	US-58	WB	17.17
15	WLB	Wilbur	I-5	SB	130
16	ASH	Ashland, SB	I-5	SB	18.08
17	KFP	Klamath Falls POE	US-97	NB	271.73
18	BND	Bend	US-97	NB	145.5
19	JBN	Juniper Butte	US-97	NB	106.9
20	KFS	Klamath Falls, SB	US-97	SB	271.41
21	UMT	Umatilla POE	I-82	EB	183.8
22	RPT	Rocky Point	US-30	WB	16.53

3.1 DATA ARCHIVE

In support of this and other research, a WIM data archive was created. This archive is housed under the Portland Transportation Archive Listing (PORTAL) umbrella at Portland State University's Intelligent Transportation Systems Lab. PORTAL is the official Archived Data User Service (ADUS) for the Portland metropolitan region as specified in the Regional ITS Architecture. PORTAL provides a centralized, electronic database that facilitates the collection, archiving, and sharing of information/data for public agencies within the region. The creation of the PORTAL data archive was supported by a CAREER grant from the National Science Foundation (NSF). In addition, the FHWA (through ODOT) has supported the purchase of hard disc storage and the Portland metropolitan regional government (Metro) has invested in the ongoing support of the archive.

The archive stores data in a PostgreSQL relational database management system (RDBMS). This archive implements a data warehousing strategy in that it retains large amounts of raw operational data for analysis and decision-making processes, and in that these data are stored independently of their operational sources, allowing the execution of time-consuming queries with no impact on critical operations uses. The database server is a Dell Server with two Quad Core Intel Xeon Processors running at 2.33 GHz with 8GB of memory. The database server runs Red Hat Linux. The RDBMS stores data physically on a 3.2 Terabyte redundant array of independent disks (RAID) providing both high-speed access and increased reliability through redundancy in the event of hardware failure. Offsite backups of the raw data are done once a week.

Monthly data are sent from ODOT via an FTP connection. These data are processed and then loaded into the WIM archive. A forthcoming OTREC report will describe the WIM data archive in detail (including data quality efforts) but a short description follows. There are four primary tables in the WIM data. A schematic of the database is shown in Figure 3.1. The truck-level observations are loaded in a table called *wimdata*. A table *stations* includes the identifying information about each station. The table *stationmap* is a list of all possible routes (i.e., upstream-to-downstream station pairs) which defines the free-flow travel time, distance, and a parameter called upper time (time to travel between stations at 50 mph). An algorithm described in Monsere et. al (2009), produces a table *linktraveltime* of all trucks matched by transponder identification number between stations. The search algorithm matches a truck with a transponder at an upstream station with the same transponder at the downstream station. All matches within the time window of 0.75*free-flow time to 2*free-flow time are recorded. Free-flow time is defined as the time to traverse the route between stations at 55 mph (the posted speed limit for trucks on Oregon roadways).This table contains the upstream and downstream station numbers, tag number, and timestamps of each observation and whether the truck has been identified as a thru vehicle.

At the time of this report's publication, data are available for every truck observed from July 2005 to October 2009 (approximately 43,053,800 observations).

WIMDATA

timestamp	timestampwithtimezone
year	integer
month	integer
day	integer
hour	integer
minute	integer
seconds	integer
lane	integer
speed	integer
type	integer
length	integer
gvw	real
esal	real
sumlen	real
numaxles	integer
axl1	real
axl2	real
axl3	real
axl4	real
axl5	real
axl6	real
axl7	real
axl8	real
axl9	real
axl10	real
axl11	real
axl12	real
axl13	real
axl14	real
spc1	real
spc2	real
spc3	real
spc4	real
spc5	real
spc6	real
spc7	real
spc8	real
spc9	real
spc10	real
spc11	real
spc12	real
spc13	real
spc14	real
tag	text
stationnum	integer
gvw_zero	boolean
gvw_50	boolean
mph_10	boolean
mph_99	boolean
length_200	boolean
axle_sum_length	boolean
axle_sum_7	boolean
axle_first_5	boolean
num_axle_13	boolean
gvw_280	boolean
axle_spc_34	boolean
gvw_diff_7	boolean
truck_table	integer

STATIONS

stationnum	integer
station_code	character(3)
longname	text
name	text
route	character(5)
direction	character(2)
hwy_no	integer
roadbed	integer
mp	doubleprecision
lrs	charactervarying(15)
lat	doubleprecision
long	doubleprecision
filename	prefixtext

STATIONMAP

linkid	integer
up_station	integer
up_stationname	character(3)
dwn_station	integer
dwn_stationname	character(3)
freeflow	real
distance	real
uppertime	real

LINKTRAVELTIME

linkid	integer
up_station	integer
up_tag	text
up_timestamp	timestampwithtimezone
dwn_station	integer
dwn_tag	text
dwn_timestamp	timestampwithtimezone
thru_truck	boolean

Figure 3.1: Key Table Definitions for PSU PORTAL WIM Archive

3.2 DATASET FOR MODEL DEVELOPMENT AND TESTING

To conduct the analysis described in the following chapters, a subset of the database was prepared. The data stored in the archive have not been processed for data quality and measurement errors are known to exist. Thus, the objective of the subset procedure was to identify a month of data from a pair of stations with a minimum of sensor error.

To indentify the best performing stations in 2007, records of the most common vehicle (five-axle semi truck (Oregon type 11, FHWA Class 9) were compared for matched transponder-equipped trucks for all station pairs. Previous work (Nichols and Cetin, 2007) has shown that the weight of the steering axle is fairly constant for any truck-loading condition and the spacing between the drive axles (axle 2 and 3) is within a small range (based on manufacture's data). Further, properties of the vehicle such as length, number of axles, and axel spacing should not vary substantially between an upstream and downstream observation of the same vehicle (assuming the transponder is on the same vehicle). Some differences could be expected, such as if a tractor changed trailers or operated a drop axle. The assumption was made that stations with good measurement quality for Class 9 trucks would measure all vehicles with similar accuracy.

This comparison was done graphically. For all station pairs, kernel density plots of the ratio of the upstream measurements to the downstream measurements for four different metrics were created: total truck length, distance in feet between axles 2 and 3 (the tandem drive axles), the total number of axles, and the steering axle weight. If the upstream and downstream sensors are calibrated in exactly the same way, a density plot of the ratio should be tightly distributed around x=1. Samples of these plots are shown for three selected station pairs in Figure 3.2-7 (plots for all stations are in Appendix B). It is clear from the figures that plot C: KFP to LWL has the data with the best upstream-downstream match (the ratio is most tightly distributed around 1). Graphical inspection of similar plots for all stations was used to select Link 234 - Klamath Falls to Lowell (KFP to LWL) as the "best" link. This 145-mile route is mostly a two-lane primary rural highway and consists of US-97 from just north of the California border north to the junction with OR-58, where it heads west over Oregon's Cascade mountains. In addition, Link 231 – Klamath Falls to Bend also exhibited good data quality and was selected for further testing of the matching algorithm. This 125-mile route is also a two-lane primary rural highway (US-97).

To further narrow the subset to one month in 2007, the above metrics and additional variables of interest for the re-identification algorithm (lengths between each axle pair and the weights for each axle) were considered. Upon inspection, there did not appear to be much month-to-month variation for these station pairs; however, October 2007 seemed to show the most consistent agreement between the upstream and downstream detectors. Plots of the metrics for each month are shown in Appendix C for Link 231 and 234.

Records of **all** vehicles for the three stations were used to test and develop the re-identification algorithm(s) described in the subsequent chapters. A total of 25,639 trucks were observed at Klamath Falls, 15,401 at Lowell and 23,609 at Bend in October 2007.

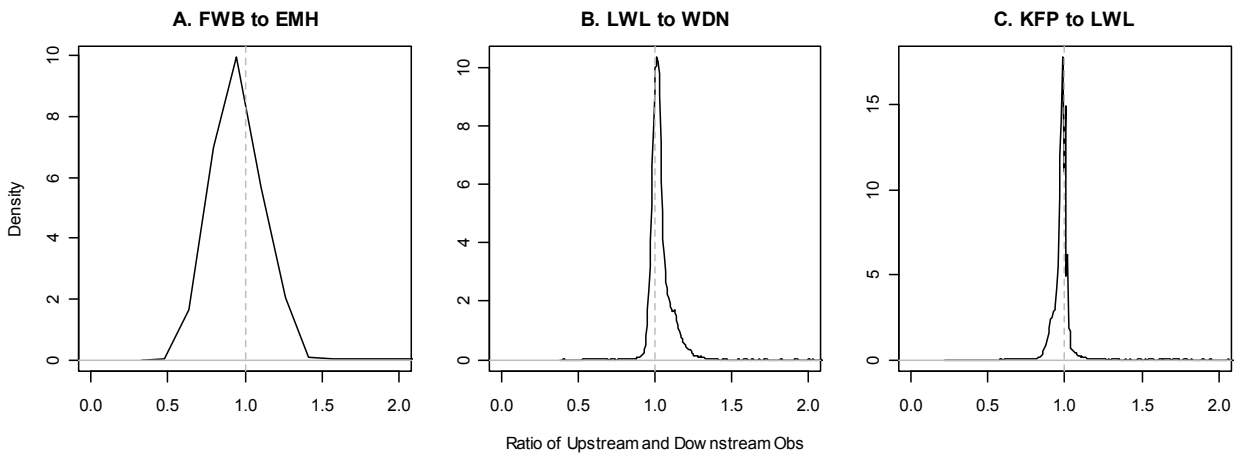


Figure 3.2 Kernel Density Plots of the Ratio of Upstream to Downstream Truck Length, Class 9 Trucks, 2007

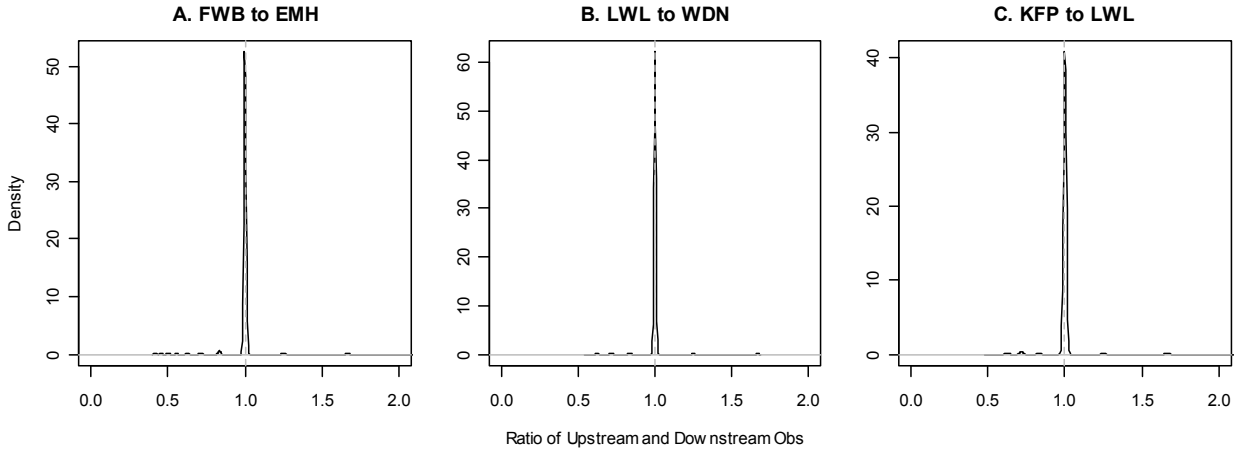


Figure 3.3 Kernel Density Plots of the Ratio of Upstream to Downstream Number of Axles, Class 9 Trucks, 2007

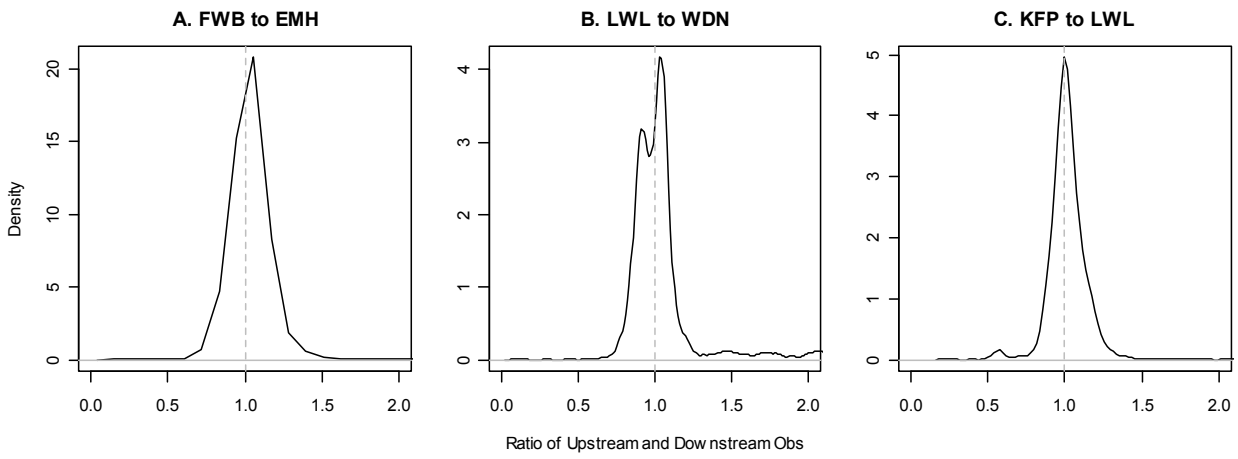


Figure 3.4 Kernel Density Plots of the Ratio of Upstream to Downstream Steering Axle Weight, Class 9 Trucks, 2007

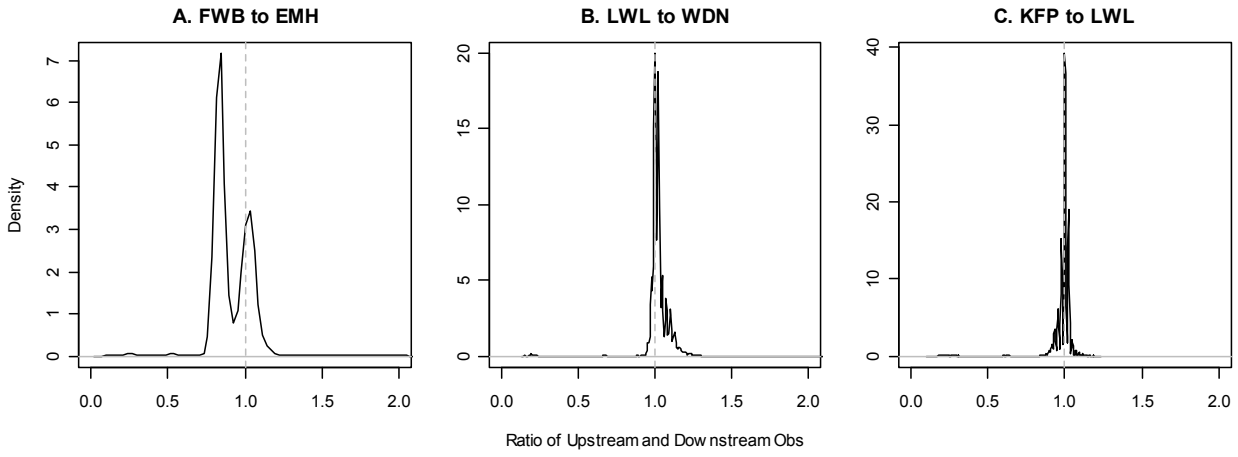


Figure 3.5 Kernel Density Plots of the Ratio of Upstream to Downstream Spacing Between Axle 2-3, Class 9 Trucks, 2007

4.0 RE-IDENTIFICATION ALGORITHMS

The re-identification problem can be described as follows. Given two separate datasets that consist of vehicle-attribute data (such as length, axle spacing, axle weights or some attributes of the magnetic signature), the re-identification algorithms attempt to match the pairs of measurements (one from each dataset) that belong to the same vehicle. These two datasets are collected at some upstream and downstream points in a transportation network. To simplify the discussion an example is given in Figure 4.1, which shows graphically two datasets for four vehicles that cross upstream and downstream stations. Each box represents a vehicle and the attribute data is indicated with horizontal bars. The actual matching is indicated with arrows in Figure 4.1. For both sites, vehicle number 3 only crosses one of the sites.

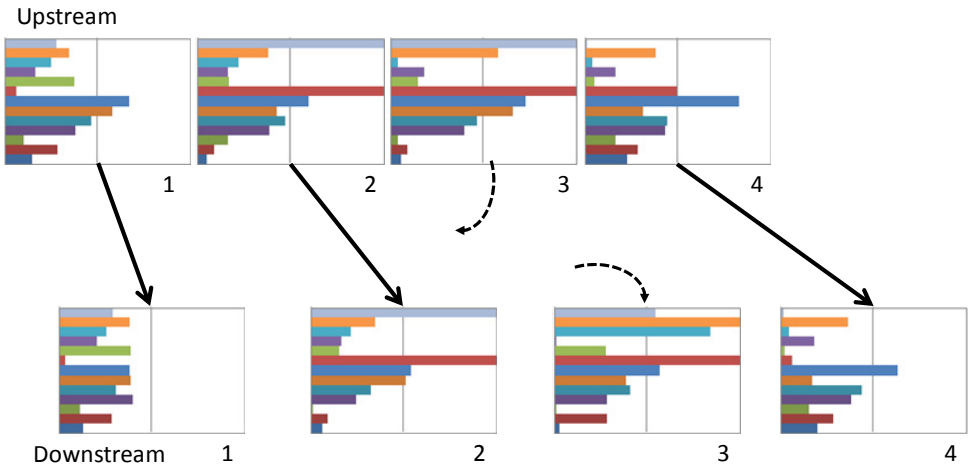


Figure 4.1: All vehicles are correctly matched while there is no match for one vehicle

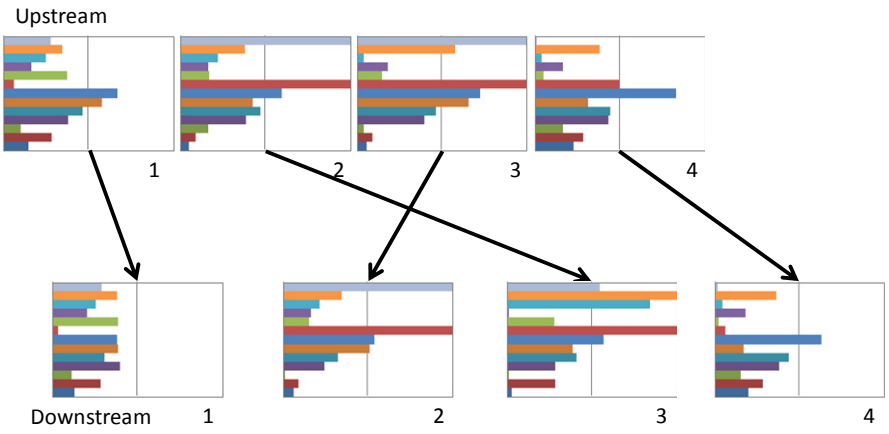


Figure 4.2: Vehicles 2 and 3 are mismatched

Vehicle re-identification algorithms attempt to match each vehicle in the downstream set to a vehicle in the upstream (or vice versa) based on some "similarity" measure that is a function of the attribute data between the two sites. Figure 4.2 shows a potential outcome from a hypothetical algorithm for the same vehicles given in Figure 4.1. In this case, vehicles numbered 1 and 4 are matched accurately, as the similarity measure is maximized for these pairs. On the other hand, vehicles 2 and 3 are mismatched. In reality, for downstream vehicle 3 there is no match at the upstream, but the matching algorithm identifies the upstream vehicle 2 as the best match among the four possibilities. Based on this simple illustration it can be observed that not only a mechanism is needed to identify the best match (in terms of the similarity in attribute data), but also there needs to be a method in place to screen out vehicles that cross one site but not the other.

Therefore, the vehicle re-identification approach developed in this research consists of two main stages. In the first stage, each vehicle from the downstream station is matched to the most "similar" upstream vehicle, as is typically done in vehicle re-identification methods. Both a distance-based method and a Bayesian method are utilized to solve the first-stage problem. These methods essentially capitalize on the variance in vehicle populations and the consistency or correlation of the measurements taken at the upstream and downstream stations. Figure 4.3 and Figure 4.4 show axle weights and axle spacing, respectively, that belong to the same vehicles measured at two stations. As it can be observed, there is high correlation between the measurements taken at these two sites. There is also significant variance in the attribute data due to the fact that physical characteristics of trucks vary significantly.

For the second stage, several methods are developed to screen out vehicles that cross only one site. These methods increase the accuracy of matching but may reduce the total number of vehicles matched. By setting a threshold value, these methods allow the user to trade off accuracy versus the total number of vehicles being matched. These methods involve calculating both the highest and the second highest similarity measures for each vehicle being matched. As demonstrated in this report, the screening approach improves the accuracy of the re-identification methods significantly.

The overall approach taken in model development can be described as follows. The WIM data from a given upstream-downstream station pair are first used to create "link data," which contain attribute data only for those trucks that cross *both* upstream and downstream sites. This is done based on the transponder data as explained before. The link data is then divided into training (about two-thirds of the data) and testing datasets. The training dataset is then utilized for model development. The performance of these models is then evaluated on the test datasets. This process is carried out for AVC data and WIM data separately. AVC data contain only vehicle length and axle spacing, whereas WIM data contain both the AVC data and axle weights. Both datasets have timestamp information.

The next subsections provide a detailed explanation of the algorithms and methods developed for solving the re-identification problem.

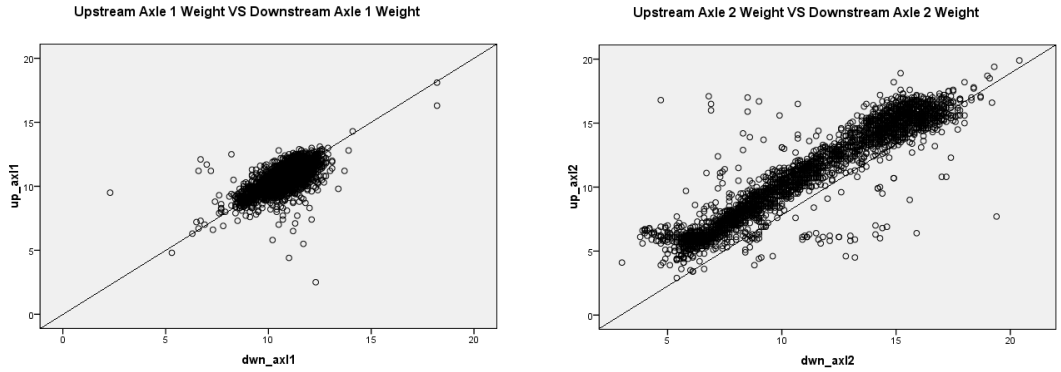


Figure 4.3 Axle 1 and axle 2 weights corresponding to the same trucks measured at upstream and downstream sites

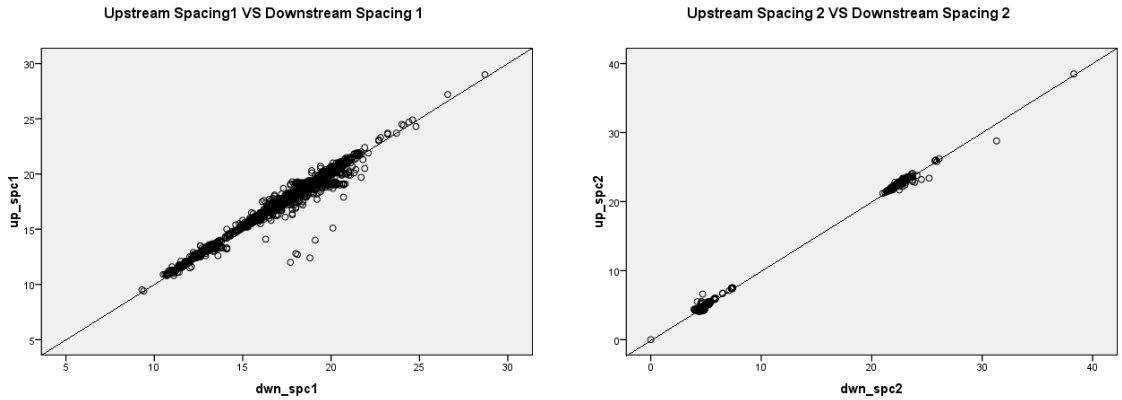


Figure 4.4 Spacing between axles 1 and 2 and 2 and 3 corresponding to the same trucks measured at upstream and downstream sites

4.1 NOTATION AND THE SEARCH SPACE

Let ***U*** and ***D*** be two non-empty sets that denote the vehicles crossing the upstream WIM station and downstream WIM station, respectively. Depending on various factors including the station locations, WIM record validity (i.e., crossed sensors properly), and types of activity between the sensors, four general cases arise:

i) $U \subset D$ and $U \neq D$ (i.e., all vehicles crossing the upstream site also cross downstream site)

ii) $D \subset U$ and $U \neq D$ (i.e., all vehicles crossing the downstream site also cross upstream site)

iii) $U = D$ (i.e., all vehicles cross both sites)

iv) $D \not\subset U, U \not\subset D,$ and $U \cap D \neq 0$ (i.e., not all vehicles in the upstream or downstream cross both sites)

Even though the fundamental re-identification problem is the same in all four cases, the search procedure in the third case is the simplest, as all vehicles cross both sites. In this case, for any selected vehicle there is a match in the other set (i.e., there is a one-to-one mapping between the members of the two sets). One can apply not only statistical matching algorithms but also assignment algorithms to assign all the members in one set to those in the other set while ensuring that each member is assigned only once. This is demonstrated in Cetin and Nichols 2009 and is shown to significantly improve the accuracy of matching vehicles.

The last case above (iv) is somewhat more difficult than the others since one needs to consider the possibility that a vehicle taken from one set might not have a match in the other set. In the first three cases, there is always a match for each vehicle in the smaller set (or in either set for case iii). The methods developed in this research can be used for any one of these four cases as the methods for screening (explained in Section 4.4) can be applied to screen out vehicles that do not cross both sites. Without loss of generality, the methods (of the first stage) will be described for case ii where for each vehicle in *D* a match will be identified in *U*, which has more samples than set *D*. Then, in the second stage, the screening methods will be applied to the results of the first stage to determine which matched vehicles will be kept and which ones should be eliminated. In Chapter 6, the models are applied to datasets that fall into both case ii and case iv.

Let $\boldsymbol{X^U}$ and $\boldsymbol{X^D}$ be two matrices with the same number of columns that denote the data collected at an upstream station and a downstream station, respectively; and $\boldsymbol{X^U}_i$ and $\boldsymbol{X^D}_j$ denote rows of these two matrices that correspond to the measurements (e.g., axle weights) taken for vehicle *i* at the upstream station and for vehicle *j* at the downstream station. Further, assume that the timestamps indicating arrival times of vehicles at each station are given and denoted by $\boldsymbol{t^U}_i$ for the upstream vehicles and $\boldsymbol{t^D}_j$ for the downstream vehicles. Given $\boldsymbol{X^U}$, $\boldsymbol{X^D}$, $\boldsymbol{t^U}_i$ and $\boldsymbol{t^D}_j$ the vehicle-matching problem involves determining $\boldsymbol{X^U}_i$ and $\boldsymbol{X^D}_j$ that are generated by the same vehicle. Let δ_{ij} be a binary variable that equals 1 if $\boldsymbol{X^U}_i$ and $\boldsymbol{X^D}_j$ belong to the same vehicle and equals zero otherwise. The main objective of the matching algorithms is to estimate all δ_{ij}'s with minimum error.

As mentioned before, a two-stage approach is proposed in this research for the re-identification problem. In the first stage, for each vehicle in $\boldsymbol{D}$ a match is found in $\boldsymbol{U}$. This is accomplished by a Bayesian method as explained below. In the second stage, a new method is proposed to screen out mismatched vehicles to improve accuracy. These two stages are explained in detail in the subsequent sections.

For the first stage of re-identification, each vehicle in $\boldsymbol{D}$ needs to be matched to the most similar vehicle in $\boldsymbol{U}$. Since timestamp information is available for each vehicle, a reasonable "search space" from the upstream vehicle records ($\boldsymbol{U}$) can be identified based on travel times. Before the search starts to match a downstream vehicle *j* to an upstream vehicle *i*, a search space for vehicle *j*, denoted by S_j, is determined based on the timestamps at two stations ($\boldsymbol{t^U}_i$ and $\boldsymbol{t^D}_j$) and some defined time window. The variability in travel time can be captured by specifying minimum and maximum values for travel times. The minimum value (*minTime*) can be easily predicted based on an assumed maximum travel speed and the distance between the two stations. The maximum value can exhibit a large variation depending on the individual vehicle speeds, travel distance, and traffic-flow interruptions between the two stations, and any pick-up, delivery, or rest stop the driver may make. The maximum value (*maxTime*) can be taken as a multiple of the minimum

time if no data exists or can be based on observations. The search space for a downstream vehicle *j* is then determined as follows:

$$S_j = \{i \in \boldsymbol{U} \mid \boldsymbol{t}^{\boldsymbol{D}}_j - maxTime \leq \boldsymbol{t}^{\boldsymbol{U}}_i \leq \boldsymbol{t}^{\boldsymbol{D}}_j - minTime\} \quad (1)$$

Depending on the difference between *maxTime* and *minTime* or simply time window, the number of vehicles among which a match can be found varies. Larger time windows will result in a larger number of vehicles in the search space, which can make the matching problem more difficult.

4.2 DISTANCE-BASED METHOD

Perhaps the simplest method to re-identify vehicles involves calculating the Euclidean distance and matching vehicle pairs that give the smallest distance. This method is generally used as a baseline and does not produce very accurate results. The distance-based method entails calculating a weighted Euclidian distance measure as shown below. If N is the number of vehicle attributes (e.g., axle weights and spacing) collected at two WIM stations, the key steps of this method can be described as follows:

For each vehicle *j* in ***D***

Identify a search space (see equation 1), $S_j \subset \boldsymbol{U}$

For each $i \in S_j$

Calculate $d_{ij} = \sum_{k=1}^{N} w_k \left(\left(x_i^k - x_j^k \right) / x_i^k \right)^2$

$i_m = \underset{i}{\operatorname{argmin}}\, d_{ij}$

Match vehicle *j* to i_m, i.e., $\delta_{ij}=1$ if $i=i_m$

The distance-based method is only used as a baseline for assessing the results of the Bayesian method. It is not suggested as a viable method for solving the re-identification problem. The weights (w_k) can be optimized by a multidimensional optimization algorithm. However, based on the analyses performed, matching accuracy does not change significantly as these weights are optimized. Therefore, all weights (w_k) are set to one in calculating the distance (d_{ij}).

4.3 BAYESIAN METHOD

The Bayesian re-identification method relies on calculating the posterior probability of a match between two vehicles given two sets of data points collected for a vehicle pair (i,j) at the upstream and downstream stations. A vehicle *j* at the downstream station is matched to the upstream vehicle *i* that yields the largest probability of a match. The steps of the Bayesian method are formally explained below.

For each vehicle *j* in ***D***

Identify a search space (see equation 1), $S_j \subset \boldsymbol{U}$

For each $i \in S_j$

Calculate $P(\delta_{ij} = 1|\text{data})$

$m = \underset{i}{\text{argmax}}\ P(\delta_{ij} = 1|\text{data})$

Match vehicle j to m, i.e., $\delta_{ij}=1$ if $i=m$

Once a search space is identified, $P(\delta_{ij} = 1 \mid \boldsymbol{x_{ij}})$, the conditional probability that $\boldsymbol{X}^U{}_i$ and $\boldsymbol{X}^D{}_j$ belong to the same vehicle given data (i.e., $\boldsymbol{x_{ij}} = \boldsymbol{x}^U{}_i \cup \boldsymbol{x}^D{}_j$), can be computed by the Bayes theorem as follows:

$$P(\delta_{ij} = 1|x_{ij}) = \frac{f(x_{ij}|\delta_{ij}=1)P(\delta_{ij}=1)}{f(x_{ij}|\delta_{ij}=1)P(\delta_{ij}=1)+f(x_{ij}|\delta_{ij}=0)P(\delta_{ij}=0)} \quad (2)$$

In order to calculate this posterior probability, both the two conditional probability density functions (i.e., $f(x_{ij}|\delta_{ij}=1)$ and $f(x_{ij}|\delta_{ij}=0)$) and the prior probabilities (i.e., $P(\delta_{ij}=0)$ and $P(\delta_{ij}=1)$) are needed. The functions $f(x_{ij}|\delta_{ij}=1)$ and $f(x_{ij}|\delta_{ij}=0)$ are the density functions that characterize the collected data at two stations when it belongs to the same vehicle and different vehicles, respectively. Figure 4.5a-b and Figure 4.5c-d illustrate how the data would distribute for observations when $\delta_{ij}=1$ and $\delta_{ij}=0$, respectively, for a simple case when only a single attribute is considered. As it can be observed from these figures, when vehicles match (i.e., upstream and downstream measurements belong to the same vehicle) there is high correlation between the measurements, which is critical for re-identification to work effectively. On the other hand, when random data for upstream and downstream measurements are plotted the correlation disappears as expected and a roughly uniform distribution of points is observed (Figure 4.5c-d). Since this amounts to an approximately uniform value for the density function, $f(x_{ij}|\delta_{ij}=0)$ in equation (2) can be replaced by some arbitrary constant (α). Furthermore, the travel-time information can be used to approximate the prior distribution $P(\delta_{ij}=1)$, as opposed to assigning a fixed value to the prior. If the probability density function for the travel time is denoted by, $f(t_{ij})$ then, the posterior probability in equation (2) can be simplified to:

$$P(\delta_{ij} = 1|x_{ij}) \sim \frac{f(x_{ij}|\delta_{ij}=1)f(t_{ij})}{f(x_{ij}|\delta_{ij}=1)f(t_{ij})+\alpha} \quad (3)$$

where α is a positive arbitrary constant accounting for $f(x_{ij}|\delta_{ij}=0)$ and $f(\delta_{ij}=0)$. Since in matching vehicles only relative magnitude of this posterior probability is important, the selected value of α is not critical. In this research the simplified version (equation 3) is used which does not require the estimation of $f(x_{ij}|\delta_{ij}=0)$, an advantage in terms of model calibration and development.

In order to use equation 3, two probability distributions, i.e., $f(x_{ij}|\delta_{ij}=1)$ and $f(t_{ij})$, are needed to calculate the posterior probability. These probability density functions are found based on fitting finite mixture models to the training dataset as explained in Chapter 5. Finite mixture modeling is a well-known, semi-parametric technique for fitting a statistical distribution that is a weighted sum of multiple distributions. A mixture model is able to model quite complex distributions and can handle situations where a single parametric family cannot provide a satisfactory model (McLachlan and Peel 2000).

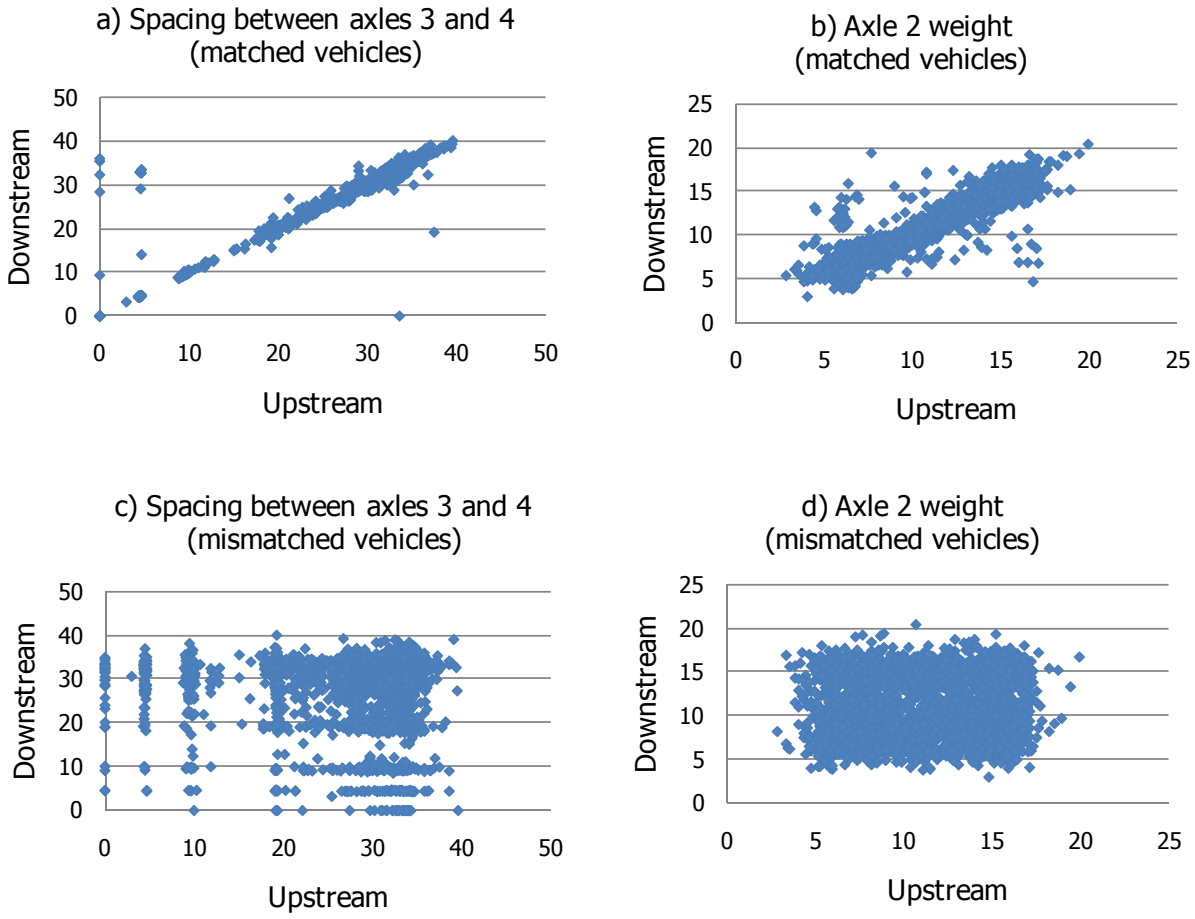


Figure 4.5 Spacing (ft) between axle 3 and 4 (a&c) and weight (kips) of axle 2 (b&d) at two stations for matched and mismatched trucks

4.4 METHODS FOR SCREENING MISMATCHED VEHICLES

When vehicle re-identification methods presented in Sections 4.2 and 4.3 are applied to a dataset, inevitably some vehicles get mismatched, especially when not all vehicles in the upstream or downstream cross both sites (case iv as explained in Section 4.1). In this section, several screening methods are proposed to control the error rate caused by mismatching. The main objective is to screen out vehicles that cross the downstream station but do not appear in the upstream. Even though these screening methods are explained in the context of the Bayesian model, they are equally applicable to any other re-identification algorithm that match vehicles based on an estimated metric such as similarity distance and probability.

The overall idea is to devise a secondary process to improve accuracy by applying a certain rule to the output of the re-identification method to determine whether or not the matched pairs of vehicles will be kept. In other words, for each paired vehicles a test will be performed to classify the matched pair either as a true match or false positive. Those classified as true match will then

constitute the total vehicles that are matched, which obviously will be less than the original number of matched vehicles. Those classified as false positive will not be matched at all. Consequently, the accuracy is improved at the expense of matching fewer vehicles than the original number matched by the re-identification algorithm.

A simple or naïve approach to screening mismatched vehicles would be to impose a threshold on the value of the posterior probability $P(\delta_{mj}=1|\, x_{mj})$ for the matched pairs and retain only those matched vehicle pairs that produce a higher probability than the threshold value. However, this method may result in eliminating too many vehicles to improve the accuracy.

Several other types of methods are proposed in this research. The proposed screening method involves examining not only the posterior probability $P(\delta_{mj}=1|\, x_{ij})$ for the matched vehicle pairs, which is the largest value for a downstream vehicle j being assigned to an upstream vehicle m among all vehicles in the search space, i.e., $i \in S_j$, but also the second largest posterior probability $P(\delta_{kj}=1|\, x_{kj})$, for the vehicle pairs j and k, where $k \in S_j$. The rationale behind this approach is as follows. If the vehicle pairs are truly matched, the difference between $P(\delta_{mj}=1|\, x_{ij})$ and $P(\delta_{kj}=1|\, x_{kj})$ is expected to be much larger as compared to the same difference for mismatched (false positive) vehicles. Because, for false positives both the largest and the second largest probabilities, i.e., $P(\delta_{mj}=1|\, x_{ij})$ and $P(\delta_{kj}=1|\, x_{kj})$, effectively measure the same thing (they are both for mismatched pairs of vehicles) and consequently are expected to have similar values . On the other hand, for truly matched vehicles the gap between $P(\delta_{mj}=1|\, x_{ij})$ and $P(\delta_{kj}=1|\, x_{kj})$ is expected to be significantly larger as each probability measures a different scenario. This is illustrated in Figure 4.6a and Figure 4.6b.

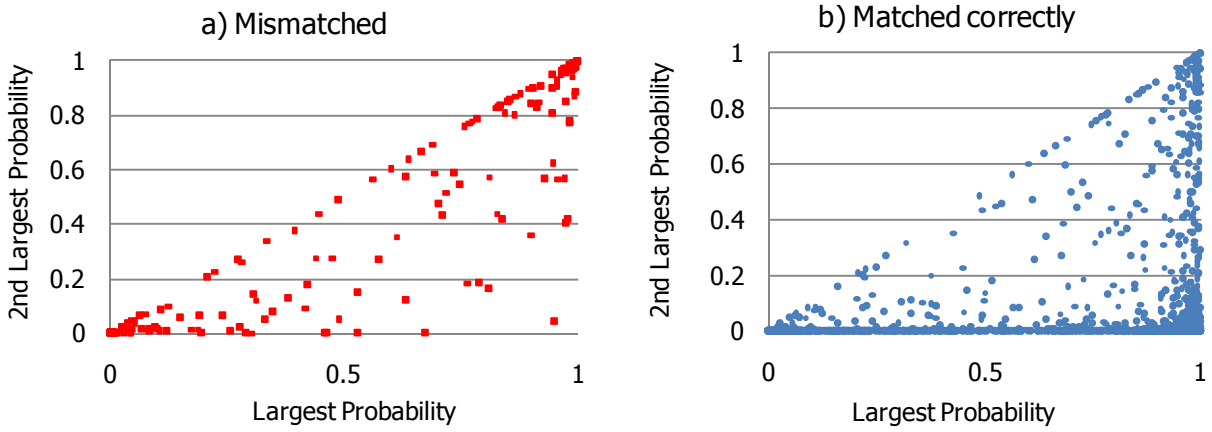


Figure 4.6 Distribution of largest and second largest probabilities from the Bayesian Model when vehicles are mismatched (a) and matched accurately (b) for the WIM scenario

To utilize both the largest and the second largest posterior probabilities (i.e., $P(\delta_{mj}=1|\, x_{ij})$ and $P(\delta_{kj}=1|\, x_{kj})$, which are denoted by P_1 and P_2, respectively, for the sake of simplification in screening mismatched vehicles three different criteria are proposed and tested. The objective of these criteria is to classify the points in Figure 4.6-a as false positive while trying to keep as many points on Figure 4.6-b as true matches.

These three criteria are explained below:

- **45^o Line**: As it can be observed in Figure 4.6-a, a significant portion of the observations for mismatched vehicles are clustered around a 45^o line that goes through the origin. Therefore, the following criterion seems to be a plausible option:

$$\begin{cases} \text{If } (P_1 - P_2) > \Delta \text{ then classify as "true match"} \\ \text{else classify as "mismatched"} \end{cases}$$

where Δ is a user-specified threshold value.

- **Ratio**: As discussed above, it is expected that P_1 and P_2 would have similar values when the vehicles are not matched correctly and the opposite would be true otherwise. For a defined threshold value Δ the criterion for ratio method is as follows:

$$\begin{cases} \text{If } (P_1 - P_2)/P_1 > \Delta \text{ then classify as "true match"} \\ \text{else classify as "mismatched"} \end{cases}$$

- **Mixture Model**: For this method, mixture models are fitted to the data of Figure 4.6-a and Figure 4.6-b separately to obtain bivariate density functions for the distribution of P_1 and P_2 in these two distinct cases. The probability distribution fitted to the data in Figure 4.6-b is denoted as pdf_1 and the one fitted to the data in Figure 4.6-a as pdf_0. The criterion for this method is as follows:

$$\begin{cases} \text{If } (pdf_1)/(pdf_1 + pdf_0) > \Delta \text{ then classify as "true match"} \\ \text{else classify as "mismatched"} \end{cases}$$

The application of these three criteria to the dataset and their performance are explained in Chapter 6.0. The next chapter presents the finite mixture modeling technique that is used to estimate the necessary probability density functions of the Bayesian model.

5.0 FINITE MIXTURE MODELS

The two probability distributions (i.e., $f(x_{ij}|\delta_{ij}=1)$ and $f(t_{ij})$) needed to calculate the posterior probability specified in equation 3 are found based on fitting statistical finite mixture models to the training datasets. Finite mixture modeling is a well-known, semi-parametric technique for fitting a statistical distribution that is a weighted sum of multiple distributions. A mixture model is able to model quite complex distributions and can handle situations where a single parametric family cannot provide a satisfactory model (McLachlan and Peel 2000). Given a random dataset with an unknown distribution, finite mixture models provide a flexible framework to estimate probability density function for each of the components composing the model. Because of their usefulness as an extremely flexible method of modeling, finite mixture models have gained attention from many disciplines including astronomy, biology, genetics, medicine, psychiatry, economics, engineering, and marketing. Besides their direct application in data analysis and providing descriptive models for probability distributions, other applications of mixture models in these disciplines include image analysis, spectral analysis, cluster and latent class analysis, discriminant analysis, and survival analysis.

In mixture modeling, the unknown density of a multivariate random vector $\boldsymbol{Y}$, $f(\boldsymbol{y})$, is assumed to be written in the form

$$f(y) = \sum_{i=1}^{g} \pi_i f_i(y), \tag{4}$$

where the $f_i(y)$'s are component densities and π_i's are nonnegative numbers that sum to one. The quantities $\pi_1, \ldots, \pi_g$ are called the mixing proportions. Even though there are various flavors of mixture modeling, mixture models with normal components, where each $f_i(y)$ is a (multivariate) normal density, are quite common and have many applications. An in-depth account of mixture modeling can be found in McLachlan and Peel 2000. In this research, mixture models with normal components are fitted to the data as any continuous distribution can be approximated arbitrarily well by a finite mixture of normal densities. In order to fit a g-component mixture model, the parameters of the normal densities (mean and covariance) and the mixing proportions need to be estimated. The number of parameters to be estimated depends on the number of components (g) and the dimension or size of the random vector $\boldsymbol{Y}$. For example, if a two-component normal mixture model is fit to a multivariate dataset with three dimensions, then there are three means and six covariates per component and one mixing proportion (π) to be estimated – for a total of 19 parameters. The estimation of these model parameters is conveniently achieved through the use of the well-known expectation maximization (EM) algorithm (Dempster et al. 1977), which is explained briefly in the next subsection.

5.1 EM ALGORITHM

The Expectation Maximization (EM) algorithm is a tool for simplifying and solving complex maximum likelihood problems, such as those encountered in mixture models (Trevor et al.

2001). The maximum likelihood function is used for the estimation of component parameters that make up the mixture model. The EM algorithm is explained below for a simple case that involves a two-component mixture model for a univariate dataset.

Let the mixture model consist of two components, each having a Gaussian univariate distribution, denoted by random variables Y_1 and Y_2. Then, each component will have two parameters that need to be estimated, the mean and the variance:

$$Y_1 \sim N(\mu_1, \sigma_1^2) \quad (1)$$
$$Y_2 \sim N(\mu_2, \sigma_2^2) \quad (2)$$

Each point in the dataset comes either from Y_1 or Y_2. Thus a random variable Y can be written as:

$$Y = (1 - \alpha)Y_1 + \alpha Y_2 \quad (3)$$

Where α is either 0 with probability 1-π, or 1 with probability π. Thus the density function f(y) of random variable Y can be written as the sum of component normal densities $\varphi_i(y)$ with parameters μ_i, σ_i^2

$$\varphi(y) = (1 - \pi)\varphi_1(y) + \pi\varphi_2(y) \quad (4)$$

Where

$$0 \leq \pi \leq 1 \quad (5)$$

The general representation of the unknown density function $\varphi(y)$ for the univariate random variable Y can be written as:

$$\varphi(y) = \sum_{k=1}^{g} \pi_k \varphi_k(y) \quad (6)$$

Where $\varphi_k(y)$ represents component densities which are normal, and π_ks are the mixing proportions which are non-negative numbers that add up to one.

$$\pi_k \geq 0 \; for \; k = 1, \cdots, g \quad (7)$$
$$\sum_{i=1}^{g} \pi_i = 1 \quad (8)$$

The density function given in equation 6 is referred to as a *g-component* finite mixture density. Let $y_1, y_2, \ldots, y_n$ be the observed variables; in other words, each point of the given data. Based on such data, parameters of the normal densities and the mixing proportions need to be estimated for specifying the *g*-component mixture model. In other words, to fit a g-component mixture model, the mean μ_k, variance σ_k^2, and the mixing proportion π_k need to be estimated for each component. In order to fit the density function model to the given data, the method that is generally used is the maximum likelihood estimation.
The parameters that need to be estimated for the example problem are:

$$\theta = (\pi, \theta_1, \theta_2) = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) \quad (9)$$

The maximum likelihood function based on *n* observations is:

$$L(\theta; \boldsymbol{Z}) = \prod_{i=1}^{n} \varphi(y_i) \tag{10}$$

where $y_i$represents the i^{th} observation, and Z is the set of given data.
Equation 10 can be expanded as:

$$L(\theta; \boldsymbol{Z}) = \prod_{i=1}^{n} [(1-\pi)\varphi_1(y_i) + \pi\varphi_2(y_i)] \tag{11}$$

Because of the complexity of solving the likelihood function in a multiplication form, the logarithms of both sides are taken. This changes the multiplication into a summation.
The log-likelihood function based on *n* number of data points becomes:

$$L(\theta; \boldsymbol{Z}) = \sum_{i=1}^{n} \log [(1-\pi)\varphi_1(y_i) + \pi\varphi_2(y_i)] \tag{12}$$

This likelihood function is still difficult to solve because of the sum of the terms inside the logarithm. However, by the help of unobserved latent variables α_i taking values 0 or 1 as mentioned in equation 3, it can be said that if α_i is equal to 1 then Y_i comes from model 2, otherwise Y_i comes from model 1. With the assumption that the values of α_i are known, the log-likelihood would be:

$$L(\theta; \boldsymbol{Z}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} [(1-\alpha_i) \log \varphi_1(y_i) + \alpha_i \log \varphi_2(y_i)] + \sum_{i=1}^{n} [(1-\alpha_i) \log \pi + \alpha_i \log(1-\pi)] \tag{13}$$

The maximum likelihood estimates of (μ_1, σ_1^2) are the sample mean and variance of those data with $\alpha_i = 0$, and likelihood estimates (μ_2, σ_2^2) are the sample mean and variance of those data with $\alpha_i = 1$.
In reality, the latent variables are unknown. In order to find these variables as well as the component parameters, an iterative approach is taken in the EM algorithm.

The first step is to make initial guesses about the parameters $\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$. Generally, when making the initial guesses the mixing proportions are given equal probability for each component; in this case, 0.5. For the estimates of the means, random variable y_is are taken randomly in two groups and their average is taken. The estimates for the variances can be assumed equal, and can be set to the overall variance which is

$$\hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n} \tag{14}$$

The second step or the expectation step (E-step), involves computing the expected value ε_i of α_i which is also called as the responsibility of model 2 for observation i. The initial or best estimates of the parameters $(\hat{\pi}, \hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2)$ are going to be used for computing ε_i.

$$\hat{\varepsilon}_i = \frac{\hat{\pi}\varphi_2(y_i)}{(1-\hat{\pi})\varphi_{\hat{\theta}_1}(y_i) + \hat{\pi}\varphi_{\hat{\theta}_2}(y_i)} \; for \; i = 1,2,\dots,n \tag{15}$$

Where $\varphi_{\hat{\theta}_k}$ is the normal probability density function with the estimated parameters for component k.
The responsibilities obtained in equation 15 are then used in the third step, or the maximization step (M-step), for computing the estimates of the parameters. The weighted maximum likelihood fits for means, variances, and mixing probability are computed as follows:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n}(1-\hat{\varepsilon}_i)\, y_i}{\sum_{i=1}^{n}(1-\hat{\varepsilon}_i)}$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i\, y_i}{\sum_{i=1}^{n}\hat{\varepsilon}_i}$$

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{n}(1-\hat{\varepsilon}_i)(y_i-\hat{\mu}_1)^2}{\sum_{i=1}^{n}(1-\hat{\varepsilon}_i)}$$

$$\hat{\sigma}_2^2 = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i(y_i-\hat{\mu}_1)^2}{\sum_{i=1}^{n}\hat{\varepsilon}_i}$$

$$\hat{\pi} = \sum_{i=1}^{n}\frac{\hat{\varepsilon}_i}{n}$$

The expectation and maximization steps are iterated until convergence. These computations in each step can be extended to include more than two components. Also, each component may be multidimensional consisting of multivariate data.

6.0 APPLICATIONS OF THE RE-IDENTIFICATION ALGORITHMS

The vehicle re-identification methods described in Chapter 3.0 are applied to selected archived vehicle datasets from WIM sites in Oregon. The locations of the selected sites and other relevant details are shown in Figure 6.1. This diagram shows three WIM sites (e.g., Klamath Falls Port of Entry (KFP), Bend Weigh Station (BND), and Lowell Weigh Station (LWL)), and the total number of transponder-equipped trucks that crossed these stations in October 2007. Based on the unique transponder numbers, the total flows between the sites are indicated on the diagram by arrows. Link 234 shows the flow of trucks from KFP to LWL whereas Link 231 represents the trucks that cross both KFP and BND. Other transponder-equipped trucks that crossed only one of the sites are indicated by dotted arrows. Table 6.1 shows the total number of all trucks (including those without transponders) that crossed all three sites.

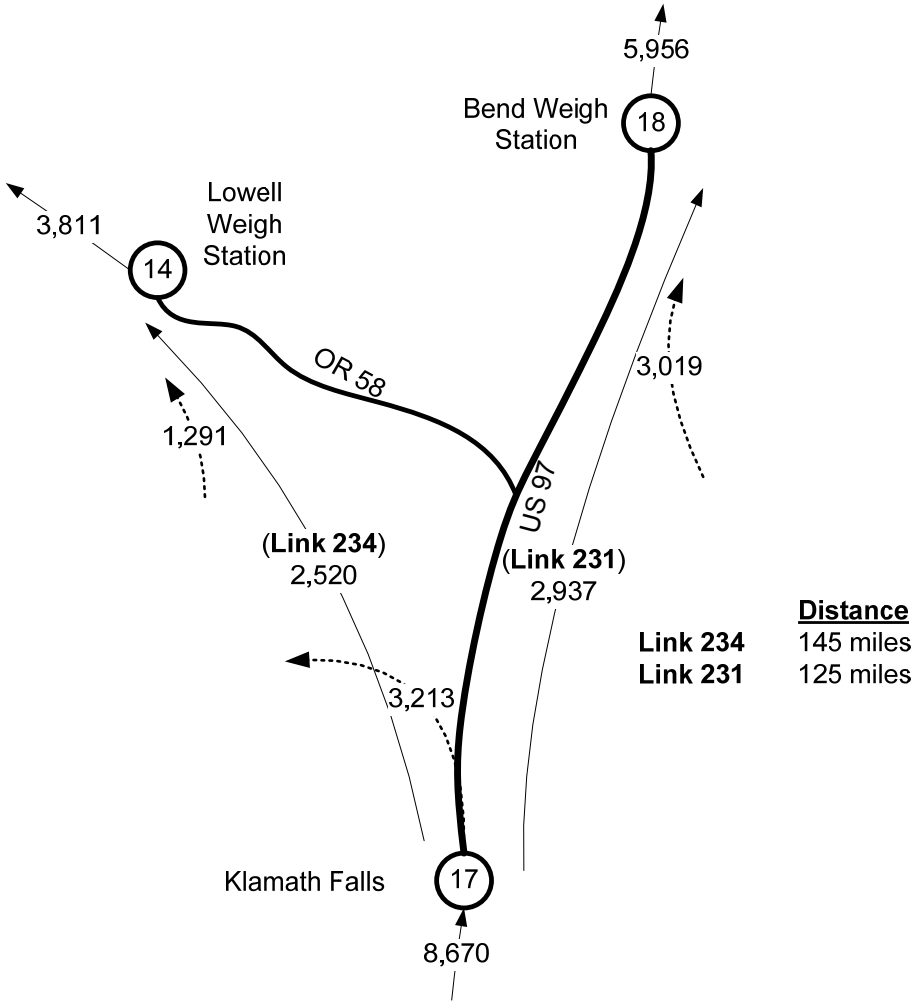


Figure 6.1 Link 231 and Link 234 and the number of trucks with transponders crossing these sites in October 2007

The re-identification methods are first applied to Link 234 data and then to Link 231 data. In each case, the total link data are split into training and testing datasets as explained before. The parameters needed for mixture models are estimated based on the training dataset. Application of the re-identification models involves matching downstream vehicles to upstream vehicles. Therefore, two datasets need to be prepared as inputs: one for the downstream and one for the upstream station.

Table 6.1 Number of trucks observed at three stations, October 2007 data

Trucks	Node 17 (KFP)	Node 14 (LWL)	Node 18 (BND)
With Transponders	8,670	3,811	5,956
Without Transponders	16,969	11,590	17,653
Total	25,639	15,401	23,609
% With Transponders	34%	25%	25%
% Class 9 (among all trucks)	67%	58%	48%

In testing the models, for Link 234 two scenarios are considered in creating the datasets for the downstream station. In the *first scenario*, only those *common* trucks that cross both LWL and KFP stations are selected as the downstream data. In other words, the downstream data only consists of a subset of the 2,520 trucks that constitute the Link 234 data. Even though this scenario may not be realistic, this simpler case is tested to see how the algorithms will perform when it is known that for every downstream vehicle there is a match in the upstream. In addition, depending on the proximity of the WIM stations and the transportation network, this first scenario may well be applicable. In the *second scenario*, an "open system" is considered where the downstream station includes both common trucks and those that cross only the downstream site. In other words, the test data for the downstream is a subset of the combined data of both link data (2,520 trucks) and those that enter at some midpoint (1,291). For Link 234, only the second scenario (open system) is analyzed. In all cases, the upstream dataset consists of *all* trucks observed at the upstream point (KFP), including those that *do not* carry transponders.

As mentioned in Chapter 3.0, the performance of the re-identification models is evaluated for AVC data and WIM data separately. AVC data contain only vehicle length and axle spacing whereas WIM data contain both the AVC data and axle weights. The results for Link 234 and Link 231 are presented below.

6.1 APPLICATION OF THE METHODS TO LINK 234 DATA

Applying the Bayesian re-identification model involves two key steps: model training and model testing. In model training, both the conditional density function, $f(x_{ij}|\delta_{ij}=1)$, and the probability distribution for travel time, $f(t_{ij})$, are obtained by fitting mixture models to a training dataset which consists of common vehicles that are correctly matched based on the tag numbers. The estimated probability density function for travel time, which has three components, is shown in Figure 6.2. Since $f(x_{ij}|\delta_{ij}=1)$ is multidimensional it cannot be drawn. Since there are two cases to be analyzed, one for AVC and one for WIM data, two different conditional density functions, $f(x_{ij}|\delta_{ij}=1)$, are needed. For the AVC data, the total vehicle length and four axle spacings (e.g., axle spacings 1-2, 2-3, 3-4, and 4-5) are used as the attribute data, which results in a five-dimensional density function for $f(x_{ij}|\delta_{ij}=1)$. For the WIM data, in addition to the five-vehicle

attributes considered in AVC case, five axle weights (e.g., axles 1 to 5) are also used as the attribute data, which results in a 10-dimensional density function for $f(x_{ij}|\delta_{ij}=1)$. These mixture models are estimated with special software called EMMIX (McLachlan and Peel 2000). Only five axles are considered in fitting the mixture distributions since the predominant vehicle type is FHWA Class 9 at both stations.

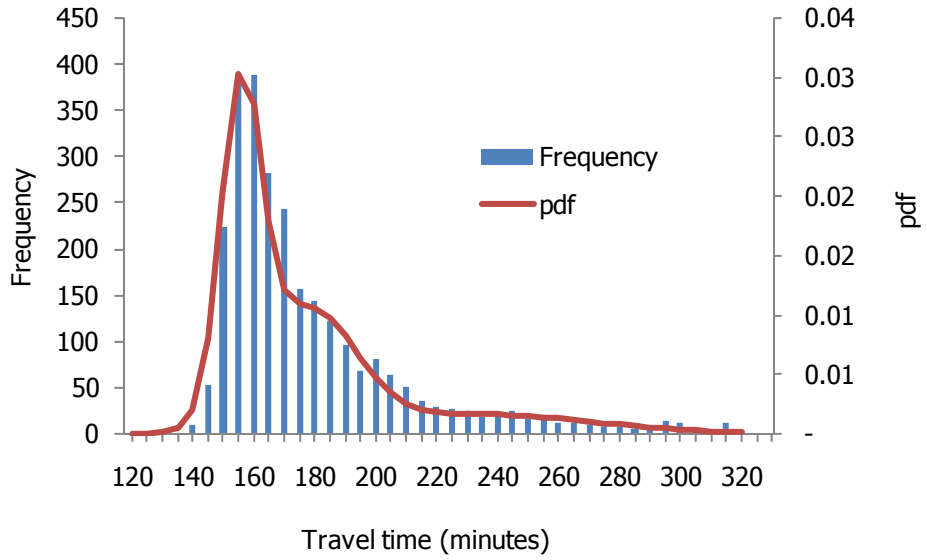


Figure 6.2 Travel-time histogram for Link 234 and a probability density function (pdf) fit by mixture distributions

The variability of travel time between the two stations plays a critical role in vehicle re-identification as the potential matches are usually identified by considering time windows for the travel time. Larger variation results in considering more samples as potential matches, which makes the problem more difficult. Figure 6.2 show the histogram of travel times for the vehicles in the training sample and a probability distribution fitted based on mixture models. The minimum travel time between the stations at the 55 mph speed limit would be about 158 minutes. Based on this figure, a travel-time window between 120 and 316 minutes is used to identify potential matches for a vehicle. Based on these values, on average about 155 vehicles need to be considered as potential candidates in finding a match. In other words, on average, there are about 155 trucks observed in the upstream within an interval of 196 minutes (316-120 = 196).

6.1.1 Testing Scenario 1: Using Only Common Trucks that Cross Both Sites

In the testing step, the models estimated based on the training dataset are applied to the testing data. Training and testing datasets are mutually exclusive. For scenario 1, 1,000 common trucks that cross both KFP and LWL stations are selected. The attribute data collected at the LWL station for these 1,000 vehicles constitute the downstream data. The upstream data, as mentioned before, includes all trucks that cross KFP station.

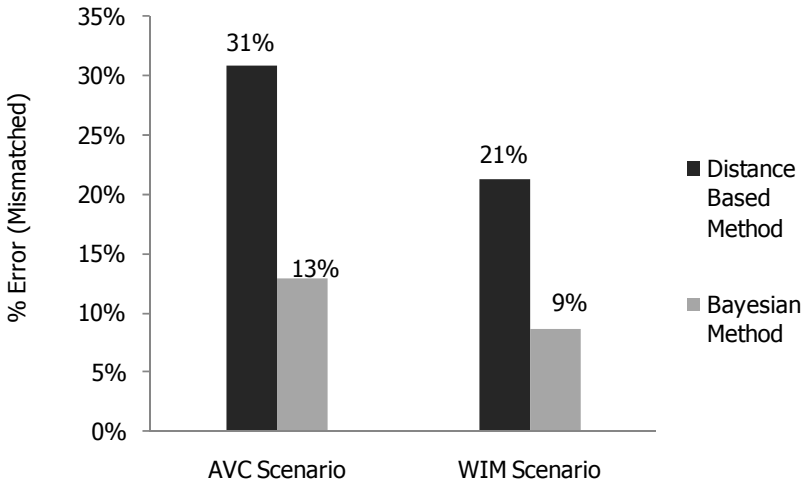


Figure 6.3 Accuracies of the two-vehicle re-identification algorithms

Figure 6.3 shows the accuracy of the Bayesian model and the distance-based model for both the AVC and WIM scenarios. As mentioned before, the distance-based method is only used for comparison purposes since its limitations are clear. As it can be observed, the Bayesian method performs very well in both scenarios and match vehicles with 91% and 87% accuracy for the WIM and AVC scenarios, respectively.

Figure 6.4 shows the results of the screening methods described previously in Chapter 4.4 when applied to the output of the Bayesian method for the WIM scenario. For the WIM scenario, when for all 1,000 vehicles a match is found by the Bayesian model, 86 of them turn out to be mismatched (hence the 9% error rate). The purpose of the screening process is to minimize the total number of mismatched vehicles by omitting some of the vehicles (in particular the mismatched ones) as unmatched or by classifying them as mismatched as explained in Section 4.4. The criteria or rules to do so are also explained previously.

Figure 6.4-a shows the result for the simple or naïve approach to screening mismatched vehicles. In this case, the matched pairs of vehicles are being classified as "true match" and "mismatch" based on a single variable (P_1). As it is evident, this method eliminates a significant number of vehicles from being matched to reduce the percent error. In these plots, the percent error is calculated by dividing the mismatched vehicles remaining after the screening step by the total vehicles matched, which varies depending on the selected threshold. This total number of vehicles matched is shown on the secondary vertical axes of the plots in Figure 6.4.

The results for three methods described above (i.e., 45^o Line, Ratio, and Mixture Model) are presented in parts b-d of Figure 6.4. Among the four methods, the Ratio and Mixture Model methods clearly perform better. To easily compare the performance of these four methods, tradeoff curves are constructed as shown in Figure 6.5-a. Since there are two conflicting objectives (i.e., minimizing percentage error and maximizing the total vehicles matched), these

tradeoff curves provide a convenient way to visualize which method is superior in achieving both objectives simultaneously. From this figure, it is apparent that the Mixture Model method dominates other options across all threshold values. However, the Ratio method closely follows the Mixture Model method, and may be the preferred option since it is simpler to implement.

A similar analysis is also carried out for the AVC scenarios to reduce the 129 (13% of 1,000 vehicles) mismatched vehicles. For brevity, only the tradeoff plots of the four screening criteria are presented, as shown in Figure 6.5-b. In this case, both Mixture Model and the Ratio methods provide very similar results. Based on these analyses, it seems that the Ratio method would be a good criterion for screening mismatched vehicles as it has good performance and is simpler to implement.

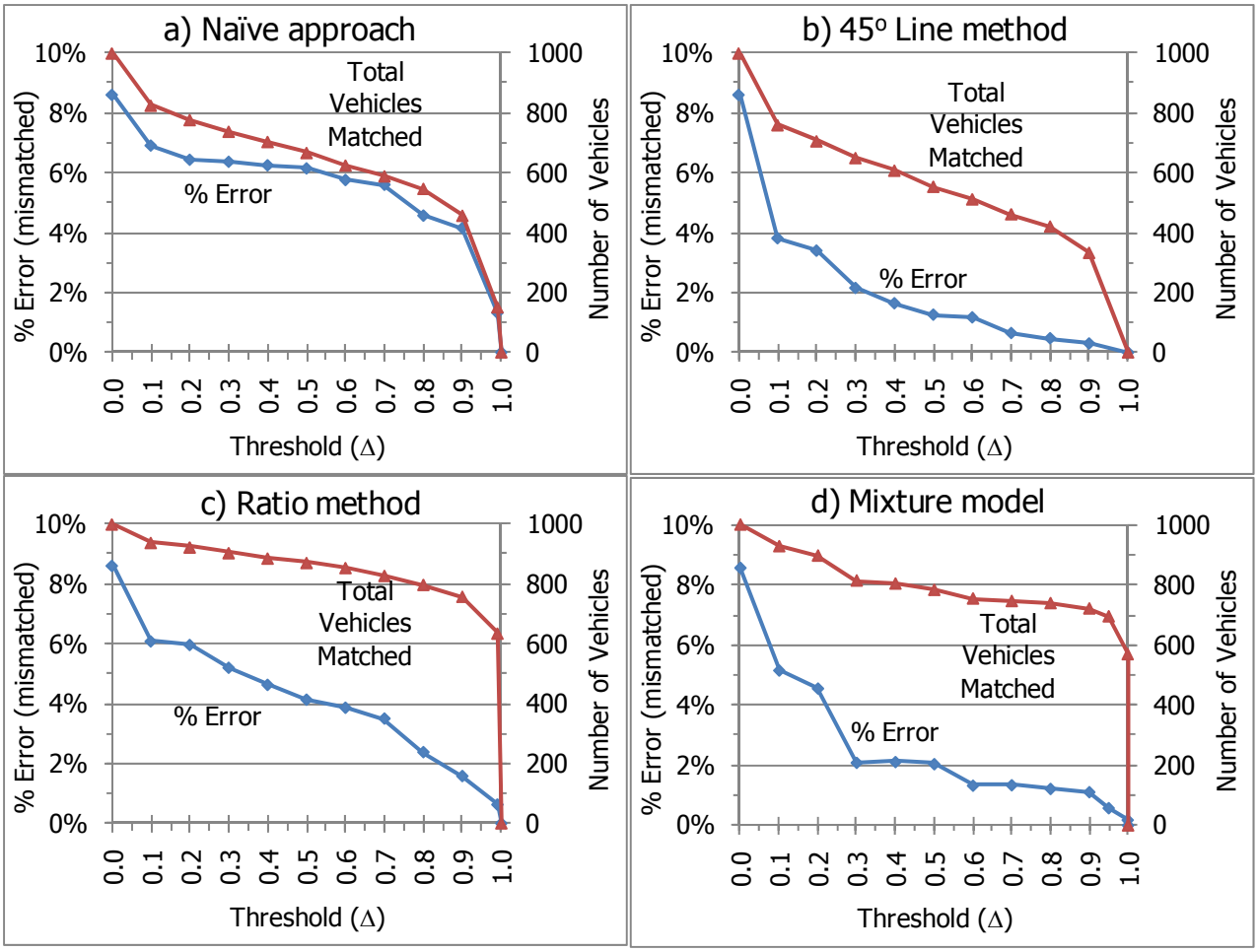


Figure 6.4 Change in error and total vehicles matched for the WIM scenario as the threshold varies for four screening criteria: (a) naïve approach; (b) 45° line; (c) ratio; and (d) mixture model.

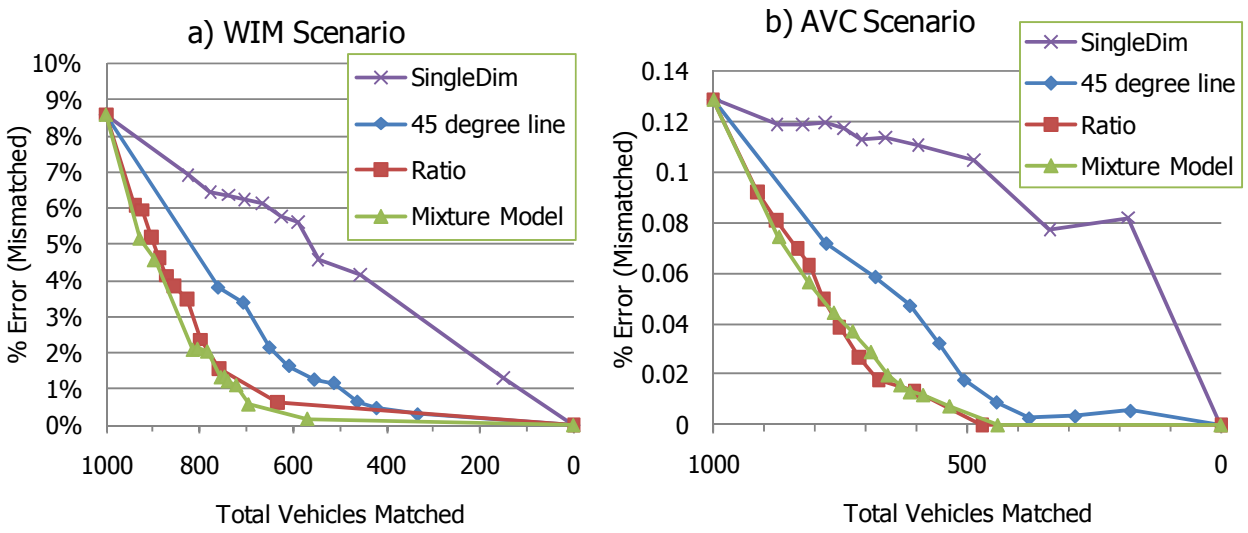


Figure 6.5 Tradeoff curves of the four screening criteria for the WIM scenario (a) and for the AVC scenario (b)

6.1.2 Testing Scenario 2: Open System

For this scenario, those trucks that cross LWL but not KFP are also included in the downstream dataset. In particular, the testing dataset for the downstream now includes a total of 1,400 trucks, out of which 405 cross only LWL station and 995 (common trucks) cross both sites. Consequently, without applying the screening methods there would be a significant percentage of error if only the Bayesian method is applied to find a best match for every downstream vehicle in the upstream station. Even if all *common* trucks are matched accurately, still there would be 29% error (29% = 405/1,400) since 405 trucks never crossed the upstream station. Therefore, the screening methods play a more critical role in accurately matching vehicles in open systems.

The results for this scenario are graphically summarized in Figure 6.6 for the AVC data, and Figure 6.7 and Figure 6.8 for the WIM data. Figure 6.8 is the same as Figure 6.7 except the axes are scaled to provide more detailed information. The vertical axis in all figures shows the percent error, which is calculated simply by dividing the total number of vehicles mismatched to the total number of vehicles matched (shown in the horizontal axis). In addition, Table 9.1 and Table 9.2 in Appendix A provide the actual numbers of vehicles that are matched or mismatched at each threshold value in each screening method.

First, it can be observed that the re-identification methods give more accurate results when axle weights are used in addition to the axle spacing since the results for the WIM data are much better as compared to those of the AVC data. This result is expected as more variables contain additional information that can be used to distinguish between vehicles.

Second, in terms of the four screening methods, the 45^{o} line method performs better than others in both cases as this method more effectively reduces the mismatch error. Even though the ratio

method was performing very well in Scenario 1, it is not effective when there are many vehicles that need to be screened out, as in the case of this open system.

Third, when the WIM data is used for re-identification, vehicles can be matched effectively with a reasonable level of error (8-10%) while the total number of vehicles being matched is kept reasonably close to the actual number of vehicles crossing both the upstream and downstream sites. In this case, there are 995 trucks that cross both sites – which is indicated by the vertical dashed line in the figures. For example, by selecting 0.01 for the threshold value in the 45° line method, 92%, or 793 trucks, are matched accurately out of 866 trucks (see Table 9.2 in Appendix A), which is about 87% of the 995 common trucks that cross both sites. In this example, of the 73 that are mismatched (866-793=73), 43 do not appear in the upstream station and the remaining 30 cross the upstream site but are not matched accurately. In other words, the 45° line method screened out 362 of the 405 (which do not cross the upstream site) but not the 43 that ended up being matched to the upstream vehicles.

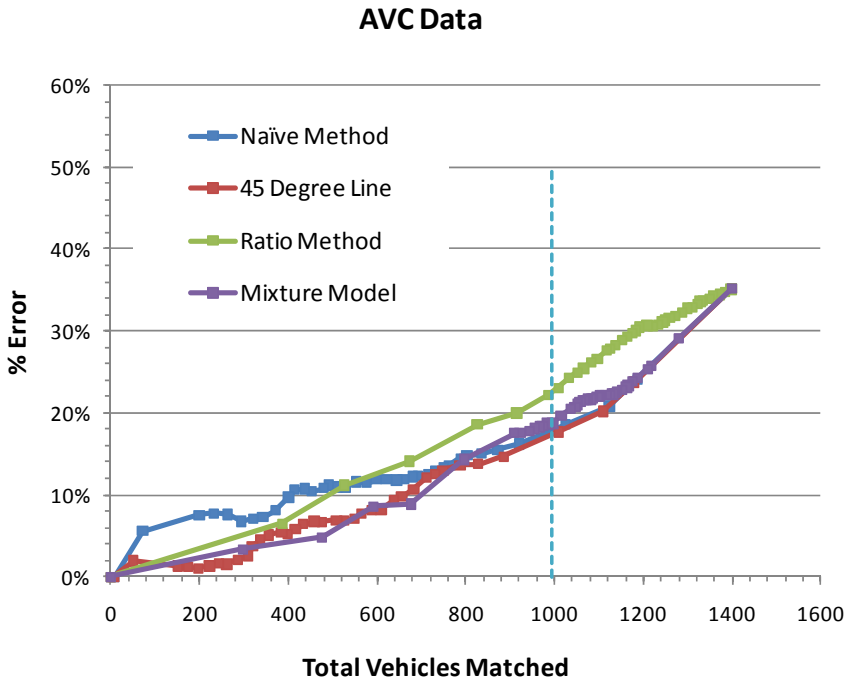


Figure 6.6 Tradeoff curves of the four screening criteria for the AVC scenario for Link 234

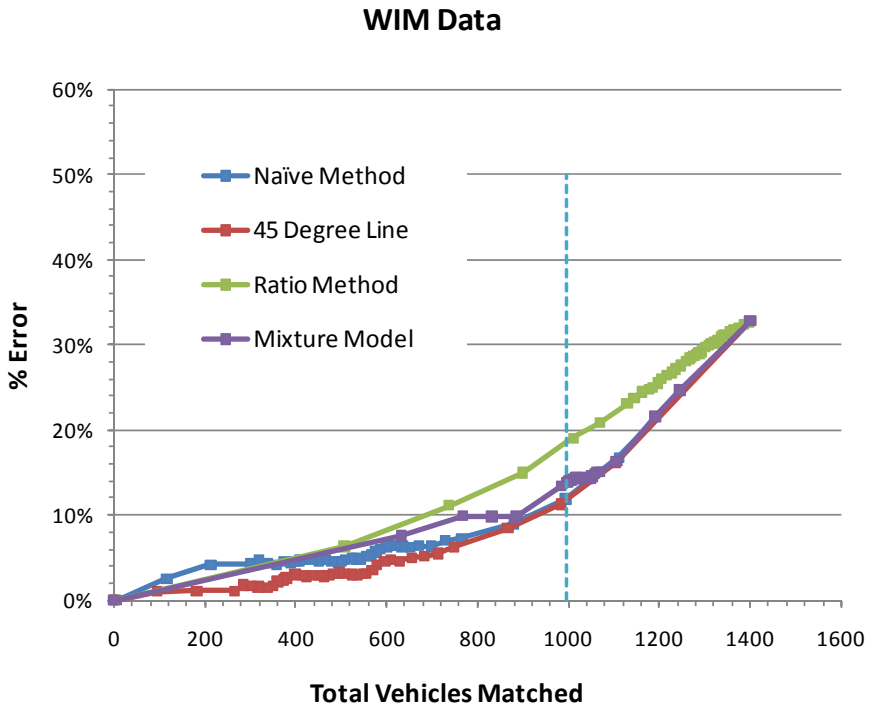


Figure 6.7 Tradeoff curves of the four screening criteria for the WIM scenario for Link 234

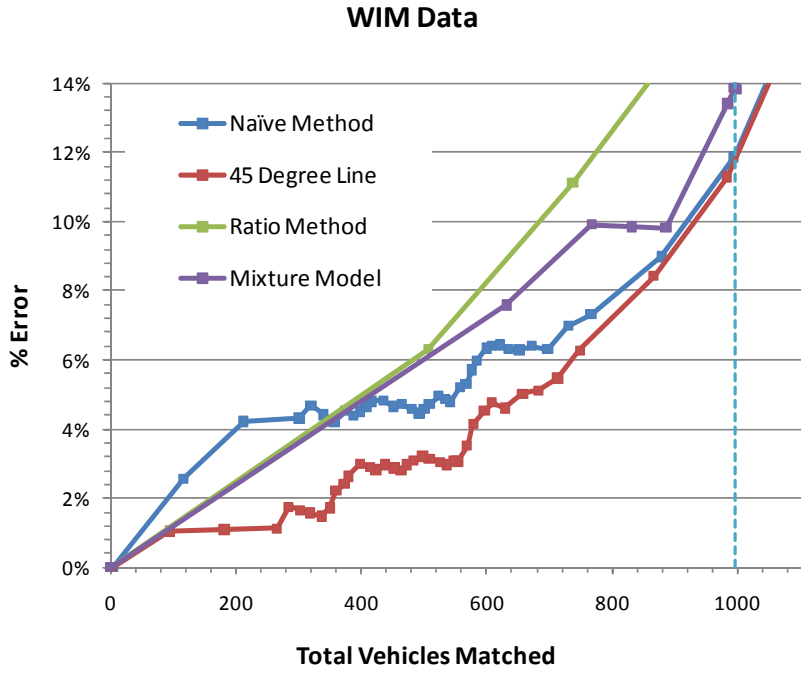


Figure 6.8 Tradeoff curves of the four screening criteria for the WIM scenario for Link 234

6.2 APPLICATION OF THE METHODS TO LINK 231 DATA

Figure 6.9 show the histogram of travel times for the vehicles in the training sample and a probability distribution fitted based on mixture models. The minimum travel time between the stations at the 55 mph speed limit would be about 158 minutes. Based on actual travel times, a travel-time window between 146 and 273 minutes is used to identify potential matches for a vehicle. Based on these values, on average, about 107 vehicles need to be considered as potential candidates in finding a match. In other words, on average, there are about 107 trucks observed in the upstream within an interval of 127 minutes (273-146 = 127).

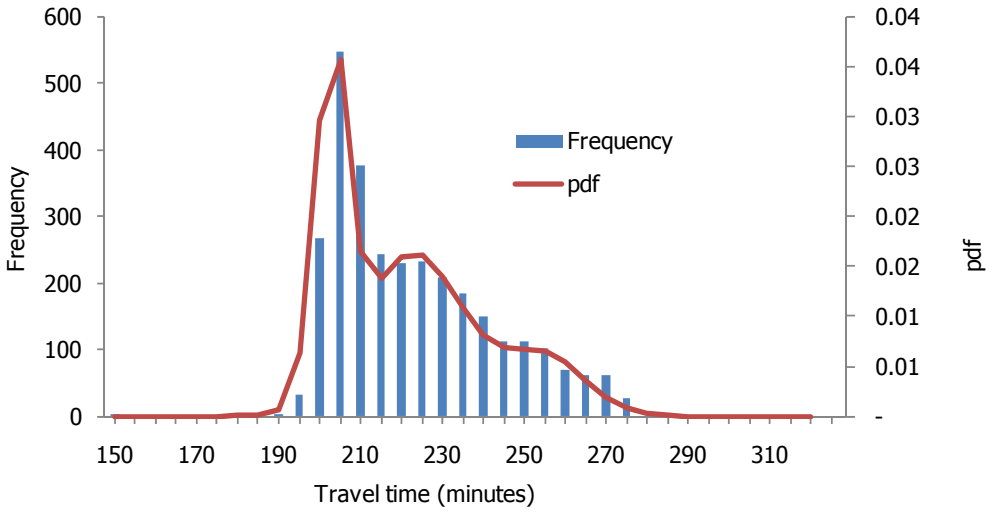


Figure 6.9 Travel-time histogram for Link 231 and a probability density function (pdf) fit by mixture distributions

For the analyses of Link 231, those trucks that cross BND (the downstream point) but not KFP (the upstream point) are also included in the downstream dataset. In particular, the testing dataset for the downstream site now includes a total of 2,000 trucks, out of which 983 cross only BND station and 1,017 (common trucks) cross both sites. Consequently, without applying the screening methods there would be a significant percentage of error if only the Bayesian method is applied to find a best match for each downstream vehicle in the upstream station. Even if all *common* trucks are matched accurately, there still would be 49% error (49% = 983/2,000) since 983 trucks never crossed the upstream station. Therefore, the screening methods play a more critical role in accurately matching vehicles in open systems.

Similar to the analysis done for Link 234, the upstream dataset (KFP) encompasses all trucks, including those do not carry transponders.

The results for Link 231 trucks are graphically summarized in Figure 6.10 for the AVC data, and Figure 6.11 and Figure 6.12 for the WIM data. Figure 6.12 is the same as Figure 6.11 except the axes are scaled to provide more detailed information. In addition, Table 9.3 and Table 9.4 in Appendix A provide the actual numbers of vehicles that are matched or mismatched at each threshold value in each screening method.

Similar to the observations made before for Link 234, it can be observed that the re-identification methods give more accurate results when axle weights are used in addition to the axle spacing since the results for the WIM data are much better as compared to those of the AVC data.

Second, in terms of the four screening methods, the 45° line method performs again better than the other methods as this method is able to more effectively reduce the mismatch error. For the WIM scenario, the naïve method closely follows the 45° line method but, in general, it does not perform better.

Third, when the WIM data is used for re-identification, vehicles can again be matched effectively with a reasonable low level of error (8-10%), while the total number of vehicles being matched is kept reasonably close to the actual number of vehicles crossing both the upstream and downstream sites. In this case, there are 1,017 common trucks that cross both sites – which is indicated by the vertical dashed line in the figures. For example, by selecting 0.01 for the threshold value in the 45° line method, 95%, or 815 trucks, are matched accurately out of 861 trucks (see Table 9.4 in Appendix A), which is about 85% of the 1,017 common trucks that cross both sites. In this example, of the 46 that are mismatched (861-815=46), 34 do not appear in the upstream station and the remaining 12 cross the upstream site but are not matched accurately. In other words, the 45° line method screened out 949 of the 983 (which do not cross the upstream site) but not the 34 that ended up being matched to the upstream vehicles.

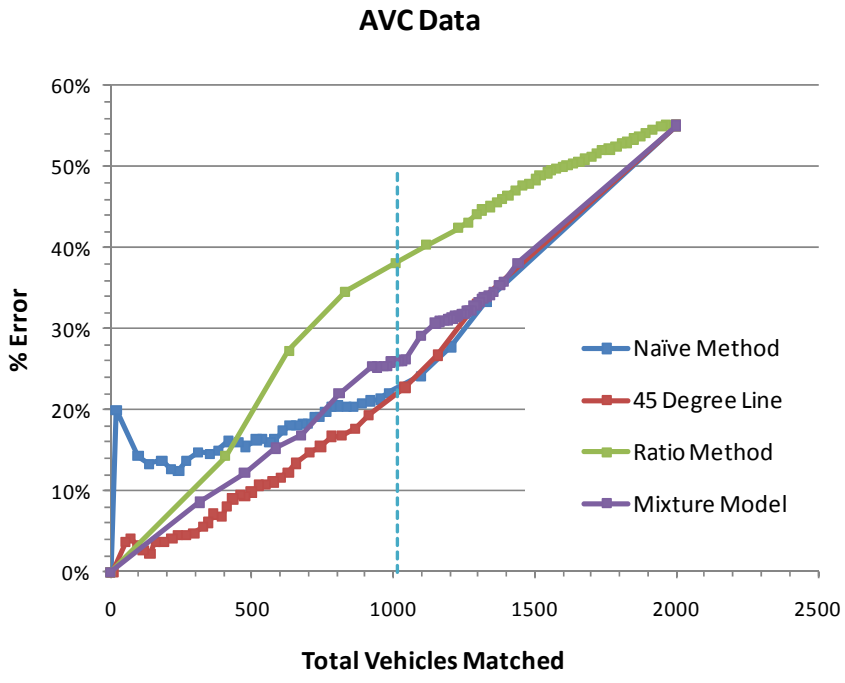


Figure 6.10 Tradeoff curves of the four screening criteria for the AVC scenario for Link 231

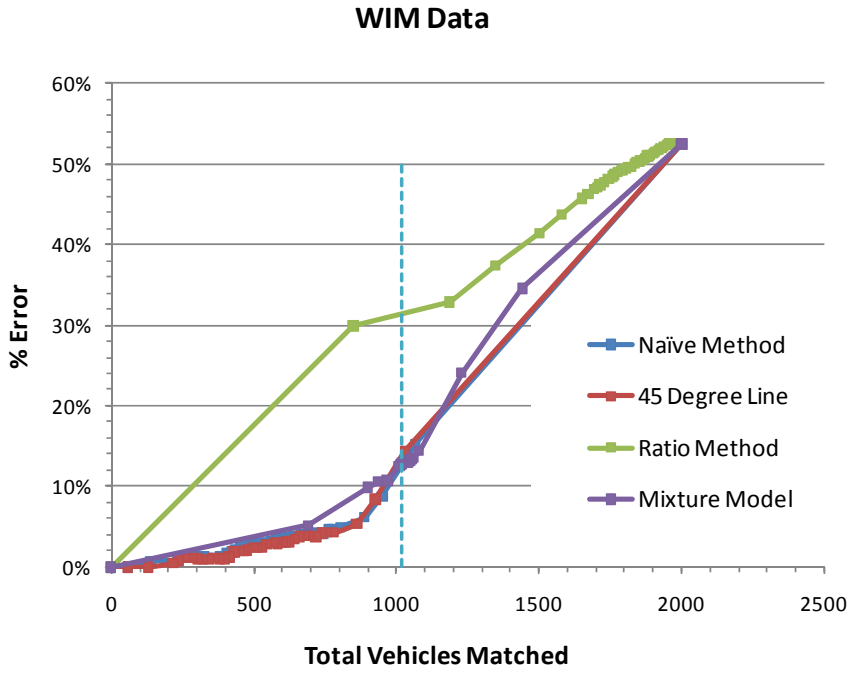


Figure 6.11 Tradeoff curves of the four screening criteria for the WIM scenario for Link 231

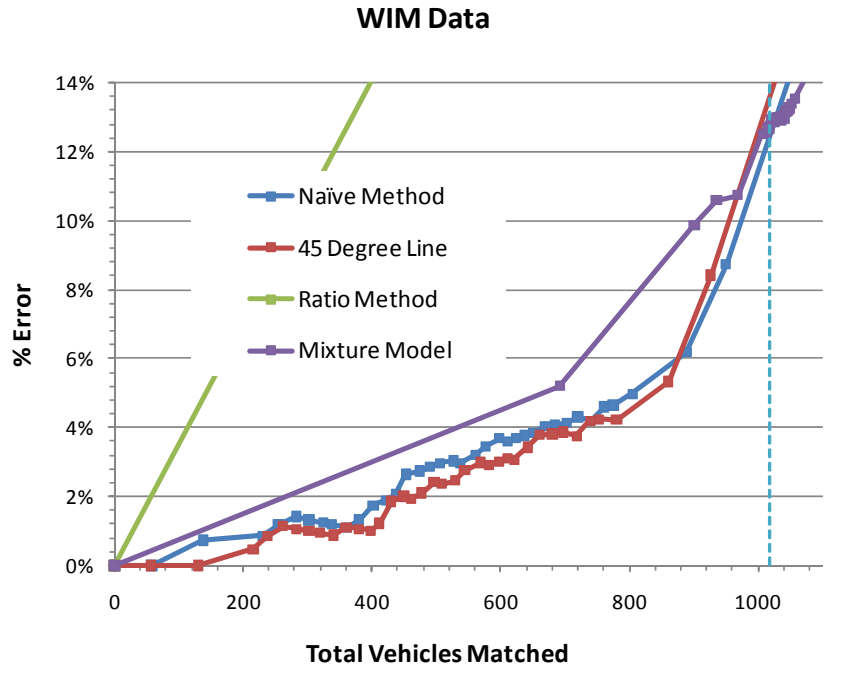


Figure 6.12 Tradeoff curves of the four screening criteria for the WIM scenario for Link 231

When these results are compared to those of Link 234, it can be observed that the overall error rates for Link 231are lower. Figure 6.13 shows the results when the 45° line method is used as the screening tool and the WIM data is used for the re-identification. The horizontal axis shows the total number of vehicles being matched as a percentage of the *common* trucks for each link (1,017 for Link 231 and 995 for Link 234). Since the total number of common trucks is different on Links 231 and 234, expressing the total vehicles matched as a percentage allows a better comparison. The results shown in Figure 6.13 can be explained by the fact that the travel times on Link 234 exhibit larger variance (see the longer tail of travel-time distribution in Figure 6.2 and compare it with Figure 6.9 for Link 231). Consequently, the average number of vehicles in the search space is larger for Link 234 (155 for Link 234 vs. 107 for Link 231). This makes it more challenging to find a correct match for vehicles on Link 234.

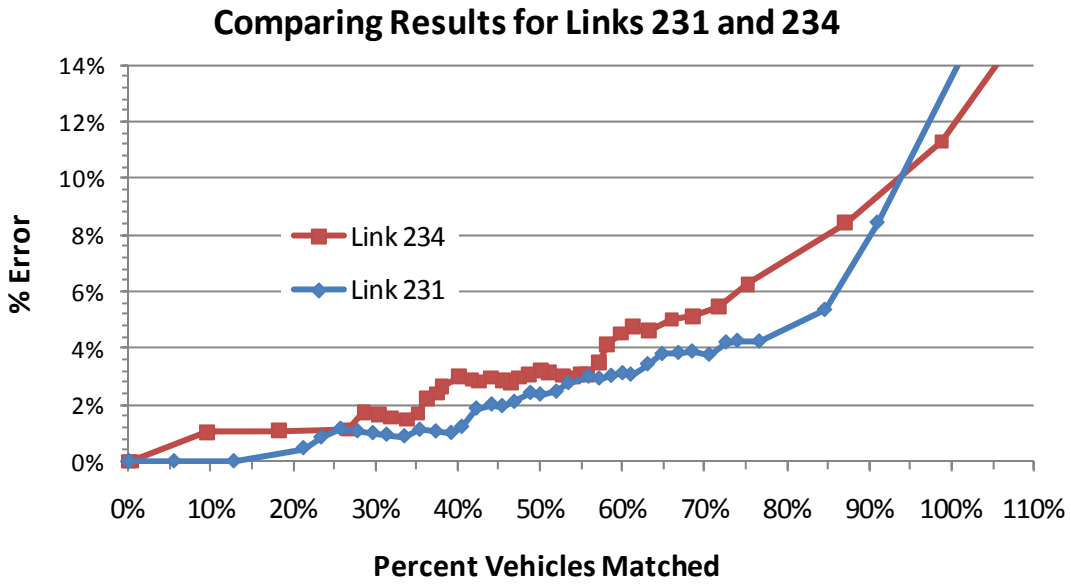


Figure 6.13 Comparing the results for Links 234 and 231 when WIM data is used for matching trucks

7.0 CONCLUSIONS

This project examined the use of vehicle-attribute data that are typically obtained from WIM and AVC sensors for anonymously re-identifying commercial vehicles so that their movements can be tracked. Tracking the movement of individual vehicles between different data collections sites provides valuable information for the estimation of travel times, travel delays, and origin-destination (OD) flows. Even though the data from transponder-equipped trucks can also be used for the estimation of travel times and OD flows, these trucks represent less than half of all trucks, or a small fraction, depending on the selected sites. For example in Oregon, on average, the rate of transponder-equipped trucks is about 40%. In addition, vehicle re-identification based on vehicle-attribute data does not raise any privacy concerns as is the case with other types of vehicle-tracking technologies (AVI, license plate recognition, etc.).

This research shows that it is feasible to re-identify trucks between WIM sites that are separated by long distances (i.e., more than 100 miles). By using the WIM data (i.e., axle weight data) and AVC (i.e., axle spacing) data from three different stations in Oregon, the research team has developed new methods to effectively re-identify trucks. Since the datasets include both the vehicle-attribute data (e.g., axle weights and axle spacing) and the corresponding unique transponder numbers, the true matching for those trucks that are equipped with transponders is known. The archived data of transponder-equipped trucks provide the needed data for model development and testing.

In this project, a new two-stage approach is developed to accurately match vehicles crossing upstream and downstream stations. For the first stage, a Bayesian method is developed where the necessary probability distributions are determined by fitting statistical mixture models to the training datasets. With the Bayesian method, for each downstream vehicle a *best* match is found in the upstream dataset. When it is known that all downstream vehicles also cross the upstream point, the Bayesian method alone can be applied to match trucks. To evaluate how this method would perform, the model is applied to test datasets taken from two sites that are separated by 145 miles. In this selected test dataset, for each one of the 1,000 downstream trucks a match needs to be found from10,581 upstream vehicles. It is found that the downstream trucks are matched to upstream trucks with 91% accuracy when both AVC and WIM data are used. This level of accuracy is significant given the fact that the upstream and downstream stations are 145 miles apart.

Even though the Bayesian method gives the best match for each truck, it does not account for the fact that some downstream vehicles may enter the road at some midpoint between the two stations and hence do not cross the upstream stations at all. Consequently, a mechanism is needed to separate out those vehicles that enter the road at a midpoint.

For the second-stage process, several methods are developed to screen out mismatched vehicles produced by the re-identification algorithm in the first stage, primarily those vehicles that enter the roadway at some midpoint between the upstream and downstream sites. These methods can be readily applied to any re-identification algorithm that computes a similarity metric. These screening models allow the user to trade off the total number of matched vehicles and the error or mismatch rate by adjusting a threshold value. When these methods are applied to sample WIM

datasets, it is observed that trucks are matched with approximately 90% accuracy while the total number of trucks being matched is about 95% of the actual common trucks that cross both upstream and downstream sites. If one is willing to match fewer trucks but improve the accuracy, the threshold value can be set to a larger value. For example, trucks can be matched with 98% accuracy if one is willing to only match about 40% of all common trucks. Depending on the application type, these methods allow the user to trade off the accuracy versus total vehicles being matched by adjusting a threshold parameter.

It is also observed that when travel times of vehicles between the upstream and downstream sites exhibit larger variation the re-identification becomes more challenging. In other words, mismatch rate increases as travel-time variance increases.

Overall, for travel-time estimation purposes, the methods presented in this report can be used effectively to match commercial vehicles crossing two data collection sites that are separated by long distances. The second phase of this research is underway to implement the methods on additional datasets and to estimate OD flows given a network created by WIM sites as nodes.

8.0 REFERENCES

Bertini, R. L., Hansen, S., Byrd, A., and Yin, T. (2005). Experience implementing a user service for archived intelligent transportation systems data. *Transportation Research Record* (1917), 90-99.

Cetin, M., and Nichols, A. P. (2009). Improving the accuracy of vehicle re-Identification by solving the assignment problem. *Transportation Research Record: Journal of the Transportation Research Board* **In press**.

Cetin, M., Nichols, A. P., and Monsere, C. M. (2010). Bayesian Models for Re-identification of Trucks over Long Distances Based on Axle Measurement Data *Journal of Intelligent Transportation Systems* **Submitted**.

Christiansen, I., and Hauer, L. 1996. Probing for Travel Time: Norway Applies AVI and WIM Technologies for Section Probe Data. *Traffic Technology International*, 41-44.

Coifman, B. (2003). Identifying the onset of congestion rapidly with existing traffic detectors. *Transportation Research Part a-Policy and Practice* **37** (3), 277-291.

Coifman, B., and Cassidy, M. (2002). Vehicle reidentification and travel time measurement on congested freeways. *Transportation Research, Part A (Policy and Practice)* **36A** (10), 899-917.

Coifman, B., and Krishnamurthy, S. (2007). Vehicle reidentification and travel time measurement across freeway junctions using the existing detector infrastructure. *Transportation Research Part C: Emerging Technologies* **15** (3), 135-153.

Dahlin, C. (1992). Proposed method for calibrating weigh-in-motion systems and for monitoring that calibration over time. *Transportation Research Record: Journal of Transportation Research Board* **1364**, 161-168.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society Series B-Methodological* **39** (1), 1-38.

Dion, F., and Rakha, H. (2006). Estimating Dynamic Roadway Travel Times Using Automatic Vehicle Identification Data for Low Sampling Rates. *Transportation Research Part B: Methodological* **40** (9), 745-766.

Elkins, L. C. Higgins, (2008) "Development of Truck Axle Spectra from Oregon Weigh-in-Motion Data for Use in Pavement Design and Analysis" Final Report FHWA-OR-RD-08-06, Oregon Department of Transportation.

Hellinga, B. (2001). Automated Vehicle Identification Tag-Matching Algorithms for Estimating Vehicle Travel Times: Comparative Assessment. *Transportation Research Record* **1774**, 106-114.

Liu, H. X., Oh, J.-S., and Recker, W. (2002). Adaptive signal control system with online performance measure for a single intersection. *Transportation Research Record* (1811), 131-138.

McLachlan, G., and Peel, D. (2000). *Finite Mixture Models*: John Wiley & Sons.

Monsere, C. M. Wolfe, H. Alawakiel, M. Stephens, Developing Corridor-Level Truck Travel Time Estimates and Other Freight Performance Measures from Archived ITS Data. Final Report SPR 304-361, OTREC-RR-09-10. August 2009.

Nichols, A. P., and Cetin, M. (2007). Numerical characterization of gross vehicle weight distributions from weigh-in-motion data. *Transportation Research Record* (1993), 148-154.

Oh, C., Ritchie, S. G., and Jeng, S. T. (2007). Anonymous vehicle reidentification using heterogeneous detection systems. *Ieee Transactions on Intelligent Transportation Systems* **8** (3), 460-469.

Oh, C., Tok, A., and Ritchie, S. G. (2005). Real-time freeway level of service using inductive-signature-based vehicle reidentification system. *IEEE Transactions on Intelligent Transportation Systems* **6** (2), 138-146.

Shuldiner, P., and Upchurch, J. 2001. Automated Travel Time Data for a Regional Traveler Information System. Institute of Transportation Engineers.

Sun, C., Arr, G., and Ramachandran, R. P. (2003). Vehicle reidentification as method for deriving travel time and travel time distributions: Investigation. *Transportation Research Record* (1826), 25-31.

Sun, C., Ritchie, S. G., Tsai, K., and Jayakrishnan, R. (1999). Use of vehicle signature analysis and lexicographic optimization for vehicle reidentification on freeways. *Transportation Research Part C: Emerging Technologies* **7** (4), 167-185.

Tawfik, A. Y., Abdulhai, B., Peng, A., and Tabib, S. M. (2004). Using decision trees to improve the accuracy of vehicle signature reidentification. *Transportation Research Record* (1886), 24-33.

TRB. 2003. A Concept for a National Freight Data Program: Special Report 276. Transportation Research Board, The National Academies.

Trevor, H., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning*: Springer.

Turner, S., Eisele, W., Benz, R., and Holdener, D. 1998. Travel Time Data Collection Handbook. Texas Transportation Institute, A&M University, College Station.

9.0 APPENDICES

APPENDIX A

RESULTS OF THE RE-IDENTIFICATION ALGORITHMS

Notation used in the next four tables:

- 0 = correctly matched;
- 1 = mismatched even though vehicle crossed the upstream WIM station
- -1 = mismatched and the vehicle never crossed the upstream WIM station

The values in each table indicate the number of trucks for which a match is found in the upstream site. The delta value is the threshold used in the screening methods.

Table 9.1 Results of the re-identification methods when applied to the Link 234 AVC data

	Naïve Method			45° Line			Ratio			Mixture Model		
Delta	0	1	-1	0	1	-1	0	1	-1	0	1	-1
0.00000	907	88	405	907	88	405	907	88	405	907	88	405
0.00001	900	87	199	900	85	195	907	88	405	907	88	284
0.00100	889	82	152	885	80	144	907	88	405	905	86	228
0.01000	834	76	114	831	72	106	907	86	404	903	86	221
0.05000	770	66	83	755	56	74	904	84	396	899	82	207
0.07500	737	63	72	714	52	62	900	82	391	896	79	201
0.10000	709	61	64	680	50	57	892	82	384	893	75	197
0.12500	684	61	58	651	45	52	891	80	378	893	73	196
0.15000	674	59	55	639	41	50	884	75	375	890	71	192
0.17500	660	54	50	625	40	46	882	72	374	883	68	190
0.20000	650	52	48	609	33	40	881	70	370	881	67	187
0.22500	638	51	44	592	28	36	879	68	363	879	66	186
0.25000	626	49	41	579	24	35	874	67	359	875	66	184
0.27500	607	47	38	559	20	30	872	64	351	869	64	183
0.30000	597	46	37	540	19	29	868	61	343	862	64	181
0.32500	581	43	36	522	16	28	861	59	338	858	63	180
0.35000	568	41	35	511	13	26	858	57	334	855	61	180
0.37500	547	40	34	491	12	24	857	55	331	850	57	179
0.40000	526	38	33	474	12	23	853	55	325	848	56	178
0.42500	511	35	31	445	11	21	848	54	319	843	56	177
0.45000	489	33	31	428	11	20	838	53	317	841	55	176
0.47500	472	29	29	406	10	18	829	52	311	836	54	175
0.50000	454	29	27	392	9	15	828	48	307	833	52	173
0.52500	437	29	26	378	8	13	827	48	301	831	51	170
0.55000	429	29	23	364	8	13	822	44	297	830	49	167
0.57500	405	26	21	338	6	12	818	42	291	824	49	163
0.60000	391	26	21	322	4	11	815	40	282	815	49	151
0.62500	371	25	19	308	2	10	812	39	275	809	49	137
0.65000	362	24	15	302	2	6	809	38	270	805	49	135
0.67500	341	20	10	280	2	4	804	36	256	800	49	135
0.70000	318	17	8	258	1	3	799	34	249	794	47	132
0.72500	299	16	7	241	1	3	794	30	241	789	46	131
0.75000	275	15	5	220	1	2	789	28	234	782	42	131
0.77500	244	15	5	195	0	2	782	23	227	776	38	129
0.80000	215	13	5	173	0	2	776	19	213	760	37	125
0.82500	184	11	4	148	0	2	768	15	204	750	36	123
0.90000	68	3	1	51	0	1	732	8	175	682	17	97
0.95000	9	0	0	6	0	0	672	6	147	616	3	57
0.99000	0	0	0	0	0	0	579	2	93	541	1	50
0.99900	0	0	0	0	0	0	468	1	58	454	1	22
0.99999	0	0	0	0	0	0	361	1	24	287	1	9
1.00000	0	0	0	0	0	0	0	0	0	0	0	0

Table 9.2 Results of the re-identification methods when applied to the Link 234 WIM data

	Naïve Method			45° Line			Ratio			Mixture Model		
Delta	0	1	-1	0	1	-1	0	1	-1	0	1	-1
0.00000	941	54	405	941	54	405	941	54	405	941	54	405
0.00001	925	49	137	925	49	130	941	54	405	938	53	254
0.00100	876	43	75	872	41	70	941	54	405	934	51	206
0.01000	800	34	45	793	30	43	941	53	403	926	49	131
0.05000	710	26	30	702	20	27	937	51	399	908	38	124
0.07500	679	24	27	674	16	23	935	47	392	903	36	124
0.10000	653	21	23	648	15	20	933	46	390	901	34	123
0.12500	629	20	23	624	13	20	931	46	387	899	31	122
0.15000	611	18	23	600	11	18	929	45	383	899	30	120
0.17500	594	18	22	580	11	18	925	45	375	895	29	120
0.20000	582	18	22	569	9	18	923	44	373	893	29	120
0.22500	569	18	21	555	7	17	923	44	368	891	29	120
0.25000	562	17	21	549	5	15	923	42	363	889	29	120
0.27500	550	16	19	537	4	13	922	41	359	888	29	120
0.30000	544	14	19	530	4	13	920	41	356	888	29	120
0.32500	537	13	17	521	4	12	917	40	355	885	29	120
0.35000	528	12	17	509	4	12	917	38	355	884	28	119
0.37500	516	10	16	493	4	12	915	36	351	883	28	119
0.40000	508	10	16	482	4	12	915	36	342	883	28	119
0.42500	498	10	16	469	4	11	914	35	338	881	28	119
0.45000	485	10	14	458	4	10	912	34	336	881	27	119
0.47500	478	10	13	450	4	9	910	34	332	877	27	119
0.50000	471	10	12	440	4	9	908	34	329	875	27	119
0.52500	458	10	12	425	4	9	906	33	326	872	27	119
0.55000	442	10	12	412	3	9	905	33	320	871	27	119
0.57500	430	9	12	403	3	9	903	32	312	871	27	119
0.60000	414	9	12	387	3	9	900	30	306	871	27	119
0.62500	396	9	11	369	2	8	899	29	299	871	27	119
0.65000	389	8	11	364	1	8	895	29	293	869	27	119
0.67500	381	8	10	352	1	7	891	29	284	866	25	119
0.70000	371	8	9	344	1	5	891	27	278	863	23	119
0.72500	359	8	9	332	1	4	889	26	271	863	23	118
0.75000	342	7	8	313	1	4	886	25	266	861	23	118
0.77500	325	7	8	297	1	4	878	24	260	860	23	118
0.80000	305	7	8	280	1	4	873	21	251	860	21	117
0.82500	288	7	6	263	0	3	868	18	243	857	21	117
0.90000	204	5	4	180	0	2	846	13	210	852	20	112
0.95000	114	1	2	94	0	1	818	9	183	799	14	73
0.99000	5	0	0	3	0	0	765	6	129	750	13	69
0.99900	0	0	0	0	0	0	655	4	78	691	12	64
0.99999	0	0	0	0	0	0	475	1	31	584	6	42
1.00000	0	0	0	0	0	0	0	0	0	0	0	0

Table 9.3 Results of the re-identification methods when applied to the Link 231 AVC data

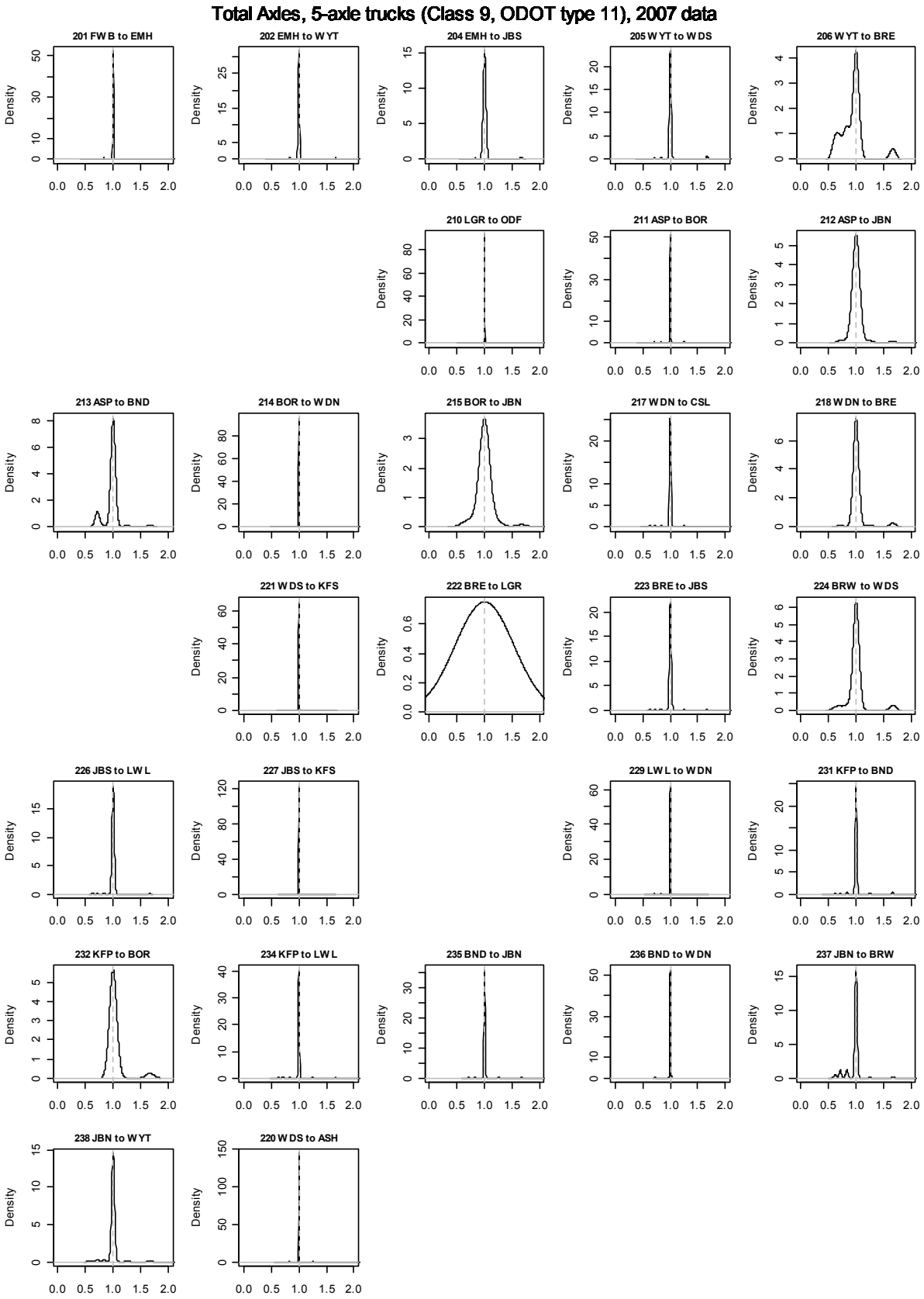
	Naïve Method			45° Line			Ratio			Mixture Model		
Delta	0	1	-1	0	1	-1	0	1	-1	0	1	-1
0.00000	898	119	983	898	119	983	898	119	983	898	119	983
0.00001	885	113	329	866	106	321	879	112	972	892	114	435
0.00100	870	108	226	848	88	221	879	111	972	892	112	384
0.01000	831	93	172	804	72	165	877	104	967	889	112	374
0.05000	767	87	130	736	57	120	870	97	949	885	110	358
0.07500	750	86	117	712	47	106	868	89	936	882	106	351
0.10000	725	85	109	679	41	96	865	85	920	878	105	347
0.12500	703	82	102	652	39	92	861	81	907	874	102	346
0.15000	683	80	95	627	33	82	858	79	888	872	101	342
0.17500	664	79	91	599	29	74	852	74	880	869	98	336
0.20000	638	77	87	568	21	67	848	69	869	869	95	330
0.22500	623	75	84	552	16	61	843	63	857	864	95	327
0.25000	611	72	78	532	14	56	834	58	846	862	91	320
0.27500	599	70	72	512	14	50	832	52	834	856	90	317
0.30000	584	69	69	488	12	47	830	50	822	853	90	314
0.32500	571	65	63	469	12	44	824	45	809	852	89	309
0.35000	556	64	60	449	9	40	818	44	794	845	87	307
0.37500	540	63	56	431	9	35	812	43	780	840	85	304
0.40000	521	60	55	415	9	35	804	41	769	838	85	300
0.42500	504	57	49	393	9	30	800	41	756	832	84	299
0.45000	485	54	41	378	8	25	791	38	747	830	81	297
0.47500	473	54	36	367	5	22	786	37	734	827	81	294
0.50000	449	53	35	338	5	21	784	35	726	823	80	291
0.52500	431	51	33	326	4	17	776	34	709	813	79	287
0.55000	405	45	29	309	3	15	775	30	697	805	78	282
0.57500	382	44	29	281	2	12	772	29	680	796	78	274
0.60000	350	41	26	255	2	10	761	29	664	778	77	242
0.62500	323	38	19	229	2	9	757	27	646	770	72	201
0.65000	299	36	15	208	2	7	752	25	628	763	70	199
0.67500	265	35	11	181	1	6	748	24	613	755	67	198
0.70000	233	31	6	154	1	5	743	23	600	747	66	196
0.72500	211	24	6	136	0	3	738	21	585	734	65	191
0.75000	186	22	5	113	0	3	727	21	565	730	61	188
0.77500	157	20	5	91	0	3	723	21	549	722	59	187
0.80000	117	14	4	70	0	3	718	20	525	704	55	183
0.82500	84	10	4	52	0	2	708	19	503	689	52	182
0.90000	16	3	1	9	0	0	666	11	439	631	30	148
0.95000	0	0	0	0	0	0	624	9	376	561	9	105
0.99000	0	0	0	0	0	0	543	6	281	495	2	87
0.99900	0	0	0	0	0	0	459	2	170	417	1	57
0.99999	0	0	0	0	0	0	349	0	58	288	0	27
1.00000	0	0	0	0	0	0	0	0	0	0	0	0

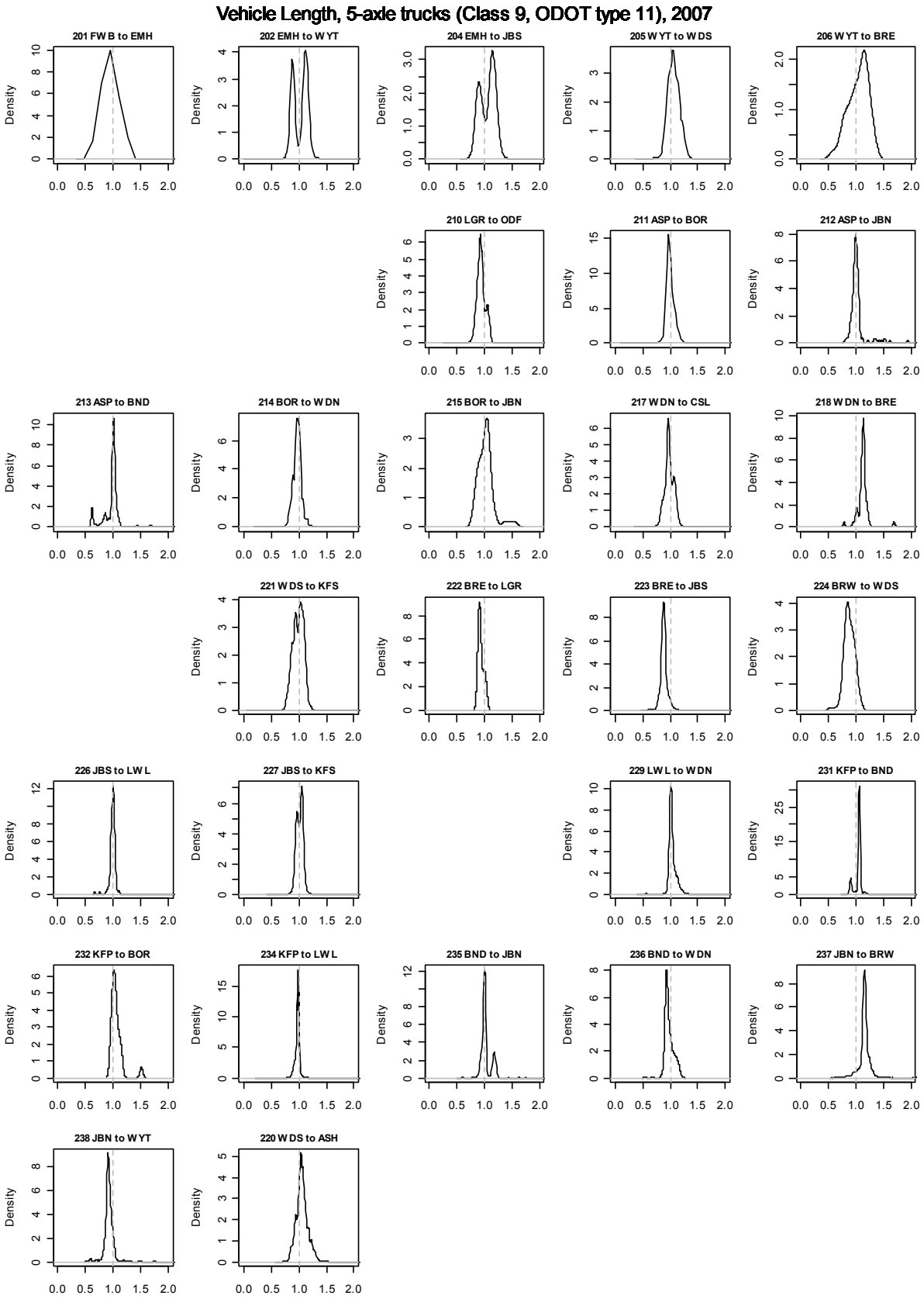
Table 9.4 Results of the re-identification methods when applied to the Link 231 WIM data

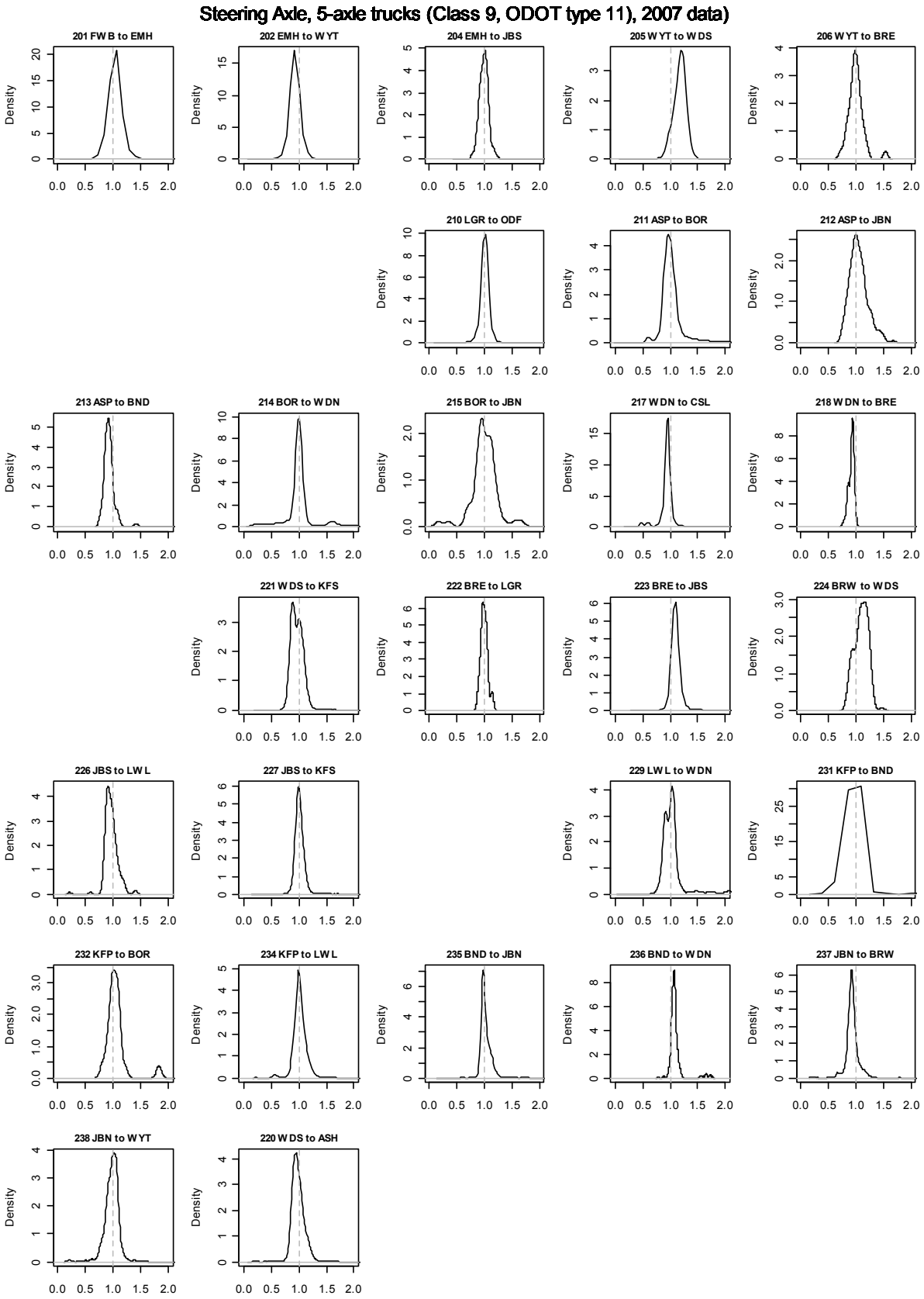
	Naïve Method			45° Line			Ratio			Mixture Model		
Delta	0	1	-1	0	1	-1	0	1	-1	0	1	-1
0.00000	951	66	983	951	66	983	951	66	983	951	66	983
0.00001	905	39	124	883	30	117	931	62	970	944	57	441
0.00100	867	24	59	848	20	58	931	62	970	933	45	251
0.01000	834	19	36	815	12	34	931	62	970	924	41	114
0.05000	765	14	26	747	9	24	930	60	965	914	35	108
0.07500	739	12	24	721	9	23	929	60	962	911	35	106
0.10000	726	12	23	708	9	22	929	60	960	909	33	106
0.12500	709	11	20	691	8	19	928	57	952	908	33	105
0.15000	689	11	20	670	8	19	927	57	945	908	33	105
0.17500	673	10	19	654	8	18	926	56	937	907	33	105
0.20000	656	10	18	635	8	17	926	56	937	906	32	105
0.22500	643	10	17	620	7	15	926	56	926	906	30	105
0.25000	625	10	15	602	6	13	926	55	921	902	30	105
0.27500	613	9	15	592	6	13	926	52	913	902	30	104
0.30000	601	9	14	579	6	12	923	51	910	900	30	104
0.32500	588	9	13	565	5	12	922	50	908	900	30	104
0.35000	576	9	13	552	5	12	922	49	899	900	30	104
0.37500	557	8	12	529	4	11	922	47	892	898	30	104
0.40000	543	8	10	516	4	9	920	46	888	897	30	104
0.42500	522	7	9	497	4	8	919	46	880	897	30	104
0.45000	510	7	9	485	4	8	918	45	872	897	30	104
0.47500	490	7	8	467	3	7	917	41	864	896	30	104
0.50000	476	7	7	453	3	6	915	40	860	894	29	104
0.52500	462	7	6	440	3	6	912	38	854	894	29	104
0.55000	441	7	5	422	3	5	911	35	848	894	29	104
0.57500	428	6	3	407	2	3	909	34	846	893	29	103
0.60000	415	6	2	395	2	2	907	33	838	892	29	103
0.62500	395	5	2	376	2	2	906	31	827	891	29	103
0.65000	375	3	2	356	2	2	906	31	821	891	29	103
0.67500	358	2	2	338	1	2	903	31	809	890	29	103
0.70000	334	2	2	316	1	2	903	28	798	889	29	103
0.72500	321	2	2	299	1	2	901	27	786	889	29	102
0.75000	298	2	2	280	1	2	901	27	777	888	29	101
0.77500	278	2	2	259	1	2	900	23	770	888	28	101
0.80000	252	2	1	236	1	1	899	22	750	886	28	101
0.82500	229	2	0	215	1	0	896	19	735	885	27	100
0.90000	137	1	0	130	0	0	889	17	673	881	26	100
0.95000	58	0	0	56	0	0	880	12	610	864	19	85
0.99000	0	0	0	0	0	0	843	7	495	836	18	81
0.99900	0	0	0	0	0	0	797	3	387	812	14	75
0.99999	0	0	0	0	0	0	595	0	254	656	2	34
1.00000	0	0	0	0	0	0	0	0	0	0	0	0

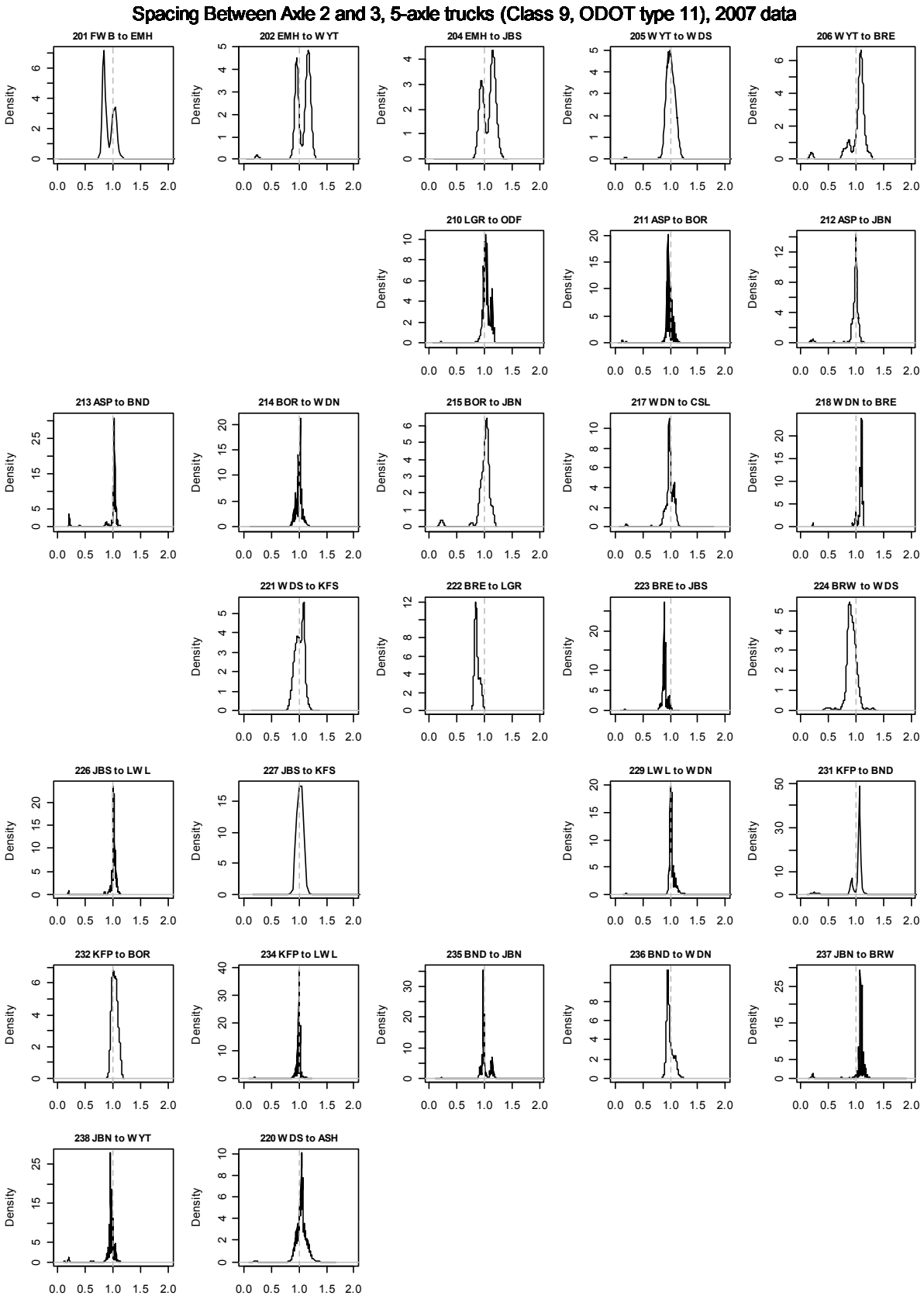
APPENDIX B

COMPARISON OF WIM MEASUREMENTS BETWEEN STATION PAIRS


Total Axles, 5-axle trucks (Class 9, ODOT type 11), 2007 data
201 FWB to EMH
202 EMH to WYT
204 EMH to JBS
205 WYT to WDS
206 WYT to BRE
210 LGR to ODF
211 ASP to BOR
212 ASP to JBN
213 ASP to BND
214 BOR to WDN
215 BOR to JBN
217 WDN to CSL
218 WDN to BRE
221 WDS to KFS
222 BRE to LGR
223 BRE to JBS
224 BRW to WDS
226 JBS to LWL
227 JBS to KFS
229 LWL to WDN
231 KFP to BND
232 KFP to BOR
234 KFP to LWL
235 BND to JBN
236 BND to WDN
237 JBN to BRW
238 JBN to WYT
220 WDS to ASH
Density

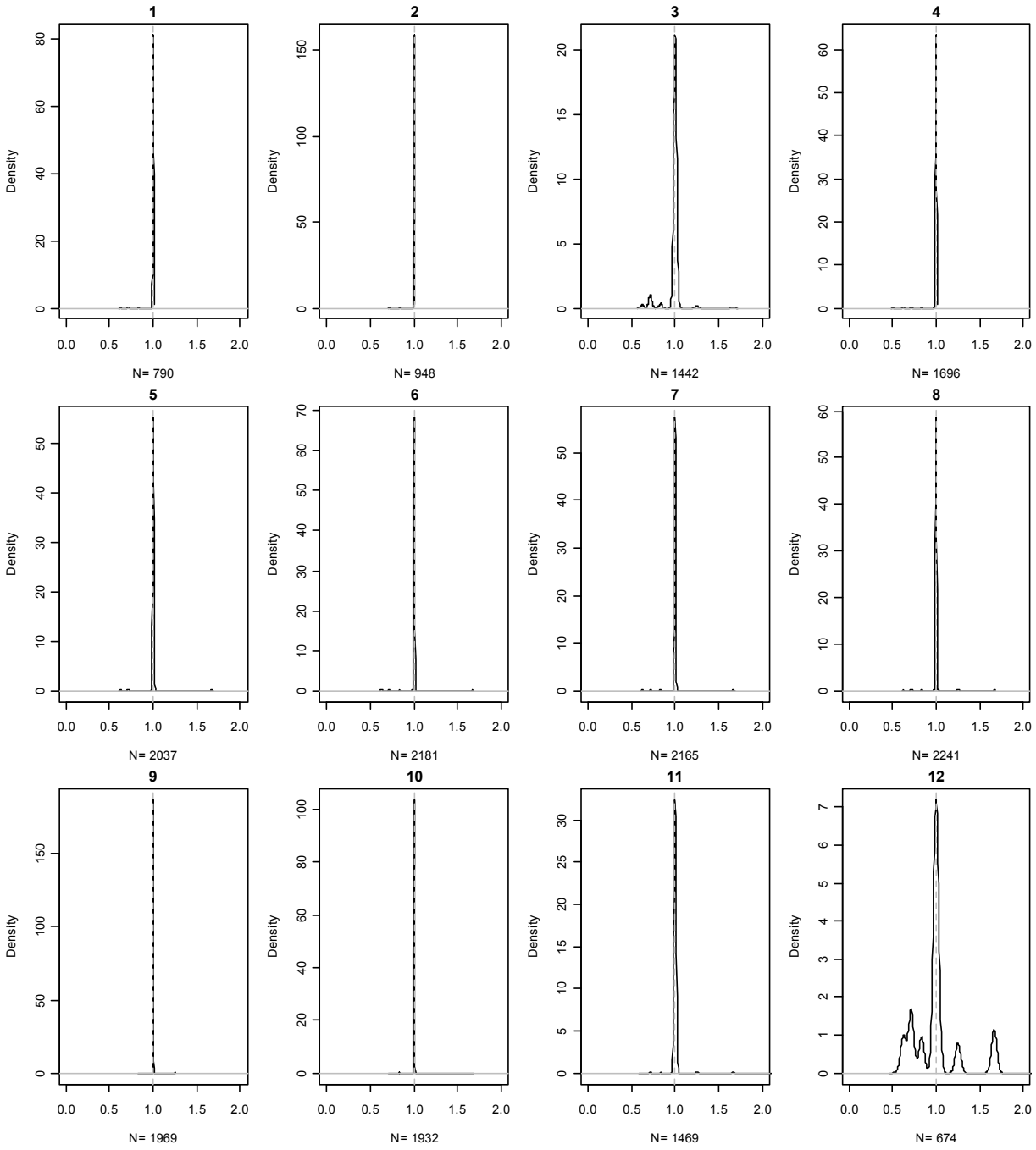

Vehicle Length, 5-axle trucks (Class 9, ODOT type 11), 2007
201 FWB to EMH
202 EMH to WYT
204 EMH to JBS
205 WYT to WDS
206 WYT to BRE
210 LGR to ODF
211 ASP to BOR
212 ASP to JBN
213 ASP to BND
214 BOR to WDN
215 BOR to JBN
217 WDN to CSL
218 WDN to BRE
221 WDS to KFS
222 BRE to LGR
223 BRE to JBS
224 BRW to WDS
226 JBS to LWL
227 JBS to KFS
229 LWL to WDN
231 KFP to BND
232 KFP to BOR
234 KFP to LWL
235 BND to JBN
236 BND to WDN
237 JBN to BRW
238 JBN to WYT
220 WDS to ASH
Density

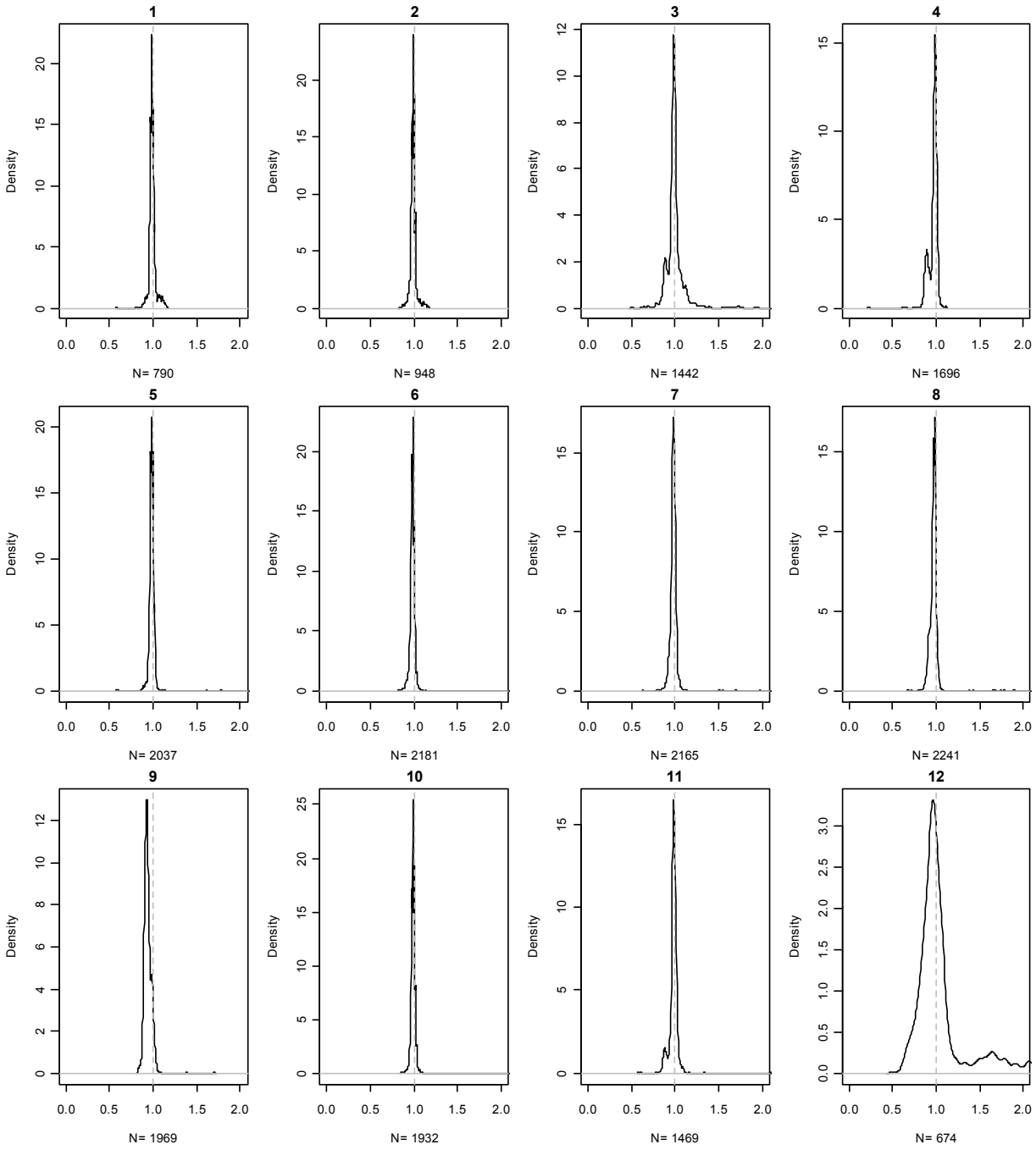

Steering Axle, 5-axle trucks (Class 9, ODOT type 11), 2007 data)
201 FWB to EMH
202 EMH to WYT
204 EMH to JBS
205 WYT to WDS
206 WYT to BRE
210 LGR to ODF
211 ASP to BOR
212 ASP to JBN
213 ASP to BND
214 BOR to WDN
215 BOR to JBN
217 WDN to CSL
218 WDN to BRE
221 WDS to KFS
222 BRE to LGR
223 BRE to JBS
224 BRW to WDS
226 JBS to LWL
227 JBS to KFS
229 LWL to WDN
231 KFP to BND
232 KFP to BOR
234 KFP to LWL
235 BND to JBN
236 BND to WDN
237 JBN to BRW
238 JBN to WYT
220 WDS to ASH
Density

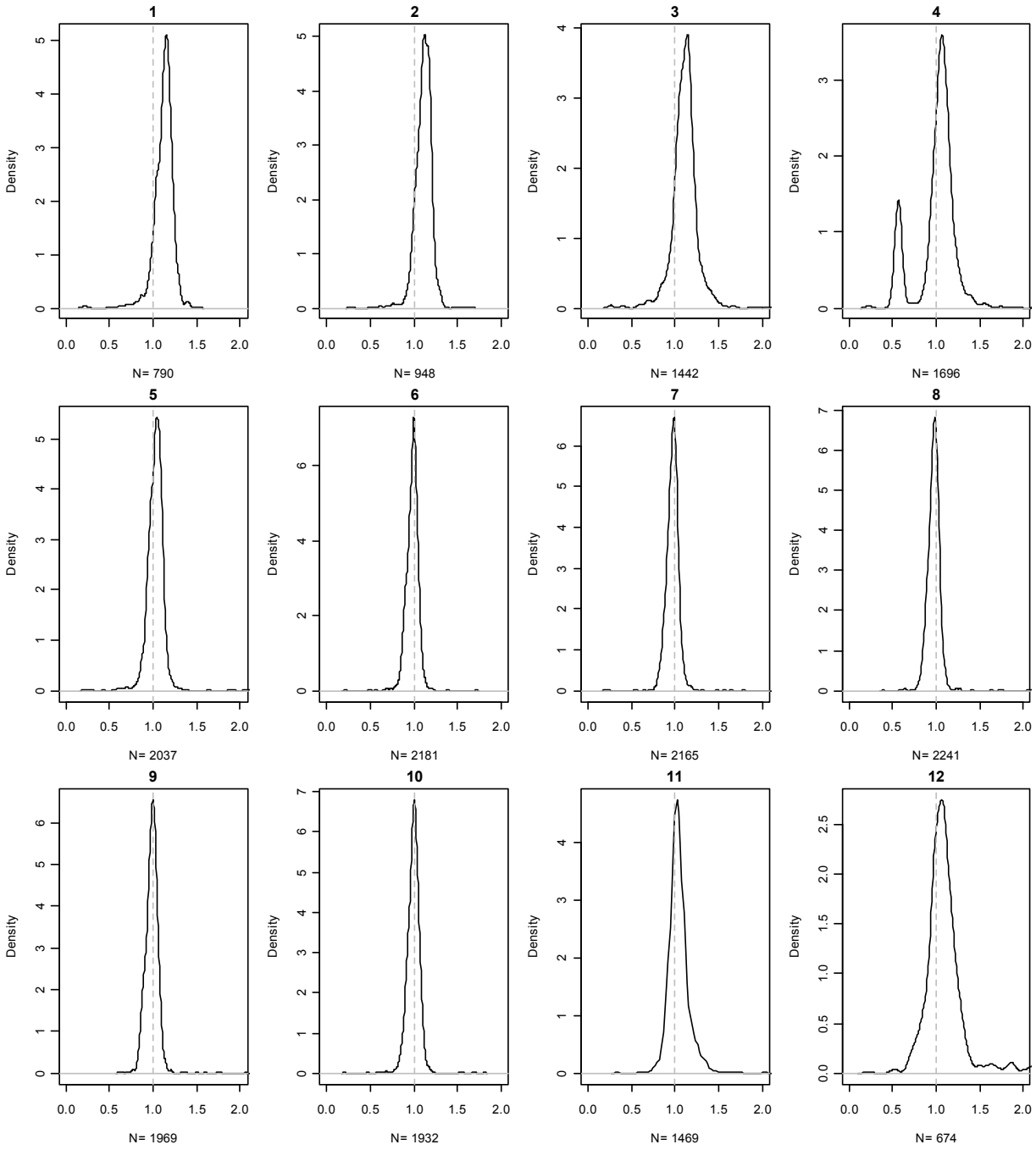

Spacing Between Axle 2 and 3, 5-axle trucks (Class 9, ODOT type 11), 2007 data
201 FWB to EMH
202 EMH to WYT
204 EMH to JBS
205 WYT to WDS
206 WYT to BRE
210 LGR to ODF
211 ASP to BOR
212 ASP to JBN
213 ASP to BND
214 BOR to WDN
215 BOR to JBN
217 WDN to CSL
218 WDN to BRE
221 WDS to KFS
222 BRE to LGR
223 BRE to JBS
224 BRW to WDS
226 JBS to LWL
227 JBS to KFS
229 LWL to WDN
231 KFP to BND
232 KFP to BOR
234 KFP to LWL
235 BND to JBN
236 BND to WDN
237 JBN to BRW
238 JBN to WYT
220 WDS to ASH
Density

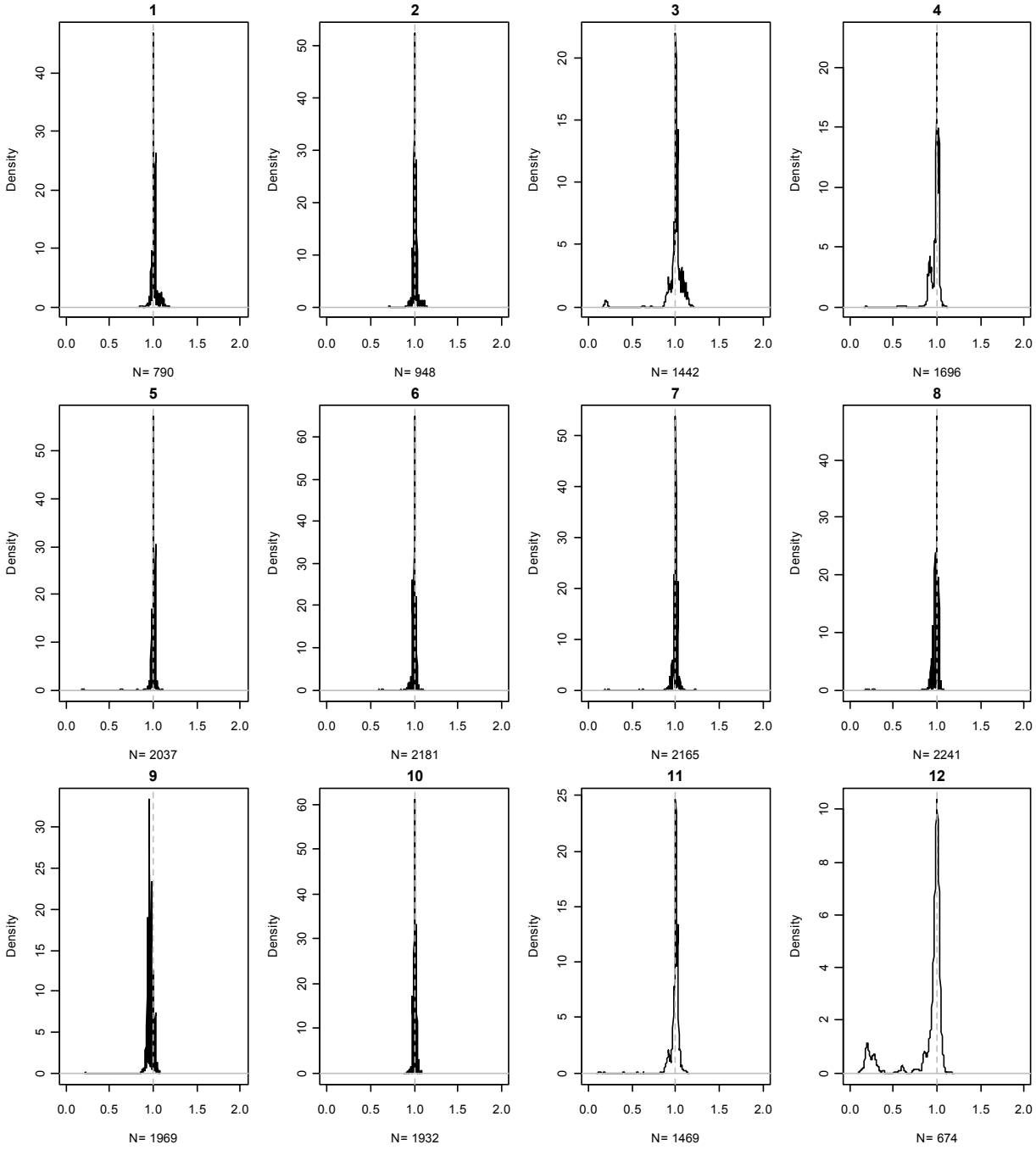
APPENDIX C

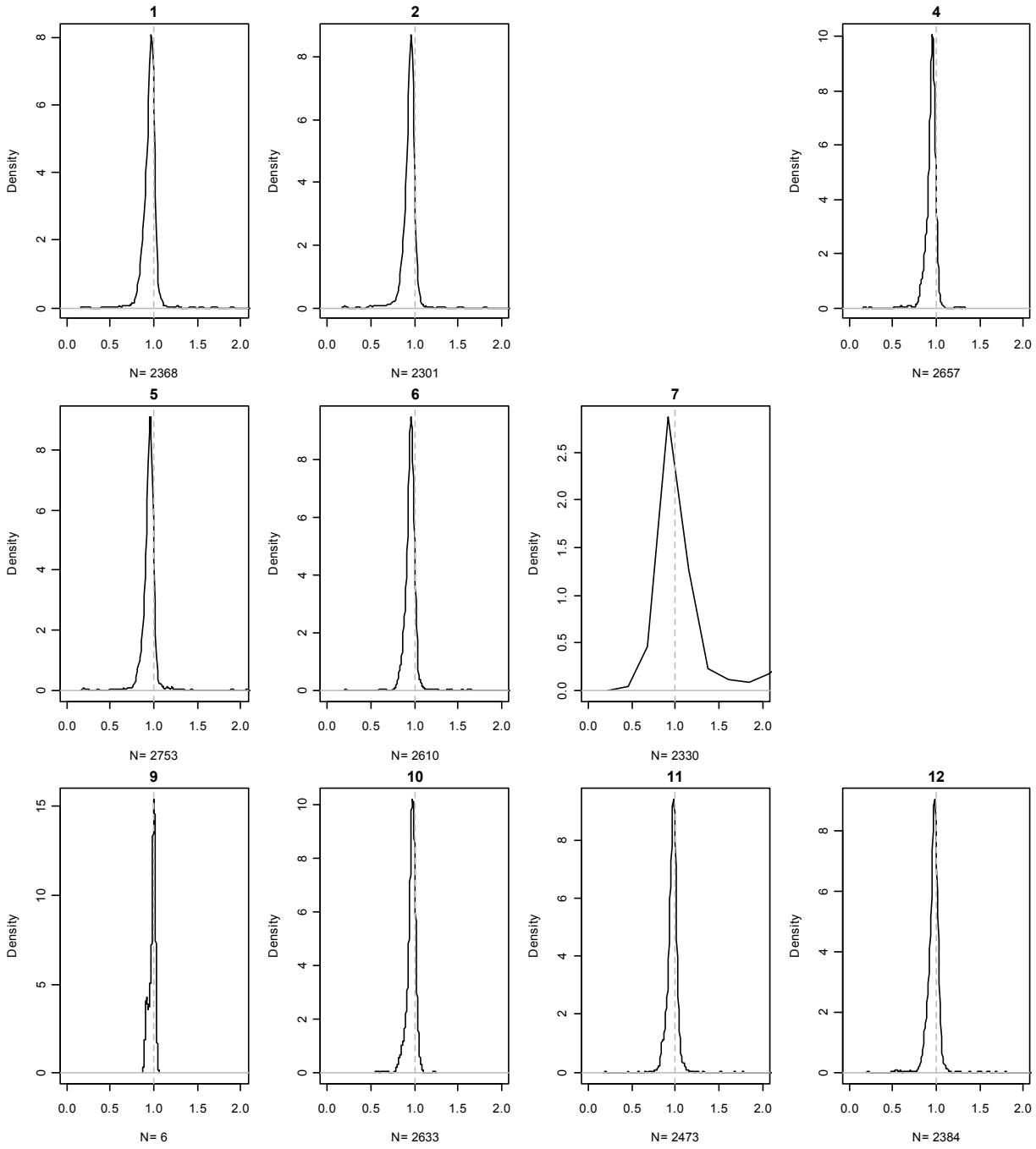
COMPARISON OF WIM MEASUREMENTS BY MONTH, LINK 231 AND 234

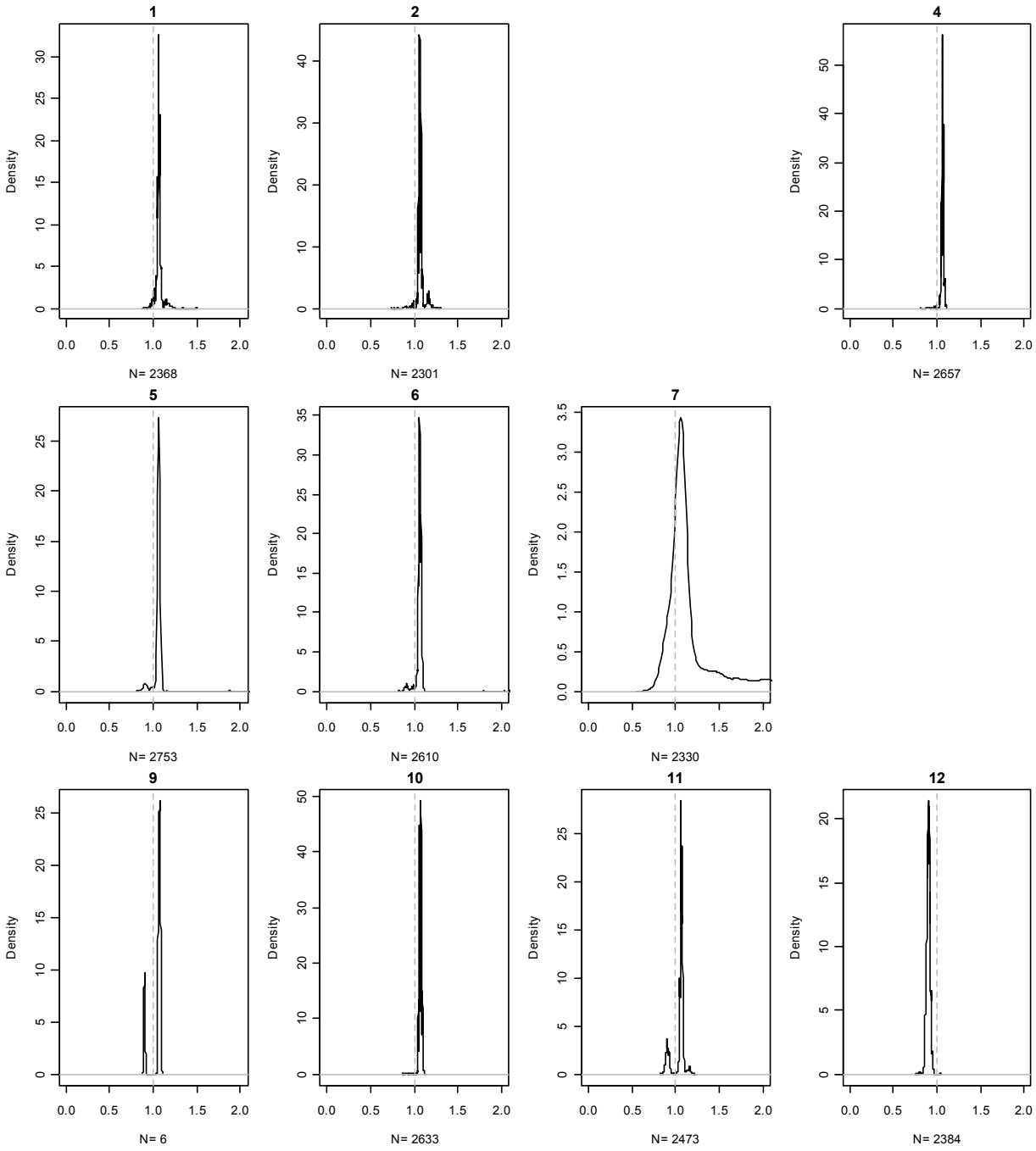

Link 234, Total Number of Axles Ratios By Month, 5-axle trucks (Class 9, ODOT type 11), 2007
1
N= 790
2
N= 948
3
N= 1442
4
N= 1696
5
N= 2037
6
N= 2181
7
N= 2165
8
N= 2241
9
N= 1969
10
N= 1932
11
N= 1469
12
N= 674
Density

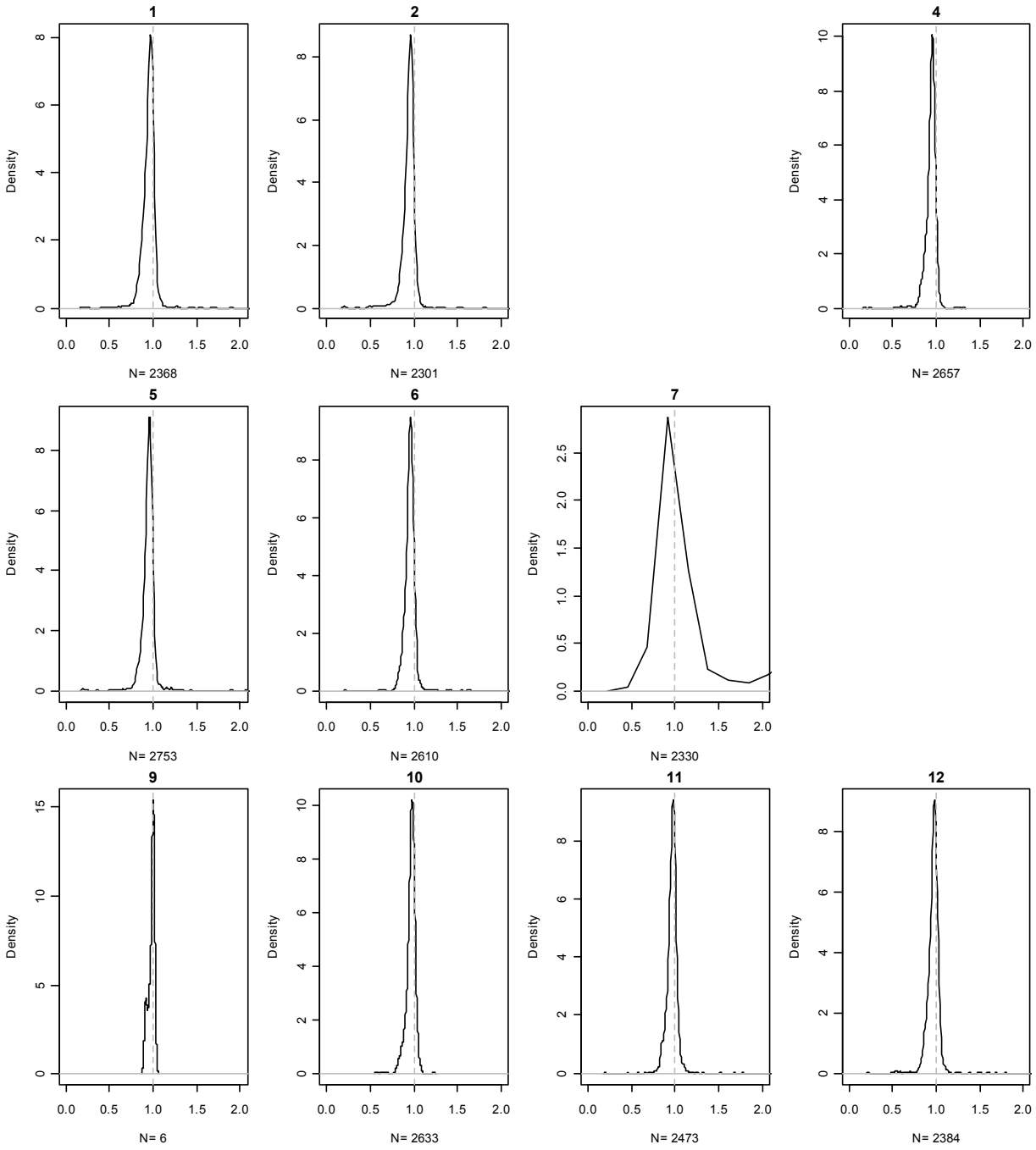

Link 234, Vehicle Length Ratios By Month, 5-axle trucks (Class 9, ODOT type 11), 2007
1
2
3
4
5
6
7
8
9
10
11
12
Density
N= 790
N= 948
N= 1442
N= 1696
N= 2037
N= 2181
N= 2165
N= 2241
N= 1969
N= 1932
N= 1469
N= 674

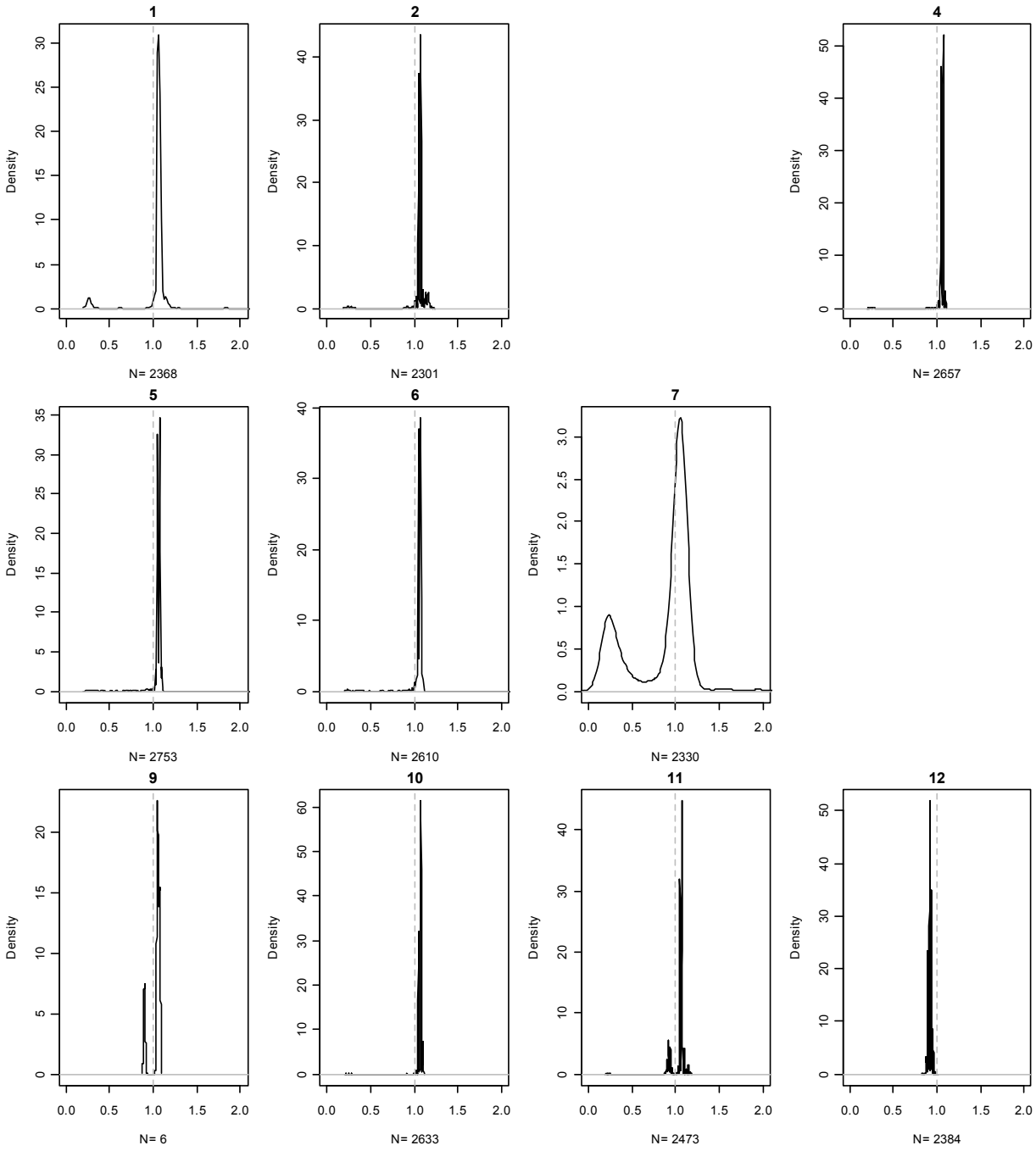

Link 234, Steering Axle 1 Weight Ratios By Month, 5-axle trucks (Class 9, ODOT type 11), 2007
1
N= 790
2
N= 948
3
N= 1442
4
N= 1696
5
N= 2037
6
N= 2181
7
N= 2165
8
N= 2241
9
N= 1969
10
N= 1932
11
N= 1469
12
N= 674
Density


Link 234, Spacing Between Axle 2-3 Ratios By Month, 5-axle trucks (Class 9, ODOT type 11), 2007
1
N= 790
2
N= 948
3
N= 1442
4
N= 1696
5
N= 2037
6
N= 2181
7
N= 2165
8
N= 2241
9
N= 1969
10
N= 1932
11
N= 1469
12
N= 674
Density


Link 231, Steering Axle 1 Weight Ratios By Month, 5-axle trucks (Class 9, ODOT type 11), 2007
1
N= 2368
2
N= 2301
4
N= 2657
5
N= 2753
6
N= 2610
7
N= 2330
9
N= 6
10
N= 2633
11
N= 2473
12
N= 2384
Density


Link 231, Vehicle Length Ratios By Month, 5-axle trucks (Class 9, ODOT type 11), 2007
1
N= 2368
2
N= 2301
4
N= 2657
5
N= 2753
6
N= 2610
7
N= 2330
9
N= 6
10
N= 2633
11
N= 2473
12
N= 2384
Density


Link 231, Steering Axle 1 Weight Ratios By Month, 5-axle trucks (Class 9, ODOT type 11), 2007
1
N= 2368
2
N= 2301
4
N= 2657
5
N= 2753
6
N= 2610
7
N= 2330
9
N= 6
10
N= 2633
11
N= 2473
12
N= 2384
Density


Link 231, Spacing Between Axle 2-3 Ratios By Month, 5-axle trucks (Class 9, ODOT type 11), 2007
1
2
4
5
6
7
9
10
11
12
Density
N= 2368
N= 2301
N= 2657
N= 2753
N= 2610
N= 2330
N= 6
N= 2633
N= 2473
N= 2384

P.O. Box 751
Portland, OR 97207

OTREC is dedicated to stimulating and conducting collaborative multi-disciplinary research on multi-modal surface transportation issues, educating a diverse array of current practitioners and future leaders in the transportation field, and encouraging implementation of relevant research results.