Volume 7 Numbers 2/3, 2004
ISSN 1094-8848

JOURNAL OF TRANSPORTATION AND STATISTICS

U.S. Department of Transportation
Research and Innovative Technology Administration
Bureau of Transportation Statistics

JOURNAL OF TRANSPORTATION AND STATISTICS

A PEER-REVIEWED JOURNAL

JOURNAL OF TRANSPORTATION AND STATISTICS

Volume 7 Numbers 2/3, 2004
ISSN 1094-8848

U.S. Department of Transportation
Research and Innovative Technology Administration
Bureau of Transportation Statistics

Bureau of Transportation Statistics

Our mission: To lead in developing transportation data and information of high quality and to advance their effective use in both public and private transportation decisionmaking.

The *Journal of Transportation and Statistics* releases three numbered issues a year and is published by the

Bureau of Transportation Statistics
Research and Innovative Technology Administration
U.S. Department of Transportation
Room 7412
400 7th Street SW
Washington, DC 20590
USA
journal@bts.gov

Subscription information

To receive a complimentary subscription:

mail Product Orders
Bureau of Transportation Statistics
U.S. Department of Transportation
Room 4117
400 7th Street SW
Washington, DC 20590
USA
phone 202.366.DATA
fax 202.366.3640
internet www.bts.dot.gov

Information Service

email answers@bts.gov
phone 800.853.1351

Cover and text design Susan JZ Hoffmeyer
Cover photo Photo by Elliott Linder
Close-up of Olivier Huc's motorcycle

The Secretary of Transportation has determined that the publication of this periodical is necessary in the transaction of the public business required by law of this Department.

JOURNAL OF TRANSPORTATION AND STATISTICS

Volume 7 Numbers 2/3
2004

Contents

Letter from the Editor-in-Chief

Dear JTS Readers,

This issue completes our seventh volume of JTS and seventh year of publication. As the new JTS Editor-in-Chief, I'd like to bring your attention to several changes in the journal that have taken place over the last few months. You can see on the front cover of this issue that the Bureau of Transportation Statistics (BTS) is now an office of the new Research and Innovative Technology Administration (RITA) within the Department of Transportation (DOT). RITA's staff come several groups within DOT: the Research and Special Programs Administration's Office of Innovation, Research, and Education, including the Volpe National Transportation Systems Center in Cambridge, Massachusetts, and the Transportation Safety Institute in Oklahoma City; the Secretary's Office of Intermodalism; and, of course, BTS. I do believe that JTS will have the opportunity to flourish within this new research environment.

In line with my vision for the journal, our editorial board endorsed a broader scope for JTS. We are just starting to include more applied papers. In the future, some of these papers will be from authors within RITA, keeping you, the readers, informed regarding DOT data and research directions. In this way, we hope to provide more information to researchers and planners, to applied and theoretical statisticians and economists, and to readers involved in numerous aspects of transportation analysis. As always, all papers we publish will undergo a thorough peer review.

You should note two new sections that make their appearance in this issue of the journal—Data Reviews and Book Reviews. Our Data Review editor, Jennifer Brady, will present synopses on new data releases from BTS. In this issue, she gives us an interesting discussion of BTS's *America's Freight Transportation Gateways* and the *Gateways Resource* CD, which contain detailed analysis and data on merchandise trade into and out of the United States.

Our second new section is Book Reviews. Vincent Yao, from the Institute for Economic Advancement at the University of Arkansas at Little Rock, joins the JTS staff as this section's editor and coordinator. Vincent's contact information appears at the beginning of the new section, so get in touch with him if you would like to author a book review in the future. We thank our editorial board for suggesting these additions—I hope you find them of value in your research.

Finally, this issue includes two new calls for papers. The first is a general call for papers reflecting our broader scope. The second is for a special JTS issue on transportation investment, which is scheduled for publication in 2007. Vincent Yao, along with Cletus C. Coughlin (Vice President and Deputy

Director of Research, Federal Reserve Bank of St. Louis), and Randall W. Eberts (Executive Director of the W.E. Upjohn Institute for Employment Research), will be guest co-editors. If you're interested in submitting a paper, check out the details on page 99.

BTS, as part of RITA, continues to pursue its mission of making the best possible transportation data available to improve the quality of transportation decisionmaking. JTS authors play a key role in this process by providing high-quality quantitative analysis as it applies to transportation issues. I hope the articles in the journal, along with the new sections that appear in this issue, will provide valuable information that you can use in your work.

PEG YOUNG, Ph.D.
Editor-in-Chief

Unregistration Rates for On-Road Vehicles in California

THEODORE YOUNGLOVE[1]
CARRIE MALCOLM[1]
THOMAS D. DURBIN[2*]
MATTHEW R. SMITH[3]
ALBERTO AYALA[4]
SANDEE KIDD[5]

[1] Statistical Consulting Collaboratory,
University of California, Riverside, CA 92521

[2] Bourns College of Engineering, Center for
Environmental Research and Technology (CE-CERT),
University of California, Riverside, CA 92521

[3] P.O. Box 38
Genesee Depot, WI 53127

[4] Research Division, California Air Resources Board,
1001 I Street, Sacramento, CA 95812

[5] Planning and Technical Support Division,
California Air Resources Board,
9500 Telstar Ave., El Monte, CA 91731

ABSTRACT

Motivated by the need to develop regional air pollution control strategies, a comprehensive field study was conducted throughout California to characterize the unregistration rate of light-duty vehicles in the state. Based on an analysis of more than 98,000 vehicle records, the average unregistered rate was found to be 3.38 ± 0.13%. This included vehicles unregistered for a period of less than three months (2.41% of the total), vehicles unregistered between three months and two years (0.95% of the total), and vehicles unregistered for more than two years (0.03% of the total). About half of the counties had unregistration rates between 2% and 4%, with most counties' rates below 5%. The unregistered fleet was more heavily weighted toward older vehicles than the registered fleet. Department of Motor Vehicles (DMV) unregistration rates were compared with the field study rates. DMV estimates ranged from 6.2% to 7.5%, which were higher than those obtained in the field study. It was also found that 1.7% of the vehicles identified in the survey were registered outside the state or the country.

Email addresses:
* Corresponding author T.D. Durbin—durbin@cert.ucr.edu
T. Younglove—tyoung@ucr.edu
C. Malcolm—ctmalcolm@sbcglobal.net
M. Smith—superengineer@netscape.com
A. Ayala—aayala@arb.ca.gov
S. Kidd—skidd@arb.ca.gov

KEYWORDS: Motor vehicle registration, motor vehicle emissions.

INTRODUCTION

Development of regional air pollution control strategies requires accurate estimation of the regional emissions inventory. Understanding and accurately portraying the in-use vehicle population is one of the most important aspects of obtaining accurate emissions inventory estimates. The registered vehicle population accounts for a majority of the vehicles on the road, however, unregistered and out-of-state vehicles represent an important proportion of the total inventory as well. Given that the unregistered vehicle population likely includes a higher percentage of older vehicles with emissions too high to meet Inspection and Maintenance (I&M) requirements, these vehicles may have a disproportionate effect on the emissions inventory.

To date, limited information is available on the contribution of unregistered vehicles to the emissions inventory. The California Air Resources Board's (CARB's) EMFAC[1] model provides estimates of current emissions for on-road motor vehicles in the state, based primarily on the population of vehicles registered with the Department of Motor Vehicles (DMV). Unregistered vehicle estimates are a recent addition to the EMFAC vehicle population. In making these estimates, CARB examined DMV records and found that approximately 7.4% of the total records for passenger cars were unregistered (CARB 2000). For the EMFAC2000 model, CARB added approximately 4.5 million vehicles to the in-use California vehicle population to account for vehicles in the process of being registered and those that were unregistered. CARB also estimated that 0.56% of the vehicle population that was chronically unregistered (unregistered for a period of more than 2 years) contribute 1% to the emissions inventory, depending on the pollutant and year being modeled (CARB 2000). CARB did not determine the total contribution from the unregistered vehicle population as a whole, however.

For modeling emissions inventories, states outside of California use the U.S. Environmental Protection Agency's MOBILE[2] model. MOBILE does not explicitly include the contribution of unregistered vehicles in its emissions rate estimates, although provisions are made for states to incorporate unregistered vehicles in their calculations.

Most previous studies of the population of unregistered vehicles have focused on California. Hunstad (1999) conducted a study to characterize uninsured motorists and provide estimates of the number of uninsured and unregistered vehicles. Hunstad examined DMV records and other estimates of unregistration, including studies by the California Energy Commission, estimates based on California Highway Patrol violations, DMV driver's license records, estimates based on surveys, and fatal accident reports. Using these collective studies, Hunstad came up with average yearly estimates of between 8.5% and 11.7% for unregistered vehicles.

Dulla et al. (1992) examined license plate number (LPN) records of vehicles in parking lots in the late 1980s and found total unregistration rates ranging from 8.3% to 9.3% with chronically unregistered vehicles (i.e., greater than 2 years) representing about 0.56% of the in-use fleet. Although the impact of unregistered vehicles on emissions inventories in other states is important, limited information exists on unregistration rates outside of California. North Carolina is one state that has developed a database from accident reports to provide estimates of the vehicle-miles traveled in urban areas by vehicles registered outside of the area (Norowzi 2003).

Data that are available on the population of unregistered vehicles still have considerable limitations. DMV database sources do not, for example, provide a good indication of whether the vehicles travel on the road. The DMV files can contain vehicles that may have become inoperative or may be located outside of the county of record. Because these vehicles are not part of the in-use fleet that would be operated in the designated area, their inclusion would result in an overestimate of the actual on-road fleet's emissions.

The most recent on-road unregistered vehicle population study was conducted over a decade ago in California (Dulla et al. 1992). Since that time, California added a requirement that vehicle owners show proof of insurance before a vehicle's registration can be issued or renewed. Furthermore, I&M

[1] EMFAC is short for Emission FACtor.
[2] MOBILE = Mobile Source Emission Factor Model.

testing procedures have been increased from an idle to a dynamometer test at 15 mph and 25 mph in areas that do not meet air quality standards. The possibility exists that these two requirements may contribute to the increased number of unregistered vehicles in the state, especially poorly maintained vehicles that cannot pass a smog test.

Given the potentially significant emissions inventory impact and the limitations of the current unregistered vehicle population estimates, improving the understanding of both the number and types of unregistered vehicles on the road is important. The objective of this work was to obtain a better understanding of the population and characteristics of unregistered vehicles. As the primary component of this study, a statewide field survey was conducted to provide an estimate of the unregistered vehicle population. As part of the survey, a database of more than 98,000 vehicle LPNs was obtained. This survey represents the most comprehensive study of vehicle unregistration rates to date and encompasses all regions of California. In addition to the total unregistration rate, the following information was collected:

- a breakdown of the time period of unregistered status into instantaneous (less than 3 months), long term (3 months to 2 years) and chronic (more than 2 years) categories by county for California;
- characteristics of unregistered vehicles including, but not limited to, model year; and
- the percentage and identity in each county of non-California vehicles or vehicles that originated out of county.

Durbin et al. (2002) presents the detailed results of this survey.

METHODOLOGY

A comprehensive field survey was conducted to determine the population of unregistered vehicles in California. The survey involved photographing license plates of vehicles that were parked in commercial parking lots to obtain registration and other information. Data were collected between June and December 2000.

A county-based stratified random sample of all California counties was conducted, with the sampling population and number of sites in each of the larger counties proportional to the county population. To ensure that the sample was demographically representative, each county was resolved to the zip code level, with zip codes selected randomly from the list of all zip codes in each county. Within the selected zip codes, as many commercial parking lots were surveyed as possible. Where security, sample team safety, or geographic size prevented complete sampling, locations were geographically balanced across the zip code. To ensure a reasonable distribution of destination types, a minimum number of sites in each county was sampled. In total, 409 zip codes were sampled during the field study, out of a total of 1,586 zip codes in California at the time of study. Because the zip codes included national parks and forests, sampling in those areas was performed as appropriate for a given county.

The sites for this study were restricted to destinations rather than residences. This provided a high probability that the vehicles captured in the survey were driven on a regular basis. Sites were also selected to represent a variety of different destinations. The primary collection sites were shopping malls, businesses, and retail stores, although the range of sites also included park and ride lots, medical facilities, and others. Overall, the different site types did not show significant variation in unregistration rates. In particular, all but one of the site types where at least 1,000 samples were collected had unregistration rates within 1% of the overall average (Durbin et al. 2002). It should be noted that this site selection and sampling methodology might not be fully representative of all vehicles distributed throughout the state. Families with multiple vehicles, for example, may use a specific vehicle for some applications such as shopping, while other vehicles may be used for longer trips or other tasks.

Consideration was given to sampling vehicles during actual driving on the road (i.e., with a license plate recorder) to obtain a broader cross section of vehicles. Because the primary purpose of this study was for support of development of the CARB EMFAC model, this approach was not chosen. In particular, it was thought that the fleet obtained

from sampling on the road would be more heavily weighted by vehicle-miles traveled and would be less consistent with the in-use population data presently used in CARB's EMFAC from the DMV.

To collect the data at each specific site, one to two photographers, using digital cameras with a reload time of less than a second between shots, took pictures while riding in a car as it slowly moved through a parking lot. Data from the photographic records were entered and compiled into a spreadsheet along with other information, such as the date and time of the site visit and a description of the site and its location (city, county, and zip code). The vehicle make and model were determined for roughly half of the vehicles photographed. The data were validated by double entry and cross-checking a 5% subsample as well as random spot checks of individual vehicles. The error rate for data entry of LPNs was consistently below 1%.

Given the nature of the rapid data collection in the field and the need to get large numbers of records, a percentage of the LPN photographs collected in the field were unreadable. The overall unreadable LPN rate averaged about 15% with a range from 1% or 2% to over 40% in some zip codes. Sites surveyed during rain events or near sundown made up the highest percentages of unreadable LPNs. Field teams generally attempted to collect 10% to 20% more photographs for each zip code to compensate for the expected unreadable percentage.

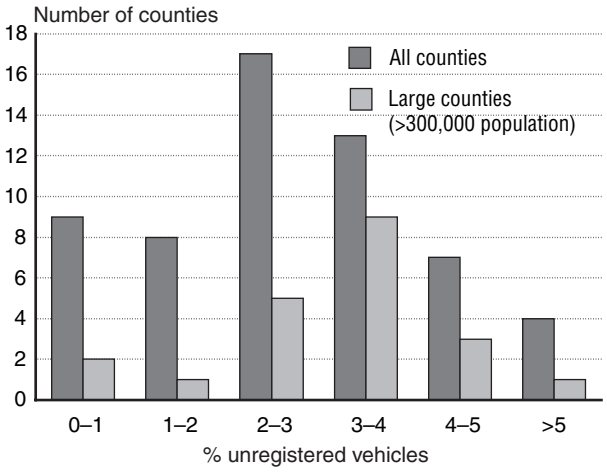
RESULTS

County Unregistration Rates

The field survey brought in more than 98,000 records. Table 1 presents data analysis for all counties with more than 1,000 samples and for the entire field survey sample population. (See Durbin et al. (2002) for detailed records for all counties).

Vehicles were considered to be unregistered if the year sticker was 1999 or older, regardless of the month, and registered if the year sticker was 2001. For vehicles with year 2000 stickers, the month of registration was evaluated against the time period when the vehicle was identified to determine the registration status. The percentage unregistered was calculated by dividing the number of unregistered vehicles by the sum of registered vehicles, unregistered vehicles, and dealer plates (registration is paid at the time of vehicle purchase, so vehicles with dealer plates were considered registered).

FIGURE 1 Percentage of Unregistered Vehicles in Large and All Counties



The overall average unregistration rate for all surveyed vehicles was 3.38 ± 0.13%, where the uncertainty represents the 95% confidence interval based on the sampling statistics (Vollset 1993). The unregistration rate ranged from 0% to 6.45% for different counties. Figure 1 presents the data for all the surveyed counties and for the most populous counties (population > 300,000). These data show that roughly 50% of the counties have unregistration rates ranging between 2% and 4%. Nearly all counties had unregistration rates below 5%. In general, larger counties had slightly higher unregistration rates than the overall distribution, with unregistration rates in larger counties generally ranging from 2% to 5%.

Counties with unregistration rates of less than 1% tended to be smaller, with sample sizes of fewer than 500 vehicles. In some small counties, the field data contained no unregistered vehicles. Alpine County has the highest rate of unregistered vehicles at 6.45%, but this figure may be due in part to the small sample size for that county. Data show that Calaveras, San Diego, and Madera counties have the next-highest unregistration rates of 5.22%, 4.99%, and 4.51%, respectively.

A one-way ANOVA test showed highly statistically significant differences in registration rates between different counties in the study ($p < 0.0001$).

TABLE 1 Registration Rates by County

County	Total	Registered	Unregistered	Instantaneous	Long term	Chronic	Dealer	Front	No plate	Out of state	Out of country	Unknown	% unreg
Los Angeles	25,835	20,153	749	535	211	3	317	2,146	15	214	3	2,238	3.53
San Diego	9,584	7,226	385	300	83	2	110	708	3	230	22	900	4.99
Orange	9,468	7,421	251	171	76	4	169	665	18	138	4	802	3.20
San Bernardino	4,417	3,115	117	72	44	1	89	314	11	148	0	623	3.52
Riverside	4,262	3,026	144	83	60	1	94	238	6	123	0	631	4.41
Santa Clara	4,109	3,084	91	55	33	3	74	271	6	43	0	540	2.80
Alameda	3,529	2,688	72	54	18	0	52	238	1	38	0	440	2.56
Sacramento	3,337	2,573	90	50	37	3	53	233	12	37	0	339	3.31
San Francisco	2,840	2,154	100	74	25	1	33	226	3	85	0	239	4.37
San Mateo	2,705	2,112	75	54	19	2	35	144	8	60	0	271	3.38
Contra Costa	2,597	1,929	42	30	12	0	44	131	3	21	0	427	2.08
Ventura	2,073	1,574	61	48	13	0	43	120	1	22	0	252	3.64
Fresno	2,059	1,652	67	53	14	0	33	146	0	2	0	159	3.82
Santa Barbara	1,464	1,144	37	27	10	0	20	74	0	32	0	157	3.08
Sonoma	1,404	1,100	28	17	11	0	7	54	1	7	0	207	2.47
Kern	1,401	1,042	46	37	9	0	16	130	1	4	0	162	4.17
Stanislaus	1,329	1,076	40	29	11	0	2	94	1	8	0	108	3.58
All counties	**98,817**	**75,168**	**2,677**	**1,906**	**749**	**22**	**1,302**	**7,091**	**108**	**1,599**	**111**	**10,761**	**3.38**

Key: *County* = the county where the data was collected, not necessarily the county where the vehicle was registered.
Registered = vehicle had a current California registration.
Unregistered = vehicle had an expired California registration.
Front = vehicle's LPN was captured from the front of the vehicle. Because California vehicles only have registration tags for rear license plates, no registration information is available from the front plate.
Dealer = vehicle's LPN was a paper plate or a dealership plate of a newly purchased vehicle being used until the issued license plate is received.
Unknown = vehicles for which either the photographic quality prevented identification of the month if the vehicle had a registration year of 2000 or the year sticker was missing.

The individual zip code data contained a wider range of unregistration rates than the overall county data. Kern, San Diego, Santa Cruz, and San Benito counties each had one zip code with an unregistration rate above 20%, although most of these zip codes had limited sample sizes. Only the San Diego zip code contained a significant sample size (507).

In analyzing the data, correlations between the unregistered vehicle population percentage and demographic variables such as household income were examined. Using data obtained from the 2000 census, regression analyses revealed the influence of household income, total county population, and renter/owner percentage on unregistration rates at both the county and zip code level, using the unregistration rate as a dependent variable. Zip codes with fewer than 25 vehicles were removed from this analysis, because the small number of samples makes estimation of the rate of registration highly variable. A statistically significant regression ($p = 0.004$, $R^2 = 0.250$) for household income was found at the county level. This regression was not statistically significant on the zip code level, however, and the low R^2 value of regression at the county level indicates that the relationship does not explain the majority of the variability observed in the unregistration rate. Statistically significant regressions of at least the 90% confidence level were found between unregistration rates and population at the county level ($p = 0.096$, $R^2 = 0.049$) and the registration zip code level ($p = 0.007$, $R^2 = 0.009$) but not the zip code where the vehicle was observed ($p = 0.658$). Again, the low R^2 values indicate that none of these relationships can account for most of the observed variability in unregistration status, although such research may be of interest in the future.

The overall 3.38% unregistered rate was broken down by the length of time a vehicle was unregistered. A total of 2.41% of the California licensed vehicles were classified as instantaneous (<3 months) unregistered. A total of 0.95% of the California licensed vehicles were classified as long-term (3 months to 2 years) unregistered. Chronic (>2 years) unregistered accounted for 0.03% of the California licensed vehicles.

A subset of vehicles operated in California were registered in other states or countries. As shown in table 1, this represents approximately 1.7% of the vehicles in this survey. In general, higher percentages of out-of-state vehicles were found in the border counties and in counties having zip codes with well-known tourist attractions. Border counties such as Del Norte, Sierra, Nevada, Alpine, Inyo, Imperial, and Kings County in the Central Valley all had relatively high proportions (>10%) of out-of-state vehicles. (Durbin et al. (2002) provide a map of out-of-state vehicle percentages by county.)

Comparisons with DMV Data

The data obtained from the field study were cross-referenced with two different DMV databases to determine the DMV registration status of the vehicles identified in the field survey. Because the field survey included only vehicles used typically on the road, this should represent a more accurate estimate of the unregistered population than a straight DMV run that could include a mixture of vehicles used infrequently or not at all. The DMV databases used for this comparison correspond to late 1998 and 2001.

Table 2 provides a comparison of the DMV results and the field survey registration status for the April 2001 DMV database. The results for the 1998 DMV database were similar and are presented elsewhere (Durbin et al. 2002). Overall, the DMV database shows a similar profile of unregistered vehicles, with roughly half as many long-term unregistered vehicles in comparison with instantaneous unregistered vehicles and smaller numbers of chronically unregistered vehicles. Differences were observed, however, for the various registration categories between the DMV database and the field study. For example, a number of vehicles identified as registered in the field were unregistered in the DMV database and vice versa.

Several possible explanations exist for the vehicles observed to be registered in the field survey, but subsequently found to be unregistered. First, differences between the time of the field survey and the time of the DMV run may account for the discrepancy. In particular, vehicles registered at the time of the field study may have subsequently fallen out of

TABLE 2 Cross-Tabulation of Registration Status with the 2001 DMV Database

Field study status	Total	DMV registered	DMV unregistered	DMV instantaneous	DMV long term	DMV chronic
Registered	70,418	65,425	4,993	3,630	1,156	207
Unregistered	2,459	2,124	335	107	201	27
Instantaneous	1,771	1,636	135	20	106	9
Long term	670	480	190	86	93	11
Chronic	18	8	10	1	2	7
Front	6,627	6,092	535	307	193	35
Unknown	9,153	8,397	756	136	573	47
Total	**88,657**	**82,038**	**6,619**	**4,180**	**2,123**	**316**

registration. The fact that a majority of the vehicles in this category were found to be instantaneously unregistered supports this conclusion. A second possibility is that some vehicles identified as registered in the field had stolen tags or switched LPNs, although this probably represents a smaller fraction of the vehicles.

Interestingly, many of the vehicles found to be long-term or chronically unregistered in the field study were also found to be registered. For vehicles found to be unregistered in the field but registered with the DMV, it is probable that the vehicle was registered subsequent to the field study. This would account for the large percentage of instantaneously unregistered vehicles in the field study that were found to be subsequently registered with the DMV. Given the small number of vehicles in this category, it is possible that the owners of these vehicles simply did not adhere their registration stickers. Another subset of vehicles were long-term unregistered in the field study but instantaneously unregistered in the DMV. This subset of vehicles could be attributed to a combination of factors including an unattached or stolen sticker in conjunction with a subsequent late registration.

Differences also exist in the total unregistration rate between the two DMV databases and the field study results. For the 2001 DMV database run, for example, the unregistration rate was approximately 7.5% compared with the 3.38% obtained from the field results. The 1998 DMV database shows a higher estimated unregistration rate of 6.2%. It should be noted that since the data entry error rate for the field study records was typically below 1%, we anticipate that this accounts for only a small portion of the observed discrepancy.

Unregistered Vehicle Model Year Distributions

The observed LPNs from the field study were also cross-referenced with DMV records to obtain the model year for all vehicles having readable LPNs. Figure 2 presents a model year distribution for all observed vehicles. The model year distribution is heavily weighted to newer vehicles, as expected.

Figure 3 shows the model year for vehicles unregistered for more than 3 months (i.e., long-term and chronic unregistered vehicles), using data from the field study. Figure 4 shows the percentage unregistered for each model year category. For model year 2000 vehicles, 1 out of 10 had a long-term unregistered status, causing the percentage for that model year to be high. When compared with figure 2, the unregistered vehicle population is more heavily weighted to the older model years.

Comparison with Inspection and Maintenance Data

To better understand the relationship between unregistration rates and the requirements that vehicles pass an I&M test for registration, field study records were cross-referenced with I&M records. To do this, the LPNs for all registered and unregistered vehicles (over 90,000) were sent to the Bureau of Automotive Repair (BAR). BAR provided I&M test results from mid-1996 to the present for all vehicles with a license plate match. The observed plates were matched with data from both the BAR90 and BAR97 programs.

FIGURE 2 Model Year of all Observed Vehicles

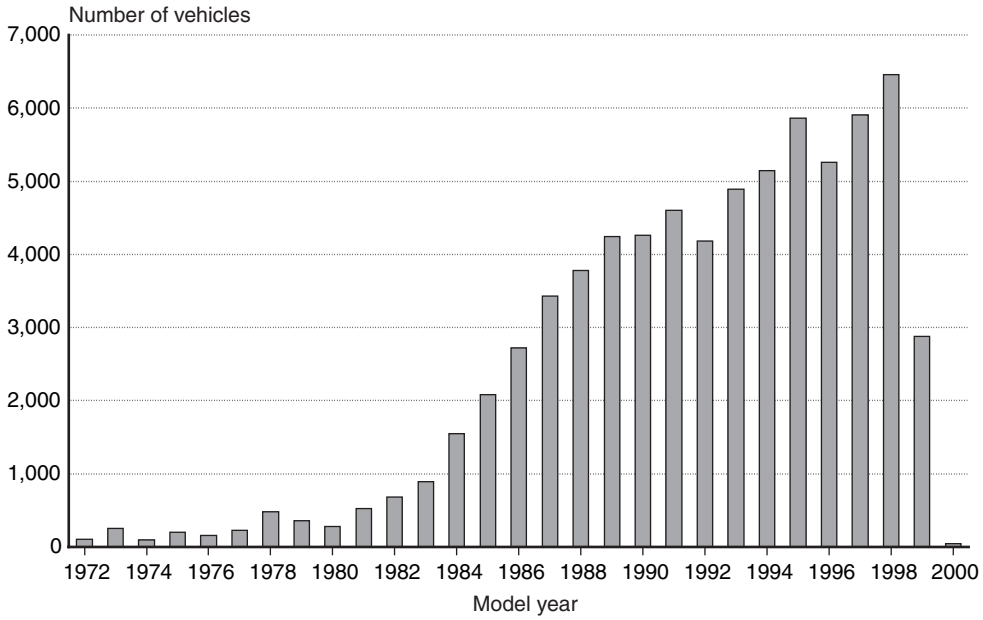
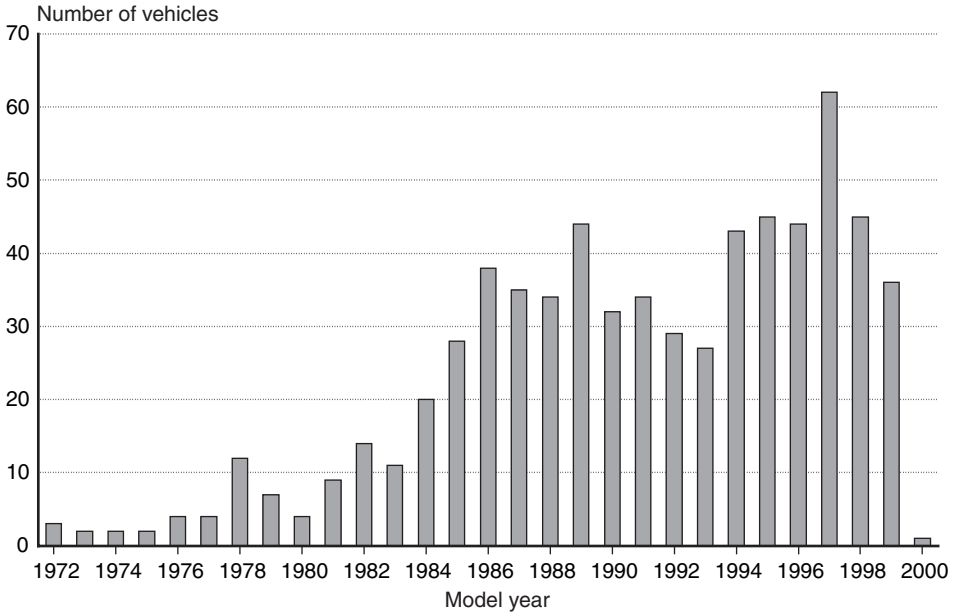


FIGURE 3 Model Year of all Unregistered Vehicles for Less Than 3 Months

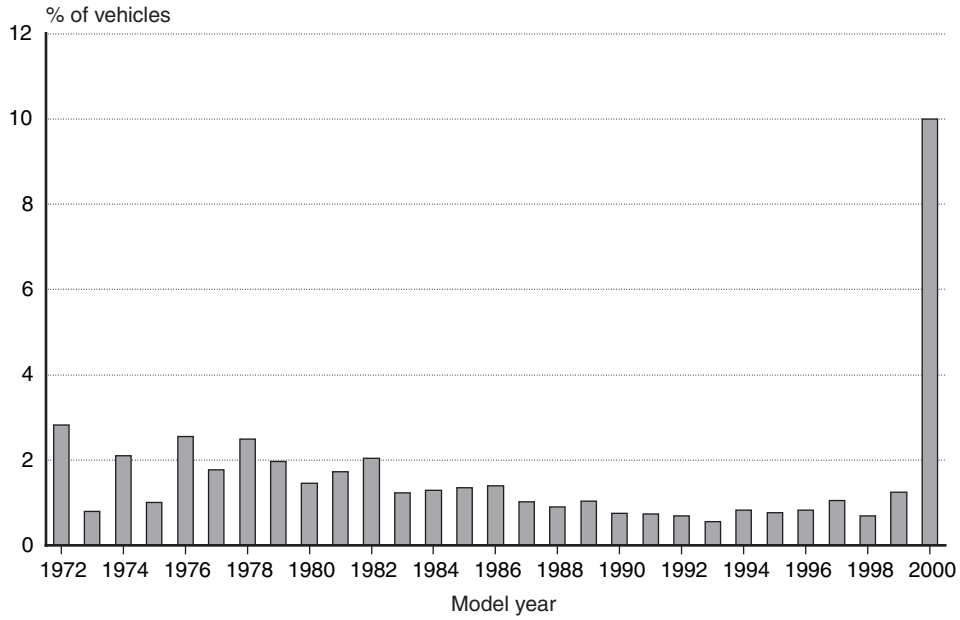


Note: An older version of the DMV database was used to determine the model year, hence, information on 1999 and 2000 model-year vehicles was limited.

The BAR90 and BAR97 programs represent two levels of I&M programs in California. The BAR90 program, an older emissions test, uses a two-speed idle test. The BAR90 program is used only in areas of California that do not require enhanced I&M. The BAR97 program, in use since 1998 in enhanced I&M areas of California, employs an accelerated simulation mode dynamometer test at 15 mph and 25 mph.

BAR rates data as "good" or "bad." Bad files can indicate unknown or aborted tests or records with incorrect entries such as the vehicle identification number (VIN). Because the data were matched based on LPNs instead of VINs, these bad records were initially included and the aborted and unknown tests were subsequently separated out. The cross-tabulation of the I&M data and the observed data yielded 66,436 matching records.

FIGURE 4 Percentage of Total Vehicles, by Model Year that are Unregistered for More Than 3 Months



Bad data accounted for only 1,400 records out of the 66,436 BAR matches.

Table 3 presents a comparison of registration status and I&M test results. The I&M test results are broken into six categories: passed, failed, gross polluter, tampered, test aborted, and unknown (where no results were recorded). The results show the majority of the vehicles in all categories passed their last I&M. A large number of unknown and aborted I&M results come from the BAR-classified bad data. Overall, the percentage of vehicles passing the last I&M drops with the length of time since the vehicle was last registered. A chi-square test of independence between smog check results and unregistration status showed that the differences in the overall failure rate for the registered vehicles and instantaneous and long-term unregistered vehicles were statistically significant at the 0.0001 level. Chronically unregistered vehicles were not included in this chi-square test due to the small number of samples. The test indicates that smog check results are not independent of observed registration status, with a higher proportion of unregistered vehicles failing the smog check test. Interestingly, three vehicles that were chronically unregistered when observed by the survey team in 2000 had since taken and passed an I&M test in 2001. One vehicle identified as unregistered in the field study was tested and found to be a gross polluter in 2000, indicating that the lack of registration for the vehicle may have been related to failing the I&M.

The I&M failures rates for the matched vehicles from the survey are less than the average failure rate for all vehicles taking the test in the state, which is slightly above 15% (CARB/BAR 2004). Because the fleet surveyed represents vehicles driven on a more regular basis, it is probable that the survey vehicles are newer than recorded during I&M. Furthermore, vehicles that are model years 1991 or newer have failure rates below the statewide average, while vehicles from the early 1980s have more than double the statewide average for failed inspections (CARB/BAR 2004). Vehicles that initially fail the I&M test can also pass a retest after repair to obtain registration.

DISCUSSION AND CONCLUSIONS

Overall, the results of this study showed a generally consistent average statewide unregistration rate of between 3.4% to 7.5%, with an additional 1.7% of the vehicles registered out of state. The DMV and field studies also yielded similar distributions of largely instantaneous unregistered vehicles, with smaller numbers of chronic and long-term unregistered vehicles. Although differences were observed in the specific unregistered populations, both show

TABLE 3 Cross-Tabulation of BAR Smog Check Test Results with Registration Status

Field status	Number	Passed		Failed		Gross polluter		Tampered		Aborted		Unknown	
		#	%	#	%	#	%	#	%	#	%	#	%
Registered	51,974	50,423	97.0	682	1.3	250	0.48	25	0.1	391	0.8	203	0.4
Instantaneous	1,449	1,385	95.6	30	2.1	13	0.90	2	0.1	12	0.8	7	0.5
Long term	601	549	91.4	18	3.0	19	3.16	2	0.3	10	1.7	3	0.5
Chronic	13	11	84.6	1	7.7	1	7.69	0	0.0	0	0.0	0	0.0
Front	4,921	4,716	95.8	91	1.9	44	0.89	4	0.1	44	0.9	22	0.5
Unknown	7,387	7,015	95.0	157	2.1	66	0.89	13	0.2	83	1.1	53	0.7
Total	**66,346**	**64,100**		**979**		**393**		**46**		**540**		**288**	

Key: BAR = Bureau of Automotive Repair.

trends that, on a continuous basis, approximately 3% to 8% percent of the on-road vehicle population is unregistered and needs to be accounted for in emissions inventory models.

The results of this study can be compared with those from a limited number of other estimates. BAR has conducted some analyses of remote sensing device (RSD) data collected throughout the state as part of various studies over the years (Amlin 2002). From this data, BAR found that vehicles unregistered for a period of more than one year comprised approximately 0.6% to 0.7% of the on-road fleet. Because the RSD data would tend to be more heavily weighted toward vehicle-miles traveled (VMT), the BAR estimate includes an adjustment to compensate for older unregistered vehicles that would be driven fewer miles. The BAR study also found that the VMT for unregistered vehicles is significantly lower than the VMT for registered vehicles.

For comparison with these results, the percentage of vehicles unregistered for more than one year was determined using the present data. Of the vehicles in the field study, 0.15% were unregistered for a period of more than one year, which was less than the BAR estimates. BAR also examined the DMV database and found that between 5.1% and 5.5% of the DMV population had delinquent registration but renewed at some point in time.

Dulla et al. (1992) conducted an earlier parking lot survey of unregistered vehicles in the South Coast Air Basin of Southern California during 1989. In their field study, these researchers found that approximately 9.3% of vehicles had expired license plates, with fewer than 2% of the vehicles unregistered for more than one year. These values are higher than the 3.38% overall unregistration with less than 1% unregistered for more than one year found in the present study.

In looking more specifically at the counties sampled in the Dulla et al. study, the unregistered rates in Los Angeles, Orange, Riverside, and San Bernardino Counties in this field study were 3.53%, 3.20%, 4.41%, and 3.52%, respectively; also lower than those found by Dulla et al. It is worth noting that the sample population in the Dulla et al. study was targeted more toward older vehicles, because the vehicles randomly sampled in parking lots are likely to be newer than the universe of vehicles found in the DMV database, and older vehicles are typically driven less. In fact, the sample population included a larger portion of older vehicles than the DMV population. Since the DMV data include a larger percentage of older vehicles than the on-road population, it is possible this estimate is biased high relative to the on-road population. In this regard, for vehicles less than 10 years of age, Dulla et al. (1992) found the unregistered population was 6.9%.

Hunstad (1999) also provided some estimates of unregistered vehicles based on available database sources including California Energy Commission data, California Highway Patrol (CHP) citations, driver's license records, surveys, and fatal accident reports. Estimates based on vehicle-not-at-fault in fatal accidents, CHP violations, and DMV records for vehicles unregistered less than one year were all in the range of 8% to 12% on an annual basis. Analysis of driver's license citations, on the other

hand, indicated a 3% to 4% unregistration rate. Citations appearing on driver's license records differ somewhat from other citation sources in that only convictions appear on an individual's driver's license record, not citations that are ruled unjustified by the court. In each of the categories examined by Hunstad, a number of factors could bias this estimate upward or downward. In general, these data should provide a rough estimate of the unregistered percentage. Not-at-fault fatal accident victims data are probably one of the less biased data sources, but include a large fraction of unknown license plates that had to be accounted for in the estimated unregistered percentage. This unknown category is for vehicles for which there is no identifying information, perhaps because the vehicle left the scene, or for which no DMV records could be found.

Surveys of unregistered ownership were also analyzed with unregistration rates found to be between 7% and 16%. For these surveys, it is important to note that estimates based on ownership of unregistered vehicles would be biased high relative to on-road population estimates if the vehicles are not driven on a regular basis. Data collected in surveys tend to support this hypothesis.

The results of the present study indicate that the population of unregistered vehicles was biased toward older vehicles relative to the total on-road fleet. To this extent, it is estimated that unregistered vehicles would make a disproportionate contribution to the emissions inventory on a population basis. Smog check records also indicate a higher percentage of failures for unregistered vehicles. The majority of the unregistered vehicles in all categories were found to pass their last smog check, however, indicating this is probably not the most significant contribution to the unregistered vehicle population.

Given the differences in unregistration rates found for various methodologies, further research should be conducted in this area. In particular, our results indicate there could be a range of as much as 3% to 8% for estimates of unregistered vehicles, with a wider range of estimates when other studies are also considered. With the higher emissions rates that can be assumed for the older unregistered vehicle fleet, this could represent a considerable uncertainty in emissions inventory estimates.

Within this study, more detailed analysis and comparison between the field study results and the DMV results could provide some important insights into this issue. In particular, for vehicles whose registration status does not match between the field study and the DMV results, the individual LPN records could be analyzed to better understand these differences. The analysis of the individual vehicle records would allow a better determination of whether the differences can be attributed to variations in the time period of the population snap shots, stolen stickers or license plates, errors in the database, or other factors.

ACKNOWLEDGMENTS

The authors wish to thank the following individuals for their valuable contributions to this project:

- Data collection—Hugo Galdamez, John Gaskins, Shayna Golbaf, Beth Lee, Don Pacocha, Alison Rowlen, Judy Swineford, and Sylvester Wheeler.
- Data entry—Jennifer Anderson, Warren Katzenstein, Lindsey Miller, Shelly Miller, Andrea Ruiz, Jessica Walters, and Ryan Wicks.
- Department of Consumer Affairs, Bureau of Automotive Repair, smog check data—David Amlin, Mark Duewel, Kathy Runkle, and Greg Sweet.

The authors also wish to thank the California Air Resources Board for their financial support under contract number 99-318 and the South Coast Air Quality Management District for its financial support.

REFERENCES

Amlin, D. 2002. California Bureau of Automotive Repair, personal communication.

California Air Resources Board (CARB). 2000. Technical Support Document for a Public Meeting to Consider Approval of Revisions to the State's On-Road Motor Vehicle Emissions Inventory. May.

California Air Resources Board and Bureau of Automotive Repair. 2004. Evaluation of California's Smog Check Program, presentation to the California Inspection and Maintenance Review Committee, January. Available at http://www.imreview.ca.gov/presentations/index.html.

Dulla, R.D., Y. Horie, and S. Sidawi. 1992. Unregistered Vehicle Study—Phase 1, report prepared for the California Air

Resources Board (Contract No. A866-163) by Sierra Research Inc., Sacramento, CA.

Durbin, T.D., T. Younglove, C. Malcolm, and M.R. Smith. 2002. Determination of Non-Registration Rates for On-Road Vehicles in California: Final Report, prepared for the California Air Resources Board (Contract No. 99-318) by the University of California, Riverside, Bourns College of Engineering, Center for Environmental Research and Technology. March.

Hunstad, L. 1999. Estimating Uninsured Vehicle and Unregistered Vehicle Rates: Sensitivity to Data and Assumptions, report by the California Department of Insurance. July.

Norowzi, B. 2003. North Carolina Department of Transportation, Office of Statewide Planning, personnel communication.

Vollset, S.E. 1993. Confidence Intervals for a Binomial Proportion. *Statistics in Medicine* 12:809–824.

A Bayesian Network Model of Two-Car Accidents

MARJAN SIMONCIC

Institute for Economic Research
Kardeljeva pl. 17
Ljubljana, Slovenia
Email: simoncicm@ier.si

ABSTRACT

This paper describes the Bayesian network method for modeling traffic accident data and illustrates its use. Bayesian networks employ techniques from probability and graph theory to model complex systems with interrelated components. The model is built using two-car accident data for 1998 from Slovenia, and inferences are made from the model about how knowledge of the values of certain variables influences the probabilities for values of other variables or outcomes (e.g., how seat-belt use affects injury severity). An advantage of the Bayesian network method presented here is its complex approach where system variables are interdependent and where no dependent and independent variables are needed.

INTRODUCTION

This paper presents a Bayesian network model of two-car accidents based on different factors that influence accident outcomes. The outcomes examined are "fatality or serious injury" and "other outcomes." Influencing factors include:

1. road characteristics (e.g., roadway, pavement),
2. traffic flow characteristics,

KEYWORDS: Road accidents, modeling, Bayesian networks, machine learning.

3. time/season factors (e.g., weather, season, weekday, daytime, rush hour),
4. characteristics of the people involved in an accident (e.g., age, sex, driving experience, health status, intoxication),
5. use of protective devices (seat belt, air bag),
6. types of vehicles (especially their crash resistance design), and
7. speed of the vehicles involved.

Besides these factors, other stochastic influences affect the likelihood of an accident and its outcome. The factors presented above are highly interrelated. For instance, road conditions are influenced by the weather. Traffic flow depends on the time of the day, whether it is a weekday or weekend, and weather conditions. The characteristics of people involved (e.g., age, sex, experience) can often be related to the speed of the vehicles in an accident and the use or non-use of seat belts. The outcome of an accident is, by and large, dependent on the speed of the vehicles involved.

A large road accident dataset was used to model the interdependence among the variables related to accidents ("knowledge of the subject") and the dependence of the outcome on the relevant variables. Bayesian networks[1] seem particularly useful for representing knowledge in domains where large sets of interrelated (and relevant) data are available. They are based on a combination of probability theory, which deals with uncertainty, and graph theory, which deals with complexity (interrelatedness). These networks are an important tool in the design and analysis of machine learning algorithms and are based on the idea of modularity whereby a complex system is built by combining simpler parts. Probability theory connects parts and ensures the consistency of the system as a whole while providing the possibility of interfacing the models with the data (see Jordan 1999). This paper aims to show that Bayesian networks can also prove their potential in modeling road accidents.

[1] Some similar or synonymous concepts are graphic models, belief networks, probabilistic networks, independence networks, causal networks, and Markov fields.

BAYESIAN NETWORKS

A Simple Example of a Bayesian Network

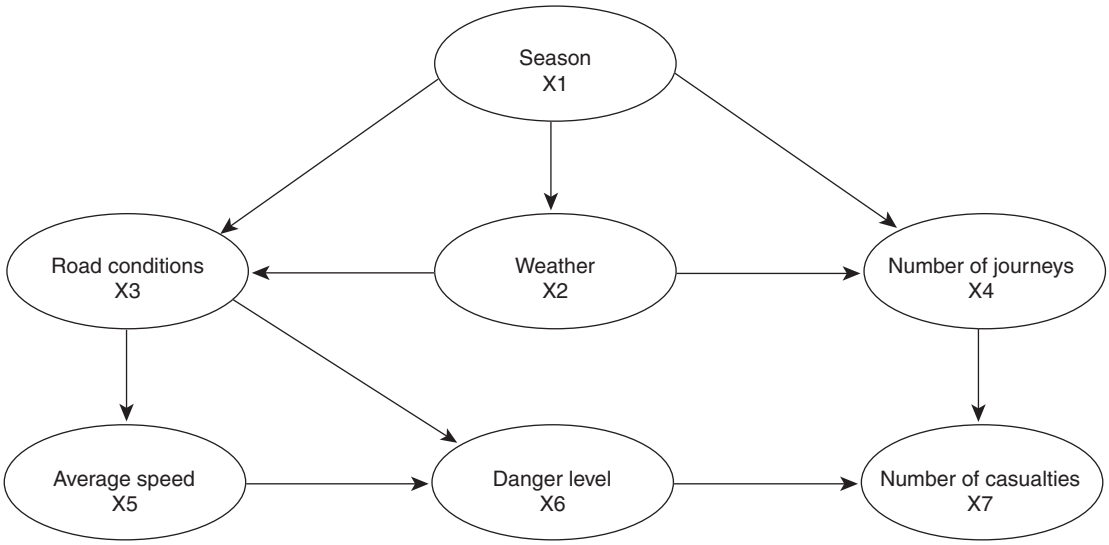
This section presents a simple Bayesian network for road accidents. The example is merely for illustrative purposes and is not intended to present a valid model. The aim is to introduce the concept of Bayesian networks by example.

Using a given geographic area, the number of road accident casualties per day can be schematically explained. Many factors are interrelated: the number of road casualties depends on how many trips car drivers took in the area and the danger level; the number of trips is related to weather conditions and the season (e.g., summer means more vacation travel); season and weather are also related; the level of danger is influenced by the average speed of vehicles on the roads and on road conditions (e.g., a slippery road); and road conditions depend on the weather and season and influence the average speed and level of danger. Figure 1 presents these relationships in a directed acyclic graph where the nodes correspond to different variables that are characteristic of the given domain under consideration. Links[2] in the graph represent dependence between variables, and acyclic means that there is no node from which it is possible to follow a sequence of (directed) links and return to the same node.

Let us suppose that all variables can only take on a finite number of discrete values. We are interested in identifying the probabilities of different events expressed in given values for all variables. This can be expressed with a joint probability distribution over all possible events in the given domain. The number of possible events grows exponentially with the number of relevant variables and, therefore, the joint probability function approach quickly becomes unmanageable. Bayesian networks can streamline the process, because they are a compact way of factoring the joint probability distribution into local, conditional distributions that reduce the number of multiplications necessary to obtain the probability of specific events.

[2] In Bayesian network literature, the terms *vertex* and *edge* are sometimes applied instead of *node* and *link*.

FIGURE 1 Example of a Bayesian Network



If we interpret the Bayesian network in probabilistic terms, the related joint distribution function over a given domain can be written (described by n variables) with the product[3]:

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | pa(x_i)) \qquad (1)$$

where X_i is the variable and x_i is its value; $Pa\ (X_i)$ is the set of variables that represents X_i's parents[4] and $pa(X_i)$ is a vector of actual values for all parents of X_i. Let us note here the general validity of the chain rule formula:

$$P(x_1, x_2, \ldots, x_n) = P(x_1)P(x_2|x_1) \bullet$$
$$P(x_3|x_1, x_2) \ldots P(x_n|x_1, x_2, \ldots, x_{n-1})$$

From our example in figure 1, we have:

$$P(x_1, x_2, \ldots, x_7) = P(x_1)P(x_2|x_1) \bullet$$
$$P(x_3|x_1, x_2)P(x_4|x_1, x_2)P(x_5|x_3) \bullet$$
$$P(x_6|x_3, x_5)P(x_7|x_4, x_6)$$

Aside from the global semantics reflected in equation (1), there is also a local meaning related to a Bayesian network. From figure 1, we see:

$$P(x_4|x_1, x_2, x_3) = P(x_4|x_1, x_2)$$

where X_4 is independent of the variable X_3 given X_1 and X_2 (reflecting the fact that X_3 is not among the parents of X_4). These local semantics are very useful for constructing a Bayesian network. Here, only direct causes (or predispositions) are selected as the parents of a given variable, which leads to the automatic fulfillment of local independence conditions.

Links in Bayesian networks may have different meanings. If we have a link from node A to node B, this could mean:

1. A causes B,
2. A partially causes or predisposes B,
3. B is an imperfect observation of A,
4. A and B are functionally related, or
5. A and B are statistically correlated.

This paper employs the second meaning of a link.

Bayesian networks for a certain domain can be used for inference purposes. With the network in figure 1, we will illustrate the meaning of inference and also show the difference between a Bayesian network model and better known classical models, such as logistic regression. After a product specification (equation (1)) of a joint probability distribution is obtained, the probability of any event in the domain can be expressed. Conditional events where certain variables have known values are especially interesting. This type of probabilistic inference is called a belief update. An example for the domain represented in figure 1 is the following:

$$P(X_3 = 'slippery' | X_7 = 'high')$$

$$= \frac{P(X_7 = 'high', X_3 = 'slippery')}{P(X_7 = 'high')} =$$

$$\sum_{X_1}\sum_{X_2}\sum_{X_4}\sum_{X_5}\sum_{X_6} P(x_1, x_2, x_3 = 'slippery', x_4, x_5, x_6, x_7 = 'high')$$

$$/ \sum_{X_1}\sum_{X_2}\sum_{X_3}\sum_{X_4}\sum_{X_5}\sum_{X_6} P(x_1, x_2, x_3, x_4, x_5, x_6, x_7 = 'high')$$

[3] The probability of the event A is denoted by $P(A)$.

[4] Node A is the parent of node B if there is a link from A toward B in the graph.

For illustrative purposes, we have assumed that one possible value of the variable X_3 (road conditions) is "slippery." This variable can also take on other values. A similar description holds for variable X_7. This expression can be further simplified, but this is unnecessary here.

Let us now illustrate the difference between the Bayesian network model and the classical logistic regression (for logistic regression see Agresti (1990) or Hosmer and Lemeshow (2000)). The most significant difference is that with logistic regression the model's dependent and independent variables must be chosen; while, with the Bayesian network model, all variables are treated equally. The logistic regression has a response (or dependent) variable Y that is a categorical variable with J ($J \geq 2$) classes and a vector X (with p components) of explanatory (or independent) variables that are also categorical[5] variables. Here, Y could be the number of casualties (with $Y = 1$ for "high" and $Y = 0$ for "other"). The components of vector X could be the six other variables from figure 1. The generalized logit model can be put in the following way:

$$\log \frac{P(Y=1|X)}{P(Y=0|X)} = \beta_0 + \sum_{k=1}^{p} \beta_k x_k \qquad (2)$$

If the attributes X are also 0/1 variables, then the following formula is valid[6]:

$$\frac{P(Y=1|x_1,x_2,\ldots,x_k=1,\ldots,x_p)P(Y=0|x_1,x_2,\ldots,x_k=0,\ldots,x_p)}{P(Y=0|x_1,x_2,\ldots,x_k=1,\ldots,x_p)P(Y=1|x_1,x_2,\ldots,x_k=0,\ldots,x_p)} = \exp(\beta_k); \qquad k = 1, 2, \ldots, p$$

The expression is called the odds ratio and allows an easy interpretation of the estimated parameters.[7] In the logit model for figure 1, $\exp(\beta_k)$ is the odds that the number of casualties will be high in the circumstances given by variable $x_k = 1$ relative to the odds that the number of casualties will not be high in the circumstances given by variable $x_k = 0$.

[5] In a general logistic regression, they are not limited to only these types of variables.

[6] A similar interpretation is possible if we have categorical variables with more than two values.

[7] Explanatory variables can be interdependent. Their interdependence plays a role in the estimation of these parameters (see chapter 2 in Hosmer and Lemeshow 2000).

It is obvious that the model shown in equation (2) does not explicitly take into account eventual interdependence between variables of X, nor does it allow for an estimation of other probabilities that could be of interest (e.g., the belief update given as an example for the network in figure 1). Interdependences among variables in a Bayesian network are explicit and represent a distinguishing feature of the method.

The general problem of computing posterior probabilities (or of a belief update) for large and structurally more complex Bayesian networks is computationally very demanding (more precisely: NP-hard). The computational burden was the reason that the inference in Bayesian networks was initially limited only to special types of structures, namely tree-structured networks. Later, efficient algorithms were proposed for more general types of network structures (Lauritzen and Spiegelhalter 1988; Zhang and Poole 1996).

Formal Definition of Bayesian Networks

Bayesian networks contain qualitative (structural) and quantitative (probabilistic) parts. The qualitative part is based on statistical independence statements and can be represented by a directed acyclic graph. The nodes are related to random variables of interest for a given domain, while the links correspond to a direct influence among the variables. The quantitative part is captured by local probability models, given by a set of conditional probability distributions. Both the qualitative and quantitative parts of the Bayesian network uniquely represent the joint probability distribution over a domain. The definitions follow.

Definition 1. A Bayesian network B is a triplet (X, A, P) where:

1. X is a set of nodes
2. A is a set of links that, together with X, represent a directed acyclic graph:
   $G = (X, A)$
3. $P = \{P(x|pa(x)):x\varepsilon X\}$

where $Pa(X)$ is the set of parents of X, and $pa(x)$ is its instantiation.[8] P stands for probability.

[8] When the state of a variable is known, we say that it is instantiated. We have an instantiation of a set of variables if each variable is instantiated (Jensen 2001).

It is clear that P is the set of conditional probabilities for all variables, given their parents. From definition 1, the conclusion can be drawn that nodes and variables are used interchangeably. Variables in a Bayesian network are called nodes when we speak about the graph.

Graph G corresponding to a Bayesian network has to be acyclic. If cycles were allowed, the feedback influence would be enabled. It is well known that feedback cycles are difficult to model quantitatively and no calculus has been developed for the Bayesian network to cope with these.

The notion of conditional independency is a basic concept of Bayesian networks. We say that (random) variables A and B are independent given the variable C if the following is true:

$P(A|B,C) = P(A|C)$

This means that if the value of variable C is known, then knowledge of B does not alter the probability of A.

The Bayesian network provides a graphic representation of many independency relationships that are embedded in the underlying probability model. No formal definitions are provided here, but it should be understood that the mathematical conception of d-separation is fundamental relative to independence (Jensen 2001).

The next definition gives the global interpretation of Bayesian networks.

Definition 2. The prior joint probability P_B of a Bayesian network B is defined by the following expression:

$$P_B(X) = \prod_{x \in X} P(x|pa(x))$$

The factorization in definition 2 rests on a set of local independence assumptions, asserting that each variable is independent of its predecessors[9] in the network, given its parents. The opposite is also true. We can use the interdependence in constructing Bayesian networks from expert opinion, because selecting as parents all the direct causes of a given variable satisfies the local conditional independence conditions (Pearl 2000).

For the Bayesian network from figure 1, the prior joint probability is equal to:

$$\begin{aligned} P_B(x) &= P_B(X_1 = x_1, X_2 = x_2, \ldots, X_7 = x_7) \\ &= P(X_1 = x_1)\, P(X_2 = x_2 | X_1 = x_1) \bullet \\ &\quad P(X_3 = x_3 | X_1 = x_1, X_2 = x_2) \bullet \\ &\quad P(X_4 = x_4 | X_1 = x_1, X_2 = x_2) \bullet \\ &\quad P(X_5 = x_5 | X_3 = x_3) \bullet \\ &\quad P(X_6 = x_6 | X_3 = x_3, X_5 = x_5) \bullet \\ &\quad P(X_7 = x_7 | X_4 = x_4, X_6 = x_6) \bullet \end{aligned}$$

When we have a joint probability distribution defined on a set of variables X, we can calculate the probability distribution of any subset S of X. This calculation is called marginalization and is very useful in inference exercises on Bayesian networks.

Definition 3. Let S be a subset of the set of variables X. The marginal probability $P_B(S)$ is defined by

$$P_B(S) = \sum_{y \in X - S} P_B(y)$$

Let us now suppose that some variables have specific values. In our example from figure 1, variables X_7 and X_3 may be observed to have values "high" (X_7) and "slippery" (X_3). If $Y \subseteq X$ is the set of variables with actual (observed) values, Y_0 is the corresponding vector of values and $X_1 \subseteq X$ is the set of variables of interest $(X_1 \subseteq X - Y)$, then the following definition of posterior probability is useful.

Definition 4. The posterior probability

$P_B(X_1 | Y = Y_0)$ of X_1

is defined by the expression

$$P_B(X_1 | Y = Y_0) = \frac{P_B(X_1, Y = Y_0)}{P_B(Y = Y_0)}$$

THE MODEL

Data

This paper focuses on road accidents in which two car drivers were involved. The empirical part is based on data from the road accidents database assembled by the Slovenian Ministry of the Interior from police reports. For the model, 1998 data containing 36,704 Slovenian police accident reports were used. From this total, 17,558 (48%) were of the selected type. To illustrate the risk of Slovenian drivers being involved in a two-car accident, some basic data show that, in 1998, 797,855 cars were registered in Slovenia (the country has 2 million inhabitants). Because we are looking at accidents involving two cars, we know that approximately

[9] A is a predecessor of B if a directed path (a sequence of links) exists from A to B.

4% of the Slovenian car fleet was involved in accidents of this type that year.

Table 1 presents data on two-car accidents for selected variables. Variables from Accident_type to Cause (the first column of table 1) are related to the accident, while variables from Age to Injury are related to the drivers.[10] The share of accidents that resulted in a fatality or serious injury of at least one person is 1.9%. Over 70% of accidents occur in built-up areas and more than half happen in good weather and under normal traffic conditions. Among participants, the lion's share corresponds to drivers 25 to 64 years old, yet the share of drivers under 25 years of age is also relatively high (23%). For drivers involved in accidents, a significant proportion has less than one year of driving experience (12.9%). Only a small share of drivers involved in accidents was intoxicated (4.3%).

Bayesian Network Estimation

A Bayesian network for a given domain can be estimated using different approaches. This paper uses a template model that should not vary from one problem to another. Our purpose here is to estimate a fixed Bayesian network over a given set of variables, obtained by a combination of expert judgment and empirical data. Specifications for some alternative possibilities for estimating a Bayesian network are presented below.

A difficult part of building a Bayesian network is quantifying probabilities, which can be derived from various sources:

1. from domain experts (subjective probabilities),
2. from published statistical studies,
3. derived analytically, or
4. learned directly from raw data.

This paper uses the last option, mainly because of the availability of a relatively large database.

Sometimes the process of learning the structure of a Bayesian network (if necessary) may be even more difficult than quantifying probabilities. According to the structure, models can be classified as those with a known structure or those with an unknown structure. We experimented with both options.

There are basically two different approaches to learning the structure of a Bayesian network from data: 1) search and scoring methods and 2) dependency analysis methods. In the first approach, different scoring criteria are used for evaluating competing structures. Two of the well-known methods of this type are the Bayesian scoring method (Cooper and Herskovits 1992) and the minimum description length method (Lam and Bacchus 1994). Because learning a Bayesian network structure by a search and score approach is NP-hard, different heuristic searches have been proposed. Algorithms from the second group try to discover the dependences among variables from data and then use them to infer the structure. During this process, a conditional independence test, usually based on the concept of mutual information of two nodes (variables), X and Y, is used

$$I(X,Y) = \sum_{x,y} P_e(x,y) \ln \frac{P_e(x,y)}{P_e(x)P_e(y)}$$

In this expression, P_e denotes the observed relative frequencies in the dataset. Conditional mutual information is defined analogously:

$$I(X,Y|Z) = \sum_{x,y,z} P_e(x,y,z) \ln \frac{P_e(x,y|z)}{P_e(x|z)P_e(y|z)}$$

Z can be a single node or a set of nodes. Mutual information I is non-negative and equal to 0 when X and Y are conditionally independent. The higher the mutual information, the stronger the dependence between X and Y. In heuristic algorithms a certain threshold ε is usually used: if $I(X,Y)$ is smaller than ε, then X and Y are taken as marginally independent. Similarly, if $I(X,Y|Z)$ is smaller than ε, we consider X and Y as conditionally independent given Z.

All these methods can be expected to find the correct structure only when the probability distribution of the data satisfies certain assumptions. But generally both types of methods find only approximations for the true structure.

According to the available data, models for learning Bayesian networks can be classified into those with complete data available or those with incomplete data available. In the first case, all variables are

[10] Passengers are taken into account only indirectly. A fatal accident may mean that both drivers were only injured, but at least one passenger was killed.

TABLE 1 Data on Two-Car Accidents in Slovenia: 1998

Variable	Values	Frequency	Relative frequency (%)
Accident_type	Fatality or serious injury	654	1.9
	Other	34,462	98.1
Alco12[1]	No	32,164	91.6
	Yes	2,952	8.4
Weekday	After weekend	5,442	15.5
	Before weekend	6,338	18.0
	Weekend	9,140	26.0
	Workday	14,196	40.4
Settlement	No	10,156	28.9
	Yes	24,960	71.1
Weather	Bright	20,088	57.2
	Cloudy	9,094	25.9
	Fog	454	1.3
	Rainy	4,264	12.1
	Snow	882	2.5
	Other	334	1.0
Traffic	Dense	5,530	15.7
	(Grid)lock	122	0.3
	Normal	20,214	57.6
	Sparse	8,830	25.1
	Unknown	420	1.2
Roadway	Dry	24,142	68.8
	Ice	546	1.6
	Other	220	0.6
	Slippery	1,938	5.5
	Snow	698	2.0
	Wet	7,572	21.6
Night	No	26,658	75.9
	Yes	8,458	24.1
Cause	HI Inappropriate speed	4,956	14.1
	OS Other	1,472	4.2
	PD Failing to give way	7,244	20.6
	PR Wrong overtaking	1,520	4.3
	PV Car maneuvers	8,296	23.6
	SV Wrong side/direction	4,754	13.5
	VR Unsuitable safety distance	6,874	19.6
Age	≤ 24	8,008	22.8
	25–64	25,538	72.7
	65–inf	1,570	4.5
Experience	0–1	4,530	12.9
	1–5	7,000	19.9
	6–10	7,235	20.6
	11–inf	16,351	46.6
Alcohol	No	33,593	95.7
	Yes	1,523	4.3
At-fault driver	No	16,651	47.4
	Yes	18,465	52.6
Sex	Female	9,036	25.7
	Male	26,080	74.3
Belt use	No	4,342	12.4
	Yes	30,774	87.6
Injury	Fatality or serious injury	247	0.7
	Other	34,869	99.3

[1] Alco12 (defined for each accident) is determined by the value for the variable Alcohol (defined for both drivers involved in the accident). Its value is "Yes" if at least one of the two drivers was intoxicated and is "No" otherwise.

Source: Database of the Slovenian Ministry of the Interior.

observed for all instances in the database while, in the second case, values for some variables may be missing or some variables may not even be observed (hidden variables). Because the available database used for this paper contains complete data, the first possibility is relevant.

Variables Considered in the Model

Some conditions of an accident may be called exogenous. They are tied to the accident and happen without the volition or action of the drivers involved. Variables from table 1 in this category are:

1. weather condition,
2. weekday,
3. settlement (whether an accident occurs in a built-up area or not), and
4. daytime (whether an accident occurs during the night or day).

These external conditions influence some internal and objective conditions also tied to the accident, such as traffic and the roadway. For each accident, these conditions are also exogenous.[11]

Besides these internal and objective conditions, there are also internal subjective (and not volitional) conditions that relate to the drivers involved:

1. age and sex,
2. driving experience,
3. intoxication (alcohol), and
4. use of a seat belt.

Objective and subjective internal conditions influence the cause of an accident. The particular cause further influences the outcome of the accident. Here, only two types of accident outcomes are considered: a fatality or serious injury, and other.[12] Subjective internal conditions and the cause of an accident influence the type of driver injury.

Different network structures can reflect these conditions. In the process of finding a suitable network structure, we experimented with PowerConstructor. PowerConstructor (Cheng et al. 2001) is a computer program that can estimate the Bayesian network structure if a database of cases is available. The method (Cheng et al. 1997) used in PowerConstructor for comparing competing structures is of the dependency analysis type and requires $O(n^4)$ conditional independence tests (n being the number of variables). The program is able to take into account additional restrictions on variables (e.g., partial ordering, forbidden links, roots, leaves, and causes and effects).

For this research, external variables and the variables related to the driver (e.g., age, sex, and experience) were among the root nodes (links can only point out of such nodes). Variables relating to the type of accident and the drivers' injuries were put among the leaf nodes (links can only point into such nodes). The variable related to the fault of the two drivers involved was also put among the leaves. PowerConstructor produced results pretty much as anticipated, except for some links that were missing.

Our anticipation was also based on some relevant findings from the literature. Kim (1996) analyzed the differences between male and female involvement in motor vehicle collisions in Hawaii and found that male drivers are:

1. 4 times more likely than female drivers to not wear a seat belt,
2. 3.6 times more likely than female drivers to be involved in alcohol-related collisions,
3. 2 times more likely than female drivers to be involved in speed-related collisions, and
4. 1.3 times more likely than female drivers to be involved in head-on collisions.

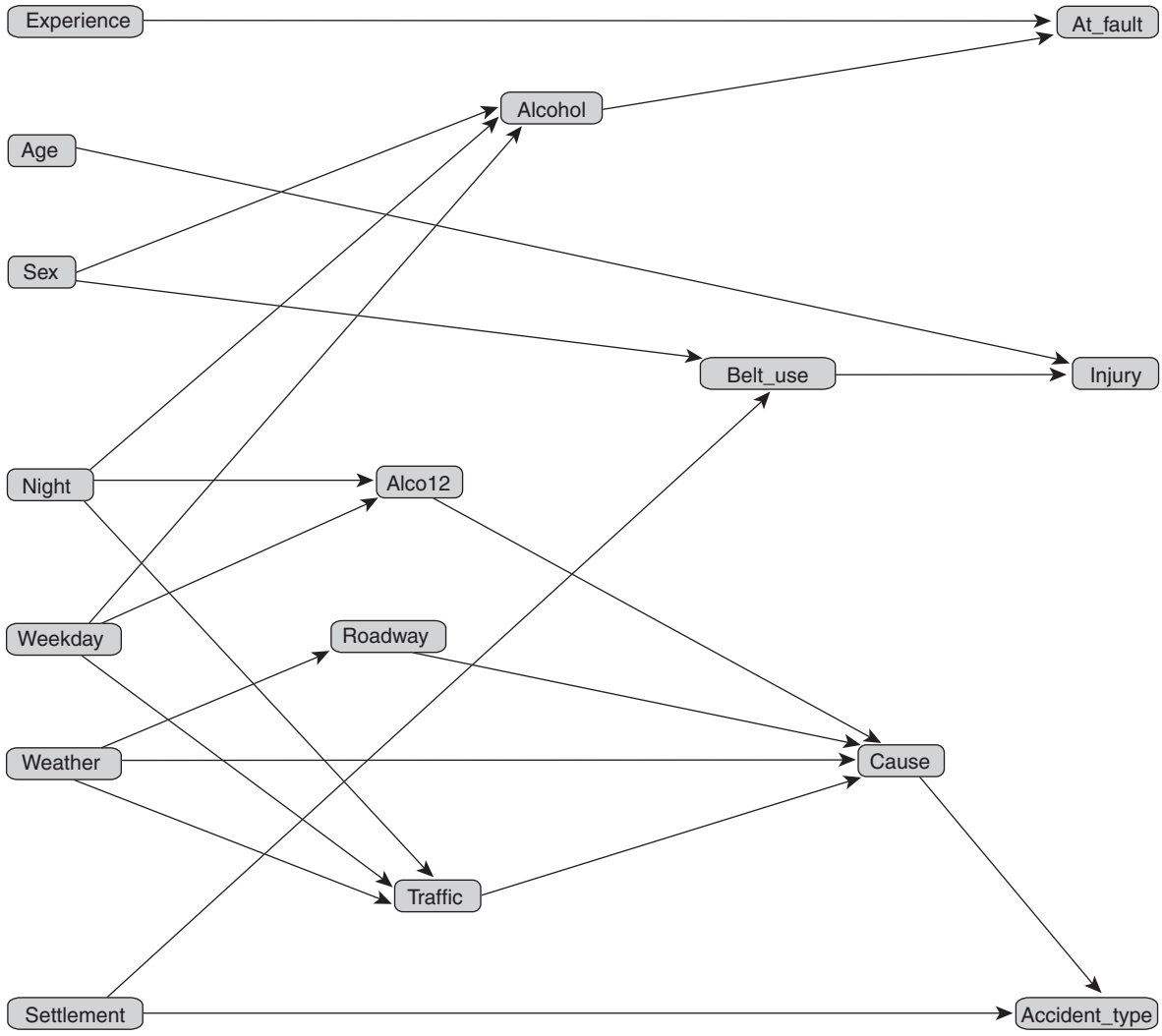
For the relationship between road accident severity and recorded weather, Edwards (1998) based her conclusions on data from police reports and found that:

1. accident severity decreases in rain as compared with good weather,
2. accident severity in fog shows geographical variation, and
3. evidence for accident severity in high winds is inconclusive.

[11] It is assumed that an individual driver does not significantly influence traffic conditions.

[12] The variable Accident_type is related to the accident, while the variable Injury is related to the driver. This presents no problem for an analysis with Bayesian networks.

FIGURE 2 The Network Structure



It is also well known that older drivers are more likely to be killed if involved in a fatal crash than younger drivers. Based on these results and common sense, additional restrictions for PowerConstructor included the following links:

1. Age → Injury (older drivers are expected to be more prone to serious injuries than younger drivers)
2. Seat belt → Injury (drivers not wearing a seat belt are likely to be more vulnerable)
3. Experience → At-fault driver (drivers with little driving experience are more likely to be at fault)
4. Sex → Seat belt use
5. Sex → Alcohol
6. Alcohol → At-fault driver

The resulting network is presented in figure 2. It is evident that only a small number of all theoretically possible interdependences was found to be important.

Weekday, daytime, and weather conditions influence traffic. An assumption was made that the share of intoxicated drivers is greater for accidents that happen at night than during the day. Only weather influences road conditions.[13] The type of accident and the use of a seat belt also depend on whether an accident happens in a built-up area or not (settlement variable). A smaller share of drivers wearing a seat belt in built-up areas was expected.

Figure 2 also takes into account the different characteristics of drivers. Drivers with little driving experience are more likely to be at fault in an accident than more experienced ones. There are also significant differences between men and women, with women being more likely to use seat belts than men. On average, older drivers are more prone to serious injuries than younger ones.

The central variable in figure 2 is the cause of an accident,[14] which is influenced by road, weather, and traffic conditions and by the variable related to driver intoxication. Finally, the outcome of an accident (defined as the most serious injury to participants in an accident) is largely conditioned by the cause of the accident.

[13] New variables could have been added here but were not in order to maintain a more manageable total number of variables.

[14] This is partly conditioned by the large number of possible states (seven) and by the method used in PowerConstructor.

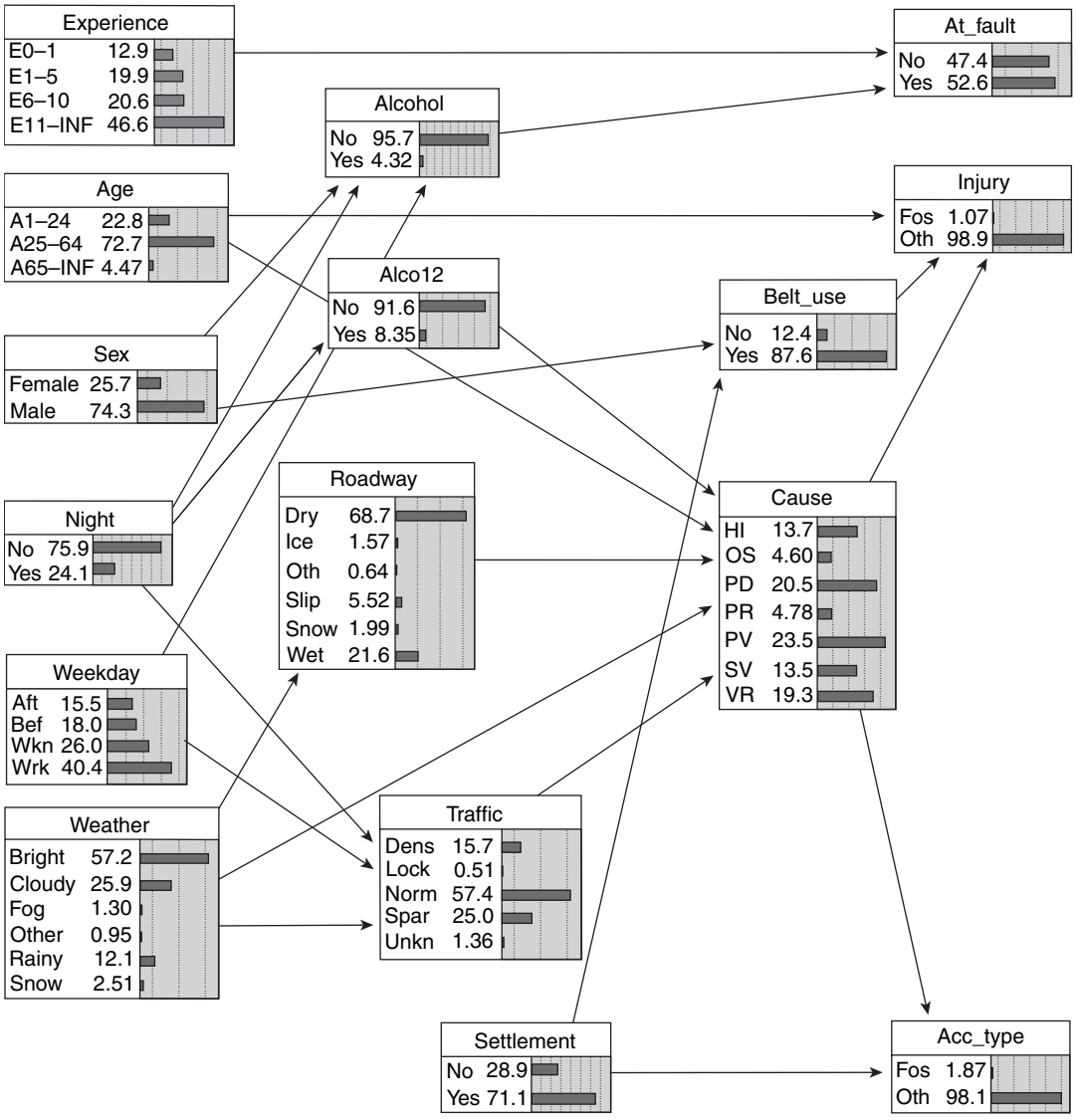
The estimated structure seems plausible, but a different one may also be acceptable. The scoring functions used in the optimizing approach could shed some light on the quality of the estimated Bayesian network. Furthermore, the Kullback-Leibler measure of divergence could be used. Its value could be computed for the structure at hand but would only be of interest when comparing two or more specific structures. By presenting the most probable explanation (MPE), the corresponding probability, and the relative frequency obtained from the database, the statistical quality of the given network can be seen. MPE is given by the most probable configuration of values for all variables in the Bayesian network. For the estimated structure, the MPE is given by the following values for variables:

Night = No; Weekday = Wrk (working day); Weather = Bright; Settlement = Yes; Experience = E11–Inf (driver's experience of 11 years or more); Sex = Male; Age = A25–64; Seat_belt = Yes; Alcohol = No; Alco12 = No; Roadway = Dry; Traffic = Norm (normal); Cause = PV (car maneuvers); At-fault_driver = No; Injury = Oth (other then fatality or serious injury); Accident_type = Oth

Given the estimated structure of the Bayesian network and the conditional probabilities for each node, the probability of the MPE can be computed as shown below.

$$
\begin{aligned}
P(MPE) = {} & P(Night = No) \bullet P(Weekday = Wrk) \bullet P(Weather = Bright) \bullet P(Settlement = Yes) \bullet \\
& P(Experience = E11\text{–}Inf) \bullet P(Sex = Male) \bullet P(Age = A25\text{–}64) \bullet \\
& P(Roadway = Dry \mid Weather = Bright) \bullet \\
& P(Traffic = Norm \mid Weather = Bright, Weekday = Wrk, Night = No) \bullet \\
& P(Belt\_use = Yes \mid Sex = Male, Settlement = Yes) \bullet \\
& P(Alcohol = No \mid Night = No, Weekday = Wrk, Sex = Male) \bullet \\
& P(Alco12 = No \mid Night = No, Weekday = Wrk) \bullet \\
& P(At\text{-}fault\_driver = No \mid Experience = E11\text{–}Inf, Alcohol = No) \bullet \\
& P(Cause = PV \mid Roadway = Dry, Traffic = Norm, Weather = Bright, Alco12 = No) \bullet \\
& P(Injury = Oth \mid Age = A25\text{–}64, Belt\_use = Yes) \bullet \\
& P(Accident\_type = Oth \mid Settlement = Yes, Cause = PV) = 0.00181
\end{aligned}
$$

FIGURE 3 Estimated Unconditional Probabilities of the Bayesian Network



Note: See table 1 for explanation of variables.

An examination of databases for 1998 and 1999 produced the following relative frequencies for MPE:

$P_e(1998) = 94 / 35116 = 0.00268$

$P_e(1999) = 103 / 39950 = 0.00258$

It is obvious that even the most likely explanation has a small probability of its appearance. A comparison of $P(MPE)$ and $P_e(MPE)$ can serve as an indication of the quality of the estimated Bayesian network.

Figure 3 presents probabilities (also called beliefs) estimated from the database of accidents for 1998 and based on the assumption of the network structure given in figure 2. Values of variables related to the different nodes are self-explanatory. Let us recall the abbreviation used for accident type and injury: 1) *Fos* means a fatality or serious injury, and 2) *Oth* means other (less serious) outcomes. (Abbreviations for values related to the variable Cause are explained in table 1.) Figure 3 shows only the unconditional probabilities that correspond to each node (and not the conditional probabilities discussed earlier).

INFERENCE IN THE BAYESIAN NETWORK

The discussion here focuses on only three tables with specific inference results. For the inference process, Netica software (Norsys 1997) was used, and it proved to be very convenient and effective. Results are presented in tables 2 to 4 where predetermined

TABLE 2 Inference Results: Evidence for Accident Type
Probabilities

Acc_ type	Cause (HI)	Cause (OS)	Cause (PD)	Cause (PR)	Cause (PV)	Cause (SV)	Cause (VR)	Settl (No)	Night (Yes)	Alco12 (Yes)
Fos	0.279	0.034	0.276	0.065	0.092	0.229	0.026	0.612	0.253	0.106
Oth	0.134	0.046	0.204	0.048	0.238	0.133	0.196	0.283	0.241	0.083

Key: Fos = fatality or serious injury, HI = inappropriate speed, OS = other, Oth = other, PD = failing to give way, PR = wrong overtaking, PV = car maneuvers, Settl = settlement, SV = wrong side/direction, VR = unsuitable safety distance.
Note: For definition of Alco12, see table 1.

TABLE 3 Inference Results: Evidence for Intoxication Variables
Probabilities

Intox	Cause (HI)	Cause (OS)	Cause (PD)	Cause (PR)	Cause (PV)	Cause (SV)	Cause (VR)	Sex (Male)	Night (Yes)	At_fault (Yes)
Yes	0.204	0.050	0.180	0.069	0.129	0.254	0.113	0.909	0.752	0.879
No	0.131	0.045	0.208	0.046	0.245	0.125	0.201	0.753	0.206	0.510

Key: HI = inappropriate speed, OS = other, PD = failing to give way, PR = wrong overtaking, PV = car maneuvers, SV = wrong side/direction, VR = unsuitable safety distance.

TABLE 4 Inference Results: Evidence for Some Exogenous Variables
Probabilities

Exogenous variables	Injury (Fos)	At_fault (Yes)	Acc_type (Fos)	Alcohol (Yes)	Alco12 (Yes)
Risky values	0.015	0.605	0.046	0.145	0.245
Nonrisky values	0.008	0.497	0.009	0.004	0.043

Key: Fos = fataility or serious injury.
Note: For definition of Alco12, see table 1.

values for a selected categorical variable (or variables) are given in the first column and probabilities for variables of interest are seen in other columns.

Table 2 shows inference results based on evidence for the variable related to the type of accident. Inference results are presented only for variables Cause, Settlement, Night, and Alco12. The probability that the cause of the accident is inappropriate speed (HI) is 0.279 in the case of accident type "Fos" (fatality or serious injury) and 0.134 for the accident type "Oth" (less severe injury). The odds ratio is therefore 2.1. Only a slightly smaller odds ratio is found for cause SV (wrong side/direction); a similar odds ratio for the Settlement variable (2.2); smaller odds ratios for variables Night and Alco12; and odds ratios smaller than 1 for cause PV (car maneuvers), OS (other), and VR (safety distance).

Table 3 reports the inference results based on the evidence for the intoxication variables (Alcohol and Alco12). The probability of an accident taking place at night is 0.752 if drivers are intoxicated and 0.206 if they are not. The odds ratio is, therefore, 3.7. Odds ratios are also high for variables Sex, At_fault, and Cause (for the values related to inappropriate speed and driving on the wrong side of the road).

Inference results based on the evidence for some exogenous variables are presented in table 4. The results shown correspond to a risky situation (driving at night, outside built-up areas, on the weekend, and in rainy weather) and to risky demographic variables (young and inexperienced drivers, i.e., males less than 25 years of age and less than 1 year of driving experience). Nonrisky values were defined with the opposite values for binary variables. For other (non-binary) variables, the following values were used: age between 24 and 65, driving experience more than 11 years, and for the weekday the working day. Odds ratios are especially high for the type of accident and intoxication variables.

While more inference results and a complete picture of the influence on all variables are available, this paper presents only the more interesting variables because the primary aim is to illustrate the capabilities of Bayesian networks in this domain of knowledge. A more indepth analysis of inference results could be used for detecting any weaknesses

in the Bayesian network and for improving its structure. By using data for more than one year, the results become more reliable. New variables can also be added, for example, actual data on traffic flows on the road sections on which accidents occur or other specific characteristics of roads and regions.

CONCLUSIONS

This paper deals with road accidents involving two car drivers. A model of such accidents is presented to capture the interrelations between different relevant variables. To this end, Bayesian networks that have proved their modeling capabilities in different knowledge domains were used. The paper first introduces Bayesian networks on a small example and then formally defines them. After presenting data on two-car accidents for Slovenia in 1998, a structure is proposed based on knowledge of the domain and on computer experiments. For this structure the corresponding probabilities were estimated from the available database. We then demonstrate how the estimated Bayesian network can be used for drawing inferences. Inference results are consistent with expectations as far as the direction of influence is concerned.

The estimated Bayesian network can be regarded as a compact and structured representation of the given database of two-car accidents. This representation relates to specific types of accidents in a given country and year. It also enables different inferences, but other methods, such as logistic regression, should also be used.

Based on the research presented here, we feel that Bayesian networks can be fruitfully applied in the domain of road-accident modeling. Compared with other well-known statistical methods, the main advantage of the Bayesian network method seems to be its complex approach where system variables are interdependent and where no dependent and independent variables are needed. The method's chief weakness is the somewhat arbitrary search for an appropriate network structure. Nevertheless, the results shown here are encouraging and point to possible directions for improvement, such as including more variables and larger datasets that cover more years. Extending the Bayesian network (with good performance results) into a decision network is another possibility.

ACKNOWLEDGMENTS

The Ministry of Science and Education of the Republic of Slovenia supported this research. Thanks go to anonymous referees for suggestions on improving this paper and to Jie Cheng for providing his PowerConstructor software for use with the data. Any errors, however, remain ours alone.

REFERENCES

Agresti, A. 1990. *Categorical Data Analysis*. New York, NY: Wiley & Sons.

Cheng, J., D.A. Bell, and W. Liu. 1997. Learning Belief Networks from Data: An Information Theory Based Approach. Proceedings of the Sixth ACM International Conference on Information and Knowledge Management.

____. 2001. Learning Belief Networks from Data: An Efficient Approach Based on Information Theory. Available at http://www.cs.ualberta.ca/~jcheng/bnpc.htm, as of January 24, 2005.

Cooper G.F. and E. Herskovits. 1992. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning* 9:309–347.

Edwards, J.B. 1998. The Relationship Between Road Accident Severity and Recorded Weather. *Journal of Safety Research* 29(4):249–262.

Hosmer, D.W. and S. Lemeshow. 2000. *Applied Logistic Regression*. New York, NY: Wiley & Sons.

Jensen, F.V. 2001. *Bayesian Networks and Decision Graphs*. New York, NY: Springer-Verlag.

Jordan, M.I., ed. 1999. *Learning in Graphical Models*. Cambridge, MA: The MIT Press.

Kim, K.E. 1996. Differences Between Male and Female Involvement in Motor Vehicle Collisions in Hawaii, 1986–1993. Proceedings from the Second National Conference. Available at http://www.durp.hawaii.edu.

Lam, W. and F. Bacchus. 1994. Learning Bayesian Belief Networks: An Approach Based on the MDL Principle. *Computational Intelligence* 10:269–293.

Lauritzen, S.L. and D.J. Spiegelhalter. 1988. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society B* 50(2):157–194.

Norsys Software Corp. 1997. *Netica Application User's Guide*. Vancouver, Canada.

Pearl, J. 2000. *Causality*. Cambridge, UK: Cambridge University Press.

Zhang, N.L. and D. Poole. 1996. Exploiting Causal Independence in Bayesian Network Inference. *Journal of Artificial Intelligence Research* 5:301–328.

Development of Prediction Models for Motorcycle Crashes at Signalized Intersections on Urban Roads in Malaysia

S. HARNEN[1]
R.S. RADIN UMAR[2,*]
S.V. WONG[3]
W.I. WAN HASHIM[4]

[1] Road Safety Research Center
Universiti Putra Malaysia, 43400 UPM Serdang,
Selangor Darul Ehsan, Malaysia;
and Department of Civil Engineering
Universitas Brawijaya, Malang, East Java, Indonesia

[2] Road Safety Research Center and
Department of Civil Engineering
Universiti Putra Malaysia, 43400 UPM Serdang,
Selangor Darul Ehsan, Malaysia

[3] Road Safety Research Center and
Department of Mechanical
and Manufacturing Engineering
Universiti Putra Malaysia, 43400 UPM Serdang,
Selangor Darul Ehsan, Malaysia

[4] School of Civil Engineering
Universiti Sains Malaysia,14300 Nibong Tebal,
Seberang Perai Selatan, Pulau Pinang, Malaysia

ABSTRACT

Because more than half of the motor vehicles in Malaysia are motorcycles, the safety of this form of transportation is an important issue. As part of a motorcycle safety program, Malaysia became the first country to provide exclusive motorcycle lanes in the hopes of reducing motorcycle crashes along trunk roads. However, little work has been done to address intersection crashes involving motorcycles. This paper provides models for predicting motorcycle crashes at signalized intersections on urban roads in Malaysia. A generalized linear modeling technique with a quasi-likelihood approach was adopted to develop the models. Traffic entering the intersection, approach speed, lane width, number of lanes, shoulder width, and land use at the approach of the intersection were found to be significant in describing motorcycle crashes. These findings should enable engineers to draw up appropriate intersection treatment criteria specifically designed for motorcycle lane facilities in Malaysia and elsewhere.

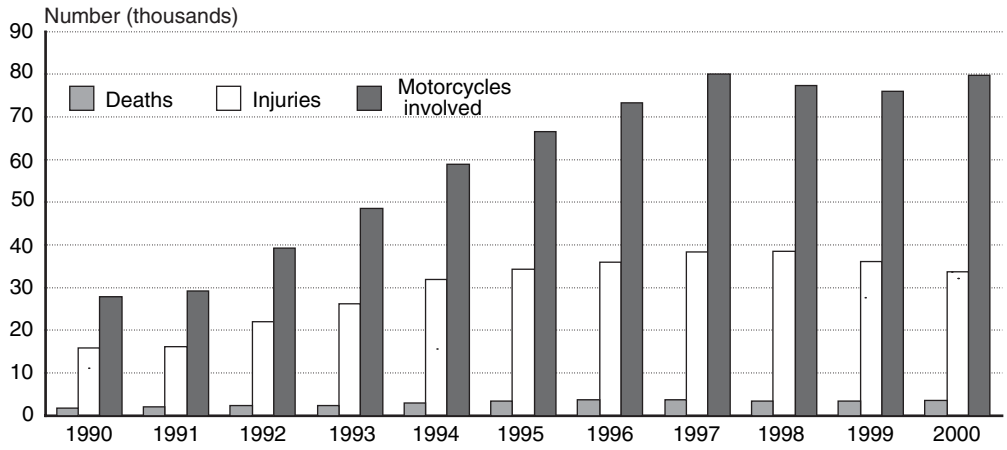
INTRODUCTION

Motorcycle crashes continue to be a problem in both developing and developed countries. Fatality rates (measured in deaths per 10,000 registered

E-mail addresses:
* Corresponding author—radinumx@eng.upm.edu.my
S. Harnen—hsulistio@telkom.net
S.V. Wong—wongsv@eng.upm.edu.my
W.I. Wan Hashim—cewhwi@eng.usm.my

KEYWORDS: Motorcycle crashes, generalized linear models, prediction model, intersection crashes, motorcycle crash model.

FIGURE 1 Motorcycle Crashes in Malaysia: 2000



	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Deaths	1,826	2,062	2,307	2,416	2,946	3,362	3,778	3,760	3,409	3,410	3,519
Injuries	15,857	16,235	21,957	26,226	31,957	34,372	35,960	38,317	38,448	36,100	33,651
Motorcycles involved	27,845	29,225	39,272	48,511	58,921	66,508	73,268	80,100	77,298	76,032	79,816

Source: Polis Di Raja Malaysia, *Statistical Report: Road Accidents, Malaysia 2000* (Kuala Lumpur, Malaysia: 2002, Traffic Branch, Royal Malaysian Police).

vehicles) in these crashes are much higher than in nonmotorcycle[1] crashes. In the United States, the National Highway Traffic Safety Administration (USDOT 2002) reported a fatality rate of 6.5 per 100 million vehicle-miles traveled, and motorcyclists were about 26.1 times as likely as passenger car occupants to die in a motor vehicle traffic crash. The Canadian rate was 4.7 in 1999, which rose to 5.1 in 2000; the Canadian nonmotorcycle fatality rate in 2000 was 0.7 (Transport Canada 2001). Similarly large rates have been reported in other developed countries: Australia's rate was 6.2 in 2001, an increase of about 9% from 2000 and more than 4 times the fatality rate of other road users (ATSB 2002); the United Kingdom's rate was 7.3 in 2000, decreasing to 6.6 in 2001, about 10 times the fatality rate for passenger car occupants (DfT 2002); Swedish (SI 2000), French, and German (OECD 2002) rates in 2000 were 4.1, 5.3, and 2.2, respectively.

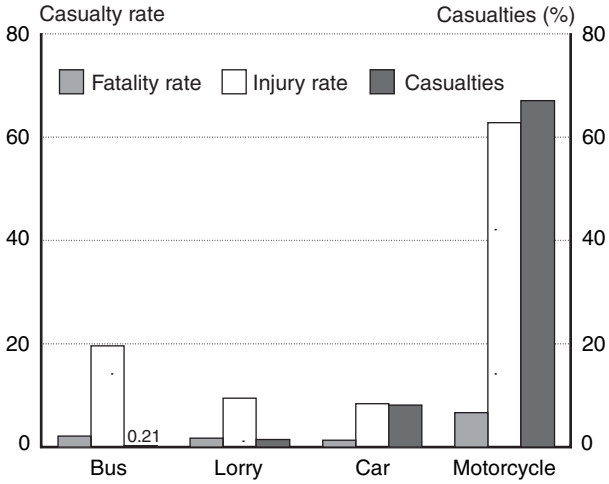
In developing countries, deaths and serious injuries from motorcycle accidents constitute a large portion of total road casualties especially in Asian countries, because motorized two-wheelers make up 40% to 95% of their vehicle fleets. As a result, more than half the road fatalities were riders or pillion passengers.

In Malaysia, motorcycles constitute more than half the total vehicle population and contribute more than 60% of the casualties (deaths and serious and slight injuries) in traffic crashes. In 2000, 79,816 crashes involved motorcycles, an increase of almost three-fold from 1990. Of these, almost 3,000 motorcyclists were killed every year during this period (figure 1). Moreover, motorcyclist casualties were much higher than those of occupants in other types of vehicles (figure 2).

In an attempt to reduce casualties, exclusive motorcycle lanes were constructed along major trunk roads in the country. Since the implementation of this initiative, a number of studies (Radin Umar 1996; Radin Umar et al. 1995, 2000) have been carried out to evaluate the impact of these lanes on motorcycle crashes on highway links. Results indicate the lanes had a significant effect ($p < 0.01$), reducing motorcycle crashes by 39% following the opening of the lanes to traffic. However, little research has been done on motorcycle crashes at intersections. Indepth studies would allow traffic engineers to establish appropriate intersection treat-

[1] Nonmotorcycle refers to all types of motorized vehicles excluding motorcycles.

FIGURE 2 Motorcycle Rider Casualties Compared with Casualties for Occupants in Other Types of Vehicles: 2000



Key: Fatality rate = fatalities per 10,000 registered vehicles; injury rate = injuries per 10,000 registered vehicles; casualties (%): as a percentage of all casualties in traffic crashes in Malaysia.

Source: Polis Di Raja Malaysia, *Statistical Report: Road Accidents, Malaysia 2000* (Kuala Lumpur, Malaysia: 2002, Traffic Branch, Royal Malaysian Police).

ment criteria specifically designed for motorcycle lane facilities.

Recent studies on traffic crash modeling have used the generalized linear modeling (GLM) approach (McCullagh and Nelder 1989) with Poisson or negative binomial error structure. This approach is widely accepted as more appropriate for the characteristics of crashes (i.e., discrete, rare, and independent) than the classical linear model based on normal error structure with a constant variance. Crashes can be characterized by their mean number per unit time and are simply represented by a Poisson random variable.

Many researchers have reported the usefulness of the GLM approach in developing predictive models for traffic crashes using either cross-sectional or time series analysis (Griebe and Nielsen 1996; Mountain et al. 1996, 1998; Tarko et al. 1999; Vogt and Bared 1998; Vogt 1999; Radin Umar et al. 1995, 2000; Radin Umar 1996; Bauer and Harwood 2000; Saied and Said 2001; Taylor et al. 2002). For example, an earlier study on crashes at intersections prepared for the Federal Highway Administration of the U.S. Department of Transportation in connection with the development of the Interactive Highway Safety Design Model (IHSDM) (Bauer and Harwood 2000) provided direct input into the Accident Analysis Module of the IHSDM.

The analysis included all collision types using three-year crash frequencies (1990 to 1992) and geometric design, traffic control, and traffic volume data from a database provided by the California Department of Transportation. The analysis was performed using the SAS GENMOD procedure. The models were developed using the GLM approach with a log-normal regression model and a loglinear regression model (a Poisson regression followed by a negative binomial regression model). In this study, the 10% significance level of the *t*-statistic of the parameter estimates was used to assess the significance of the fitted model. The explanatory variables (continuous and categorical) that follow were found to be significant in explaining crashes at intersections:

- major road ADT (average daily traffic) and minor road ADT,
- average lane width on major roads,
- number of lanes on major and minor roads,
- design speed of major roads,
- major-road right-turn and left-turn channelizations,
- access control on major roads,
- functional class of major roads,
- outside shoulder width on major roads,
- terrain,
- road lighting,
- minor-road right-turn channelization,
- major-road left-turn prohibition, and
- median on major roads.

As an extension to our earlier analysis (Harnen et al. 2003a, 2003b), this paper presents the development of prediction models for motorcycle crashes at signalized intersections along both the exclusive and non-exclusive motorcycle lanes on urban roads in Malaysia. We used the GLM approach with Poisson error structure to develop our models. The parameter estimates and tests of their significance were carried out using GLIM 4 statistical software (NAG 1994), which is specifically designed for fitting generalized linear models.

THE DATA

Selected Intersections

The intersections studied were located on urban roads in four districts of the state of Selangor, Malaysia. The data collected covered motorcycle crashes, traffic and pedestrian flow, approach speed, intersection geometry, number of legs, and land use. The intersections were selected based on the following conditions between 1997 and 2000: a) only marginal change in land use; b) no major modifications or upgrading; c) an equal number of lanes on the corresponding major and minor roads; d) only marginal change of signal characteristics, for example, signal timing and signal phasing; e) no access road within a 50-meter distance from the intersection stop lines; and f) intersections must have had fatalities and/or serious and slight injuries in crashes. Please note that while data were collected on signal characteristics they are not analyzed here. However, they will be included in future work. Based on the intersection files (142 signalized intersections with motorcycle crashes in the period 1997 to 2000) extracted from the Microcomputer Accident Analysis Package (MAAP) database and visits to the sites to ensure that they met the requirements, 51 intersections were chosen. In this study, motorcycle crashes occurring within 50 meters of the corresponding stop lines of the intersection were classified as intersection crashes.

Motorcycle Crash Data

Four-year's worth of motorcycle crash data on the selected intersections, from 1997 through 2000, were collected from the police crash record form, POL 27 (Pin 1/91). The POL 27 is designed for easy completion (Radin Umar et al. 1993) and is fully compatible with the MAAP database developed by the Transport Research Laboratory (Hills and Baguley 1993). Data were extracted from two complementary sources: the MAAP database for fatal and serious injury crashes, and the Computerized Accident Recording System (CARS 2000) database for slight injury crashes. Both databases are based on the POL 27 record form.[2]

[2] The MAAP database is located at the Road Safety Research Center, Universiti Putra Malaysia, while the CARS 2000 database is located at the Traffic Branch, Royal Malaysian Police Headquarters.

Traffic Flow Data

In this study, the estimated annual average daily traffic (AADT) defines the traffic flow on each selected intersection. Hourly traffic volume (disaggregated by nonmotorcycles and motorcycles) was counted on major- and minor-road approaches and then converted to AADT by using hourly, daily, and monthly factors. These factors were determined based on 24-hour permanent traffic count station and traffic census data, available from the Highway Planning Unit, Ministry of Works in Malaysia (HPU 2001a, 2001b) and were developed using the method proposed by McShane et al. (1998). The AADT is expressed in terms of the number of nonmotorcycles per day and motorcycles per day.

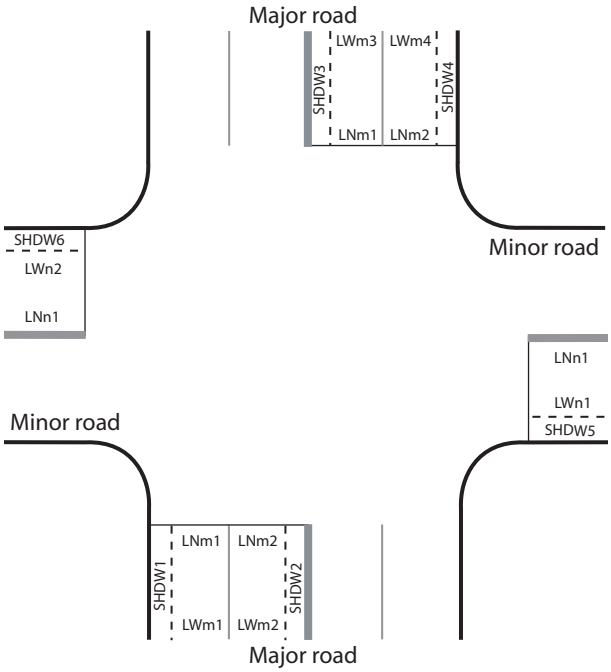
Other Data Used

Approach speed and pedestrian flow were also considered in this study. However, while these data were not available in the database, they were collected onsite following criteria used by Golias (1997) in an earlier study. The 85th percentile approach speed on major and minor roads was used to represent the approach speed on each intersection. Arndt and Troutbeck (1998) also considered this characteristic in an earlier study on traffic crashes. The approach speeds were measured at a 50-meter distance upstream from the corresponding stop lines of the intersection and were counted for all vehicles moving during the time the signal was green.

Pedestrian flow at each intersection was defined as the total number of pedestrian crossings per hour counted on major- and minor-road approaches. It should be noted that pedestrians per hour rather than pedestrians per day was used because there was no supporting data to convert hourly pedestrian flow to annual average daily pedestrians (the AADT for pedestrians).

Intersection geometry, number of legs, and land use for each selected intersection were also observed onsite. Of the 51 selected intersections, 27 were three-legged while 24 were four-legged. The land use adjacent to the intersection was classified into two categories: commercial and noncommercial areas. A commercial area was defined as an area with a concentration of offices, shops, and railway and bus stations, while residential areas and unused land come under the category of noncommercial

FIGURE 3 Typical Layout of Intersection Geometry



Key: LWm1, 2, 3, 4 = lane width on major road approaches;
LWn1, 2 = lane width on minor road approaches;
LNm1, 2 = number of lanes on major road approach;
LNn1 = number of lanes on minor road approach;
SHDW1, 2, 3, 4, 5, 6 = shoulder width on major- and minor-road approaches.

area. Of the 51 intersections, 33 were located in commercial areas and 18 were in noncommercial areas. Figure 3 shows a typical layout of intersection geometry considered in the study.

MODEL DEVELOPMENT

Prior to carrying out the statistical modeling, we did some preliminary work to facilitate the modeling process. This included formulating the theoretical models, specifying the error structure and link function, identifying the model variables, and defining the goodness-of-fit and significance tests.

Using our earlier analysis of motorcycle crashes at intersections (Harnen et al. 2003a, 2003b) and studies of traffic crashes at intersections (Griebe and Nielsen 1996; Vogt and Bared 1998; Vogt 1999; Bauer and Harwood 2000; Saied and Said 2001), we defined the model structure and the variables included.

Two separate models (Models 1 and 2) were proposed. These models used the same data and structure but employed different explanatory variables. In Model 1, the response variable was the number of motorcycle crashes and the explanatory variables were traffic flow (disaggregated by nonmotorcycles and motorcycles both for major and minor roads), pedestrian flow, approach speed, lane width, number of lanes, number of legs, shoulder width, and land use. The continuous variables were identified as traffic flow, pedestrian flow, approach speed, lane width, and number of lanes, while the categorical variables were number of legs with two-factor levels, shoulder width with three-factor levels, and land-use with two-factor levels. In Model 2, the response variable was motorcycle crashes, while the explanatory variables were traffic flow and shoulder width. Both traffic flow and shoulder width were continuous variables.

The main differences in these two models are the explanatory variables included. Model 2, which has three continuous variables, is simpler than Model 1 and can be used further to establish major- and minor-road flow criteria for intersection treatment. This can be done by using the design curves relating major- and minor-road flows and shoulder widths developed based on Model 2.

Model 1, which has 13 variables (combination of continuous and categorical), was aimed at giving more room to engineers for analyzing the variables contributing to motorcycle crashes. Software that is specifically designed for Model 1 application could make it easier and faster to analyze the variables and estimate motorcycle crashes.

Taking the earlier studies on intersection crash modeling into consideration, the theoretical models containing all terms used in this study were formulated as follows:

Model 1

$$MCA = k_1 QNMm^{\alpha_1} \bullet QNMn^{\alpha_2} \bullet QMm^{\alpha_3} \bullet QMn^{\alpha_4} \bullet QPED^{\alpha_5} \bullet \mathrm{EXP}^{z} \quad (1)$$

where

$$z = \beta_1 SPEED + \beta_2 LWm + \beta_3 LWn + \beta_4 LNm + \beta_5 LNn + \beta_6 NL + \beta_7 SHDW + \beta_8 LU + e$$

Model 2

$$MCA = k_2 Qmajor^{\delta_1} Qminor^{\delta_2} \mathrm{EXP}^{(\lambda_1 SHD + e)} \quad (2)$$

where *MCA* is motorcycle crashes per year. Descriptions of all the explanatory variables are presented

in table 1. The k_1, k_2, α_1, α_2, α_3, α_4, α_5, β_1, β_2, β_3, β_4, β_5, β_6, β_7, β_8, δ_1, δ_2, and λ_1 are the parameters to be estimated and the (e) term is the error representing the residual difference between the actual and predicted models.

Using a logarithmic transformation, the loglinear version of the model is:

Model 1

$$\begin{aligned} Ln(MCA) &= Ln(k) + \alpha_1 Ln(QNMm) + \alpha_2 Ln(QNMm) \\ &+ \alpha_3 Ln(QMm) + \alpha_4 Ln(QMn) + \alpha_5 Ln(QPED) \\ &+ \beta_1(SPEED) + \beta_2(LWm) + \beta_3(LWn) + \beta_4(LNm) \\ &+ \beta_5(LNn) + \beta_6(NL) + \beta_7(SHDW) + \beta_8(LU) + e \end{aligned} \quad (3)$$

Model 2

$$\begin{aligned} Ln(MCA) &= Ln(k) + \delta_1 Ln(Qmajor) \\ &+ \delta_2 Ln(Qminor) + \lambda_1(SHD) + e \end{aligned} \quad (4)$$

To allow direct interpretation of the parameter estimates produced by GLIM 4, the flow functions in equations (3) and (4) need to be transformed using a natural logarithmic (*Ln*), while the others do not. It should be noted that the total four-year crash frequencies were used to fit the models. However, by introducing an offset variable in the fitting process, the final model would be able to estimate the number of crashes per year. This approach has also been implemented in earlier studies on traffic crashes at intersections (Mountain et al. 1998) and motorcycle crashes at intersections (Harnen et al. 2003a, 2003b).

We based the model on the Poisson error structure and used the quasi-likelihood approach (McCullagh and Nelder 1989) to overcome the dispersion problem. A loglinear cross-sectional model was employed with the link function specified as the log (NAG 1994). This approach has been used in earlier studies on motorcycle crashes on highway links (Radin 1996; Radin et al. 1995, 2000) and in our earlier analysis of motorcycle crashes at intersections (Harnen et al. 2003a, 2003b).

Using the quasi-likelihood approach, the dispersion parameter was estimated from the mean deviance (scaled deviance over its degrees of freedom). This may result in a model where the scaled deviance is equal to its degrees of freedom. The final model was based on the goodness-of-fit and significance tests carried out on the models such as the change in scaled deviance from adding or removing the terms, the ratio of scaled deviance to its degrees of freedom (mean deviance), and the 5% significance level of *t*-statistics of the parameter estimates.

Both multivariate and univariate analyses were conducted for Model 1, while only multivariate analysis was undertaken for Model 2. We used multivariate analysis to assess which of the variable(s) had the most effect on the probability of motorcycle crashes. The univariate analysis was employed to obtain a complete picture of the effect of all explanatory variables on motorcycle crashes. It should be noted that only those variables found significant at the 5% level in the univariate analysis were subsequently included in the multivariate analysis.

RESULTS

Model 1

Table 2 presents the results of the univariate analysis for Model 1. It can be seen that all terms, except *QPED*, *LNn*, and *NL*, were significant at the 5% level. The respective scaled deviance was equal to its corresponding degrees of freedom, as the quasi-likelihood approach had been introduced in the fitting process. Because the terms *QPED*, *LNn*, and *NL* were not significant at the 5% level, they were then excluded from any further analysis.

The multivariate analysis (table 3) shows that all explanatory variables were significant at the 5% level. The scaled deviance was equal to its degrees of freedom, changing from 15,022.0 to 39.0 with a loss of 11 degrees of freedom. The mean deviance changed from 300.4 to 1.0.

On the basis of the multivariate analysis, the final model is:

$$\begin{aligned} MCA = 0.002822\ QNMm^{0.3241} \bullet QNMn^{0.0835} \bullet \\ QMm^{0.0683} \bullet QMn^{0.1296} \bullet \mathrm{EXP}^{z} \end{aligned} \quad (5)$$

where

$$\begin{aligned} z = {} & 0.02602\ SPEED - 0.0727\ LWm - 0.0718\ LWn \\ & - 0.01758\ LNm - \beta_7 SHDW + \beta_8 LU \end{aligned}$$

and where *MCA* is motorcycle crashes per year, β_7 = 0.0, 0.01755, and 0.02554 for *SHDW* = 1, 2, and 3, respectively, β_8 = 0.0 and 0.01591 for *LU* = 1 and 2,

TABLE 1 Description, Factor Levels, Coding System, and Basic Statistics of the Explanatory Variables

Explanatory variables	Description	Factor levels	Coding system in GLIM	Min	Max	Mean	Median
MODEL 1							
QNMm	Nonmotorcycle flow on major road (nmpd)		*QNMm*	14,527	50,529	31,389	32,354
QNMn	Nonmotorcycle flow on minor road (nmpd)		*QNMn*	2,133	20,129	11,276	11,129
QMm	Motorcycle flow on major road (mpd)		*QMm*	5,510	21,899	12,228	10,792
QMn	Motorcycle flow on minor road (mpd)		*QMn*	1,752	4,771	3,183	3,142
QPED	Pedestrian flow (pedestrians/hour)		*QPED*	0	235	36	19
SPEED	Approach speed (km/hour)		*SPEED*	53.00	68.00	59.57	59.50
LWm	Average lane width on major road (m)		*LWm*	3.30	4.00	3.58	3.60
LWn	Average lane width on minor road (m)		*LWn*	3.40	4.00	3.69	3.60
LNm	Number of lanes on major road (lanes/traffic direction)		*LNm*	2	5	2.6	2.0
LNn	Number of lanes on minor road (lanes/traffic direction)		*LNn*	1	3	1.6	2.0
NL	Number of legs	2	(1) 3-legged (2) 4-legged	1	2	1.5	1.0
SHDW	Average shoulder width on major and minor roads	3	(1) *SHDW* = 0.00 m (2) 0.00 < *SHDW* ≤ 1.00 m (3) *SHDW* > 1.00 m	1	3	1.7	2.0
LU	Land-use category	2	(1) Noncommercial area (2) Commercial area	1	2	1.7	2.0
MODEL 2							
Qmajor	Traffic flow on major road (vehicles/day)		*Qmajor*	20,043	72,428	43,617	42,258
Qmajor	Traffic flow on minor road (vehicles/day)		*Qminor*	4,504	24,900	14,459	14,293
SHD	Average shoulder width on major and minor roads (m)		*SHD*	0	1.3	0.5	0.9

Key: km = kilometers; m = meters; mpd = motorcycles per day; nmpd = nonmotorcycles per day.

respectively (table 1). Figure 4 shows the actual and predicted motorcycle crashes.

Model 2

Table 4 presents the results of the multivariate analysis of Model 2. All terms were found to be significant at the 5% level. The scaled deviance was equal to its degrees of freedom, because the quasi-likelihood approach had also been introduced in the fitting process. The scaled deviance changed from 854.8 to 47.0 with a loss of 3 degrees of freedom and the mean deviance changed from 17.1 to 1.0.

TABLE 2 Univariate Analysis of Model 1

Explanatory variables	Estimates	Standard errors	Degrees of freedom	Scaled deviance	*t*-statistics	Sig. at 0.05
Constant	–9.2260	0.3480	49	49	–26.55	Yes
QNMm	0.9835	0.0334			29.42	Yes
Constant	–1.2210	0.2160	49	49	–5.64	Yes
QNMn	0.2490	0.0243			10.26	Yes
Constant	–0.7520	0.2580	49	49	–2.92	Yes
QMm	0.1943	0.0288			6.76	Yes
Constant	–2.0910	0.3790	49	49	–5.51	Yes
QMn	0.3877	0.0478			8.12	Yes
Constant	0.8636	0.0748	49	49	11.54	Yes
QPED	0.0357	0.0237			1.51	No
Constant	–3.6760	0.1090	49	49	–33.63	Yes
SPEED	0.0771	0.0018			42.83	Yes
Constant	3.2900	1.1800	49	49	2.79	Yes
LWm	–0.6510	0.3290			–1.98	Yes
Constant	3.0200	1.0500	49	49	2.88	Yes
LWn	–0.5800	0.2950			–1.97	Yes
Constant	1.1960	0.1260	49	49	9.47	Yes
LNm	–0.1023	0.0519			–1.97	Yes
Constant	1.0780	0.1200	49	49	8.97	Yes
LNn	–0.0744	0.0697			–1.07	No
Constant	1.0020	0.1280	49	49	7.83	Yes
NL (2)	–0.0294	0.0826			–0.36	No
Constant	1.0578	0.0524	48	48	20.17	Yes
SHDW (2)	–0.1812	0.0856			–2.12	Yes
SHDW (3)	–0.2750	0.1190			–2.32	Yes
Constant	0.8316	0.0752	49	49	11.05	Yes
LU (2)	0.1774	0.0885			2.01	Yes

Note: Estimates for factors (2) and (3) are the differences compared with the reference level (1).

The final model developed in this analysis was:

$$MCA = 0.0004693\ Qmajor^{0.5948} \bullet Qminor^{0.2411} \bullet \mathrm{EXP}^{-0.0589\ SHD} \quad (6)$$

Tables 2, 3, and 4 show that the variables have a consistent effect on motorcycle crashes. This is indicated by the sign (plus or minus) of the parameter estimates for each of the corresponding variables that are identical.

DISCUSSION

Model 1

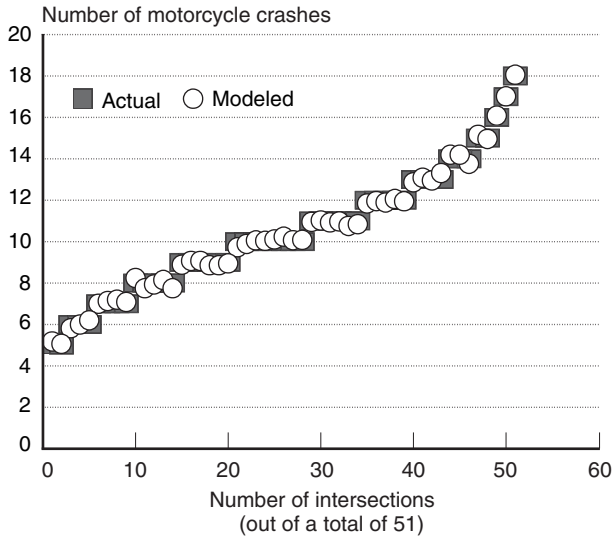
The final Model 1 reveals that the number of motorcycle crashes per year is proportional to the

TABLE 3 Multivariate Analysis of Model 1

Explanatory variables	Estimates	Standard errors	Degrees of freedom	Scaled deviance	*t*-statistics	Sig. at 0.05	Mean deviance
Constant	–5.8700	0.4580	50	15,022.0	–12.81	Yes	300.4
QNMm	0.3241	0.0297	49	748.6	10.91	Yes	15.3
QNMn	0.0835	0.0183	48	483.6	4.57	Yes	10.1
QMm	0.0683	0.0188	47	241.5	3.64	Yes	5.1
QMn	0.1296	0.0230	46	142.8	5.63	Yes	3.1
SPEED	0.0260	0.0033	45	75.1	7.79	Yes	1.7
LWm	–0.0727	0.0320	44	70.7	–2.27	Yes	1.6
LWn	–0.0718	0.0305	43	69.1	–2.35	Yes	1.6
LNm	–0.0176	0.0044	42	55.0	–3.97	Yes	1.3
SHDW (2)	–0.0176	0.0069	40	47.5	–2.55	Yes	1.2
SHDW (3)	–0.0255	0.0100	40	47.5	–2.56	Yes	1.2
LU (2)	0.0159	0.0055	39	39.0	2.91	Yes	1.0

Note: Estimates for factors (2) and (3) are the differences compared with the reference level (1).

FIGURE 4 Actual and Modeled Motorcycle Crashes: 1997–2000 (Model 1)



traffic flow entering the intersection. The estimates of *QNMm*, *QNMn*, *QMm*, and *QMn* indicate that an increase in nonmotorcycle and motorcycle flows on major and minor roads is associated with more motorcycle crashes (figure 5). For instance, doubling nonmotorcycle flow on a major road (*QNMm*) is expected to cause an increase of about 25% in motorcycle crashes. If all traffic entering the intersection is doubled, an increase of about 45% in motorcycle crashes would result. We also found that nonmotorcycle flows on major roads (*QNMm*) was the most important variable for the probability of motorcycle crashes. The results support the findings of earlier studies on traffic crashes at intersections (Summersgill 1991; Mountain et al. 1998; Rodriguez and Sayed 1999; Vogt and Bared 1998; Vogt 1999; Bauer and Harwood 2000).

The *SPEED* estimate shows that an increase in approach speed is associated with a rise in motorcycle crashes. For instance, if the approach speed goes up by 10 kilometers per hour, 30% more motorcycle crashes can be expected. Our findings support earlier studies on the relationship of traffic speed to crashes (Griebe and Nielsen 1996; Vogt and Bared 1998; Bauer and Harwood 2000; Lynam et al. 2001; USDOT 2002; Taylor et al. 2002).

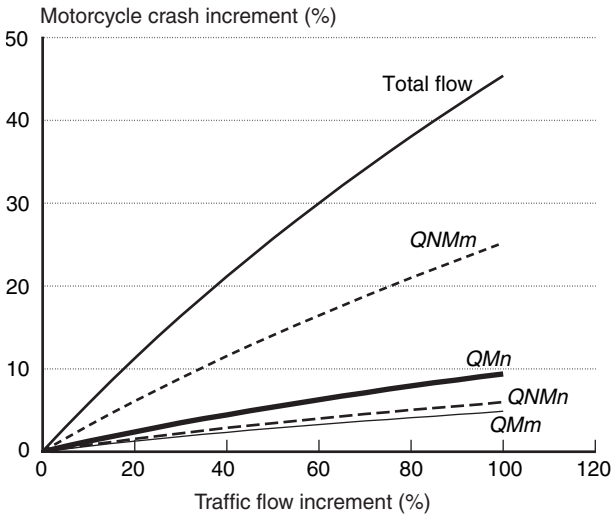
The estimates of *LWm* and *LWn* imply that a wider lane is associated with a reduction in motorcycle crashes. For instance, widening the lane on major and minor roads by 0.50 meters is expected to reduce motorcycle crashes by some 3.6% and 3.5%, respectively. This result is in line with the finding reported in an earlier study on traffic crashes at intersections (Bauer and Harwood 2000).

Meanwhile, the estimate of *LNm* indicates that an increase in the number of lanes on a major road is associated with a reduction in motorcycle crashes. However, the effect of this variable is marginal (1.7%). The result seems to be in line with the finding reported by Bauer and Harwood (2000). This reduction was probably the result of the presence of

TABLE 4 Multivariate Analysis of Model 2

Explanatory variables	Estimates	Standard errors	Degrees of freedom	Scaled deviance	*t*-statistics	Sig. at 0.05	Mean deviance
Constant	–7.6640	0.4650	50	854.8	–16.49	Yes	17.1
Qmajor	0.5948	0.0707	49	65.4	8.41	Yes	1.3
Qminor	0.2411	0.0640	48	52.1	3.77	Yes	1.1
SHD	–0.0589	0.0261	47	47.0	–2.25	Yes	1.0

FIGURE 5 Effects of Traffic Flow on Motorcycle Crashes: Model 1



an exclusive right turn lane on the major road. Of the 51 intersections we studied, 48 had an exclusive right turn lane on each major road approach. The presence of such lanes may reduce rear-end crashes for motorcycles. It should be mentioned that an exclusive turning lane was counted as a lane in our measurements of *LNm*. Earlier studies confirmed the benefit provided by such lanes for crash reduction at intersections (Kulmala 1992; Vogt 1999; Bauer and Harwood 2000) and at links (Tarko et al. 1999). However, for a better explanation, a separate model should be developed to explain the effects of an exclusive left, exclusive right, and short turning lanes on all types of motorcycle crashes at intersections.

The *SHDW* estimates indicate that a wider paved shoulder is associated with fewer motorcycle crashes. The result seems to be in line with the finding reported by Bauer and Harwood (2000). For instance, 25% more motorcycle crashes occur at intersections without a shoulder than at intersections with a shoulder wider than 1.0 meters. When we compare motorcycle crashes at intersections without a shoulder with crashes where the shoulder width is between 0.0 meters and ≤1.0 meters, the difference is smaller, only 1.7% more crashes occur when there is no shoulder. This finding seems reasonable because motorcyclists use the available shoulder width when approaching an intersection, and the rates of rear-end and sideswipe crash types between motorcycles on the shoulder and other vehicles on the adjacent lane should be lower if the shoulder is wider. This situation is common in countries like Malaysia with a high population of motorcycles. However, a better explanation can be provided, and a separate model was developed to explain the effect of shoulder width on all types of motorcycle crashes at intersections.

The estimate of *LU* shows that signalized intersections located within commercial areas are associated with increased motorcycle crashes. The result confirms the findings of an earlier study on traffic crashes at four-legged signalized intersections (Wang and Ieda 1997). However, the difference in the estimation of motorcycle crashes between commercial and noncommercial areas is marginal (1.6%). As explained earlier, this study includes only those intersections located within commercial areas having no access road to the adjacent land use within 50 meters of the intersection stop lines. As such, the number of conflicts between vehicles entering or leaving the intersection and vehicles turning into or out of the adjacent land use may be reduced, hence fewer crashes. The effect of access control or the number of accesses on traffic crashes has also been reported in earlier studies (Vogt 1999; Bauer and Harwood 2000).

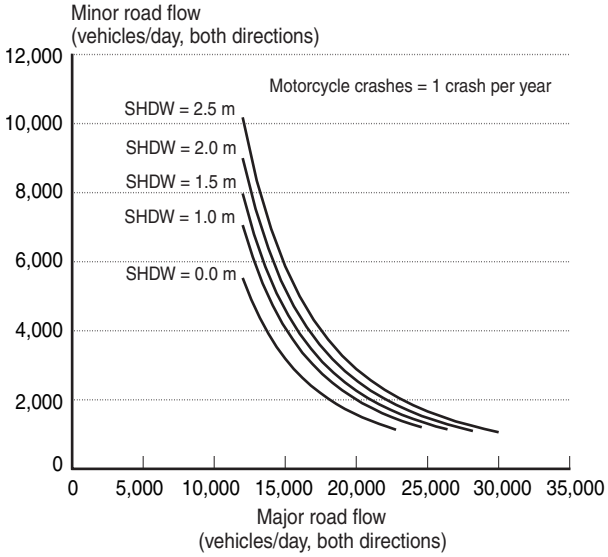
Model 2

Model 2 results verify the contribution of traffic flow, both on major roads (*Qmajor*) and minor roads (*Qminor*), to motorcycle crashes. The estimates of the variables show that an increase in traffic flow on

major and minor roads is associated with a greater number of motorcycle crashes, and an increase in shoulder width (*SHD*) is associated with a reduction in these crashes. For example, widening the shoulder by 1.0 meters is expected to reduce the number of motorcycle crashes by about 6%. In this model, the effect of shoulder width on motorcycle crashes can be directly quantified when the width is changed, and this is one of the main differences between Model 1 and Model 2.

As described earlier, design curves relating major- and minor-road flows for different shoulder widths can be developed based on Model 2 (figure 6). As discussed, wider shoulders at intersections offer higher levels of safety to motorcyclists approaching the junction. Based on the relationships among the variables developed based on Models 1 and 2, future work includes carrying out an indepth analysis of whether intersection treatments that have nonexclusive motorcycle lane facilities could reduce motorcycle crashes.

FIGURE 6 Relationship of Major- and Minor-Road Flows with Differing Shoulder Widths (based on Model 2)



CONCLUSIONS

This paper presents motorcycle crash prediction models for signalized intersections on urban roads in Malaysia. The models reveal that traffic flow, approach speed, intersection geometry, and land use are significant factors in explaining motorcycle crashes at signalized intersections. The number of crashes is proportional to the level of traffic entering the intersections. An increase in motorcycle crashes is associated with a larger total vehicle flow on major and minor roads. Nonmotorcycle flows on major roads had the most effect on the likelihood of motorcycle crashes.

An increase in approach speed is associated with more motorcycle crashes, while wider lanes, a greater number of lanes, and wider shoulders bring a reduction in these crashes. Furthermore, more motorcycle crashes occur at signalized intersections located within commercial areas than at intersections located outside of commercial areas.

The models developed in this study present information to aid traffic engineers in deciding the appropriate level of intervention for intersection treatment with respect to motorcycle crashes. Using our models, design parameters for intersections may be changed to achieve appropriate safety levels. Decisions on whether special treatment to minimize motorcycle conflicts is needed at intersections can be objectively carried out based on the models. However, the models might only be valid for a typical traffic environment in developing countries like Malaysia, where the proportion of motorcycles is 20% to 40% of all vehicles at signalized intersections.

For design options, further investigation of the role of parameters of traffic flow by time periods (hourly, peak hour, peak periods) and categorizing the models by time period(s) is suggested, and the need for further categorization of model structure by different intersection geometric configurations (e.g., intersections with and without exclusive motorcycle lanes) is also advised.

ACKNOWLEDGMENTS

This paper reports findings of part of a study conducted for the Intensified Research Priority Area (IRPA) project, *Development of Design Criteria and Standards for Malaysian Motorcycle Lanes*. We gratefully acknowledge the financial support from the Ministry of Science, Technology and Environment Malaysia. The authors would like to thank the Royal Malaysian Police and the Highway Planning Unit, Ministry of Works, Malaysia, for providing the data.

REFERENCES

Arndt, O.K. and R.J. Troutbeck. 1998. Relationship Between Roundabout Geometry and Accident Rates, presented at the International Symposium on Highway Geometric Design Practices, Boston, MA.

Australian Transport Safety Bureau (ATSB). 2002. *Road Fatalities Australia: 2001 Statistical Summary.* Canberra, Australia: Department of Transport and Regional Services.

Bauer, K.M. and D.W. Harwood. 2000. *Statistical Models of At-Grade Intersection Accidents—Addendum,* Publication No. FHWA-RD-99-094. Washington, DC: U.S. Department of Transportation, Federal Highway Administration.

Department for Transport (DfT). 2002. *Road Accidents Great Britain: 2001, The Casualty Report.* London, England.

Golias, J.C. 1997. Effects of Signalization on Four-Arm Urban Junction Safety. *Accident Analysis and Prevention* 29(2):181–190.

Griebe, P. and M.A. Nielsen. 1996. Safety at Four-Armed Signalized Junctions Situated on Roads with Different Speed Limits. *Proceedings of the Conference on Road Safety in Europe,* VTI konferens No.7A, Part 2. Birmingham, England: VTI, Swedish National Road and Transport Research Institute. 151–163.

Harnen, S., R.S. Radin Umar, S.V. Wong, and W.I. Wan Hashim. 2003a. Motorcycle Crash Prediction Model for Non-Signalized Intersections. *IATSS Research* 27(2):58–65.

____. 2003b. Predictive Models for Motorcycle Accidents at Three-Legged Priority Junctions. *Traffic Injury Prevention* 4:363–369.

Highway Planning Unit (HPU). 2001a. *Golden River Permanent Count Station: Annual Report.* Kuala Lumpur, Malaysia: Ministry of Works.

____. 2001b. *Traffic Volume Malaysia: Biannual Report.* Kuala Lumpur, Malaysia: Ministry of Works.

Hills, B.L. and C.J. Baguley. 1993. Accident Data Collection and Analysis: The Use of the Microcomputer Package MAAP in Five Asian Countries. *Proceedings of the Conferences on Asian Road Safety (CARS'93),* Kuala Lumpur, Malaysia, April 6–31.

Kulmala, R. 1992. Prediction Model for Accidents at Highway Junctions. *ITE Compendium of Technical Papers,* 302–305.

Lynam, D., J. Broughton, R. Minton, and R.J. Tumbridge. 2001. *An Analysis of Police Reports of Fatal Accidents Involving Motorcycles,* Report TRL 492. Berkshire, England: TRL Limited.

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models,* 2nd edition. London, England: Chapman and Hall.

McShane, W.R., R.P. Roess, and E.S. Prasas. 1998. *Traffic Engineering,* 2nd edition. Upper Saddle River, NJ: Prentice Hall.

Mountain, L., B. Fawaz, and D. Jarret. 1996. Accident Prediction Models for Roads with Minor Junctions. *Accident Analysis and Prevention* 28(6):695–707.

Mountain, L., M. Maher, and B. Fawaz. 1998. The Influence of Trends on Estimates of Accidents at Junctions. *Accident Analysis and Prevention* 30(5):641–649.

Numerical Algorithm Group (NAG). 1994. *The GLIM System: Release 4 Manual,* 2nd edition. Edited by B. Francis, M. Green, and C. Payne. Oxford, England: Clarendon Press.

Organization for Economic Cooperation and Development (OECD). 2002. *Fatalities by Traffic Participation, International Road Traffic and Accident Database (IRTAD).* Paris, France.

Radin Umar, R.S. 1996. Accident Diagnostic System with Special Reference to Motorcycle Accidents in Malaysia, Ph.D thesis, University of Birmingham, England.

Radin Umar, RS., M. Ahmad Rodzi, and A. Aminuddin. 1993. Model Diagnosis dan Rawatan Kemalangan Jalan Raya di Malaysia. *Pertanika Journal of Science and Technology* 1(1):125–151.

Radin Umar, R.S., G.M. Mackay, and B.L. Hills. 1995. Preliminary Analysis of Exclusive Motorcycle Lanes Along the Federal Highway F02 in Shah Alam, Malaysia. *IATSS Research* 19(2):93–98.

____. 2000. Multivariate Analysis of Motorcycle Accidents and the Effect of Exclusive Motorcycle Lanes in Malaysia. *Crash Prevention and Injury Control* 2(1):11–17.

Rodriguez, L.P. and T. Sayed. 1999. Accident Prediction Models for Urban Unsignalized Intersections in British Columbia, presented at the 78th Annual Meetings of the Transportation Research Board, Washington, DC, January.

Saied, A.M. and G.M. Said. 2001. A General Linear Model Framework for Traffic Conflicts at Uncontrolled Intersections in Greater Cairo. *Proceedings of the Conferences on Traffic Safety on Three Continents, Moscow, Russia,* VTI konferens 18A Part 3. Birmingham, England: VTI, Swedish National Road and Transport Research Institute.

Summersgill, I. 1991. *What Determines Accident Risk? Papers on Vehicle Safety, Traffic Safety and Road User Safety Research, Safety 91.* Berkshire, England: TRL Limited.

Swedish Institute (SI). 2000. Road Safety in Sweden. *Fact Sheets on Sweden.* Stockholm, Sweden.

Tarko, A.P., S. Eranky, K.C. Sinha, and R. Scinteie. 1999. An Attempt to Develop Crash Reduction Factors Using Regression Technique, presented at the 78th Annual Meetings of the Transportation Research Board, Washington, DC, January.

Taylor, M.C., A. Baruya, and J.V. Kennedy. 2002. *The Relationship Between Speed and Accidents on Rural Single-Carriageway Roads,* Report TRL 511. Berkshire, England: TRL Limited.

Transport Canada. 2001. *Canadian Motor Vehicle Traffic Collision Statistics*. Available at www.tc.gc.ca/roadsafety/stats/menu.htm.

U.S. Department of Transportation (USDOT), National Highway Traffic Safety Administration, National Center for Statistics and Analysis. 2002. *Traffic Safety Facts 2001: A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System*, Report No. DOT HS 809 484. Washington, DC.

Vogt, A. 1999. *Crash Models for Intersections: Four-Lane by Two-Lane Stop-Controlled and Two-Lane by Two-Lane Signalized*, Report Number FHWA-RD-99-128. Washington, DC: U.S. Department of Transportation, Federal Highway Administration.

Vogt, A. and J.G. Bared. 1998. *Accident Models for Two-Lane Rural Road: Segments and Intersections,* Report Number FHWA-RD-98-133. Washington, DC: U.S. Department of Transportation, Federal Highway Administration.

Wang, Y. and H. Ieda. 1997. Effects of Drivers' Age, Flow Rate and Some Other Road Environment Related Factors on Traffic Accidents at Four-Legged Signalized Intersections, presented at the 2nd Conference of the Eastern Asia Society for Transportation Studies, Seoul, Korea.

Effects of Extreme Values on Price Indexes: The Case of the Air Travel Price Index

JANICE LENT

Bureau of Transportation Statistics
Research and Innovative Technology Administration
U.S. Department of Transportation
400 7th Street, SW
Washington, DC 20590
Email: janice.lent@dot.gov

ABSTRACT

This paper examines the effects of extreme price values on the Fisher and Törnqvist index formulas. Using a simple model, we first consider the impact of outliers on the unweighted arithmetic, harmonic, and geometric means of a collection of values. Then, under the same model, we investigate the effect of a single extremely high or low price on the price index formulas (weighted means). Further investigation using Taylor series approximations leads to some general conclusions regarding the relative robustness of the Fisher and Törnqvist indexes. These are illustrated with empirical results based on airfare data from the U.S. Department of Transportation's Origin and Destination Survey.

INTRODUCTION

Many economists have come to favor the "superlative" Fisher and Törnqvist price indexes over the more traditional Laspeyres formula (see, e.g., Diewert 1976; Aizcorbe and Jackman 1993). The U.S. Bureau of Labor Statistics recently began publishing a new price index series targeting the Törnqvist formula. The choice between the Fisher and Törnqvist formulas may be based on a variety of factors, including other price index formulas currently in use by the

KEYWORDS: Price index, extreme value, Taylor series. JEL Categories: C43, C13, E31.

organization producing the index and the relative sensitivity of the two formulas to extreme values. In this study, we compare the Fisher and Törnqvist formulas with respect to the latter criterion—sensitivity to extreme values.

Extreme-valued price ratios often occur as a result of deep discounts or "free" promotional goods or services. Such extreme-valued ratios can be either large or small, depending on whether the discounted price appears in the numerator or denominator of the price ratio. Less often, extremely high prices appear with converse effects. The Laspeyres formula is sometimes criticized as sensitive to extreme values, because it is based on an arithmetic mean of the price ratios. We will see, however, that such sensitivity depends on the direction of the outlying value (high or low), as well as on the weights used in the selected mean.

In the next section, we consider the effect of an extreme value on the unweighted arithmetic, harmonic, and geometric means. The third section contains a discussion of the corresponding effects on the Fisher and Törnqvist index formulas under differing assumptions regarding the correlation between the expenditure-share weights and the prices. This correlation is related to the elasticity of substitution (i.e., the extent to which consumers shift their purchases toward lower priced items when relative prices change).

The fourth section presents an empirical example: the case of air travel index estimates computed using data from the Passenger Origin and Destination Survey collected by the Bureau of Transportation Statistics. The extreme-valued price ratios in this application resulted from a change in data-collection procedures and are in this sense artificial. They do, however, provide an opportunity to compare the performances of the different index formulas under the conditions represented by the application. We summarize our conclusions in the final section.

EFFECTS OF EXTREME VALUES ON UNWEIGHTED MEANS

The following simple model shows the effects of an extreme value on three types of unweighted means. Let $x_1, ..., x_n$ be a collection of nonnegative values, where $x_i = \mu$ for $i = 1, ..., n-1$, while $x_n = y\mu$ for some factor $y > 0$; that is, x_n is an outlier in the collection. We define the unweighted arithmetic, harmonic, and geometric means, respectively, as follows:

$$A = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$H = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1},$$

and

$$G = \prod_{i=1}^{n} x_i^{1/n}.$$

For $M \in \{A, H, G\}$, let

$$f_M = \frac{M}{\mu}.$$

Then

$$f_A = \frac{n-1+y}{n},\ f_H = \frac{n}{n-1-y^{-1}},\ \text{and}$$

$$f_G = y^{1/n}.$$

We first consider the rate at which the various means approach μ as n approaches infinity. For fixed y, we have

$$|1-f_A| = \left|\frac{1-y}{n}\right| = O\left(\frac{1}{n}\right)$$

as $n \to \infty$. For the harmonic mean also,

$$|1-f_H| = \left|\frac{1-y^{-1}}{n-\left(1-y^{-1}\right)}\right| = O\left(\frac{1}{n}\right)$$

as $n \to \infty$, and, similarly,

$$|1-f_G| = \left|1-e^{\frac{1}{n}\ln y}\right| = \left|-\frac{1}{n}\ln y - O\left(\frac{1}{n^2}\right)\right| = O\left(\frac{1}{n}\right)$$

as $n \to \infty$. Thus, as n becomes large, all three of the means approach μ at approximately the same rate. Their behavior in the presence of an outlier differs, however, under various assumptions about the

TABLE 1 Limits and Orders of Magnitude for the Unweighted Mean Factors

	$y \to \infty$		$y \to 0$	
	Limit	Order of magnitude	Limit	Order of magnitude
$f_A(y)$	∞	$\Omega(y)$	$\frac{n-1}{n}$	$O(y)$
$f_G(y)$	∞	$O(y^{(1/n)})$	0	$\Omega(y^{(1/n)})$
$f_H(y)$	$\frac{1}{n-1}$	$\Omega(y)$	0	$O(y)$

outlier itself. If we suppose that n is fixed, we may follow derivations similar to those above to arrive at the results, which are summarized in table 1.

The results shown in table 1 for $f_G(y)$ may lead us to conclude that price index formulas based on the geometric mean are, overall, the most robust formulas available; at the very least, they represent a sensible choice when both high and low outliers are expected to occur. By contrast, while A is robust to low outliers, it is sensitive to high outliers; similarly, H is robust to high outliers but sensitive to low ones. In most applications, however, price indexes are not computed as unweighted means. In the next section, we examine the effect of expenditure-share weights on the Laspeyres, Paasche, Fisher, and Törnqvist indexes, with special emphasis on the latter two.

EFFECTS OF EXTREME VALUES ON PRICE INDEXES

Price Index Formulas

We begin by presenting several population index formulas. The Laspeyres index measuring price change between time periods 1 and 2 is defined as

$$L = \frac{\Sigma_{j=1}^{N} q_{j1}p_{j2}}{\Sigma_{j=1}^{N} q_{j1}p_{j1}} = \sum_{j=1}^{N} w_{j1}\left(\frac{p_{j2}}{p_{j1}}\right),$$

where p_{jt} denotes the price of item j at time $t \in \{1,2\}$, q_{jt} denotes the quantity of item j purchased at time t,

$$w_{jt} = p_{jt}q_{jt}/\Sigma_{k=1}^{N} p_{kt}q_{kt},$$

and N denotes the number of items in the target population. The weight w_{jt} is the *expenditure share* for item j in period t; the price ratios p_{j2}/p_{j1} are often called *price relatives*. Clearly L is the arithmetic mean of the price relatives with weights representing first period expenditure shares. The Paasche index is a harmonic mean of the price ratios, with second period expenditure-share weights:

$$P = \frac{\Sigma_{j=1}^{N} q_{j2}p_{j2}}{\Sigma_{j=1}^{N} q_{j2}p_{j1}} = \frac{1}{\Sigma_{j=1}^{N} w_{j2}\left(p_{j2}/p_{j1}\right)^{-1}}.$$

The Fisher index is simply defined as $F = \sqrt{LP}$, while the Törnqvist is a geometric mean of the price ratios with weights representing the averages of the period 1 and period 2 expenditure shares, shown as

$$T = \prod_{j=1}^{N}\left(\frac{p_{j2}}{p_{j1}}\right)^{w_{j,1,2}},$$

where $w_{j,1,2} = (w_{j1} + w_{j2})/2$.

Extreme Values and the Elasticity of Substitution

To examine the effects of an outlier on the indexes described above, suppose we have a collection of n items priced in time periods 1 and 2. Suppose further that for $j = 1, ..., n$ we have $p_{j1} = q_{j1} = 1$ and that for $j = 1, ..., n-1$ we also have $p_{j2} = 1$, while $p_{n2} = y$ (i.e., we assume for simplicity that the μ above is 1.) For $t \in \{1,2\}$, let $x_{jt} = p_{jt}q_{jt}$, the expenditure level for item j in period t. We wish to allow the quantity of an item purchased to vary in response to price change and an assumed elasticity level. When $p_{j2} = p_{j1}$, we assume that $q_{j2} = q_{j1}$. Otherwise, let

$$q_{j2} = q_{j1}\left(\frac{p_{j2}}{p_{j1}}\right)^{-\tau}$$

where $0 \le \tau \le 1$, and τ is assumed constant. Then

$$x_{j2} = x_{j1}\left(\frac{p_{j2}}{p_{j1}}\right)^{1-\tau}.$$

We define the elasticity τ in this way, because τ provides a convenient means of examining the effects of extreme values under conditions of high and low elasticity, defined relatively. Note that higher values of τ indicate less impact of price change (represented by the price ratios) on second period item-level expenditure levels.

For $j = 1, ..., n-1$, we have $q_{j2} = q_{j1} = 1$; and $q_{n2} = y^{-\tau}$.

The resulting first and second period expenditure-share weights are as follows:

$$w_{j1} = \frac{1}{n}, j = 1, ..., n \ ;$$

$$w_{j2} = \frac{1}{n-1+y^{1-\tau}}, j = 1, ..., n-1 \ ;$$

and

$$w_{n2} = \frac{y^{1-\tau}}{n-1+y^{1-\tau}}.$$

The "average weights" used in the Törnqvist index are

$$w_{j,1,2} = \frac{1}{2}\left(\frac{1}{n} + \frac{1}{n-1+y^{1-\tau}}\right), j = 1, ..., n-1,$$

and

$$w_{n,1,2} = \frac{1}{2}\left(\frac{1}{n} + \frac{y^{1-\tau}}{n-1+y^{1-\tau}}\right).$$

Note that when τ is small (low or zero elasticity) and y is large,

$$w_{n1} < w_{n,1,2}, \tag{1}$$

so the Laspeyres index gives less weight to high outliers than does the Törnqvist index. Similarly, when τ and y are both small,

$$w_{n2} < w_{n,1,2}, \tag{2}$$

indicating that the Paasche index gives less weight to low outliers than the Törnqvist. Under conditions of low elasticity, we therefore observe the following phenomena: although the Laspeyres index, based on the arithmetic mean, is sensitive to high outliers, it assigns them weights that are low relative to the Törnqvist weights, while the Paasche index, a harmonic mean, assigns lower weights to low outliers. The weights in the Laspeyres and Paasche indexes can therefore be expected to compensate, at least partially, for the sensitivity of the arithmetic and harmonic means to high and low outliers, respectively.

Under this simple model, the values of the Laspeyres, Paasche, Fisher, and Törnqvist indexes are as follows:

$$L(n, y) = \frac{n-1+y}{n};$$

$$P(n, y, \tau) = \frac{n-1+y^{1-\tau}}{n-1+y^{-\tau}};$$

$$F(n, y, \tau) = \left[\left(\frac{n-1+y}{n}\right)\left(\frac{n-1+y^{1-\tau}}{n-1+y^{-\tau}}\right)\right]^{1/2};$$

and

$$T(n, y, \tau) = \exp\left[\frac{1}{2}\left(\frac{1}{n} + \frac{y^{1-\tau}}{n-1+y^{1-\tau}}\right)\ln y\right].$$

Both the Fisher and Törnqvist indexes are known as superlative indexes, because economic theory suggests that they approximate a true cost of living index under relatively weak assumptions regarding economic conditions (Diewert 1987). (In the application considered in the next section, these indexes should be viewed as cost of flying indexes rather than as cost of living indexes.) We, therefore, focus on the relative robustness of $F(n, y, \tau)$ and $T(n, y, \tau)$ under the assumptions $\tau = 1$ and $\tau = 0$. The value $\tau = 1$ indicates that consumers shift their purchases toward items (or item categories) whose relative prices have decreased between periods 1 and 2, while τ close to zero represents the case of little or no change in buying behavior in response to price change.

First consider the case $\tau = 1$, where a value of τ represents the assumption that consumers alter the quantities of the items they purchase so as to maintain the same level of expenditure on each item—a situation corresponding to a fairly high level of elasticity. In this case, we have, for fixed n and large y,

$$F(n, y, 1) = \left[\left(\frac{n-1+y}{n}\right)\left(\frac{n}{n-1+y^{-1}}\right)\right]^{1/2}$$

$$= \left(\frac{n-1+y}{n-1+y^{-1}}\right)^{1/2}$$

$$\approx \left(1+\frac{y}{n}\right)^{1/2}, \tag{3}$$

while

$$T(n, y, 1) = y^{1/n}. \tag{4}$$

So, for reasonably large n, T is more robust than F in the presence of high outliers. For the case of low outliers, we have

$$F(n, y, 1) = \left(\frac{n-1+y}{n-1+y^{-1}}\right)^{1/2} = O\left(y^{1/2}\right) \quad (5)$$

and

$$T(n, y, 1) = \Omega\left(y^{1/n}\right) \quad (6)$$

for fixed n as y approaches 0. Under the simple model, we may therefore conclude that, with regard to robustness, conditions of high elasticity favor the Törnqvist index over the Fisher.

With $\tau = 0$, we have

$$F(n, y, 0) = \left(\frac{n-1+y}{n}\right),$$

and

$$T(n, y, 0) = \exp\left[\frac{1}{2}\left(\frac{1}{n} + \frac{y}{n-1+y}\right)\ln y\right].$$

Note that $F(n, y, 0) = L(n, y, 0) = P(n, y, 0)$. For fixed n and large y,

$$F(n, y, 0) \approx 1 + \frac{y}{n}, \quad (7)$$

while

$$T(n, y, 0) \approx y^{(n+1)/2n}. \quad (8)$$

As a rough rule of thumb, the above approximations suggest that T is likely to outperform F whenever outliers are as large as n^2. The relative robustness of T and F thus depends on the relative values of y and n, which may, in turn, depend on the aggregation level being considered. Equations (4) and (8) also indicate that, for large values of n, T is much more robust to high outliers under high elasticity than it is under low elasticity. For low outliers, however, the elasticity assumption has less impact on T. With n fixed and y small, we have

$$F(n, y, 0) = \Omega\left(\frac{n-1}{n}\right), \quad (9)$$

and

$$T(n, y, 0) = O\left(y^{1/2n}\right), \quad (10)$$

revealing that, under conditions of low elasticity, T is more sensitive to low outliers than F. Equations (5), (6), (9), and (10) suggest that T is somewhat more robust to low outliers for $\tau = 0$ than for $\tau = 1$, while F is much more robust.

The above results lead us to conclude that, under conditions of low elasticity, the Fisher index may often be more robust to outliers than the Törnqvist: the Fisher is more robust to low outliers and, when n is sufficiently large relative to any price ratios in the dataset, the Fisher is also more robust to high outliers. Conditions of higher elasticity (τ close to 1) render both indexes more robust to extremely high values. Under conditions of high elasticity, the Törnqvist is preferable to the Fisher, because it is less sensitive to both high and low outliers.

The numerical examples shown in appendix A illustrate these conclusions. Tables A1 and A2 give values of the Fisher and Törnqvist indexes under the single outlier scenario described above. (Note that these are not random values produced by a Monte Carlo simulation but simply the values of the functions $F(n, y, \tau)$ and $T(n, y, \tau)$ for the given parameters.) Table A1 gives index values under the assumption that $\tau = 1$ (high elasticity). Under this assumption, the Törnqvist is clearly more robust than the Fisher to both high and low outliers.

Table A2 shows values of the indexes under the assumption that $\tau = 0$. The bold numbers in this table highlight points at which y becomes large enough, relative to n, to render the Törnqvist index better than the Fisher for approximating the mean $\mu = 1$ in the presence of a high outlier. As expected, the turning points occur as y approaches n^2. The examples in table A2 also illustrate that, under low elasticity, both indexes are more sensitive to high outliers and less sensitive to low outliers than they are under high elasticity.

Taylor Series Results

The single outlier model employed in the previous subsections does not, of course, account for the data complexity often encountered in practical applications. Here, we look beyond the single outlier model to examine Taylor series expansions that shed further light on the relative robustness of the Fisher and Törnqvist indexes under the general assumption of low elasticity. Following and expanding on the development of Lent and Dorfman (2004a), we assume that the price indexes are computed from a

collection of expenditure share weights and *sub-indexes* I_g, which here take the place of the price ratios p_{j2}/p_{j1} in the previous subsection. Each $\hat{I}_g$ is an aggregate of the ratios p_{j2}/p_{j1} for all items j in a particular item category g. In practice, the standard formulas are often applied in this two-step fashion. The categories into which we divide the items may be defined according to item characteristics, geographic area of purchase, or both.

We begin by defining some notation. For $t \in \{1,2\}$ and for each item category g, let

$$w_{gt} = \frac{\Sigma_{j \in g} p_{jt} q_{jt}}{\Sigma_g \Sigma_{j \in g} p_{jt} q_{jt}},$$

and let

$$\mu_t = \sum_g w_{gt} I_g, \qquad \sigma_t^2 = \sum_g w_{gt} (I_g - \mu)^2,$$

and

$$\gamma_t = \sum_g w_{gt} (I_g - \mu)^3.$$

Next, with w_g defined as the Törnqvist weights,

$$w_g = \frac{w_{g1} + w_{g2}}{2},$$

let

$$\mu = \sum_g w_g I_g, \qquad \sigma^2 = \sum_g w_g (I_g - \mu)^2,$$

and

$$\gamma = \sum_g w_{gt} (I_g - \mu)^3.$$

We expand each of the superlative indexes about the point at which all of the sub-indexes I_g equal the mean μ. The relevant partial derivatives are given in appendix B. From the general form of the third-order approximation given by Lent and Dorfman (2004a) for a geometric mean, we have the following approximation of the Törnqvist index:

$$T_I = \prod_g I_g^{w_g} \approx \mu - \frac{\sigma^2}{2\mu} + \frac{\gamma}{3\mu^2}. \tag{11}$$

The second-order approximation for the Fisher index is

$$F_I = \left(\frac{\Sigma_g w_{g1} I_g}{\Sigma_g w_{g2} / I_g} \right)^{1/2} \approx \mu - \frac{\sigma_2^2}{2\mu} \tag{12}$$

Thus, to the second order, we have

$$T_I - F_I \approx \frac{\sigma^2 - \sigma_2^2}{2\mu} = \frac{1}{4\mu} \sum_g \left(w_{g2} - w_{g1} \right) \left(I_g - \mu \right)^2. \tag{13}$$

Consider the relative values of w_{g1} and w_{g2} in the presence of high outliers among the I_g and high correlation between the I_g and the w_{g2} (the case of low elasticity). Under these conditions, we are likely to have $w_{g2} > w_{g1}$ for large values of $(I_g - \mu)^2$ and thus $T_I - F_I > 0$. Similarly, in the presence of low outliers, we are likely to have $w_{g2} < w_{g1}$ for large values of $(I_g - \mu)^2$, resulting in negative values of $T_I - F_I$. Thus, the approximation shown in equation (13) indicates that, under conditions of low elasticity, the Fisher index may be more robust than the Törnqvist to both high and low outliers.

AN EMPIRICAL EXAMPLE

For the air travel price index series, the apparent elasticity of substitution is low—in some cases, even negative. The series, therefore, exemplify only the behavior of the different indexes under conditions of low elasticity (τ close to 0). Note that the elasticity reflected in the data, rather than the actual elasticity, is the quantity that affects the performance of the indexes; Dorfman et al. (1999) showed that the elasticity reflected in sample survey data need not always equal the true population elasticity.

The air travel price index series shown in figures 1 through 6 are based on data from the Bureau of Transportation Statistics' quarterly Origin and Destination (O&D) Survey. The sample for the O&D Survey comprises about 10% of all passenger itineraries having some U.S. component (i.e., itineraries that include at least one flight arriving at or departing from a U.S. airport) and includes about 6 to 7 million itineraries per quarter. Data items collected include trip route, class of service (e.g., coach, first class), and transaction fare including taxes. Note that the scales differ across figures, so comparisons across figures are distorted in some cases.

When goods and services are sampled for the purpose of estimating a price index, the sample items generally remain in the sample over an extended time period (e.g., two years) unless they are taken off the market by the retailer. The stable

sample allows comparison of prices across time for identical items. Ratios of prices in different time periods for individual items are the building blocks of the traditional price index estimators.

In the O&D Survey, however, the sampling is performed independently for each reference quarter. Since the itineraries selected in a given quarter may not match those selected for a previous or subsequent quarter, we developed and tested a two-stage process for matching categories of itineraries across quarters and comparing average prices within categories across time. The ratio of average prices for different time periods is called a *unit value index*. These subindexes are then aggregated by the Fisher, Törnqvist, and other index formulas. The index series are based only on data from sample itineraries flown on domestic carriers and are chained quarterly.[1] Lent and Dorfman (2004b) provide a more detailed description of the index estimation methodology.

The figures show the Laspeyres, Paasche, Fisher, and Törnqvist index series for various classes of service and for all classes combined. Note that, in all of the figures, the Paasche series runs either slightly below the Laspeyres series or even (for business class service) above the Laspeyres, indicating low or negative elasticity of substitution. Lent and Dorfman (2004a) describe a method of estimating the elasticity of substitution; elasticity estimates computed by their method, measuring elasticity of substitution between unit value categories as described above, run close to 0 for these data. Although air travel passengers readily substitute one carrier for another in response to fare changes, little substitution across trip origin/destination pairs occurs. Since origin/destination pairs far outnumber carriers, this substitution behavior leads to low overall estimates of elasticity of substitution between the unit value categories.

In examining figures 1 through 6, it is important to note that the Class of Service variable in the O&D Survey was redefined and standardized in 1997–98. (Formerly, air carriers had used a variety of service classifications in reporting this information, so the variable values had to be recoded by the Bureau of Transportation Statistics.) We therefore expect some unusual data values to affect the index series during this period; indeed, many of the series display a visible break between the fourth quarter of 1997 and the first quarter of 1998. These breaks may be exacerbated, because a lower percentage of the O&D Survey observations were "matched" across time during 1997–98 (see Lent and Dorfman 2004b for a description of the across-time matching method), resulting in lower than usual effective sample sizes.

Figures 1 and 2 show the series for all classes of service combined and for restricted coach class (by far the largest class), respectively. The series in figure 2 behave in typical fashion: the Laspeyres series runs just above the others, displaying a slight upward drift, while the Paasche shows a similar downward drift, and the two superlative series run between them, closely tracking each other. This type of behavior results from the large number of observations and because the 1997–98 break has relatively little impact on these series. Figure 1 is similar to figure 2, except for the noticeably larger effect of the 1997–98 change, which lifts the Törnqvist series slightly above the others. Recall that, under conditions of low elasticity, the Törnqvist index is often more sensitive to outliers than the Fisher.

Index series for other classes of service (categories comprising fewer observations) are shown in figures 3 through 6. For the unrestricted first and restricted first class indexes (figures 3 and 4), the Laspeyres series runs very slightly above the Paasche, indicating low but positive elasticity. For the unrestricted first class series, the 1997–98 break sends the Törnqvist above the other series, while the Törnqvist for restricted first class is "bumped down" and runs well below the others for 1998 and subsequent years. In both cases, the Törnqvist continues to roughly parallel the Fisher after the break, indicating that unusual data values generated the level shifts. Note also that the Törnqvist's upward shift for unrestricted first class is noticeably less severe than its downward shift for restricted first class, perhaps due to its greater robustness to high outliers than to low ones.

[1]Price index chaining is done by estimating long-term price changes as products of shorter term changes (links). Quarterly chaining can lead to "chain drift," as seen in the Laspeyres and Paasche series in the figures in this section. For more information on chain drift in the airfare indexes, see Lent 2003.

FIGURE 1 All Classes of Service: Quarterly Chained Preliminary Series

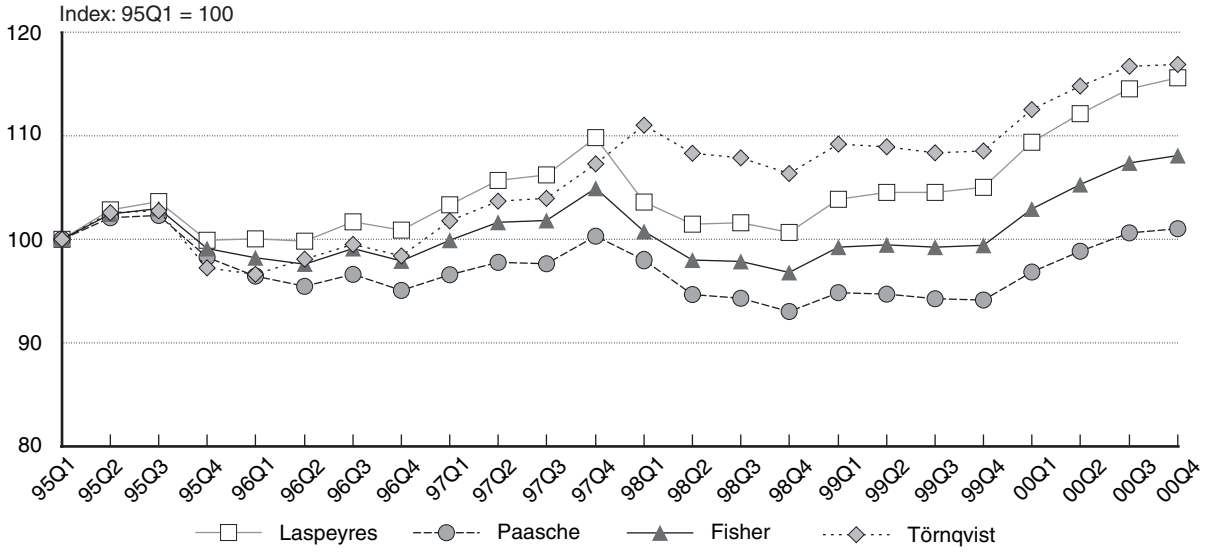
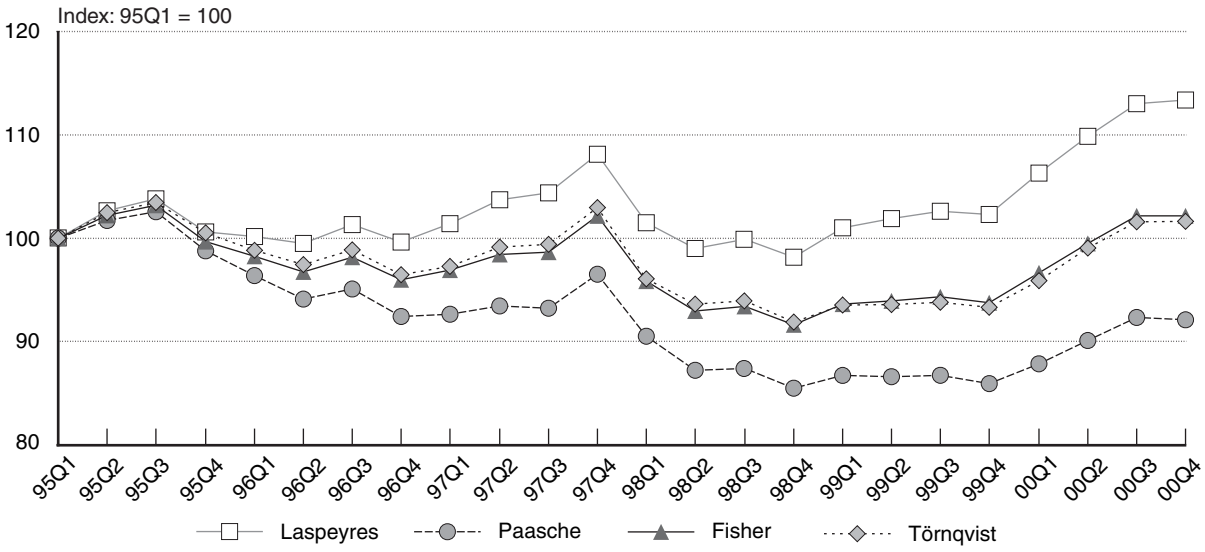


FIGURE 2 Restricted Coach Class: Quarterly Chained Preliminary Series



The business class index series (figures 5 and 6) display the relatively rare phenomenon of negative elasticity. The Paasche series runs above the Laspeyres, indicating that consumers are shifting their purchases toward *higher* priced services as relative prices change. It is important to note that sample survey data may not always reflect true population elasticity; in this case, the class-of-service categories are coarsely defined, and many different types of restrictions may apply to tickets in the restricted categories. (Restrictions may include, for example, a requirement of advance ticket purchase or, in the case of roundtrip itineraries, a Friday or Saturday night stay at the destination.) Elasticity estimates based on these data reflect substitution *between* these categories but not *within* them (for the same route and carrier) and may therefore suffer a downward bias.

On the other hand, since business class service is typically paid for by a third party (i.e., the passenger's employer), very low elasticity is expected. Some business class passengers may even choose higher priced tickets assuming that "you get what you pay for," and such behavior could also explain the negative elasticity indicated. Under negative elasticity, quantities purchased are positively correlated with

FIGURE 3 Unrestricted First Class: Quarterly Chained Preliminary Series

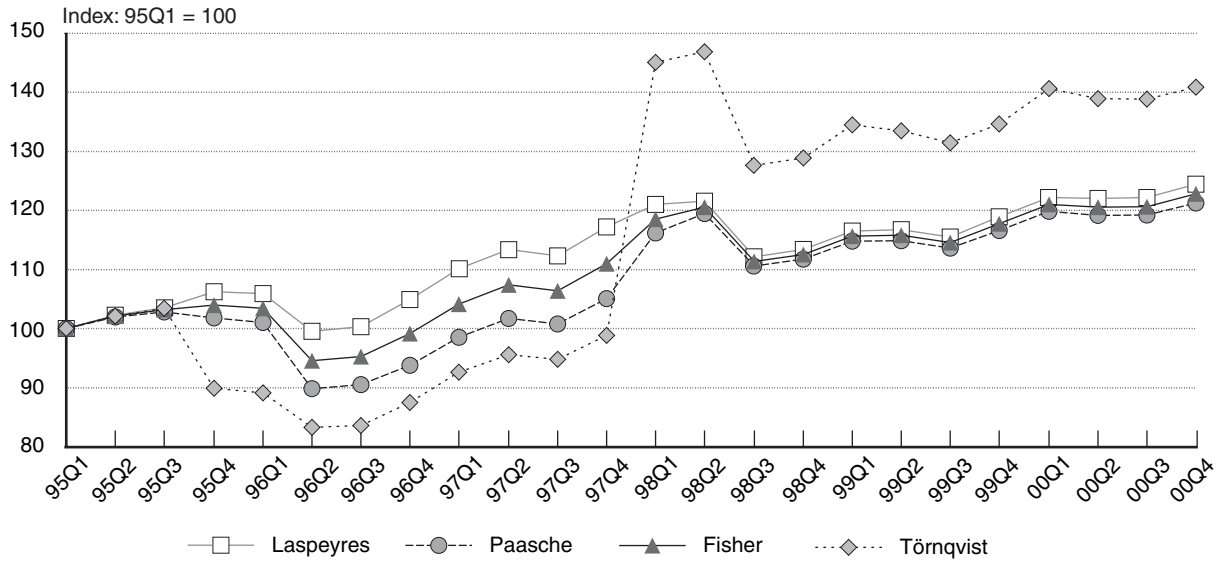
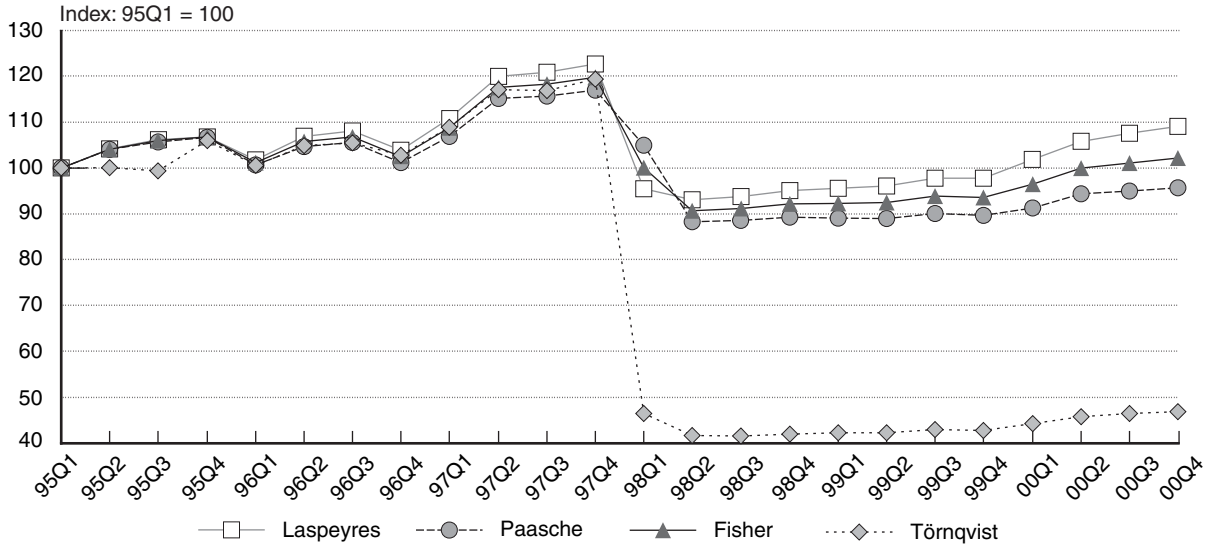


FIGURE 4 Restricted First Class: Quarterly Chained Preliminary Series



price change, and this correlation may cause expenditure shares to increase dramatically when prices increase. The Törnqvist index, whose weights are average expenditure shares, therefore assigns large weights to some high price ratios. Apart from the negative elasticity, the movements of the business class series appear similar to that of the first class series, that is, the Törnqvist index is shifted up or down during the 1997–98 period, while the other series are less affected by the unusual values. Table A3 in appendix A shows unweighted percentiles of the distributions of the unit value indexes for the first class and business class categories over the crucial period. The outliers are clearly sufficient in number and severity to impact the tails of the sample distributions.

CONCLUSIONS

Practitioners may often consider robustness to outliers an important criterion in selecting a price index formula, especially for item categories such as airfares, in which extreme prices may regularly result from frequent flyer awards and other price discriminatory discounts. Although price index formulas based on different types of means inherit the

FIGURE 5 Unrestricted Business Class: Quarterly Chained Preliminary Series

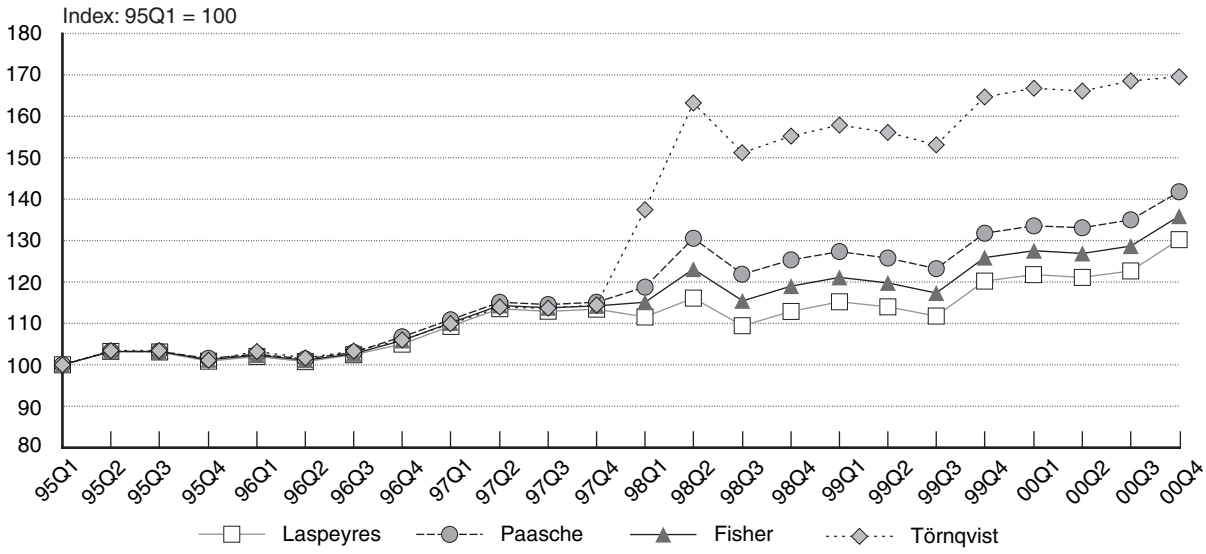
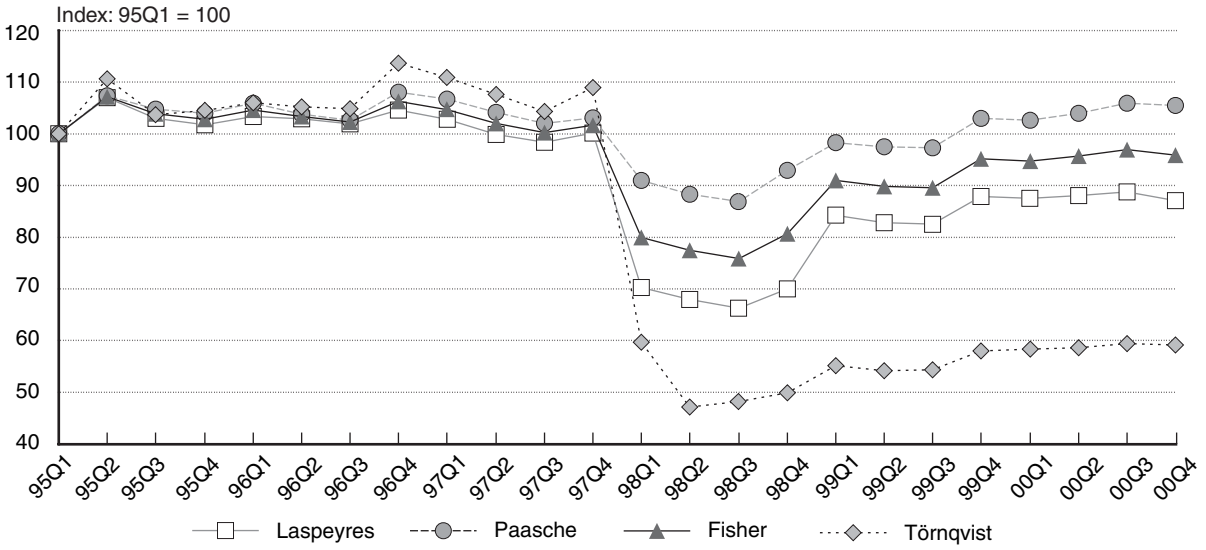


FIGURE 6 Restricted Business Class: Quarterly Chained Preliminary Series



relative robustness of these means, the weights applied in price index calculation also play a crucial role. This paper shows that, under conditions of low elasticity of substitution, the high correlation between the weights and the price ratios may offset the sensitivity of the Laspeyres and Paasche indexes, making the Fisher a more attractive option than the Törnqvist. The choice between index formulas is therefore more complex than the mere selection of an arithmetic, harmonic, or geometric mean. It requires information on the elasticity of substitution reflected in the data as well as an estimate of the magnitude of outliers (high or low) that can be expected.

ACKNOWLEDGMENT

Comments from Alan Dorfman of the Bureau of Labor Statistics resulted in significant improvements to this paper.

REFERENCES

Aizcorbe, A. and P. Jackman. 1993. The Commodity Substitution Effect in CPI Data, 1982–1991: Anatomy of a Price Change. *Monthly Labor Review,* December 1993:25–33. Washington, DC: U.S. Government Printing Office.

Diewert, W.E. 1976. Exact and Superlative Index Numbers. *Journal of Econometrics* 4:115–145.

_____. W.E. 1987. Index Numbers. *The New Palgrave: A Dictionary of Economics*. Edited by J. Eatwell, M. Milate, and P. Newman. London, England: MacMillan.

Diewert, W.E., R. Feenstra, and W. Alterman. 1999. *International Trade Price Indexes and Seasonal Commodities*. Washington, DC: U.S. Department of Labor, Bureau of Labor Statistics.

Dorfman, A., S. Leaver, and J. Lent. 1999. Some Observations on Price Index Estimators. *Proceedings of the Federal Conference on Survey Methodology,* Arlington, VA, November 1999.

Lent, J. 2003. Chain Drift in Experimental Price Index Series for Air Travel, *Proceedings of the 2003 Joint Statistical Meetings, Section on Survey Research Methods,* CD. Alexandria, VA: American Statistical Association.

Lent, J. and A. Dorfman. 2004a. Using a Weighted Average of the Jevons and Laspeyres Indexes to Approximate a Superlative Index. Working paper.

_____. 2004b. A Transaction Price Index for Air Travel. Working paper.

APPENDIX A

TABLE A1 Values of Fisher and Törnqvist Indexes with a Single Outlier
$\tau = 1, \mu = 1$

Outlier	*n* = 5		*n* = 10		*n* = 15	
y	*F*	*T*	*F*	*T*	*F*	*T*
0.010	0.18	0.40	0.27	0.63	0.34	0.74
0.025	0.27	0.48	0.41	0.69	0.49	0.78
0.050	0.37	0.55	0.53	0.74	0.62	0.82
0.100	0.48	0.63	0.66	0.79	0.74	0.86
10	1.65	1.58	1.37	1.26	1.26	1.26
20	2.18	1.82	1.70	1.35	1.50	1.35
50	3.28	2.19	2.43	1.48	2.06	1.48
80	4.09	2.40	2.98	1.55	2.50	1.55
90	4.33	2.46	3.14	1.57	2.63	1.57
100	4.56	2.51	3.30	1.58	2.76	1.58

TABLE A2 Values of Fisher and Törnqvist Indexes with a Single Outlier
$\tau = 0, \mu = 1$

	n = 5		*n* = 10		*n* = 15	
Outlier *y*	*F*	*T*	*F*	*T*	*F*	*T*
0.010	0.80	0.63	0.90	0.79	0.93	0.86
0.025	0.81	0.68	0.90	0.83	0.94	0.88
0.050	0.81	0.73	0.91	0.85	0.94	0.90
0.100	0.82	0.77	0.91	0.88	0.94	0.92
10	**2.80**	**2.87**	1.90	2.06	1.60	1.74
20	**4.80**	**4.70**	2.90	3.26	2.27	2.67
50	10.80	9.05	5.90	6.38	4.27	5.25
80	16.80	12.49	**8.90**	**8.92**	6.27	7.47
90	18.80	13.52	**9.90**	**9.68**	6.93	8.14
100	20.80	14.51	10.90	10.41	7.60	8.79

Note: Bold indicates the points at which *y* becomes large enough, relative to *n*, to render the Törnqvist index better than the Fisher for approximating the mean $\mu = 1$ in the presence of a high outlier.

TABLE A3 Unweighted Percentiles of the Unit Value Sub-index Distributions for the Airfare Index Application

	1997		1998	
	Quarter 3	Quarter 4	Quarter 1	Quarter 2
		Unrestricted business class		
No. of sub-indexes	2,944	1,739	1,800	2,985
50th percentile	0.999	1.004	1.000	1.008
90th percentile	1.286	1.292	1.376	1.565
95th percentile	1.636	1.765	1.970	3.671
99th percentile	50.464	189.000	76,283.000	10,421.000
		Unrestricted first class		
No. of sub-indexes	8,552	7,313	2,888	5,636
50th percentile	0.992	1.051	1.000	1.000
90th percentile	1.231	1.462	3.947	1.387
95th percentile	1.696	2.222	7,500.000	2.644
99th percentile	14,850.000	9,700.000	148,400.000	52,700.000
		Restricted business class		
No. of sub-indexes	2,617	2,600	2,011	1,898
50th percentile	1.000	1.005	0.982	1.000
10th percentile	0.588	0.589	0.102	0.383
5th percentile	0.245	0.297	0.009	0.100
1st percentile	0.000	0.003	0.000	0.000
		Restricted first class		
No. of sub-indexes	9,524	9,428	3,655	3,756
50th percentile	1.000	1.015	1.000	1.000
10th percentile	0.840	0.860	0.270	0.646
5th percentile	0.639	0.671	0.000	0.339
1st percentile	0.000	0.000	0.000	0.000

APPENDIX B

Partial Derivatives

To derive equations (11) and (12), the function F_I is expanded around the point $\underline{\mathbf{I}} = \underline{\boldsymbol{\mu}} = (\mu, ..., \mu)$. The general formula for the third-order Taylor expansion is

$$f(\underline{\mathbf{I}}) = f(\underline{\boldsymbol{\mu}}) + \sum f'_g(\underline{\boldsymbol{\mu}})(I_g - \mu) + \frac{1}{2}\sum\sum f''_{g1,g2}(\underline{\boldsymbol{\mu}})(I_{g1} - \mu)(I_{g2} - \mu) + \frac{1}{6}\sum\sum\sum f'''_{g1,g2,g3}(\underline{\boldsymbol{\mu}}) \bullet (I_{g1} - \mu) \bullet (I_{g2} - \mu) \bullet (I_{g3} - \mu).$$

For a derivation of the third-order expansion of T_I (equation (11)), see Lent and Dorfman (2004a). The first- and second-order partial derivatives of F_I evaluated at $\underline{\mathbf{I}} = \underline{\boldsymbol{\mu}}$ (used in the derivation of equation (12)) are as follows:

$$\left.\frac{\partial F_I}{\partial I_g}\right|_{\underline{\mathbf{I}} = \underline{\boldsymbol{\mu}}} = w_g$$

$$\left.\frac{\partial^2 F_I}{\partial I_g^2}\right|_{\underline{\mathbf{I}} = \underline{\boldsymbol{\mu}}} = \mu^{-1}\left(w_g^2 - w_{g2} + \frac{w_{g2}^2 - w_{g1}^2}{2}\right)$$

$$\left.\frac{\partial^2 F_I}{\partial I_{g_1} \partial I_{g_2}}\right|_{\underline{\mathbf{I}} = \underline{\boldsymbol{\mu}}} = \mu^{-1}\left(\frac{w_{g12}w_{g22} + w_{g11}w_{g21}}{2} + w_{g_1}w_{g_2}\right)$$

Estimating Link Travel Time Correlation: An Application of Bayesian Smoothing Splines

BYRON J. GAJEWSKI[1,*]
LAURENCE R. RILETT[2]

[1] Assistant Professor
Schools of Allied Health and Nursing
Biostatistician
Center for Biostatistics and Bioinformatics
Mail Stop 4043
The University of Kansas Medical Center
3901 Rainbow Blvd.
Kansas City, KS 66160

[2] Keith W. Klaasmeyer Chair in
Engineering and Technology,
and Director
Mid-America Transportation Center
University of Nebraska-Lincoln
W339 Nebraska Hall
P.O. Box 880531
Lincoln, NE 68588-0531

ABSTRACT

The estimation and forecasting of travel times has become an increasingly important topic as Advanced Traveler Information Systems (ATIS) have moved from conceptualization to deployment. This paper focuses on an important, but often neglected, component of ATIS—the estimation of link travel time correlation. Natural cubic splines are used to model the mean link travel time. Subsequently, a Bayesian-based methodology is developed for estimating the posterior distribution of the correlation of travel times between links along a corridor. The approach is illustrated on a corridor in Houston, Texas, that is instrumented with an Automatic Vehicle Identification system.

INTRODUCTION

Estimating and forecasting link travel times has become an increasingly important topic as Advanced Traveler Information Systems have moved from conceptualization to deployment. Sen et al. (1999) proposed estimating the correlation of travel times between various links of a corridor as an open problem for future research. In this paper, we assume that instrumented vehicles are detected at discrete

E-mail addresses:
[*] Corresponding author—bgajewski@kumc.edu
L.R. Rilett—lrilett@unl.edu

KEYWORDS: Automatic vehicle identification, Gibbs Sampler, intelligent transportation systems, Markov Chain Monte Carlo.

points in the traffic network, and links are defined as the length of roadway between adjacent detection points. The set of contiguous links forms a corridor. The link travel time for a given instrumented vehicle is calculated based on the times at which each of these vehicles passes a detection point.

Using these observations, link summary statistics, such as travel time mean and variance as a function of time of day, can be obtained. The travel time statistics for the corridor may be obtained directly or be based on the sum of the individual link travel times. In the latter case, a covariance matrix often is required, because link travel times are rarely independent.

This paper focuses on estimating the correlation of link travel times using Bayesian statistical inference. While the problem is motivated and demonstrated using vehicles instrumented with Automatic Vehicle Identification (AVI) tags, the methodologies developed can be generalized to any probe vehicle technology. In addition, while the AVI links have fixed lengths, the procedure can be applied to links of any length.

The mean link travel time is a key input for estimating the link-to-link travel time correlation coefficient. A continuous estimate of the mean link travel time as a function of the time of day is an important input to this process. In this paper, we use a natural cubic splines (NCS) approach to estimate the mean travel time as a function of time. The difference between each individual vehicle travel time and the corresponding estimated mean travel time is used, along with standard correlation equations, to obtain a point estimate of the correlation coefficient. A technique for calculating the variability of the estimate is also developed in order to make inferences about the statistical significance of this correlation coefficient.

Traditionally, variability is estimated using asymptotic theory. However, for the travel time estimation problem, this approach is complicated because of the nonparametric nature of the estimator and the covariance between links. Consequently, we adopted a Bayesian approach, which had a number of benefits in terms of interpretation and ease of use. An added benefit to this approach is that the actual distribution of the parameter is provided, which allows a much broader range of statistical information, and consequently better results, to be obtained. Further, we hypothesized that the distribution of the correlation coefficient could be used by traffic operations staff to help characterize the corridor in terms of the consistency of individual vehicle travel times relative to the mean travel time. As such, it may be considered a performance metric for traffic operations.

In this paper, an 11.1 kilometer (km) (7.0 mile) test bed located on U.S. 290 in Houston, Texas, was used to demonstrate the procedure. AVI data were obtained from the morning peak traffic period. We chose this time period because U.S. 290 experiences the highest levels of congestion in the morning than at any other time, and because estimating and forecasting travel times during congested periods are considerably more complex than during noncongested periods.

This paper is divided into four sections. First we present a traditional approach to correlation coefficient estimation with a special focus on the inherent complexities and difficulties. Next we provide detailed discussion of the proposed Bayesian approach. The third section demonstrates the methodology using AVI data observed from the test bed and compares the Bayes approach to a more traditional approach for estimating correlation. We found that the estimates and their intervals can be calculated using the proposed approach. Then, the estimated correlation coefficients are examined from the viewpoint of traffic flow theory. The last section gives concluding remarks.

We hypothesized that the positive correlation indicates the links can be categorized as consistent in that drivers who wish to drive faster (or slower) than the mean travel time can do so. Conversely, if the correlation between links is negative, then the links can be categorized as inconsistent. In this situation, drivers who are slower (or faster) than average on one link are more likely to be faster (or slower) than average on the other link. Finally, when the correlation coefficient is at or near zero, then the system is operating between the two extremes. Here, the drivers are unable to maintain consistently lower or higher travel times between links, again in relation to mean travel time, and the link travel times may be considered independent.

This latter case is often assumed in corridor travel time forecasting and estimation, although the assumption is rarely tested.

TRADITIONAL APPROACH AND SMOOTHING SPLINES

Over the last 10 years, most urban areas of North America have seen extensive deployment of intelligent transportation system (ITS) technologies. ITS traffic monitoring capabilites can be categorized based on whether they provide point or space information. For instance, inductance loop detectors provide point estimates of speed and volume. Conversely, AVI systems provide space mean speed estimates of instrumented vehicles. The focus of this paper is on AVI-equipped systems. Note that even though the focus is on AVI systems, the procedure can be readily generalized to other systems that provide space information such as those that utilize global positioning satellites or cell phone location technology.

Because of the nature of AVI systems, the speed of any one vehicle at a given time is unknown; instead, the travel time (or space mean speed) of each vehicle on each link is calculated based on the time stamps recorded at each AVI reader. The travel time of vehicle i along link l on any given day is defined as Y_{il}. Because the relationship between travel time variability and time of day may be considered unstable, a natural log transformation $z_{il} = ln(Y_{il})$ is used to stabilize the relationship. Assume that g_{il} is the expected value of z_{il}. It is assumed that the distribution of this transformation has a multivariate normal (*MVN*) distribution as shown below:

$$[z_{il}] \overset{iid}{\sim} MVN([g_{il}], \Sigma) \qquad \forall i, l \qquad (1)$$

where

$\mathbf{g}_{il}$ is a smooth function representing the mean log travel time for link l; and

$\Sigma = [\sigma_{ll'}]$, where $[\sigma_{ll'}] = \sigma_l^2$ when $l = l'$ is the variance-covariance matrix of the log travel time between links l and l'.

The normality assumption can be checked globally by inspecting the residuals. In this paper $\mathbf{g}_l$, the column vector of g_{il}'s from $i = 1,\ldots,n$, will be estimated using NCS, an approach that is discussed in detail elsewhere (Green and Silverman 1995; Eubank 1999). The fundamental calculation of an NCS is linear in nature. For example, $\hat{\mathbf{g}}_l = A(\alpha)z_l$, where z_l is the column vector of z_{il}'s, gives the mean log travel time profile for a particular day on link l, where α is the tuning parameter. The tuning parameter is discussed later, and details for the calculation of $A(\alpha)$ are shown in the appendix under "NCS Algorithm."

The test bed for this study is a three-lane section of U.S. 290 located in Houston. It has a barrier-separated high-occupancy vehicle (HOV) lane that runs along the centerline of the freeway, but the data utilized are from the non-HOV section of the freeway. Eastbound (inbound) travel time data were collected over a 11.1 kilometer (7.0 mile) stretch of U.S. 290 from 4 AVI reader stations (yielding 3 links). The lengths of links were 2.5 (1.6), 4.6 (2.9), and 4.0 (2.5) kilometers (miles), respectively. The data were collected over 20 weekdays in May 1996 from 6:00 a.m. to 8:00 a.m.

Figure 1 and table 1 outline an example of the above calculation for a subset of the data (i.e., 18 observations on day 1 that begin at 7 seconds and run to 6,822 seconds). In general, a tuning parameter, α, that is too large produces a mean estimate

FIGURE 1 Logarithm of Travel Time for Link 1 v. Time of Day
Seconds from 6 a.m. on day 1

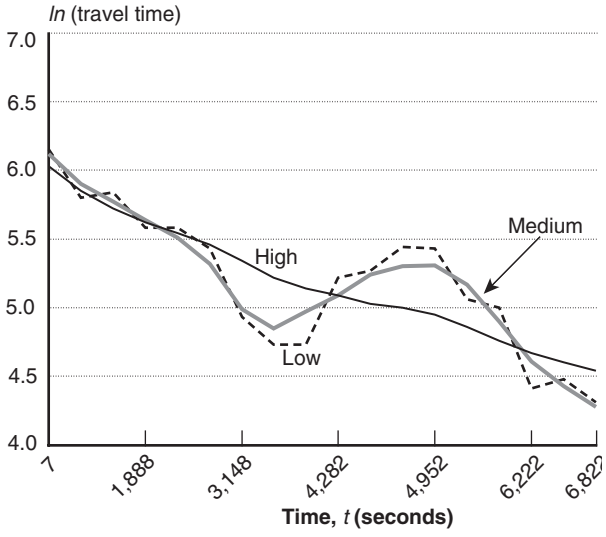


TABLE 1 Logarithm of Travel Time for Link 1 v. Time of Day
Seconds from 6 a.m. on day 1

		Estimated mean *ln* (travel time)		
Time of entry *t* (seconds)	**Observed *ln* (travel time)**	**α = 8x10^3 (Low)**	**α = 3x10^5 (Medium)**	**α = 1x10^{11} (High)**
7	6.15	6.15	6.12	6.03
816	5.80	5.80	5.90	5.85
1,415	5.85	5.84	5.77	5.72
1,888	5.58	5.58	5.64	5.62
2,252	5.58	5.58	5.51	5.54
2,607	5.43	5.43	5.32	5.46
3,148	4.93	4.93	4.99	5.34
3,691	4.74	4.73	4.85	5.22
4,086	4.69	4.73	4.97	5.14
4,282	5.27	5.22	5.09	5.09
4,569	5.23	5.27	5.24	5.03
4,731	5.47	5.44	5.30	5.00
4,952	5.43	5.43	5.31	4.95
5,356	5.05	5.06	5.17	4.86
5,811	5.02	5.00	4.90	4.76
6,222	4.38	4.41	4.61	4.67
6,531	4.52	4.48	4.43	4.60
6,822	4.29	4.31	4.28	4.54

that is too smooth and does not follow the pattern laid out by the data. For example, it can be seen that when α = 1x10^{11}, the NCS is basically a decreasing straight line and captures none of the traffic dynamics. Conversely, a tuning parameter that is too small yields a rough NCS. It may be seen that when α = 8x10^3, the function essentially runs through each observation point and does not provide adequate smoothing. However, for the intermediate value α = 3x10^5, there is an adequate tradeoff between the travel time dynamics and the smoothing. Therefore, it is important to identify a tuning parameter that is smooth but appropriately follows the dynamic trend. This can be accomplished using visual inspection or by automatic techniques.

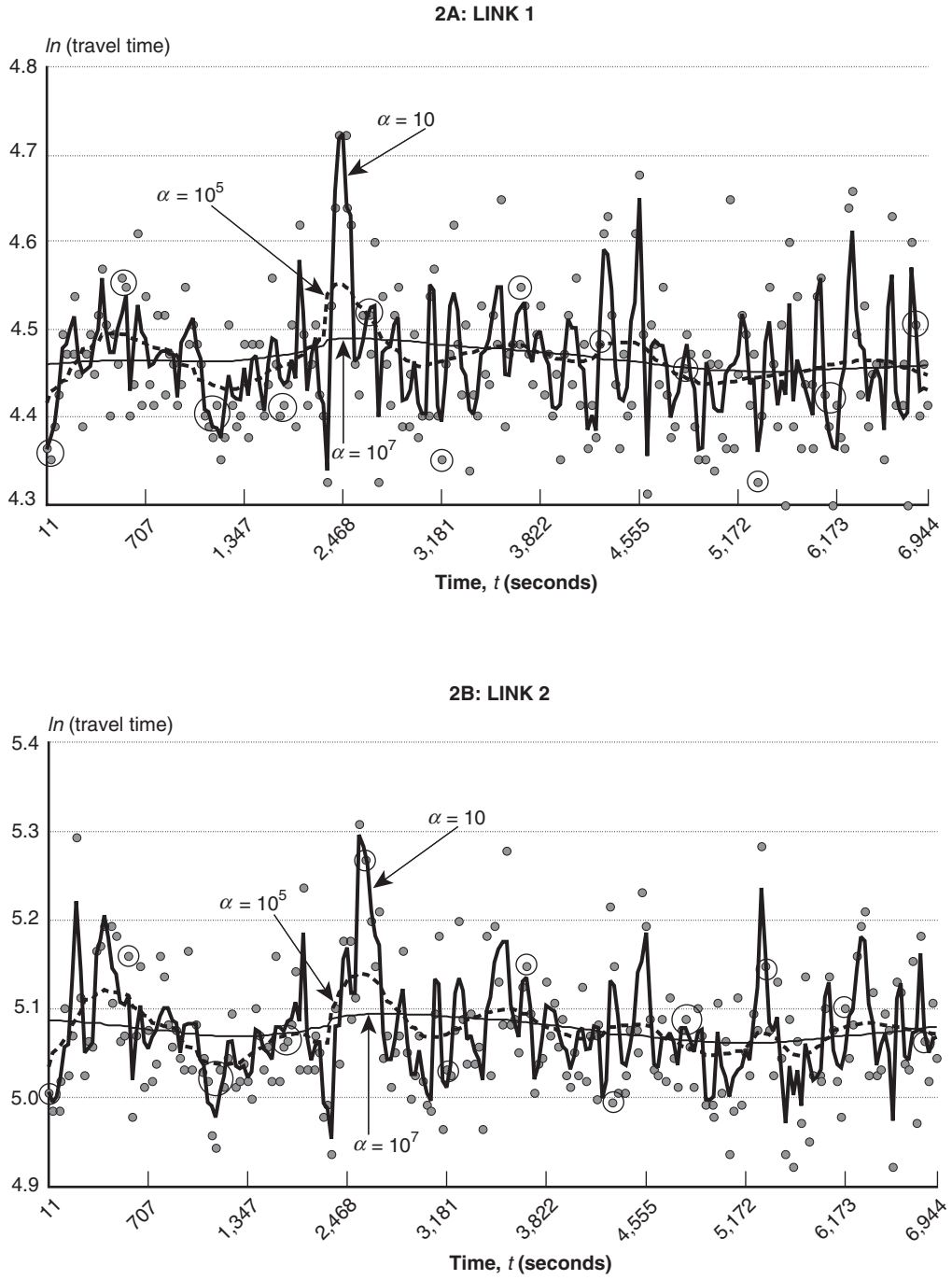
A popular choice for automating the selection of the tuning parameter is using the Generalized Cross Validation (GCV) method (Green and Silverman 1995, p. 35). In this method, the smoothing curve for one choice of tuning parameter is calculated without the first value. Subsequently, the average square error is calculated using the remaining values. This is repeated for all times during the day. A modified version of the process eliminates the need to remove each value by using the following formula:

$$GCV(\alpha) = \frac{n^{-1}\sum_{i=1}^{n}\{z_{il} - g_{il}\}^2}{\left\{1 - n^{-1}\sum_{i=1}^{n} A_{ii}(\alpha)\right\}}.$$

In essence, a convex function of the tuning parameter is drawn and the choice of tuning is the minimum of this function. The advantage of using this procedure is that the process is automated.

Figure 2 shows the log travel time as a function of time of day (6 a.m.–8 a.m.) for test bed links 1 and 2. For illustration purposes, a subset of the vehicles has its log travel times highlighted with a circle around the observation. In this particular example, the log travel time experiences only slight changes during this period of time: it begins relatively flat,

FIGURE 2 Log Travel Time v. Time of Day for Links 1 and 2
Seconds from 6 a.m. on day 19



experiences an increase at around 2,000 seconds and decreases starting at 4,000 seconds. Figure 2 also shows three NCS where each one has a different tuning parameter. For this example, a tuning parameter of $\alpha = 1\text{x}10^5$ was chosen based on visual inspection and $\alpha = 0.065\text{x}10^5$ was identified based on the GCV process. For a particular day, the same tuning parameter is applied to all links along the corridor.

To illustrate the correlation between travel times on the two links, consider the first highlighted vehicle that begins to traverse link 1 at approximately $t_1 = 11$ seconds. Note that this vehicle has a

lower than average travel time on both links 1 and 2. The second highlighted vehicle begins to traverse link 1 at approximately t = 548 seconds, and it can be seen that its observed link travel times are above the mean travel time on links 1 and 2. Eight of the 12 highlighted vehicles in figure 2 show evidence of positive correlation. Notice that this method requires that the vehicles traverse both links, and vehicles entering after the beginning of the first link or exiting before the end of the second link are not included in the correlation calculation. Later we employ these calculations for three links where we include only vehicles that traverse the entire three-link corridor.

To quantify the above relationship, we calculated the cross product of the residuals. More specifically the covariance,

$$\sigma_{12} = E\left[(z_{i1} - g_{i1})(z_{i2} - g_{i2})\right] = E[\varepsilon_{i1}\varepsilon_{i2}] \text{ and}$$

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

where σ_1 as defined earlier, was obtained using the following procedure. The mathematical details of the procedure are in the appendix under "Classic Estimation of Correlation."

Step 1: Transform the data to logarithms.
Step 2: Estimate the mean function by using NCS.
Step 3: Calculate residuals by subtracting the mean function from the logarithm travel times.
Step 4: Estimate the variance and covariance using the procedure outlined in the appendix.
Step 5: Calculate confidence intervals of the correlation using standard asymptotic distribution theory (Wilks 1962).

For the example data shown in figure 2, the above procedure resulted in an estimated correlation of $\hat{\rho}_{12}$ = 0.6918. This relatively high value reflects the positive correlation for the log travel time of vehicles between the links.

Note that the above approach is problematic because it results only in a partial solution. First, it would be desirable to have the correlation coefficient for the untransformed data Y_{il} and not the natural log of the data, $ln(Y_{il})$. Because the above correlation coefficient reflects the transformed, rather than untransformed, scale, interpretation is difficult. Secondly, and more importantly, in order to make statistical inferences regarding the correlation coefficient, the distribution of the untransformed correlation coefficient is required. Identifying the *distribution* of the untransformed correlation coefficient is equivalent to finding the standard error of this estimate in the normal distribution case using large sample theory. Additionally, because the correlation calculation requires an estimate of the mean function using NCS, this stage of uncertainty should be incorporated into the estimate of the standard error. The entire process is very difficult to accomplish, because the NCS and the sum of squares residuals need to be calculated simultaneously. The traditional or classic approach outlined in the appendix yields an approximation that does not account for the uncertainty in the NCS. To overcome these difficulties, an approach for obtaining the distribution of the correlation coefficient on the untransformed scale using Bayesian methodology is developed.

BAYESIAN APPROACH

To address the covariance problems identified in the preceding section a Bayesian methodology is employed.

General Background

In Bayesian inference, the unknown parameters of the probability distributions are modeled as having distributions of their own (Gelman et al. 2000). Generally, the identification of the distribution of the parameters, or prior distribution, is done before the data are collected. Suppose that $\underset{\sim}{\theta}$ is a vector containing unknown parameters with a prior distribution $\pi(\underset{\sim}{\theta})$. The observed data are used to update this prior distribution. The data are stored in the vector $\underset{\sim}{y}$ and its distribution, conditional on the parameter vector $\underset{\sim}{\theta}$, is the likelihood $f(\underset{\sim}{y}|\underset{\sim}{\theta})$. The parameters' distribution is updated using the Bayes theorem as shown below:

$$h(\underset{\sim}{\theta}|\underset{\sim}{y}) = \frac{f(\underset{\sim}{y}|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})}{\int_{\underset{\sim}{\theta}} f(\underset{\sim}{y}|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})d\underset{\sim}{\theta}} . \quad (2)$$

Once the posterior distribution, $h(.,.)$, is identified, it can be used to make inferences about the model parameters and to identify the percentiles, the mean, and/or the standard deviation of the distribution of the parameter.

Because the distribution shown in equation 2 is very difficult to solve, a simulation method known as Gibbs Sampler or Markov Chain Monte Carlo (MCMC) is used to approximate the distribution. This approach has become increasingly popular over the last 10 years for Bayesian inference (Gelfand 2002). The Gibbs Sampler is generally constructed of univariate pieces of the posterior distribution. (For more on the this topic, see the appendix under "Gibbs Sampler.") Note that the Gibbs Sampler requires a number of simulation replications that we denote as *nreps*.

The procedure is best illustrated by a simple example. Consider a travel time/time-of-day relationship where the mean travel time does not fluctuate and there is no need for an NCS (shown in figure 3). In this situation, it is reasonable to treat this distribution as being normally distributed, $y_i \sim N\left(\mu, \sigma^2\right)$. In this case, the parameter vector is

$$\underset{\sim}{\theta} = (\mu \ \sigma^2)'$$

where μ is the mean travel time and σ^2 is the variance of travel time. When choosing the prior distributions, it is convenient to choose a distribution of a conjugate form (Gelman et al. 2000). Because the posterior distribution is of the same family as the prior distribution, it leads to a straightforward complete conditional distribution. The prior distributions of conjugate form that we adopted in this paper are $\mu \sim N(a, p^2)$ and $\sigma^2 \sim IG(c, d)$, where *IG* is the inverse gamma distribution. (A more detailed discussion of the technical reasons for choosing conjugate prior distributions can be found in Gelman et al. (2000). The details of the Gibbs Sampler algorithm for the example are in the appendix under "Example Gibbs Sampler.")

For the simple example, *nreps* was set to 2,000 and the distributions are summarized with histograms as shown in figure 3. In figure 3B, the mean parameter is summarized. The 5th and 95th percentiles of $\mu|\underset{\sim}{y}$ are 5.068 and 5.086 seconds, respectively. Because of the simple nature of the example, it is possible to use standard methods to calculate the 90% credible intervals. Note that the classical *t*-distribution-based 90% confidence interval, which would be 5.069 to 5.085 seconds, is comparable to the percentiles of the Bayes approach because of the diffuse priors. An added advantage of the simulation is that any function of the distribution can be summarized. For example, figure 3C displays the distribution of $ln\left(\sigma^2\right)$ in the form of a histogram. It shows that, like the mean, the log variance tends to have a normal distribution. This is similar to the normal distribution properties associated with maximum likelihood estimators.

Natural Cubic Spline Bayesian Method

While the idea of Bayesian NCS has been used in other applications (Berry et al. 2002), here we expand the concept in order to calculate the covariance function for the travel time for vehicles between links. The travel time along link l when starting at time t_i is defined as Y_{il}. Because travel time variability is unstable as a function of time, the variance is stabilized using a natural log transformation $z_{il} = ln(Y_{il})$. It is assumed that this distribution will have a multivariate normal (*MVN*) distribution with a smooth mean function and a fixed covariance matrix as shown in equation 3 where:

$$[z_{il}] \overset{iid}{\sim} MVN([g_{il}], \Sigma) \qquad \forall i,l \qquad (3)$$

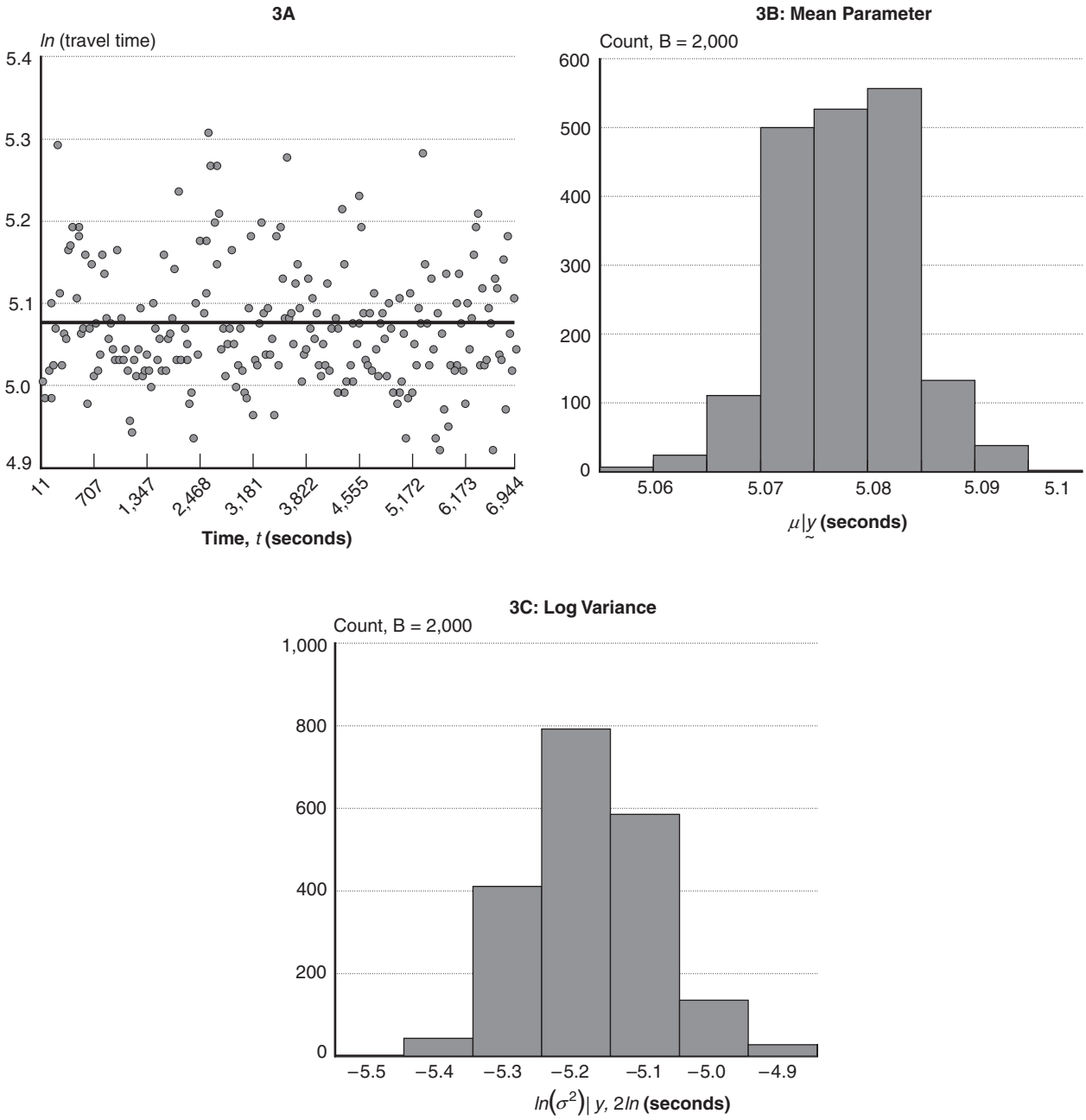
g_l is a smooth function representing the mean log travel time for link l; and

$\sum = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ & \sigma_2^2 & \sigma_{12} \\ \text{sym} & & \sigma_3^2 \end{pmatrix}$ is the variance-covariance matrix of the log travel time.

This assumption can be checked globally in the model using Bayes *p*-values (Gelman et al. 2000).

As discussed earlier, the area of focus is on the untransformed space and, therefore, standard techniques are used to calculate expectations of exponential space random variables (Graybill 1976). The moment generating function (MGF) for the multivariate normal distribution is $m(t) = \exp(\mathbf{t}'\boldsymbol{\mu} + \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2)$. Using this MGF, it can

FIGURE 3 Example Problem with Static Travel Time



be shown that the covariance for an individual vehicle between two links as a function of time is

$$\operatorname{cov}(y_{il}, y_{il'}) = \exp(g_{il} + g_{il'} + \sigma_l^2/2 + \sigma_{l'}^2/2) \bullet (\exp(\sigma_{ll'}) - 1). \quad (4)$$

The correlation coefficient of the untransformed data is shown in equation (5), and it is important to note that it is not a function of time. This allows a point estimate of the covariance between links to be specifically obtained for any given day. This is shown below:

$$corr(y_{il}, y_{il'}) = \frac{\left(\exp(\sigma_{ll'}) - 1\right)}{\sqrt{\left(\exp\left(\sigma_l^2\right) - 1\right)\left(\exp\left(\sigma_{l'}^2\right) - 1\right)}}. \quad (5)$$

The prior distribution of the smooth mean for link l is

$$g_l \sim MVN\left(\mathbf{0}, \left(\alpha K/\sigma_l^2\right)^{-1}\right)$$

where $K = \alpha n Q B^{-1} Q'$ and α are used in calculating NCS and $\Sigma \sim \text{Inv-Wishart}(\Lambda_0^{-1})$. The matrices Q and B are defined in the appendix. "Inv-Wishart" denotes the inverse Wishart distribution. If a noninformative version of the inverse Wishart distribution is used, the following posterior distribution is obtained:

$$g_l | z_l, \sigma_l^2 \alpha \sim MVN\left(A(\alpha) z_l, \sigma_l^2 A(\alpha)\right) \quad (6)$$

where

$\Sigma | \varepsilon \sim \text{Inv-Wishart}_n(S)$

$\varepsilon_l = z_l - g_l$, and

$$S = \sum_{i=1}^{n} \varepsilon_i \varepsilon_i'$$

The above posterior distributions are extensions of the multivariate calculations found in many Bayesian texts (e.g., see Gelman et al. 2000). The distributions can readily be calculated with a two-step Gibbs Sampler. The approach is summarized in the appendix under "NCS Bayesian Algorithm."

The steps can be calculated easily in any matrix-based programs that can simulate a multivariate normal distribution and an inverse Wishart. For example, S-plus, MATLAB, R, or SAS-IML would be appropriate.

DATA ANALYSIS

The methodology is illustrated on the test bed using three days' data representing three different traffic conditions: moderate, heavy, and light congestion. In all cases sampled, *nreps* is set to 500.

An assessment of this log fit can be found using Bayes factors or Bayes *p*-values (Gilks et al. 1996). In this paper, Bayes *p*-values were used to verify model goodness-of-fit and to check the validity of the underlying assumption. Two steps are involved in this calculation. First, the predictive values of the MCMC output are calculated using the MCMC model parameters. In this paper, the predictive distribution for all links at t_i is

$$\underset{\sim}{z}_{i,p}^{(b)} = \underset{\sim}{g}_i^{(b)} + e_i^{(b)} \Sigma^{(b)1/2}$$

where

$e_i^{(b)} \sim MVN(\mathbf{0}, I)$ and

p stands for predictive distribution with the "b"th iteration of the MCMC.

From the output, a χ^2 discrepancy function is calculated between the observed data and the parameters, as well as between the predicted data and the parameters from the MCMC. The discrepancy functions are calculated for each of the iterations of the MCMC. In addition, the proportion of iterations from the MCMC for which the data discrepancy is larger than the predicted discrepancy is enumerated.

The flexibility in the choice of discrepancy function allows the user to test many alternatives. The average within-link auto-correlation per day is a relevant and interesting criterion to test. This discrepancy function uses the standardized dataset's one lag auto-correlation. The standardized data are

$$\underset{\sim}{u}_i^{(b)} = \left(\underset{\sim}{z}_i - \underset{\sim}{g}_i^{(b)}\right) \Sigma^{(b)-1}$$

and the one lag correlation for the data and the predictive data is

$$T^{(b)} = \frac{\sum_{i=1}^{n-1} \underset{\sim}{u}_i^{(b)} \underset{\sim}{u}_{i+1}^{(b)}}{n-1} \text{ and}$$

$$T_p^{(b)} = \frac{\sum_{i=1}^{n-1} \underset{\sim}{e}_i^{(b)} \underset{\sim}{e}_{i+1}^{(b)}}{n-1}.$$

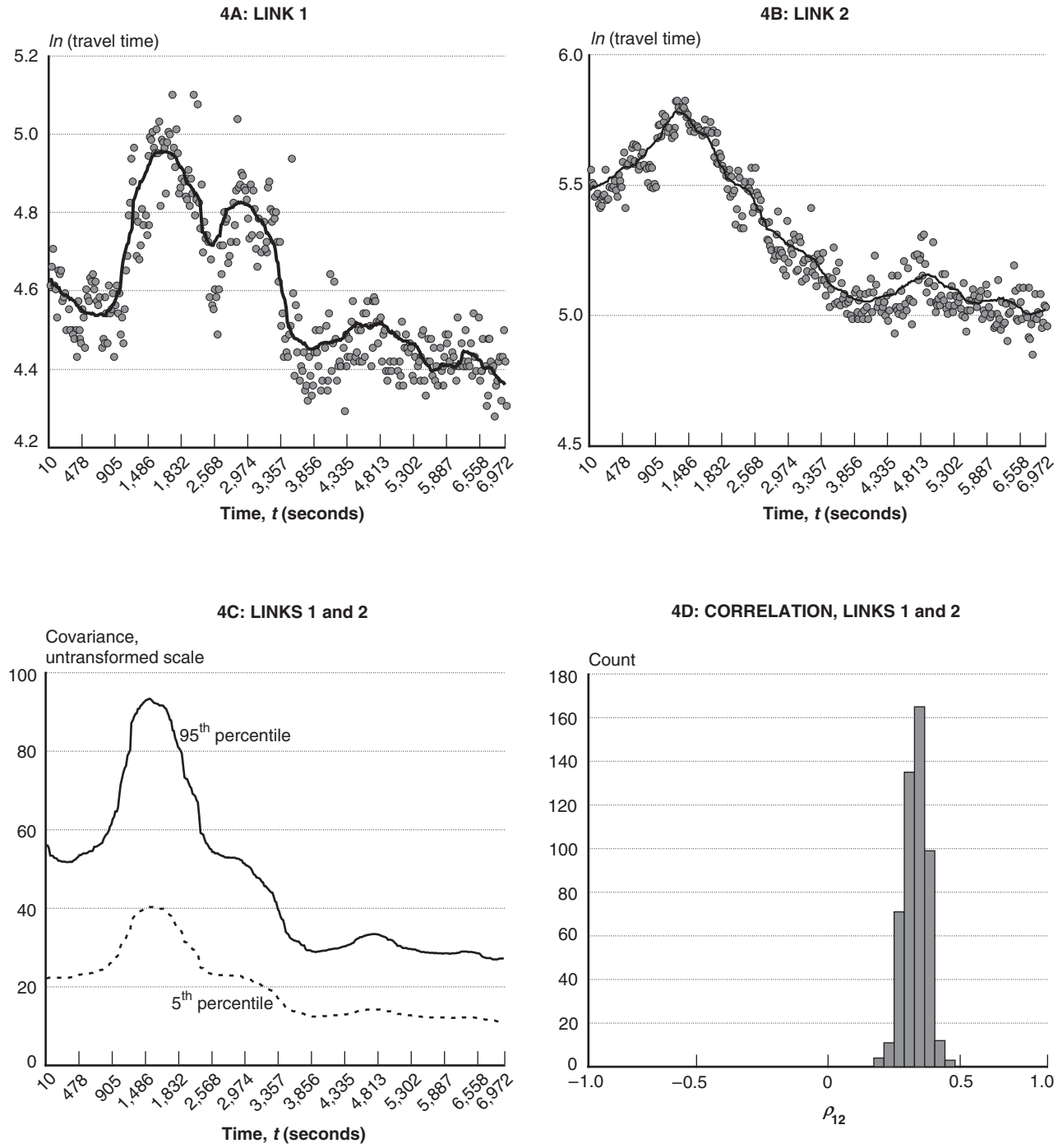
From this, the

$$p\text{-value} = \left\{ \#\left(T^{(b)} - T_p^{(b)}\right) > 0 \right\} / nreps > 0.$$

A test using this choice of discrepancy function, on all days, found that 75% of the days have a *p*-value within an acceptable range of 0.01 to 0.99. The other 25% of the days report a *p*-value less than 0.01. These latter *p*-values come from days that have predominately free-flow traffic conditions. These results indicate that while the model fits for traffic that is dynamic, it needs additional work for free-flowing traffic days. It is hypothesized that an extra parameter that accounts for auto-correlation may reduce this discrepancy. Because free-flow traffic conditions are not as interesting from a traffic monitoring center (TMC) point of view, this extension is not performed here.

Figure 4 presents data for a day with moderate traffic congestion. Figures 4A and 4B show the

FIGURE 4 Illustration of Covariance Function and Correlation for a Moderately Congested Two-Link Example: Day 10



relationship between the log travel time and time of day for links 1 and 2, respectively. In both instances, the natural log travel time fluctuates between six (high) and four (low). Using the Gibbs Sampler, the 5th and 95th percentile values of the covariance of the travel times were calculated and are shown in figure 4C. Note that the covariance is positive and fluctuates proportionally to the mean travel times of the links. Because the correlation result in equation 5 is time-independent, figure 4D can be used to show the distribution of the correlation coefficient. The distributions for the Bayes approach are summarized with the 5th and 95th percentile values. We refer here to this region as the 90% Bayes credible

region (BCR). The classic approach utilizes a 90% confidence interval (CI) based on normal asymptotic theory. This corresponds to a 90% BCR of (0.26, 0.42). The classic 90% CI was slightly narrower (0.30, 0.45).

Figure 5 shows an analysis similar to that in figure 4 but for a day in which the congestion is much greater. It can be seen that the correlation is negative between adjacent links. The 90% BCR of the correlation coefficient is (–0.39, –0.12). The classic 90% CI was narrower and had a shift of (–0.42, –0.17).

FIGURE 5 Illustration of Covariance Function and Correlation for a Very Congested Two-Link Example: Day 7

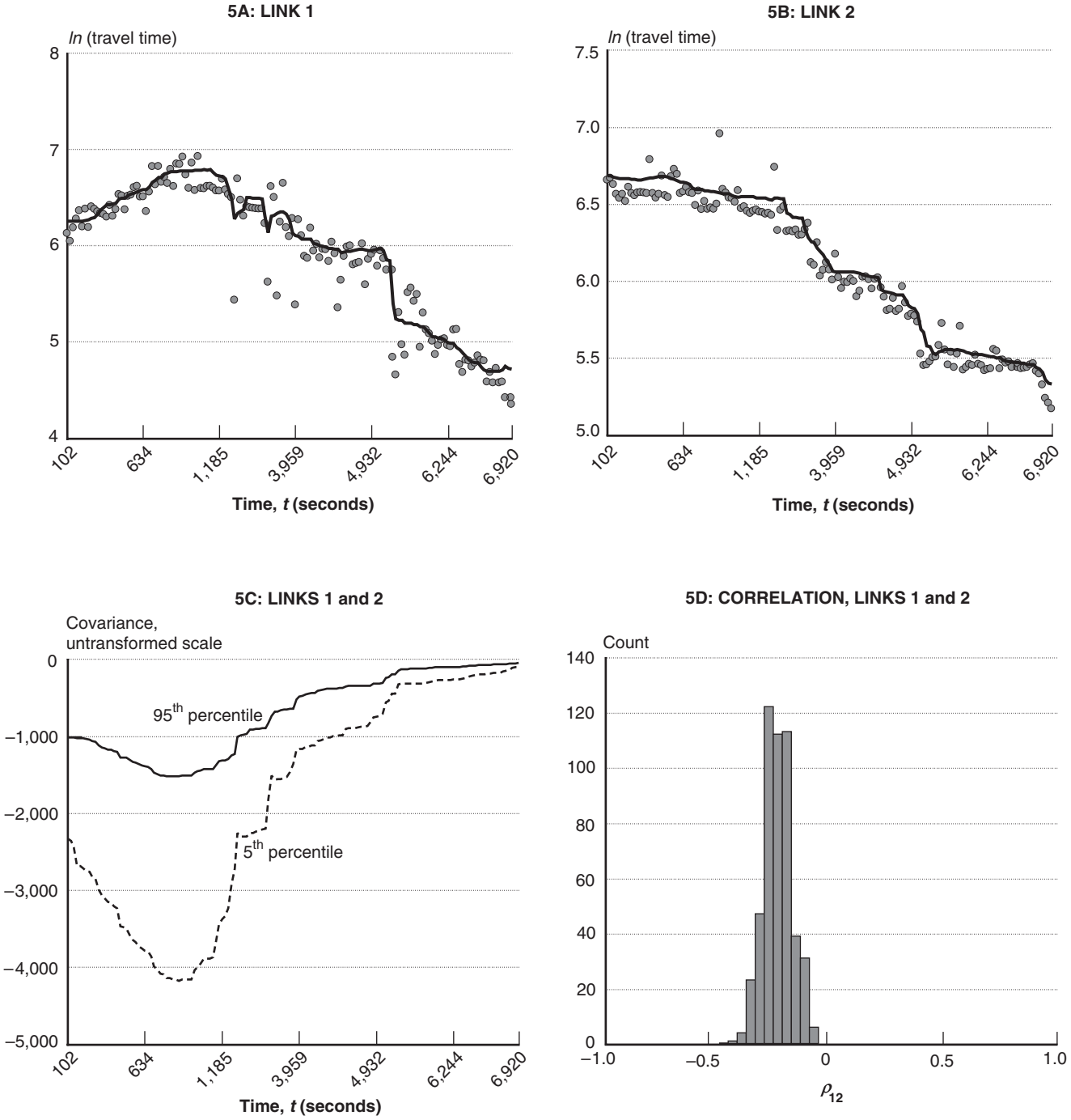
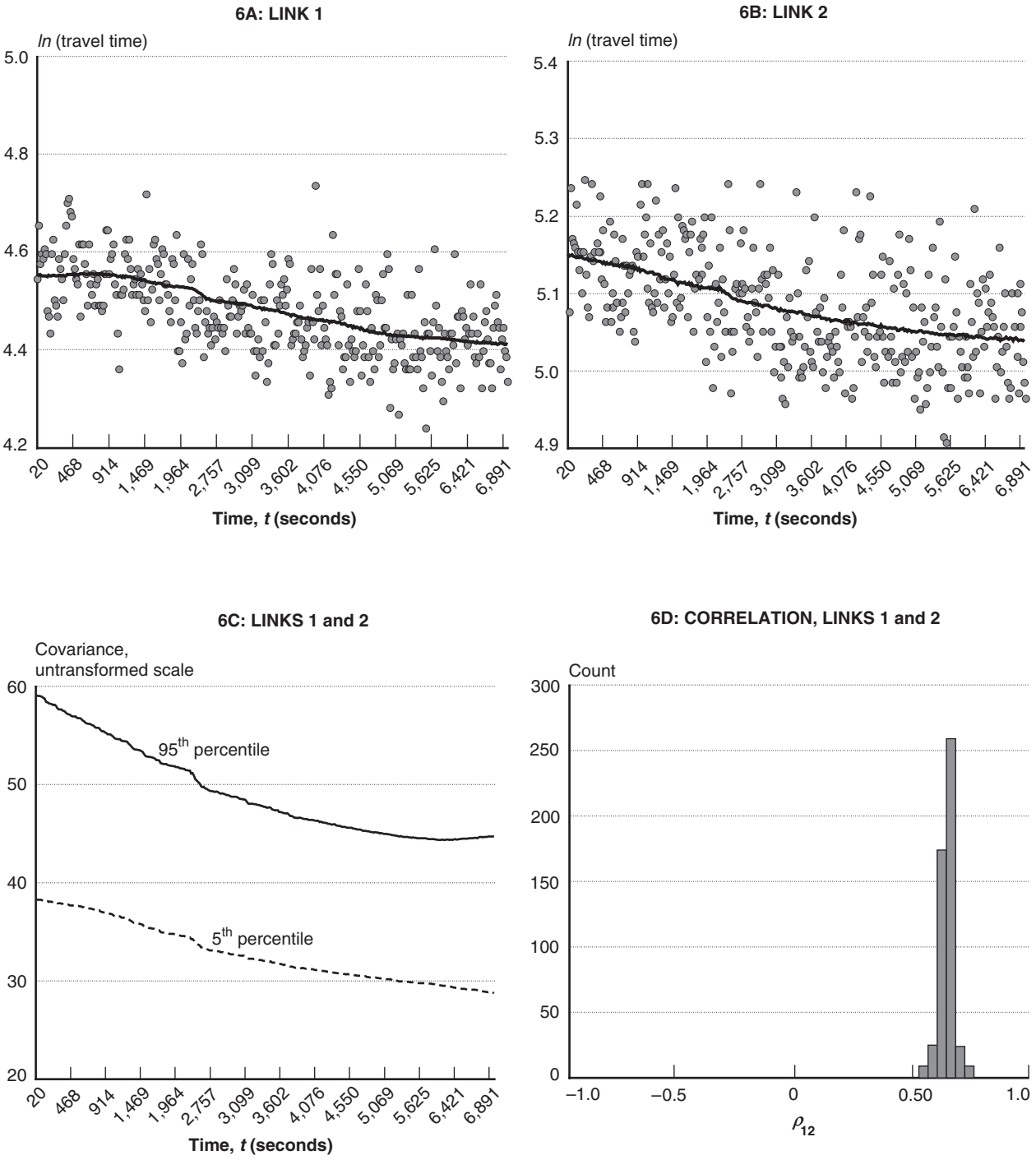


FIGURE 6 Illustration of Covariance Function and Correlation for a Less-Congested Two-Link Example: Day 20



Lastly, figure 6 shows a situation where the travel times are less dynamic with high levels of free-flowing traffic, reflected by the positive correlation. The 90% BCR of the correlation coefficient is (0.59, 0.69). The classic 90% CI is (0.61, 0.71). In summation, this correlation coefficient reflects the amount of freedom an individual vehicle has in traveling at a consistent speed relative to the overall average travel time. Figure 6 shows a high positive correlation, while figures 4 and 5 show correlation nearer zero or negative correlation, respectively.

The "C" component of all three figures reports the covariance as a continuous function of the time of day (equation (4)). It is the pretransformed space that allows for interpretations on the original scale. Notice that the fluctuations are proportional to the mean travel time from the NCS. This result illustrates the distinct advantage of the Bayes approach over the frequentist approach. In order to derive a confidence interval when using a frequentist approach, a large sample size is required to be able to apply the linear assumption used in asymptotic theory. In addition, there is a propagation of the uncertainty in the covariance because there are several steps in its estimation. For example, there is a logarithm transformation and an adjustment for the smooth mean via NCS.

In contrast, the Bayesian approach does not rely on the large sample size assumption. In addition, the nature of the MCMC iterations implicitly accounts for the transformed space. Specifically, the covariance function's actual distribution is calculated while ensuring that all forms of error are propagated. However, given the large number of probe vehicles observed, it seems conceivable that the large sample properties hold. Therefore, we compare the frequentist-based approach (or classic approach) to the Bayesian method.

One interesting result is that for any given day equation (5) summarizes the correlation between pairs of links. Because the equation relies on the variance and covariance between links, which requires the estimate of the mean travel times, the distribution still needs a method that includes the error propagation mentioned above.

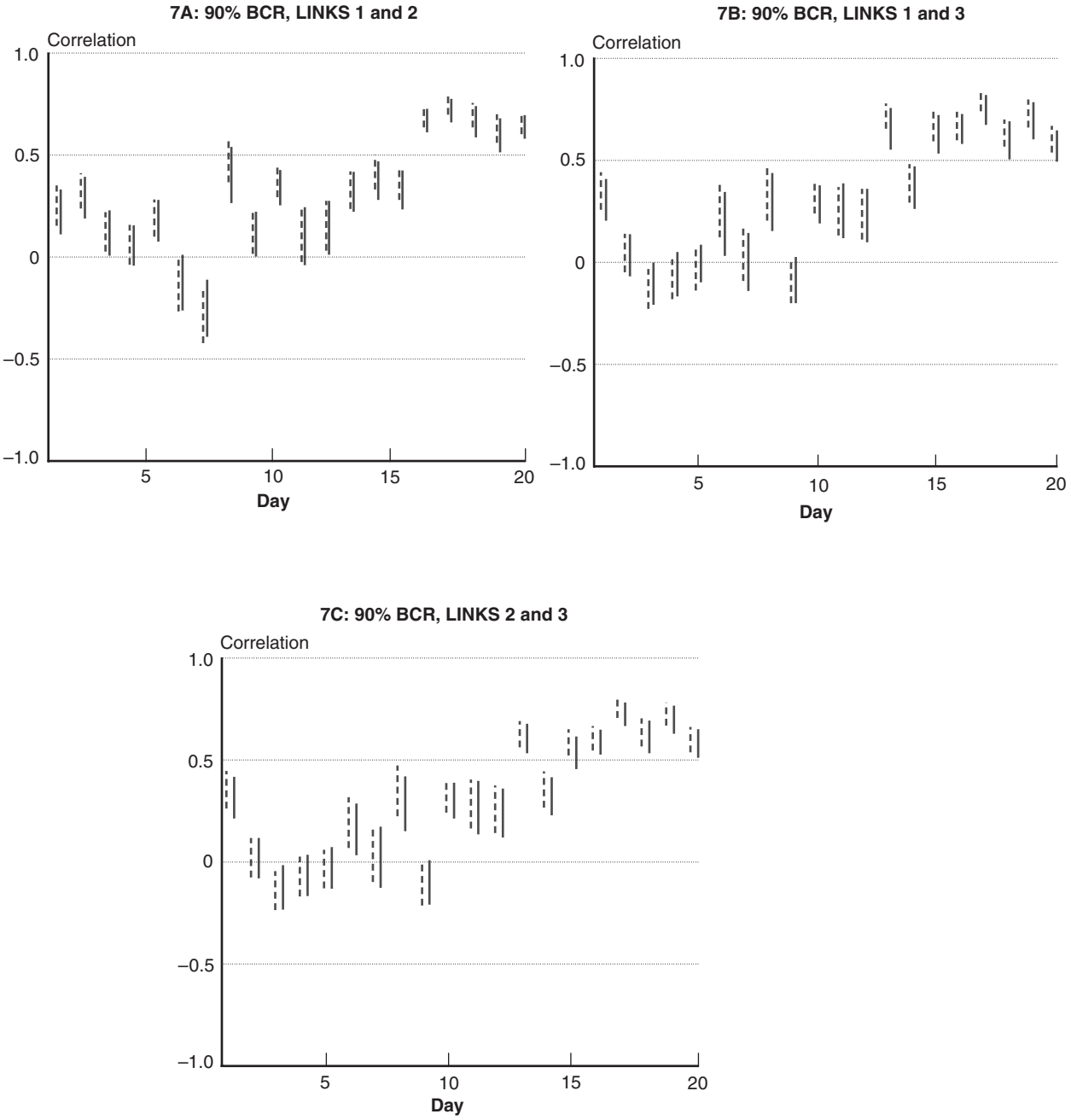
Figure 7 summarizes the correlation coefficients for all 20 days. The correlation BCR percentiles for all three links and their pairs are shown along with the classic 90% CI that appears next to the BCR for comparison purposes. For the adjacent links 1 and 2, there are four days that either have a negative correlation or a correlation where the 90% BCR covers zero. There are six such days between the adjacent links 2 and 3. The results are consistent for the non-adjacent links (i.e., links 1 and 3). This illustrates that the nonpositive correlation remains constant from link to link and seems to be a within-day characteristic.

Also, in terms of space mean speed, this correlation measure can be compared with traditional traffic congestion measures. Suppose a link is considered congested when the speed falls below 56 km/hr (35 mph). For the first 12 days, all links had a minimum space mean speed that ranged from 8 km/hr (5 mph) to 32 km/hr (20 mph). This latter case corresponds to days when the 90% BCR was below 0.5. In contrast, the last eight days have a minimum space mean speed ranging from 73 km/hr (45 mph) to 105 km/hr (65 mph) and at least one link pair with a 90% BCR covering 0.5. This demonstrates that the single correlation measure matches traditional measures of congestion that are based on speed. In general, the heavier the congestion, the lower the correlation of travel time between links.

Indeed, an obvious question raised by figure 7 is to ask whether the MCMC method is necessary or if the classic approach is an adequate approximation. The lengths of their respective 90% intervals indicate that on average the Bayes intervals are 10.5% longer than the classic intervals. The MCMC approach has longer intervals because it accounts for the uncertainty in the estimate of the smoothing spline. The user of our algorithm may want to balance the gain in the MCMC approach with the loss in time it takes to implement the algorithm. For the data from day 1, the 500 MCMC iterations take 36 seconds to implement using MATLAB on a 2.00 GHz processor with 1.00 GB RAM, whereas the classic approach takes less than 1 second to implement. This difference in implementation time might be different but is well worth the effort for those users who wish to account for all of the uncertainty generated in the estimate of interlink correlation. Given the rapid increase in computational abilities, it is our belief that computation concerns will not be a deciding factor.

The NCS smoothing technique is commonly used in statistics but not extensively in transportation engineering. There is an explicit tradeoff between the tuning parameter and the fitted curve, and it is important that the tuning parameter be selected in an appropriate and consistent manner. Figure 8A

FIGURE 7 90% Bayes Credible Regions for the Correlation for 20 Days

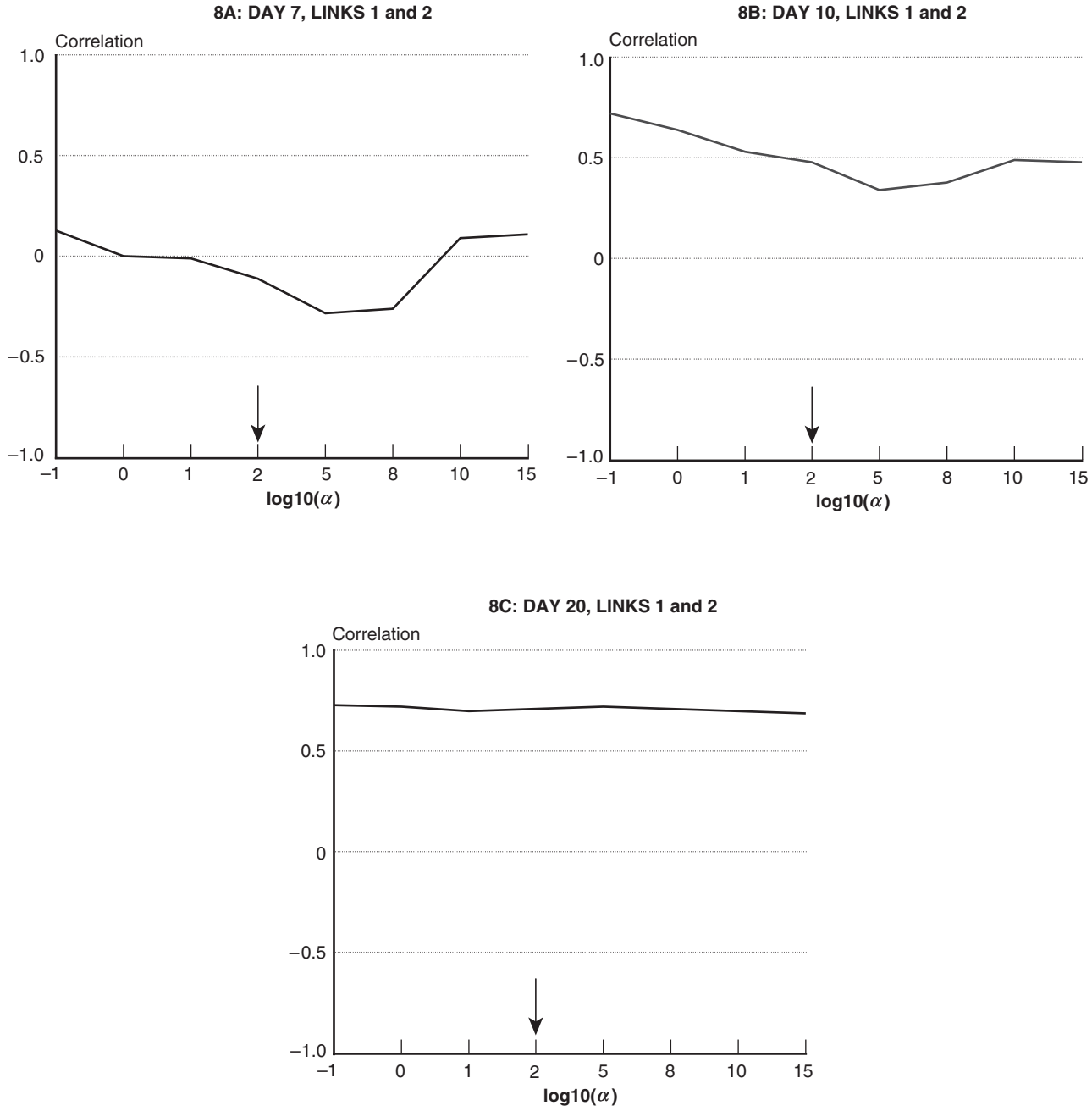


Note: Dashed lines indicate the classic interval approach and solid lines denote the Bayes interval approach.

shows the correlation between links 1 and 2 (i.e., $\hat{\rho}_{12}(\alpha)$) as a function of the logarithm of the tuning parameter value for day 7. The arrow represents the "optimal" tuning parameter based on the Generalized Cross Validation method. However, note that in that figure (8A), the sign of the correlation is positive for tuning values but negative for others. Therefore, it is important that the tuning parameter be identified correctly, otherwise the correlation value might be not only erroneous but of a different sign.

Figures 8B and 8C show the correlation between links 1 and 2 (i.e., $\hat{\rho}_{12}(\alpha)$) as a function of the logarithm of the tuning parameter value for days 10 and 20, respectively. The link-to-link correlation seen in these figures is relatively high and stable. For these

FIGURE 8 Correlation Coefficient for Links 1 and 2 as a Function of the Tuning Parameter



examples, the travel time fluctuation is relatively smooth, and the large values of the tuning parameter safeguard the dynamic trend.

CONCLUDING REMARKS

This paper demonstrates that for the travel time estimation problem the traditional approach is complicated because of the nonparametric nature of the estimator and the covariance between links. We adopted a Bayesian approach that had a number of benefits in terms of interpretation and ease of use. As an added benefit, this approach provided the actual distribution of the parameter, which allowed a much broader range of statistical information, and consequently better results, to be obtained. We found that, contrary to a common assumption used in many transportation engineering applications, the link covariance is non-zero. Furthermore, the

distribution of the correlation coefficient has the potential to be used as a performance metric for traffic operations.

From a transportation engineering perspective, this work is important for two reasons. First, this paper shows that the common assumption that link travel time covariance is zero is erroneous. More importantly, we developed a method for calculating the covariance with appropriate intervals. This technique can be readily incorporated for calculating travel time variance and the associated interval. This will have relevance in a wide range of applications including route guidance and traffic system performance measurement. Secondly, correlation coefficients have the potential for categorizing the performance of the traffic system, because they are a direct measure of how constrained drivers are with respect to traveling at their desired speed. The use of the proposed technique in the above transportation applications will be the focus of the next step in the research.

Two caveats to our study are as follows. First, the results depend on the length of the links in which the vehicles traverse. Suppose that the travel times for the three links have a positive correlation. However, if for that distance the links were shorter (i.e., six links over the same length) there is no guarantee the relationship will remain the same (e.g., all six links are positively correlated). No study finds the extent to which this occurs. This issue could be addressed in future research by utilizing a vehicle simulation program such as TRANSIMS (2003). With this simulation program, the researcher can examine these types of issues by playing "what if" scenarios with variations of the length of the links under assorted dynamic and complex traffic conditions.

The second caveat to our study is that during severely congested traffic, link travel times are essentially constant. In this case, researchers will find it difficult to utilize link travel time correlation as a congestion measure. This case is similar to the free-flow case where drivers can go at the speed they wish. In both situations, sophisticated congestion measures are not needed. However, when things are rapidly changing, this approach would be very useful. We show that the method is appropriate under several dynamic conditions, where the speed ranged from 8 km/hr (5 mph) to 105 km/hr (65 mph). These would be the most interesting traffic conditions (e.g., where the travel time fluctuates in and out of free-flow and congested traffic conditions) from a traffic management perspective.

ACKNOWLEDGMENTS

This research was funded in part by the Texas Department of Transportation through the TransLink® Research Center. The TransLink® partnership consists of the Texas Transportation Institute, Rockwell International, the U.S. Department of Transportation, the Texas Department of Transportation, the Metropolitan Transit Authority of Harris County, and SBC Technology Resources. The support of these organizations, as well as other members and contributors, is gratefully acknowledged. We thank two referees and the editor for helpful and insightful reviews that greatly improved this article. The authors would also like to thank Mary Benson Gajewski and Beverley Rilett for editorial assistance in the preparation of this article.

REFERENCES

Berry, S.M., R.J. Carrol, and D. Ruppert. 2002. Bayesian Smoothing and Regression Splines for Measurement Error Problems. *Journal of the American Statistical Association* 97(45):160–169.

Eubank, R.L. 1999. *Nonparametric Regression and Spline Smoothing,* 2nd edition. New York, NY: Marcel Dekker.

Gelfand, A.E. 2002. Gibbs Sampling. *Statistics in the 21st Century,* Edited by A.E. Raftery, M.A. Tanner, and M.T. Wells. Chapman & Hall and American Statistical Association, Washington, DC.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 2000. *Bayesian Data Analysis*. London, England: Chapman & Hall.

Gilks, W.R., S. Richardson, and D.J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. London, England: Chapman & Hall.

Graybill, F.A. 1976. *Theory and Application of the Linear Model*. Belmont, CA: Wadsworth.

Green, P.J. and B.W. Silverman. 1995. *Nonparametric Regression and Generalized Linear Models*. London, England: Chapman & Hall.

Ruppert, D., M.P. Wand, and R.J. Carroll. 2003. *Semiparametric Regression*. Cambridge, England: Cambridge University Press.

Sen, A., P. Thakuriah, X.Q. Zhu, and A. Karr. 1999. Variances of Link Travel Time Estimates: Implications for Optimal Routes. *International Transactions in Operational Research,* (6)75–87.

TRANSIMS. 2003. Available at http://transims.tsasa.lanl.gov/, as of May 2004.

Wilks, S.S. 1962. *Mathematical Statistics*. New York, NY: Wiley.

APPENDIX

NCS Algorithm

The travel times for each of the individual vehicles are y_{l1}, y_{l2}, y_{l3},..., y_{ln} and are recorded at times t_1,..., t_n. The steps for calculating $A(\alpha)$ are as follows:

1. Let Q be a matrix of zeros of dimension n–2 by n. For i from 1 to n–2 let

$$Q_{ii} = \frac{1}{(t_{i+1} - t_i)},$$

$$Q_{ii+1} = -\left(\frac{1}{(t_{i+2} - t_{i+1})} + \frac{1}{(t_{i+1} - t_i)}\right), \text{ and}$$

$$Q_{ii+2} = \frac{-1}{(t_{i+2} - t_{i+1})}$$

2. Let B be a matrix of zeros of dimension n–2 by n–2.
   For i from 2 to n–3 let $B_{ii-1} = t_{i+1}-t_i$, $B_{ii} = 2(t_{i+2}-t_i)$ and $B_{ii+1} = t_{i+2}-t_{i+1}$.
   Let $B_{11} = t_2-t_1$, $B_{12} = t_3-t_2$, $B_{n-2n-3} = t_{n-1}-t_{n-2}$ and $B_{n-2n-2}= t_n-t_{n-1}$.

3. Set
   B = B/6.

4. $A(\alpha) = I_n - n\alpha Q'(n\alpha QQ' + B)^{-1}Q$.

5. The quantity being analyzed is $A(\alpha)\boldsymbol{z}_l$, where $z_{il} = \log(y_{il})$.

Classic Estimation of Correlation

The correlation between two links,
$\hat{\sigma}_{12} = E[(z_{i1} - g_{i1})(z_{i2} - g_{i2})] = E[\varepsilon_{i1}\varepsilon_{i2}]$,
is estimated using the following procedure:

1. Given the tuning parameter α, use an NCS to derive a continuous estimate of the mean travel time over the analysis period for the two links and call them $\hat{g}_1$ and $\hat{g}_2$.
2. For each link estimate the residuals for each vehicle: $\hat{\varepsilon}_1 = \boldsymbol{z}_1 - \hat{g}_1$ and $\hat{\varepsilon}_2 = \boldsymbol{z}_2 - \hat{g}_2$.
3. Calculate the equivalent degrees of freedom (EDF):
   $EDF = n - \sum_{i=1}^{n} A(\alpha)$.
4. Estimate the covariance between links 1 and 2:
   $\hat{\sigma}_{12} = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_{i1}\hat{\varepsilon}_{i2}}{EDF} = \frac{\hat{\varepsilon}_1'\hat{\varepsilon}_2}{EDF}$.
   Estimate the variance for links 1 and 2 respectively:

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_{i1}\hat{\varepsilon}_{i1}}{EDF} = \frac{\hat{\varepsilon}_1'\hat{\varepsilon}_1}{EDF} \text{ and } \hat{\sigma}_2^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_{i2}\hat{\varepsilon}_{i2}}{EDF} = \frac{\hat{\varepsilon}_2'\hat{\varepsilon}_2}{EDF}.$$

5. Calculate the estimated correlation:
   $\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1\hat{\sigma}_2}$

Note that the basic concept of *EDF* is to penalize the estimation of the covariance matrix using the proper equivalent of degrees of freedom. The *EDF* for splines is discussed in Green and Silverman (1995) and Ruppert et al. (2003).

Inferences utilizing the above approach can be accomplished with large sample distribution theory based on Wilks (1962, p. 594). The result indicates that as the number of vehicles approaches infinity the statistic $1/2 \log[(1+\hat{\rho}_{12})/(1-\hat{\rho}_{12})]$ has an approximate normal distribution with mean $1/2 \log[(1+\hat{\rho}_{12})/(1-\hat{\rho}_{12})]$ and variance $1/n$. Thus, this distribution is used to calculate a 90% confidence interval with the formula

$$\left[\frac{(\exp(2r_L)-1)}{(1+\exp(2r_L))}, \frac{(\exp(2r_U)-1)}{(1+\exp(2r_U))}\right]$$

where

$$r_L = \frac{1}{2}\log\left(\frac{1+\hat{\rho}_{12}}{1-\hat{\rho}_{12}}\right) - 1.64\left(\frac{1}{\sqrt{n}}\right) \text{ and}$$

$$r_U = \frac{1}{2}\log\left(\frac{1+\hat{\rho}_{12}}{1-\hat{\rho}_{12}}\right) + 1.645\left(\frac{1}{\sqrt{n}}\right).$$

The 1.645 corresponds to the 90th percentile of the standard normal.

Gibbs Sampler

The approach is simulation-based where *nreps* is the number of simulations performed on the parameter vector and $b = 1,2,3,\ldots,nreps$. The Gibbs Sampler begins with a reasonable starting value $\underset{\sim}{\theta}^{(1)}$ (i.e., estimates of the parameters from a traditional approach). From this starting value, the k^{th} component of $\underset{\sim}{\theta}$ is updated conditional on the data and all the other components of the parameter vector, $h\left(\theta_k^{(b+1)} \middle| \theta_1^{(b)},\ldots,\theta_{k-1}^{(b)}, \underset{\sim}{y}\right)$. The next step is to simulate the subsequent component of the parameter vector $h\left(\theta_{k-1}^{(b+1)} \middle| \theta_1^{(b)},\ldots,\theta_{k-1}^{(b)}, \theta_k^{(b+1)}, \underset{\sim}{y}\right)$. This is repeated for all unknown parameters until *nreps* simulations for each component of the parameter vector have been completed.

Example Gibbs Sampler

The following steps are used to perform the Gibbs Sampler simulation:

1. Set the prior distribution parameters to be diffuse
   $a = 0$, $p^2 = \infty$, $c = 0$ and $d = 0$.

2. Set the starting values for the unknown parameters
   $\mu^{(1)} = \bar{y}$ and $\sigma^{2(1)} = s^2$.

3. Generate the mean portion
   $$\mu^{(b)} \sim N\left(\bar{y}, \frac{\sigma^{2(b-1)}}{n}\right),$$
   which is of conjugate form (see Gelman et al. 2000 for the derivation).

4. Generate the variance portion
   $$\sigma^{2(b)} \sim IG\left(\frac{n}{2}, \frac{\sum_{i=1}^{n}\left(y_i - \mu^{(b)}\right)^2}{2}\right),$$ which is of conjugate form.

5. Repeat steps 3 and 4 *nreps* times.

NCS Bayesian Algorithm

For convenience, the approach is summarized in algorithmic form:

1. Calculate $z_{il} = ln(Y_{il})$.

2. Obtain the starting values $g_l^{(1)}$ and $\Sigma^{(1)}$ using techniques previously discussed in the section, "Traditional Approach and Smoothing Splines."

3. Simulate
   $g_l^{(b)} \left| z_l, \sigma_l^{2(b-1)}, \alpha \sim MVN\left(A(\alpha) z_l, \sigma_l^{2(b-1)} A(\alpha)\right)\right.$.
   Note that the same tuning parameter will be used throughout the algorithm. The g's and Σ are defined in equation (3).

4. Calculate
   $\Sigma^{(b+1)} \left| \varepsilon^{(b)} \sim \text{ Inv-Wishart}_n\left(S^{(b)}\right)\right.$,
   where
   $\varepsilon_l^{(b)} = z_l - g_l^{(b)}$ and $S^{(b)} = \sum_{i=1}^{n} \varepsilon_i^{(b)} \varepsilon_i^{(b)'}$.

5. Summarize the function(s) of the unknown parameters.

Speeds on Rural Interstate Highways Relative to Posting the 40 mph Minimum Speed Limit

VICTOR MUCHURUZA*
RENATUS N. MUSSA

Department of Civil Engineering
and Environmental Engineering
Florida A&M University—Florida State University
2525 Pottsdamer Street
Tallahassee, FL 32310

ABSTRACT

The relevance of posting the 40 mile per hour (mph) minimum speed limit on the Interstate Highway System has been increasingly called into question since the National Highway System Designation Act of 1995 repealed the federally sanctioned maximum speed limit. In this study, data were collected on major interstate highways in Florida to evaluate speed distribution relative to the 40 mph posted minimum speed limit. The data revealed that the 15th percentile speed at all sites was 60 mph or above on both four-lane and six-lane highway sections. The analysis showed that the average speed at all sites was approximately 5 standard deviations above the 40 mph minimum. The coefficient of variation ranged from 7% to 11%, while the trimmed variance analysis showed that vehicles traveling below 55 mph contributed insignificantly to the variation in traffic speeds. A comparison of data collected before the speed limit rose from 65 mph to 70 mph showed that the average speed increased by 5 mph, while the variances did not change significantly. The coefficients of variation, however, increased significantly. The results reported here suggest that speed variability at the lower end of the distribution is not

Email addresses:
*Corresponding author V. Muchuruza—vmuchuruza@eng.fsu.edu
R. Mussa—mussa@eng.fsu.edu

KEYWORDS: Speed limit, speed variation, highways.

a significant factor in traffic operating characteristics on Florida rural interstate highways.

BACKGROUND

The decades-old practice of posting minimum speed limits on rural interstates and other limited access highways is predicated on the desire to reduce vehicle conflicts caused by speed variability in a traffic stream. The relevance of the 40 mile per hour (mph) posted minimum speed limit found on the Interstate Highway System is increasingly being called into question in light of the National Highway System Designation Act of 1995, which repealed the federally sanctioned maximum speed limit of 65 mph on rural highways. Most states, including Florida, then raised the maximum speed limits, and by the end of 1997, most parts of Interstates 4, 10, 75, and 95 in Florida posted 70 mph, which is the maximum speed allowed by the Florida state statutes. While the maximum speed limit fluctuated over time, the minimum did not and, in Florida, the 40 mph limit was in effect and posted across many sections of rural interstate highways, even when the U.S. Congress required states to lower the speed limit to 55 mph in 1974.

With such a wide (30 mph) gap between maximum and minimum speed limits, it is logical to question the relevance of the 40 mph posted minimum. If the review of the current speed distribution shows that the 15th percentile speed is much higher than the 40 mph posted minimum, perhaps the minimum speed needs to be increased or rescinded. Also, it is important to know if the continued posting of the 40 mph minimum speed limit results in the increase in speed variability on rural interstate highways. A review of traffic operations on sections of Florida highways may provide answers to these questions.

HISTORICAL PERSPECTIVE

The promise behind the posting of minimum speed limits on interstate highways was to reduce interactions between fast and slow moving vehicles. Many states based their minimum speed limits on the Uniform Vehicle Code (UVC) published by the National Committee of Uniform Traffic Laws and Ordinances (National Committee 1954). The UVC stipulated that minimum speed limits be established on highways whenever traffic and engineering investigations concluded that slow-moving vehicles consistently impeded the normal flow of traffic on the highways.

Studies showed that, by 1962, many states had adopted slow speed laws in their statutes in compliance with the UVC (National Committee 1964). Florida was among the states adopting slow-speed provision, making 40 mph the minimum on the four-lane interstate system, the Turnpike, and defense highways. Basically, the Florida statutes made it illegal to drive at a slow speed that impedes the normal and reasonable flow of traffic on rural highways.

The literature reveals that, in the early 1960s, 41 states and the District of Columbia instituted slow-speed laws in verbatim or significantly conforming with the UVC, while the remaining 9 states did not add minimum speed regulations in their codes. Like Florida, Georgia and South Dakota statutes explicitly stated that the minimum speed limit was 40 mph, while Michigan and North Carolina maintained a 45 mph minimum speed rule on their interstate highways.

A 2003 survey of minimum speed practices in different states conducted for the Florida Department of Transportation showed that, following the 1995 National Highway System Designation Act, 43 states raised the maximum speed limit on their Interstate Highway System roads (Mussa 2003). However, the posted minimum speed on these systems did not change. In fact, the survey showed that 14 states still use 40 mph minimum speed limit signs, 10 states use 45 mph, and 1 state uses 55 mph. Furthermore, the survey showed that 25 states do not post minimum speed limit signs. Some respondents in states that do not post minimum speed limit signs indicated that slow driving is not a big problem on their highways and if a need arose for enforcement, various rules in their state statutes, such as "impeding traffic flow," can be used to warn or cite slow drivers.

UNDESIRABLE EFFECTS OF SPEED VARIABILITY

Posting a minimum speed limit was and still is motivated by the desire to reduce speed variability in a traffic stream and its attendant consequences in effi-

ciency and safety of traffic operations. Numerous studies have documented the negative effects of speed variability.

In determining the extent to which the 55 mph federally sanctioned maximum speed limit affected safety, a Transportation Research Board (TRB) study found that the probability of crashes occurring increases as the speed variance rises. The study showed that speed variation causes significant lane changing and passing maneuvers, which are known to be potential sources of conflicts and crashes (TRB 1984). The significance of speed variance was observed by developing a fatality model that included highway safety characteristics such as traffic density, percentage of vehicles exceeding 65 mph, percentage of teenagers, and enforcement activity, as well as speed variance and average speeds. The TRB model revealed that speed variance had a statistically significant effect on fatality rates—states with wider variances in vehicle speed on the highway tended to have higher fatality rates. The study further found that the mean speed only affected the severity of crashes. Holding the effect of speed variance constant in the model presented no statistically significant relationship between the fatality rate and any other speed variables. The study concluded that controlling speed variance could be an effective tool in improving highway safety.

Another study of 36 crashes that occurred on Indiana highway 37 indicated that the crash involvement rates per million vehicle-miles of travel were higher for vehicles whose speeds were below and above the mean speed (West and Dunn 1971). After removing data on all crashes related to turning maneuvers, the authors found that the crash risk associated with vehicles traveling faster or slower was more than six times the involvement rates at the mean speed. The West and Dunn findings were supported by Hauer (1971) who developed a mathematical model to correlate accident involvement rates and vehicle travel speeds. Hauer found that the imposition of a minimum speed limit on highways was two to three times as effective as an equivalent maximum speed limit in reducing the frequency of overtaking and thereby crash involvement rates. Hauer suggested that the relationship between vehicle speed deviations and crashes might be due to a higher incidence of passing maneuvers from which the vehicle passes or is passed by another vehicle—a situation caused by the presence of slower vehicles impeding fast vehicles in the traffic stream.

Lave (1985) found that the major highway safety benefits obtained after the enactment of the 1974 National Maximum Speed Limit Act—which reduced the maximum speed limit on interstate highways to 55 mph—were due to the reduction of speed variance rather than average speed. The author argued that a reduction in speed variance was realized because speed differences between slow and fast moving vehicles were reduced enough to cause a uniform flow of traffic on interstate highways. Thus, with small speed variances there are fewer passing and overtaking maneuvers, eventually leading to the reduction in the potential for conflicts and crashes. Lave concluded that slow drivers are just as dangerous as fast drivers and thus posting minimum speed limits is desirable so as to reduce speed variance in a traffic stream.

RESEARCH AGENDA FOR A MINIMUM SPEED LIMIT

The posting of higher maximum speed limits on rural interstate highways necessitates an evaluation of the relevance of posted minimum speed limit signs that existed prior to raising the maximum speed. Some studies (e.g., West and Dunn 1971; Hauer 1971; and Lave 1985) documented that posting the minimum speed limit has the beneficial effect of smoothening traffic flow by removing perturbations caused by speed differences.

While evidence obtained from past research shows that vehicle speed variability contributes to crashes, it is a big and unsubstantiated leap to say that posting 40 mph minimum speed limit signs on a highway with a 70 mph maximum speed limit, as is the case on the Florida rural Interstate Highway System, contributes to large differences in vehicle speeds. The effect of the 40(min)/70(max) seeming mismatch can be evaluated through a carefully designed field study in which driver characteristics and the resulting operating speeds are observed over a long period of time on highway sections with similar geometrics and traffic characteristics but with some having the 40 mph minimum posted and oth-

ers not having the minimum posted. Furthermore, knowing whether the minimum speed limit should be increased above 40 mph and by how much, given that the maximum speed limit has been raised from 65 mph to 70 mph, would also be useful. To obtain this information, a study would require experimental highway sections with the desired minimum speed limit signs posted.

This study aimed at evaluating operating speed characteristics on the Florida Interstate Highway System where 40 mph minimum speed limit signs are posted. It would have been desirable to conduct a study designed as described above but control sites with no minimum speed limit signs were not available. An experimental site with 40 mph minimum speed signs removed or covered can be created for conducting a longitudinal study where both operational and traffic crash data are collected and later compared with the current conditions. However, creating such sites has legal implications that are difficult to resolve at this time. Thus, this study was limited to the following: determining how speed characteristics deviate from the 40 mph limit, and determining the speed variability that resulted before and after the limit was raised.

Note that the relevance of the 40 mph minimum speed limit is analyzed in this paper from the operational standpoint only. Certainly, law enforcement personnel would prefer to have these signs erected to provide support for warning or citing slow moving drivers; the "impeding traffic flow" criterion may be less useful for enforcement purposes.

STUDY SITES

There are four interstate highways in Florida, Interstates 4 and 10, oriented in the east-west direction, and Interstates 75 and 95, which go in a north-south direction. In addition, the Florida Turnpike is a tollway from central to south Florida oriented in the north-south direction.

Site selection targeted rural sections of these roads where minimum speed limit signs are posted. The established site selection criteria required choosing sites where the geometric characteristics produced the highest free-flow speed possible, that is, sites devoid of horizontal and vertical curves, sustained grades, or other geometric constraints. Another criterion used was to select sections with telemetered traffic monitoring stations that collect traffic flow data on volume, occupancy, and individual vehicle speeds on a 24-hour basis throughout the year. The Florida Department of Transportation operates and maintains these sites. We could not select a site on Interstate 4 because the Tampa-Orlando-Daytona Beach corridor, through which this highway runs, is heavily congested throughout the week with few periods in which free-flow speeds are attainable. Table 1 shows the study sites selected based on the established criteria discussed above.

DATA COLLECTION

As part of the data collection strategy, the project team drove through the entire Interstate Highway System to observe geometrics and traffic operating conditions. In addition, the project team evaluated over 320 telemetered traffic monitoring sites to determine their locational suitability in relation to the research objective of evaluating speed characteristics. The field review resulted in choosing sites described in table 1. The elements of the data-collection plan including the data quality checks are explained below.

Individual Vehicle Records

Telemetered traffic monitoring stations use loop detectors that provide individual vehicle records composed of the exact time of passage of a vehicle, its speed, the lane of passage, the number of axles and axle spacing, vehicle length, and, in some stations, an individual vehicle's axle weight. A cursory review of the speed characteristics at most sites indicated that there were minor differences between weekend and weekday traffic speed distribution. Thus, data from all sites were collected on weekdays in good weather conditions and dry pavement. The integrity of the data was verified by checking for errors.

Data Error Checks

Logic checks on the recorded data elements were applied to the raw data files downloaded from the count stations. Typical data errors included improper recording of speeds and loop failures on some lanes. We set an initial criterion that if more

TABLE 1 Description of the Study Sites

Site code	Highway	Number of lanes	FDOT district	County	Milepost	Geographical location
320	I-75	6	2	Columbia	22.4	Between I-10 and US 90
9904	I-75	6	2	Alachua	3.0	3 miles north of the Marion County line
9905	I-95	6	2	Duval	4.4	2 miles south of the I-295 interchange
9901	I-10	4	3	Jefferson	18.2	1 mile east of County Road 257
9928	I-10	4	3	Walton	10.3	1.3 miles west of Boy Scout Road
351	I-75	4	1	Collier	41.5	At Everglades Boulevard overpass
9919	I-95	4	7	Brevard	9.9	3.5 miles south of State Route 514
9932	Florida Turnpike	4	8	Osceola	30.2	North of the County Road 525 underpass

Key: FDOT = Florida Department of Transportation.

than 5% of the data were bad, the dataset for the whole day was discarded and another day's data were downloaded. The accuracy of individual vehicle speeds was checked by relating vehicle length and its corresponding recorded speed. When the length of the vehicle was missing, it was assumed that the vehicle did not cross both loops in the speed trap thus suggesting that the recorded individual vehicle speed was erroneous. The number of records with missing speed or length was used to check the percentage of usable counts with respect to raw data elements and finally to decide whether the data for that particular day was within acceptable limits needed for further analysis.

Next, outliers were removed prior to performing the statistical analyses. All outliers, defined as data points that were inconsistent with the general trend of the data elements, were eliminated by a computer program developed for this purpose. The computer program discarded data points showing zero speed or speed greater than 120 mph (the maximum speed value the equipment can record). The time slots in which data were discarded were coded as missing data. In all datasets used for further analysis, the percentages of data coded as missing were less than one. All individual vehicle records were then summarized per hour and per lane and the required volume, speed, and headway statistics were calculated. Analyses proceeded only after assuring the data quality through these error checks.

VOLUME ANALYSIS

Under low to moderate traffic congestion, as demand on travel lanes increases so does the need of fast moving vehicles to pass slow moving vehicles. The combination of passive and active passing maneuvers creates the potential for conflicts in the traffic stream. Higher operating speeds are generally attainable at level of service (LOS) A[1] and continually decrease as the speed-volume relationship moves toward congested flow conditions. An hour-by-hour volume analysis of the 24-hour dataset was conducted at all eight sites to determine the volume distribution across the travel lanes, the percentage of trucks on each lane, and the minimum and maximum volumes and their hour of occurrence. The traffic volumes were expressed on a per-lane basis, because, in general, volume varies by lane. The average annual daily traffic, which is the gross indicator of traffic activity, usage, and need, was estimated by multiplying the 24-hour recorded volume with the

[1] LOS classifies the quality of operation provided by the roadway from A through F, with "A" representing the most favorable driving conditions and "F" the worst, measured at the peak hour period of the day (USDOT 2000).

adjustment factors developed by the Florida Department of Transportation Statistics Office. Table 2 shows the results of the volume analysis.

Table 2 presents the results categorized by the number of lanes on the highway (i.e., four or six lanes) and by direction of travel. Examination of the hourly variation at each site showed that the demand volumes were at their lowest from midnight to dawn hours, while the peak-hour demand occurred in the afternoon, typically from 3 p.m. to 5 p.m. with a few exceptions. The lane distribution analysis for the six-lane highway sections showed that flow rates in the middle lane were typically higher than on shoulder and median lanes. On four-lane sections, the flow rates on the shoulder lanes were higher than on the median lanes.

We also analyzed the distribution of trucks in each lane. Vehicles traveling at the low end of the speed distribution tended to be trucks, recreational vehicles, and vehicles towing trailers. Table 2 shows that truck percentages are higher on the shoulder lanes in both four-lane and six-lane sections. Note that on Sites 320 and 9904 on Interstate 75 in north Florida trucks are not allowed to travel on the median lane of three-lane (in one-direction) sections (i.e., they can only use the two outermost lanes).

A comparison of the peak-hour and 24-hour truck percentages suggests that more trucks travel during the offpeak hours. The need to change lanes and to pass some slow moving vehicles—typically trucks and RVs—is high during offpeak hours. The LOS in most of the sections was B or better during these time, thus operating speeds tend to be high due to fewer traffic interactions. With trucks and RVs typically among the slower moving vehicles, changing lanes and passing, resulting from the speed variances, might be a concern.

SPEED ANALYSIS

The analysis of speed is presented in two parts. The first part of the analysis details the central tendency of the speed data while the second part looks at the speed variability in the traffic stream. The analysis of both measures of center and dispersion takes into account the demand volume, lane of travel, and the type of vehicles—passenger cars or trucks—in the traffic stream.

Central Tendency Analysis

Figure 1 shows the 24-hour mean speed of all vehicles categorized by facility type (i.e., four-lane or six-lane highway). Examination of the graphs in figure 1 reveals that average speeds of vehicles vary from shoulder to median lanes with median lanes experiencing higher average speeds. At four-lane sites, the average speeds ranged from 66 mph to 74 mph in shoulder lanes and 67 mph to 85 mph in median lanes. At six-lane sites, the average speeds of the vehicles on the shoulder, middle, and median lanes ranged from 67 mph to 70 mph, 72 mph to 75 mph, and 75 mph to 81 mph, respectively.

Pairwise comparisons of the average speeds using a *t*-test showed that, on four-lane sections, average speeds differed significantly between shoulder and median lanes ($p = 0.0002$). Further analysis showed that the average speeds were significantly different between shoulder-middle lanes and middle-median lanes on six-lane sections ($p \leq 0.0001$ and $p \leq 0.0001$). These results confirm that slow-moving vehicles generally use the shoulder lanes while fast moving vehicles use the median lanes. At the prevailing LOS, it seems that the influence of traffic intensity was not a significant factor, because at six-lane sites the middle lanes carried higher volumes than shoulder lanes yet they had higher average speeds. To further understand the profile of speeds at these highway sections, table 3 presents the overall 24-hour mean speeds by lane and vehicle type. Table 3 also shows the harmonic mean speeds weighted by lane volumes and by vehicle type. The harmonic mean speeds were calculated as follows:

$$\bar{u}_1 = \frac{\sum v_{l_i}}{\sum \frac{v_{l_i}}{\bar{u}_{l_i}}} \text{ and } \bar{u}_2 = \frac{\sum v_{t_j}}{\sum \frac{v_{t_j}}{\bar{u}_{t_j}}} \quad (1)$$

where

$\bar{u}_1$ = the harmonic mean speed weighted by the 24-hour lane volume in lane *i*,
$\bar{u}_2$ = the harmonic mean speed weighted by 24-hour vehicle type *j* volume,
$\bar{u}_{l_i}$ = the 24-hour mean speed of all vehicles in lane *i*,
v_{l_i} = the total 24-hour volume in lane *i*,
$\bar{u}_{t_j}$ = the 24-hour mean speed of all vehicles of type *j*, and
v_{t_j} = the total 24-hour volume of vehicle type *j*.

TABLE 2 Results of the Volume Analysis

Site	Direction of travel	Lane of travel	Min. hourly volume and time of occurrence		Max. hourly volume and time of occurrence		% trucks (peak hr)	% trucks (24-hr)	Peak hour LOS	AADT
Six-lane sites										
320 (I-75)	NB	Shoulder	195	4 – 5 a.m.	458	3 – 4 p.m.	41	51	A	51,065
		Middle	144		761		17	21		
		Median	17		389		3	3		
	SB	Shoulder	119	3 – 4 a.m.	496	2 – 3 p.m.	41	47	A	
		Middle	99		851		17	19		
		Median	12		459		2	3		
9904 (I-75)	NB	Shoulder	128	2 – 3 a.m.	682	3 – 4 p.m.	38	51	B	64,172
		Middle	109		1,015		11	18		
		Median	13		576		1	2		
	SB	Shoulder	145	1 – 2 a.m.	490	11 a.m. – 12 p.m.	52	54	B	
		Middle	74		733		23	20		
		Median	9		376		2	2		
9905 (I-95)	NB	Shoulder	162	2 – 3 a.m.	640	7 – 8 a.m.	21	38	B	64,284
		Middle	102		1,198		12	21		
		Median	12		601		2	6		
	SB	Shoulder	177	3 – 4 a.m.	744	5 – 6 p.m.	18	36	B	
		Middle	95		1,179		7	18		
		Median	5		630		3	6		
Four-lane sites										
9901 (I-10)	WB	Shoulder	120	2 – 3 a.m.	756	5 – 6 p.m.	30	37	A	25,627
		Median	12		224		15	15		
	EB	Shoulder	109	3 – 4 a.m.	234	3 – 4 p.m.	28	39	A	
		Median	16		523		9	16		
9928 (I-10)	WB	Shoulder	73	2 – 3 a.m.	479	3 – 4 p.m.	25	34	A	18,728
		Median	9		224		15	16		
	EB	Shoulder	82	2 – 3 a.m.	468	1 – 2 p.m.	28	32	A	
		Median	8		184		10	19		
351 (I-75)	WB	Shoulder	36	1 – 2 a.m.	424	11 a.m. – 12 p.m.	15	20	A	19,047
		Median	3		165		2	5		
	EB	Shoulder	38	3 – 4 a.m.	439	11 a.m. – 12 p.m.	14	20	A	
		Median	2		149		4	5		
9919 (I-95)	NB	Shoulder	65	1 – 2 a.m.	564	4 – 5 p.m.	45	27	A	33,917
		Median	11		451		7	12		
	SB	Shoulder	79	1 – 2 a.m.	628	3 – 4 p.m.	30	32	A	
		Median	7		457		11	14		
9932 (TNPK)	NB	Shoulder	61	4 – 5 a.m.	486	11 a.m. – 12 p.m.	15	16	A	27,163
		Median	11		262		5	9		
	SB	Shoulder	119	3 – 5 a.m.	616	1 – 2 p.m.	12	21	A	
		Median	20		399		6	10		

Key: AADT = average annual daily traffic; EB = eastbound; LOS = level of service; NB = northbound; SB = southbound; WB = westbound.

FIGURE 1 Hourly Variations of Average Lane Mean Speeds

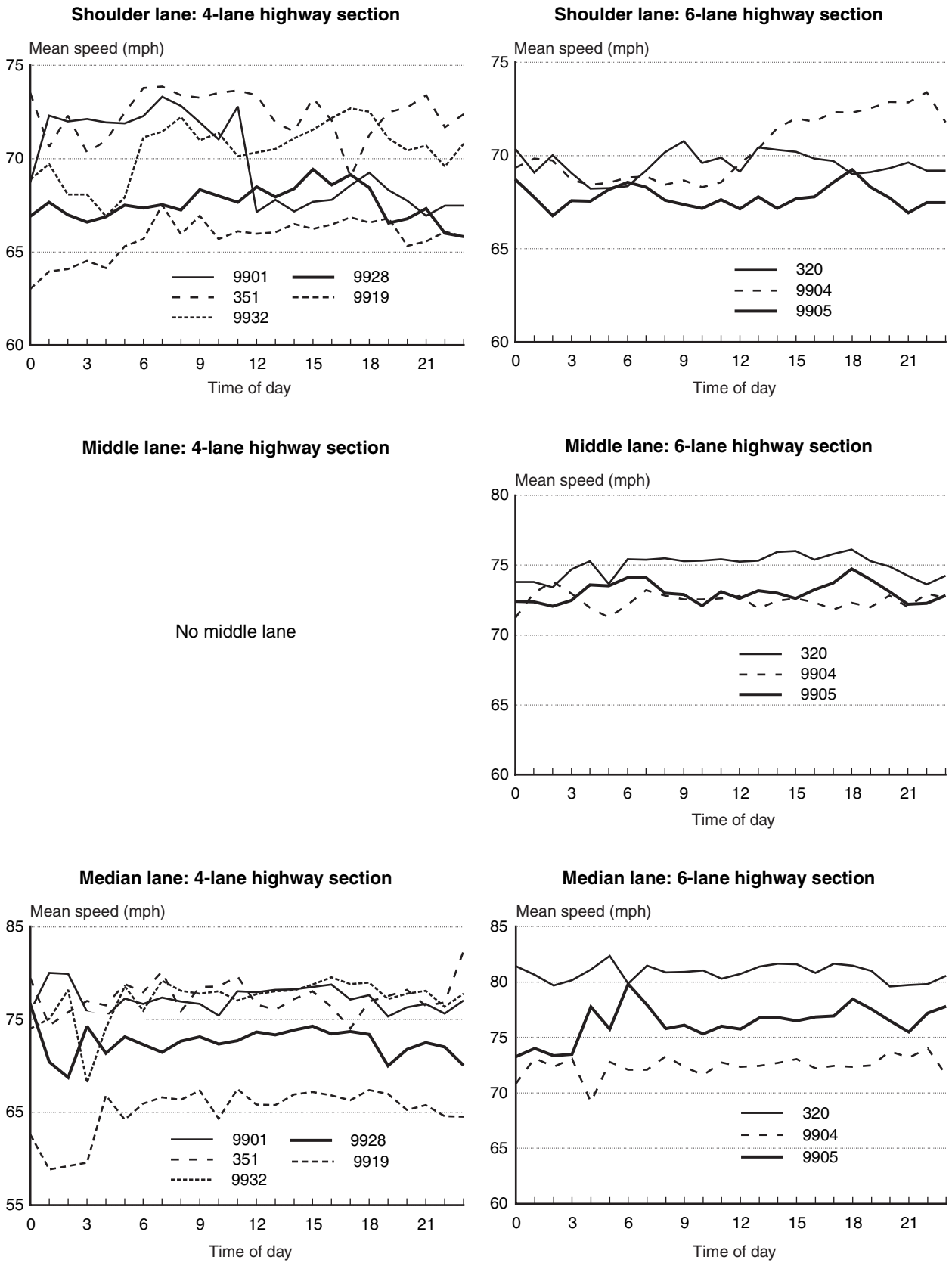


TABLE 3 Mean Speed Characteristics

Site code (highway)	Direction of travel	Travel lane	24-hour passenger car lane volume	24-hour truck volume	Lane-based mean speed $\bar{u}_1$	Vehicle type-based mean speed $\bar{u}_2$	Trimmed mean speed	Average mean speed
Six-lane highway sections								
320 (I-75)	NB	Shoulder	4,003	4,074			70	70
		Middle	8,173	2,202			73	74
		Median	3,609	99	73	73	79	79
	SB	Shoulder	4,350	3,850			69	70
		Middle	9,109	2,175			75	75
		Median	4,976	140	74	74	80	81
9904 (I-75)	NB	Shoulder	4,614	4,436			69	67
		Middle	10,606	2,227			73	72
		Median	6,512	107	71	73	76	76
	SB	Shoulder	3,814	4,184			69	68
		Middle	8,899	2,058			75	74
		Median	4,658	78	73	72	77	78
9905 (I-95)	NB	Shoulder	5,685	3,209			68	68
		Middle	12,288	3,072			73	73
		Median	5,711	316	72	72	78	77
	SB	Shoulder	7,004	3,766			68	68
		Middle	10,992	2,279			73	73
		Median	4,471	280	72	72	77	77
Four-lane highway sections								
9901 (I-10)	WB	Shoulder	4,959	2,729			73	73
		Median	2,356	367	74	74	78	78
	EB	Shoulder	4,843	2,844			73	73
		Median	2,254	407	74	74	78	78
9928 (I-10)	WB	Shoulder	4,551	2,122			68	68
		Median	2,081	266	69	69	73	73
	EB	Shoulder	4,894	2,107			70	70
		Median	1,867	257	70	71	72	72
351 (I-75)	WB	Shoulder	4,522	1,097			72	72
		Median	2,087	122	73	73	77	77
	EB	Shoulder	4,399	132			74	74
		Median	2,297	1,103	78	79	85	85
9919 (I-95)	NB	Shoulder	3,649	3,574			70	68
		Median	5,118	2,640	69	69	75	70
	SB	Shoulder	5,427	3,730			68	66
		Median	3,642	1,440	66	67	69	67
9932 (Turnpike)	NB	Shoulder	6,479	1,142			70	70
		Median	3,074	192	72	72	75	77
	SB	Shoulder	7,733	1,400			70	71
		Median	4,415	361	73	73	76	78

Key: EB = eastbound; NB = northbound; SB = southbound; WB = westbound.

Table 3 also includes the straightforward average speeds of all vehicles and the trimmed mean speeds. The trimmed mean speeds were calculated by discarding the lowest 15% and the highest 15% of vehicle speeds. We statistically analyzed the significance of the difference between the speed types displayed in this table. Pairwise comparisons of $\bar{u}_1$ and $\bar{u}_2$ showed no and slightly significant differences (p = 0.7 and p = 0.08) between lane-based and vehicle type-based mean speeds on both six-lane and four-lane highway sections, respectively. Statistical comparisons between trimmed mean speed and average speeds in each lane indicated lack of a discernible difference in both six- and four-lane sections (p = 0.57 and p = 0.40). The non-existence of the difference between trimmed speed and mean speed shows that the presence of fast and slow moving vehicles in the speed distribution has no significant effect on the average speeds on these facilities. The average speed of the bottom 15th percentile of the vehicles was 62 mph on both facility types, while in the upper 15th percentile, the average speed of vehicles was 81 mph and 83 mph on six-lane and four-lane sections, respectively.

Speed Dispersion Analysis

The dispersion of speeds was analyzed by lane and vehicle type using the standard deviation, coefficient of variation, and 10-mph pace, which is the 10 mph speed range with the highest number of observations of vehicles in the speed distribution. In addition, as is the case in most traffic engineering design and operational analyses, the 85th and 15th percentile speeds were also calculated. The results follow.

Facility Type Speed Distribution

We computed the standard deviations of vehicle speeds and the corresponding coefficient of variation. The results showed that their values varied depending on facility type. On six-lane sections, the standard deviation of speeds ranged between 4 mph and 6 mph, while on four-lane sections the standard deviations were as high as 10 mph. Specifically, Sites 351 and 9919 showed high values of standard deviations—9 mph and 10 mph on the median lanes, respectively. The field review revealed that these two sites are on highway stretches that are longitudinally straight for at least 10 miles.

The coefficient of variation, which measures relative dispersions of vehicle speeds from the average speed, was also calculated by lane for each site. This statistic was necessary to compare speed variations by examining the magnitudes of deviation relative to the magnitude of the mean given that there were different mean speeds grouped by lane. The analysis of the coefficients of variation in each lane showed that they ranged from 5% to 14%. When coefficients of variation for adjacent lanes on each site were compared, the results showed that the differences were less than 2%. These results suggest that the scatters of the vehicle speeds from the average speed are small. Therefore, the traffic speeds are very closely clustered about the mean speeds in all sections analyzed.

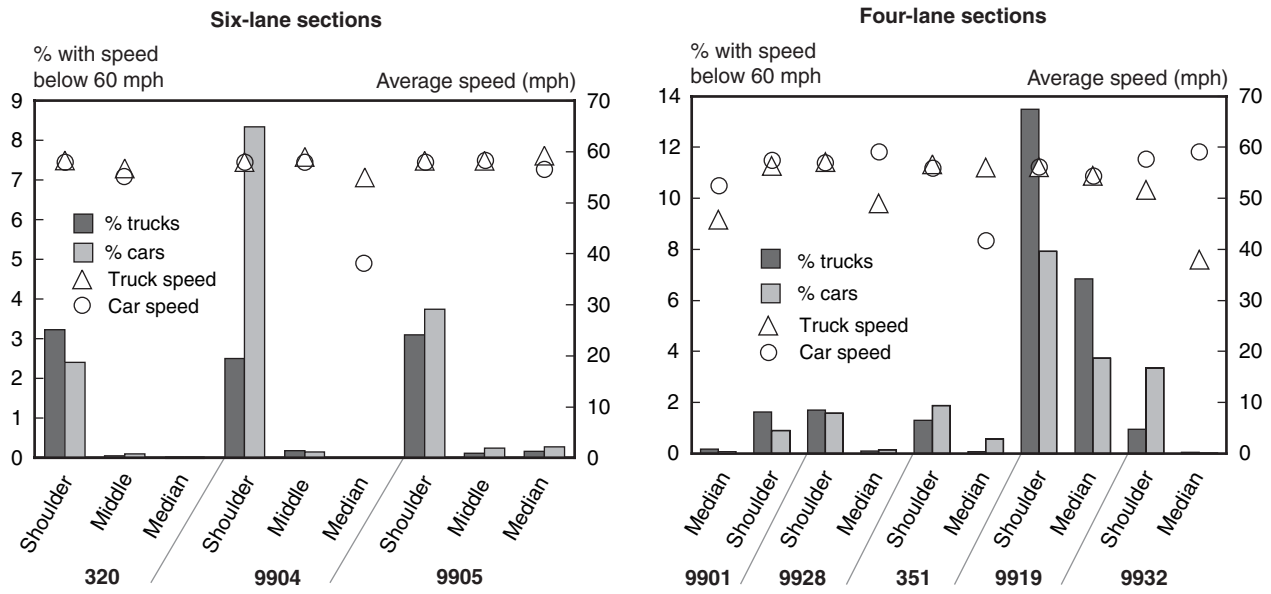
Speed Distribution by Vehicle Type

On average, the results of the speed distribution analysis by vehicle type showed that passenger car speeds were higher than truck speeds by at least 1 mph on both six-lane and four-lane sections. The results further showed that the coefficients of variation did not differ significantly between passenger cars and truck speeds for six-lane highway sections but were significant on four-lane sections (p = 0.027). Figure 2 displays the results of the speed distribution analysis at the lower end of distribution.

With respect to the vehicles traveling at the lower end of speed distribution (i.e., less than 60 mph), we found that more trucks on four-lane sections traveled below 60 mph than passenger vehicles at Sites 9901, 9919, and 9928, while more passenger cars traveled below 60 mph at Sites 351 and 9932. The results were also mixed on six-lane highway sections. At Sites 320 and 9904, which are on the same stretch and approximately 70 miles apart, different patterns of vehicles traveling below 60 mph were observed. While at Site 9904 more passenger cars traveled at speeds below 60 mph, at Site 320 more trucks traveled below 60 mph. At Site 9905, more passenger cars than trucks had speeds below 60 mph in all lanes.

The results further showed that on both four-lane and six-lane sections the percentage of vehicles at

FIGURE 2 Analysis of Vehicles Traveling at Speeds Below 60 mph



Note: In the median lane on highway 320, no vehicles traveled below 60 mph.

each site traveling below 40 mph (the posted minimum speed limit) was approximately zero. In fact, the results showed that at all sites only 1% of the vehicles traveled below 55 mph. Both passenger cars and trucks averaged speeds below 60 mph but above 55 mph on six-lane sections. On four-lane sections, the speed of vehicles traveling below 60 mph averaged above 54 mph.

Percentile and Pace Characteristics

Table 4 displays the 15th and 85th percentile speeds in each lane, 10 mph pace speeds, and the percentages of vehicles within the pace. Analysis of percentile speeds showed that, in four-lane and six-lane sections, the 85th percentile speeds ranged from 71 mph to 94 mph and 73 mph to 86 mph, respectively, while the 15th percentile speeds ranged from 60 mph to 77 mph and 62 mph to 76 mph, respectively, depending on the lane of travel (i.e., median lanes had higher percentile speeds than shoulder lanes). Of significant interest was the 15th to 85th percentile range, because it represents the proportion of vehicles traveling close to the mean speed. At the six-lane sites, the percentile speeds ranged from 7 mph to 10 mph, 8 mph to 10 mph, and 10 mph to 12 mph on the median, middle, and shoulder lanes, respectively. The ranges for four-lane sites were 7 speed was about 65 mph, 25 mph above the posted mph to 11 mph and 11 mph to 12 mph on the median and shoulder lanes, respectively.

Note that the results from Sites 351 and 9919 do not particularly follow the trend of other sites because of the somewhat large differences between percentile speeds at the two sites—14 mph and 19 mph, respectively. These differences could result from the straightness of the segments as well as a low volume of traffic that induces high-speed travel by some drivers. Furthermore, these two sites also showed the highest values of standard deviations. Table 4 further details that the paces ranged from the mid-60s to the mid-80s on both facility types with shoulder lanes experiencing lower pace speeds. The results in table 4 show that there is no direct relationship between the number of lanes on a highway and pace speeds.

Trimmed Variance Analysis

A trimmed variance analysis determined the contribution of slow- and fast-moving vehicles on overall speed variation. Using five different scenarios, vehicles traveling slower than 40 mph, 45 mph, 50 mph, 55 mph, and 60 mph were removed from the dataset when calculating the variance. The resulting speed variances from these trimming processes were then compared. At all sites, the 15th percentile minimum speed of 40 mph.

TABLE 4 Percentile and Pace Speed Characteristics

Site (highway)	Direction	Lane	15th percentile speed (u_{15})	85th percentile speed (u_{85})	10 mph pace	Vehicles in pace (%)
Six-lane freeway sections						
320 (I-75)	NB	Shoulder	64	76	66 – 76	66
		Middle	69	79	70 – 80	76
		Median	75	83	74 – 84	76
	SB	Shoulder	63	75	66 – 76	66
		Middle	70	80	71 – 81	78
		Median	76	86	76 – 86	70
9904 (I-75)	NB	Shoulder	62	73	65 – 75	68
		Middle	69	77	67 – 77	77
		Median	73	80	71 – 81	76
	SB	Shoulder	64	74	65 – 75	77
		Middle	71	79	68 – 78	75
		Median	74	82	70 – 80	70
9905 (I-95)	NB	Shoulder	63	74	63 – 73	68
		Middle	68	78	68 – 78	72
		Median	72	83	74 – 84	60
	SB	Shoulder	63	73	63 – 73	70
		Middle	69	78	70 – 80	78
		Median	73	81	74 – 84	72
Four-lane freeway sections						
9901 (I-10)	WB	Shoulder	67	78	69 – 79	69
		Median	73	82	73 – 83	75
	EB	Shoulder	67	78	69 – 79	70
		Median	73	82	73 – 83	72
9928 (I-10)	WB	Shoulder	62	73	64 – 74	66
		Median	69	78	68 – 78	77
	EB	Shoulder	65	75	66 – 76	69
		Median	68	77	67 – 77	77
351 (I-75)	WB	Shoulder	65	79	69 – 79	62
		Median	70	84	72 – 82	59
	EB	Shoulder	67	81	70 – 80	57
		Median	77	94	82 – 92	52
9919 (I-95)	NB	Shoulder	63	74	67 – 77	60
		Median	60	79	72 – 82	49
	SB	Shoulder	62	72	64 – 74	63
		Median	64	71	64 – 74	63
9932 Turnpike	NB	Shoulder	64	76	65 – 75	66
		Median	72	81	72 – 82	70
	SB	Shoulder	65	77	68 – 78	60
		Median	73	84	73 – 83	70

Key: EB = eastbound; NB = northbound; SB = southbound; WB = westbound.

The results showed no discernable contribution to speed variance for vehicles with a speed of less than 55 mph, primarily because very few vehicles at each site traveled at speeds less than 55 mph. In fact, at each site vehicles with speeds under 55 mph made up 1% of those recorded, while the percentage of vehicles with speeds of less than 40 mph was negligible (i.e., 0.15%). Although the contribution to the standard deviation of vehicles with speeds less than 55 mph is very minor, the safety implications of the presence of vehicles with very low speeds cannot be ignored. Even though only a few vehicles cause speed differential conflicts, these vehicles could be a contributory factor in crashes.

PLATOON ANALYSIS

Highway travel is generally composed of free-flowing and platooned vehicles. In free-flowing traffic, drivers can choose their speeds as they desire as long as conditions are such that slow-moving vehicles do not impede their ability to change lanes at will. Platooned vehicles travel close to each other mostly because of lack of passing opportunities, thus causing other vehicles to be trapped behind the lead vehicle. No definition exists in the literature of a headway value below which vehicles are considered to be moving in a platoon. Thus, in this study, four definitions were considered—less or equal to 1, 2, 3, and 4 seconds.

The analysis showed that six-lane highway sections carried larger proportions of platooned vehicles than four-lane sections. Further, the middle lanes of six-lane sections carried more platoons than the shoulder and median lanes. To study the effect of platooned vehicles on the distribution of speed, the mean speeds of platooned vehicles were compared with the mean speeds of nonplatooned (or free-flowing) vehicles. The statistical analysis here uses a *t*-test in which platooned and nonplatooned vehicles were paired by site and by lane of travel. The results showed that the difference between the speeds of platooned and nonplatooned vehicles were insignificant for both four- and six-lane highway sections regardless of whether the cut-off point was 1, 2, 3, or 4 seconds of time headway. These results indicate that platooned vehicles are not slow moving and thus do not create a need for free-flowing vehicles catching up behind them to pass. However, it should again be noted that the highway sections analyzed were relatively uncongested, operating at levels of service B or better for a majority of the hours in a year.

BEFORE AND AFTER COMPARISON

To understand the change in speed characteristics following the increase in the speed limit, table 5 presents a comparison of before-and-after data. In 1996, the speed limit was 65 mph at all the sites indicated in the table. Data-collection sites for both 1996 and 2002 were physically very close, and the field review of the sites indicated that for all practical purposes the geometric characteristics prevailing at these sites would produce similar driver behavior.

The results in table 5 show that the average speeds across all sites increased by 5 mph to 72 mph. The 15th percentile speed also showed a significant increase of 3 mph when averaged across all sites ($p \leq 0.0001$). A statistical *F*-test comparison of the variances indicated no significant difference between the 1996 and 2002 data ($p = 0.50$). However, significant differences were found in the variances on four-lane sections ($p = 0.0003$). Further analysis indicated that in 1996, the average speed on six-lane sections was 4.75 standard deviations above the 40 mph minimum posted speed limit. In 2002, it was 5 standard deviations above the 40 mph minimum. In four-lane sections, the results show that the average speeds were 6 and 5 standard deviations above 40 mph in 1996 and 2002, respectively. Examination of the coefficients of variation between the two datasets indicated that 2002 data show significant large variations compared with 1996. However, the coefficients of variation are still below 10%, indicating a reasonable equity in travel speeds.

DISCUSSION OF RESULTS

This paper presents a review of traffic operating characteristics on rural interstate highways in Florida. Using various analytical techniques, we determined speed characteristics in relation to the posted minimum speed limit of 40 mph. Our intent was to examine the relevance of the 40 mph minimum

TABLE 5 Comparison of Before-and-After Speed Data

Highway	Location, direction, and year	Mean speed (mph)	Standard deviation (mph)	Coefficient of variation (%)	15th percentile speed (mph)
Six-lane highway sections					
I-75	Between I-10 & CR136, NB, 1996	66	4	6	63
	Site 320, NB, 2002	73	6	8	67
	Between I-10 & CR136, SB, 1996	66	5	8	61
	Site 320, SB, 2002	74	7	9	66
	Between CR234 & SR21, NB, 1996	68	5	7	63
	Site 9904, NB, 2002	71	6	8	64
	Between CR234 & SR21, SB, 1996	67	5	7	63
	Site 9904, SB, 2002	71	6	8	64
I-95	Between CR210 and I-295, NB, 1996	67	4	6	64
	Site 9905, NB, 2002	72	7	10	65
	Midpoint CR210 and I-295, SB, 1996	63	6	10	60
	Site 9905, SB, 2002	72	6	8	65
	Near Flagler CL, NB, 1996	69	4	6	65
	Site 9905, NB, 2002	72	7	10	65
	Near Flagler CL, SB, 1996	64	5	8	63
	Site 9905, SB, 2002	72	6	10	65
Four-lane highway sections					
I-75	At mile marker 89, WB, 1996	66	4	6	61
	Site 351,WB, 2002	74	7	9	66
	At mile marker 89, EB, 1996	68	6	9	63
	Site 351, EB, 2002	78	9	11	68
I-10	Overpass E. of SR 85, WB, 1996	67	4	6	64
	Site 9901, WB, 2002	74	6	8	68
	Overpass E. of SR 85, EB, 1996	69	4	6	64
	Site 9901, EB, 2002	74	6	8	68
	C-280 overpass, WB, 1996	68	5	7	65
	Site 9901, WB, 2002	74	6	8	68
	C-280 overpass, EB, 1996	67	4	6	64
	Site 9901, EB, 2002	74	6	8	68
	Between SR257 & US221, WB, 1996	67	5	7	62
	Site 9928, WB, 2002	70	6	9	66
	Between SR257 & US221, EB, 1996	69	5	7	65
	Site 9928, EB, 2002	71	5	7	68
	East end of Aucilla River, WB, 1996	67	5	7	62
	Site 9928, WB, 2002	70	6	8	66
	East end of Aucilla River, EB, 1996	69	4	6	64
	Site 9928, EB, 2002	71	5	7	68

Key: EB = eastbound; NB = northbound; SB = southbound; WB = westbound.

speed limit in light of the increase in the maximum speed from 65 mph to 70 mph.

It is clear from the analysis that raising the speed limit increased average speeds on rural interstate highways. The comparison of 1996 data with 2002 data showed that average speeds rose by 5 mph, which is the same amount of the speed limit increase. The comparison further showed a slight increase in the coefficient of variation after the maximum speed went up; however, the increase is statistically insignificant and under 10%, a threshold that can be considered to indicate uniform operations. In addition, the 15th percentile speed showed an increase of 3 mph when averaged across all sites. In relation to the 40 mph posted minimum speed, the 2002 average speed on all sections was 5 standard deviations above this minimum speed, compared with 5.4 standard deviations for the 1996 data.

In light of the above data and analyses, from a traffic operations standpoint, several questions arise: Is the practice of posting the 40 mph minimum speed irrelevant or is it successful in ensuring that vehicles do not travel below 40 mph? Should the 40 mph posted minimum speed limit be scrapped or should it be raised to a higher value? What should that value be? These are important questions that could not be adequately answered through the research paradigm reported here. However, the data reveal a few pointers.

First, the 40 mph posted minimum speed limit probably does not have a significant influence on driver behavior given that the number of vehicles traveling below 55 mph at all sites was negligible (i.e., 1%). If these signs influenced drivers, we would expect a higher percentage of vehicles to travel at speeds in the 40 mph to 50 mph range, as is the case on the higher side of the speed distribution where a large percentage of drivers maintain speeds between 70 mph and 80 mph.

It has been suggested in the past (e.g., McShane et al. 1998) that the 15th percentile speed may be used as a measure of the minimum reasonable speed for the traffic stream. (This suggestion mirrors the attempt to use the 85th percentile speed as a measure for setting the maximum speed limit). The data reported here indicate that, in all sections studied, the 15th percentile speeds on the aggregate ranged from 60 mph to 70 mph, which is 20 mph to 30 mph above the posted minimum speed limit value. Does this mean that the minimum speed limit should be set at 60 mph? There are number of concerns that would need to be addressed before a change like this could be made. First, Florida statutes (Florida Statutes 2002) state that "no school bus shall exceed the posted speed limit or 55 mph." Second, as a tourist state, some Florida visitors drive recreational vehicles (sometimes towing a trailer) or motor homes, and field review indicated that these are the vehicles that tend to make up the lowest 15% of the speed distribution at all sites. Third, a safety analysis would be needed to fully justify any change in the minimum highway speed.

Instead of increasing the minimum speed, should it be eliminated? After all, the results of a survey conducted as part of this research showed that 25 states do not post minimum speeds on interstate highways. Currently, Florida statutes state that: "The minimum speed limit on interstate and Defense Highways, with at least 4 lanes, is 40 mph." The Florida Highway Patrol in the context of this research study indicated that such a statute is required to enable law officers to issue citations. A question was raised that in the absence of the minimum speed rule, can the law officers use another Florida statute that states "No person shall drive a motor vehicle at such a slow speed as to impede or block the normal and reasonable movement of traffic" to warn or issue citations to slow moving vehicles? One police officer pointed out that if a vehicle is alone on the highway traveling at, say 25 mph, what traffic is the driver impeding?

RECOMMENDATIONS

Further research is needed to ascertain the effect of the current posted minimum speed limit on driver behavior. While the data seem to indicate that the 40 mph minimum speed might not be that relevant based on prevailing operating speed distributions, it is not clear what the effect would be if the signs were removed from rural interstate highways. The answer to most of the questions raised above requires field evaluation, as simulation analysis would not appropriately depict driver behavior on

roadways with and without posted minimum speed limit signs.

Additional research that is planned includes collecting data on interstate highway sections in states that do not have minimum speed limits posted but have similar geometric and driver characteristics. A comparison of multistate data might shed some light on the relevance of posting minimum speed limit signs. Multistate data would also be of interest to traffic engineers who want to compare safety characteristics on sites with and without posted minimum speed limits.

REFERENCES

2002 Florida Statutes. Title XXIII, Chapter 316, section 183(3).

Hauer, E. 1971. Accidents, Overtaking, and Speed Control. *Accident Analysis and Prevention* 3:1–13.

Lave, C. 1985. Speeding, Coordination, and the 55 mph Limit. *American Economic Review* 75:1159–1164.

McShane, W.R., R.P. Roess, and E.S. Prassas. 1998. *Traffic Engineering*. Upper Saddle River, NJ: Prentice-Hall.

Mussa, R. 2003. Nationwide Survey of the Practice of Posting Minimum Speed Limit Signs on Interstate Highways, manuscript.

National Committee on Uniform Traffic Laws and Ordinances. 1954. *Uniform Vehicle Code*. Washington, DC.

_____. 1964. A Comparative Survey Based on the Uniform Vehicle Code. *Traffic Laws Annual* 1.

Transportation Research Board (TRB). 1984. *Special Report 204: 55: A Decade of Experience*. Washington, DC: National Research Council.

U.S. Department of Transportation (USDOT), Federal Highway Administration. 2000. *Highway Capacity Manual*. Washington, DC.

West, L.B. and J.W. Dunn. 1971. Accidents, Speed Deviation and Speed Limits. *Traffic Engineering* 41:52–55.

Book Reviews

Reviews that appear in JTS describe and assess books about new developments in statistics, economics, the environment, or engineering research that focus on transportation issues. The topics can be theoretical, empirical applications, or methodological innovations.

Suggestions of books for review are welcomed, as are requests to become a book reviewer. Please contact the Book Review Editor:

Prof. Vincent W. Yao
2801 S. University Avenue
Institute for Economic Advancement
University of Arkansas at Little Rock
Little Rock, AR 7204-1099
USA
(501) 569-8453; wxyao@ualr.edu

Transportation Labor Issues and Regulatory Reform

James Peoples and Wayne Talley, editors
Elsevier
November 2004, 179 pages
ISBN 0-7623-0891-5
www.elseviersocialsciences.com
$95 € 86 £57.50

Transportation Labor Issues and Regulatory Reform, volume 8 in Elsevier's Research in Transportation Economics series, is a pleasant surprise in a couple of ways. First, edited volumes often end up being mélanges of contributions united by little more than their bindings. Articles in this volume, however, cover the topic well and in a logical order. In a refreshingly concise introduction, the editors lay out the rationale for the book, its organization, and the main points of the chapters and how they relate to one another. Second, the book is worthwhile. As one of the semi-unemployed "soldiers" in the struggle that led to transport deregulation, I appreciate efforts to breathe life into the subject, although I begin to question the relevance of an event older than most of my students and some of my clothes. But the volume analyzes a range of issues that, in many cases, have continuing importance within transportation. Of perhaps even more significance, the volume is a rich case study of the impacts of regulatory changes on labor and management compensation and employment levels, as well as working conditions. As such, it would be of value to researchers and practitioners in transportation, human resource management, safety, and industrial organization.

The book may be thought of as being in three sections: 1) safety; 2) employment, productivity, and working conditions; and 3) compensation. All three sections have merit. If they had to be ranked, safety is the strongest and compensation the weakest.

Safety: Chapters 2 and 3

In a 23-page tour de force, Ian Savage examines trends in injuries since deregulation in trucking, railroads, and airlines. His command of the issues and data is impressive. Of particular importance, he makes the almost always neglected point that crash data alone provide an incomplete picture of worker safety. In trucking and railroads, transportation accidents account for only 12% of lost workdays and a mere 4% of those in the airline industry. Safety trends are examined using a variety of measures, such as lost workdays per full-time employee and per unit of output. Comparisons are provided with either all private industry or manufacturing.

In the next chapter, Daniel Rodriguez et al. explore relationships between, on the one hand, motor carrier financial performance, firm size and type of operations, and driver payment method with, on the other hand, driver safety. Not surprisingly, the sometimes confounding effects of these relationships, as well as data limitations, sap the strength of the results. But the authors do reach

some important conclusions and the study is of value as an example of a very competent empirical investigation of complex issues.

Employment, Productivity, and Working Conditions: Chapters 4, 5, and 6

Kristen Monaco and Dale Belman examine the impacts of technology on working conditions for truck drivers. The study presents considerable information about the types of technologies in use, who uses them, and to what extent. Most interesting are their findings with regard to the ability of satellite-based systems to substitute for driver experience and to increase revenue-miles.

Next, Nancy Johnson and Jonathan Anderson explore employment, productivity, and working conditions in airlines following deregulation. The piece is a rich source for data about airline profitability, bankruptcies, and output. Some of this could have been relegated to appendices, though all of us enjoy checking out when our favorite or most hated carrier went belly up. The authors clearly lay out how swings in employment and productivity relate to technological innovations, such as the hub-and-spoke system, and the health of the overall economy.

The analysis of changes in working conditions, though, seems to be trying a bit too hard to make a case for deterioration. For example, Johnson and Anderson show that between 1970 and 2001, average weekly hours worked by pilots increased from 29 to 42. While acknowledging that airline accident rates steadily declined over this period and that 42 hours may not seem a very high number, they point out that pilots may suffer from jet lag and that even slight fatigue may result in catastrophic miscalculations. They do not mention, however, that aircraft in 2001 automatically took care of many more pilot tasks than was the case in 1970. Suggestive of increasing strain on airline attendants, the authors point out that between 1995 and 2000 incidences of crews having to enforce discipline on unruly passengers rose from 140 to 266. They fail, however, to put these statistics into any context, that in 1995 there were 0.00002 such incidents per plane departure, versus 0.00003 in 2000.

Daniel Rich's investigation of productivity, technology changes, and labor relations is the best of this section. Incorporating rail, air, truck, and water transport, Rich explores the interplay of labor-saving and labor-using technologies on the sources of productivity changes and employment and compensation levels. The discussions of the setting, relevant theories, data, empirical approaches, and results are all first rate. It is simply an excellent treatment of an extremely worthwhile, but complex, topic. I recommend it to you and will insist on it for my students.

Compensation: Chapters 7, 8, and 9

Though not without considerable merit, for two reasons the weakest contribution in the volume is Stephen Burks et al.'s study of executive earnings in trucking in the era of post-deregulation. The first is not the authors' fault. Data problems limited the investigation to 1977 to 1986, too short a period to see the full effects, if any, of deregulation.

The second reason is another matter. After a brief dip past 1980, the authors find that executive compensation recovered and, thereafter, increased continually, despite much less rosy compensation trends for their subordinates. In their conclusions they state: "From 1985 onwards they [i.e., trucking executives] received pay increases in line with the wider boom in executive pay." I believe there is a simple explanation that did not enter into the empirical investigation nor discussion given by the authors. Joe, out on the loading dock or driving a truck, may have few alternatives outside the industry and his salary is pegged to variations in the fortunes of his company. But many of the skills needed for managing a firm are not industry-specific and, as such, executive compensation in trucking would be influenced by compensation levels for executives throughout the economy. The immediate post-deregulation dip in compensation might have reflected a transition from executives with skills of specific value under regulation to those better suited to lead firms in unregulated environments.

John Bitzan presents an analysis of compensation levels for low- and mid-level managers in airlines, trucking, and rail. Similar to Burks et al., Bitzan does not account for earning levels outside of a manager's industry when examining the impacts of deregulation on their earnings and, like Burks et al., he finds little or no effect. As if to atone for these "sins," the bulk of the chapter addresses whether the qualifications of managers in transport industries changed

after deregulation, how these changes compared with those in nontransportation industries, and how managers were compensated for their qualifications. These are important questions. Bitzan deals with them well and gets some intriguing results.

The editors, John Peoples and Wayne Talley, finish off the batting order with an examination of motor carrier owner-operators serving port cities. It is a perfect finale for the volume. They examine how changes in regulations affecting maritime shipping and advances in containerization impacted one segment of the motor carrier industry. The study is competently done and its results of interest. Of perhaps more significance, their work challenges readers to consider the broader impacts of policy changes in one industry throughout the economy. This may be a self-serving attempt to set the stage for their next volume. At least that is something to be wished for.

Reviewer address: Richard Beilock, University of Florida, PO Box 110240, Gainesville, FL 32611-0240, USA.
Email: rpbeilock@ifas.ufl.edu.

Statistical and Econometric Methods for Transportation Data Analysis

Simon P. Washington, Matthew G. Karlaftis, and Fred L. Mannering
Chapman & Hall/CRC Press
April 2003, 425 pages
ISBN 1584880309
www.crcpress.com
$89.95

Transportation analysis has changed dramatically in the last 50 years. For example, the sequential four-step model of travel forecasting is slowly being replaced by activity-based analysis in addition to changes in the types of techniques and methods used. We have seen cross-classification analysis replaced by linear regression, and statistical and econometric methods such as hazard-based duration models and structural equations have become an integral part of the methodological framework of travel forecasting. With this evolution in methods comes a tremendous need for books that synthesize and extend the usual statistical theory presentation to one suitable for application-oriented audiences. Washington et al.'s book provides an excellent and needed addition to this genre of texts.

The book catalogs many of the major modeling techniques used in practice, most of which are also an important springboard for more advanced theoretical and largely academic transportation modeling. In this sense, the book is an excellent addition to a practicing transportation analyst's library as well as a perfect companion to a first year graduate modeling or methods course including, for example, travel forecasting, safety, and traffic engineering. One of the book's most useful features is its singular focus on transportation. All of the examples relate to, and are focused on, transportation problems. As an added benefit, the datasets used to develop the examples can be accessed via the publisher's website (http://www.crcpress.com/e_products/downloads/).

Washington et al.'s book is organized into three major sections: the fundamentals, continuous dependent variable models, and count and discrete dependent variable models. The fundamentals section presents basic statistical theory and includes topics such as central tendency, variability, hypotheses testing, and nonparametric tests. Part 2 includes discussion of regression, simultaneous equations, latent variables, and duration models, as well as panel data and time series analysis. In Part 3, count data models and the now familiar discrete choice and discrete/continuous models are presented.

Deciding what to include in and what to leave out of a methods book aimed at a transportation audience is often very difficult. Practitioners may not have the appropriate statistical background to immediately grasp the main concepts without some review of basic theory, yet too much foundation material overlaps with many statistical texts already available to graduate students. In Part 1, much of the material covered is readily available in most introductory statistical texts. It is useful material even for some first year graduate students; however, I would have liked to have seen this material divided into two sections, with most of the basic material going into an appendix. Part 1 could then be expanded to include many of the new modern

graphical display techniques (e.g., Wilkenson 1999; Heiberger and Holland 2004) and perhaps some discussion about software and statistical computing (e.g., Gentle 2002). Despite this relatively minor caveat, the material presented in Part 1 is well done, with examples that clearly link theory to practice.

Part 2 is where the book really begins to distinguish itself. The first third deals with the basic linear regression model and includes a fairly comprehensive presentation of regression theory, assumptions, departures from assumptions, and the practical aspects of manipulating variables and estimating elasticities. The remaining two-thirds of Part 2 is devoted to fairly contemporary modeling techniques, at least in terms of transportation practice.

Of the various chapters, those dealing with time series and panel data analysis are perhaps the weakest. The remaining chapters in Part 2 are well written and the authors have done an excellent job summarizing the major modeling approaches and the main assumptions for each of the modeling techniques. For example, in their chapter on duration models, they begin with a brief discussion of the Kaplan-Meier method (the predominant nonparametric model used in survival analysis); the authors then turn to a longer exposition on semi-parametric and fully parametric models. Each section begins with a presentation of the basic model, followed by an example, which helps to motivate the method's application. In the chapter on duration models, all of the modeling approaches are tied together with a brief discussion comparing the different techniques. Finally, the chapter ends with a discussion of the modeling assumptions, which very cleverly motivates several more complicated modeling approaches addressing, in part, violations of the basic modeling assumptions.

In Part 3, the authors limit their coverage to three very important modeling approaches. The first approach is used for response variables that are considered count data and include the family of Poisson and negative binomial models. The second approach focuses on discrete choice models and the final section on discrete/continuous models. All of these modeling approaches are accessible to practitioners and increasingly form the foundation for handling many types of transportation problems. As with Part 2, each chapter begins with a presentation of the model structure and concludes with an example highly relevant to transportation planning and engineering practice.

Overall, this text adroitly fills a very important niche between practice and theory. Although I would liked to have seen a few additional topics—for example, more on contemporary graphical techniques and some elaboration on simulation methods, which are playing an increasingly important role in travel forecasting—in general, the book is very well written. I recommend it for most transportation analysts and believe it to be a good, solid addition to the libraries of transportation graduate students.

References

Gentle, J. 2002. *Elements of Computational Statistics: Springer Texts in Statistics and Computing*. New York, NY: Springer.

Heiberger, R.M. and B. Holland. 2004. *Statistical Analysis and Data Display, An Intermediate Course with Examples in S-PLUS, R, and SAS: Springer Texts in Statistics*. New York, NY: Springer.

Wilkenson, L. 1999. *Statistics and Computing, The Grammar of Graphics: Springer Texts in Statistics*. New York, NY: Springer.

Reviewer address: Debbie A. Niemeier, Professor, Department of Civil and Environmental Engineering, One Shields Avenue, Davis, CA 95616. Email: dniemeier@ucdavis.edu.

Transportation After Deregulation

B. Starr McMullen, editor
Elsevier Science
2001, 140 pages
ISBN: 0-7623-0780-3
$97.95 € 97.95 £64.95

The U.S. Railroad Revitalization and Regulatory Act of 1976 funded the reorganized bankrupt northeast and midwest railroads that formed Conrail. After the act went into effect, subsequent legislation initiated deregulation across the transportation industry and lifted most of the remaining motor carrier restrictions, including those imposed by the states. The six papers in this volume in the series, *Research in Transportation Economics,* all deal

with the theme of transportation deregulation and regulatory reform and can be classified under one of the following topics: timing and impacts of deregulation, technological change issues, safety, and railroad mergers.

In the paper by Wesley Wilson and William Wilson, the authors discuss the impact on the marketing patterns of railroads of the Staggers Rail Act, legislation that deregulated the railroad industry. Implementation of the act resulted in lower rates for shipment of goods, especially grain products. The authors present an econometric analysis of rail rates between 1972 and 1995 that focuses on five grain commodities. Their empirical model includes demand, cost, and price relationships based on the New Empirical Industrial Organization (NEIO) models and a specification for regulatory regimes that uses a dummy intercept and a time trend for regulatory reform. Their results indicate the prices of the five commodities decreased over time, but the magnitude across these commodities differed, ranging from 40% to 71%.

Lawrence Wong's study assesses the effects of deregulation in the motor carrier sector by sorting out the independent impact of deregulation from those changes caused by the interaction of deregulation and new technologies. To do this, Wong used a translog cost function model with a time trend incorporated as a third-order truncated Taylor series expansion for data from 1976 through 1987. Additionally, the modeling efforts allowed for the decomposition of the technological change into three components: input bias, output bias, and characteristic bias. The author attempts to test the Schumpeter hypothesis for the less-than-truckload (LTL) sector, but the evidence to support this hypothesis in the LTL sector was not present. The empirical results show that although implementation of new technologies provided labor-saving advantages, they required greater expenditures and induced input biases because of shifts in the output level. The results also showed that the LTL sector of the motor carrier industry is capital-intensive, which creates higher barriers for entry.

The paper by Kristen Monaco and Taggert Brooks presents a unique analysis of how deregulation affected wages in the motor carrier sector. The authors used a time series approach because prior transportation wage studies that employed cross-sectional methods could not control for macroeconomic effects. Motor carrier wages are modeled as a function of manufacturing wages, and the relationships between the two series are then assessed. Their results show that the effects of deregulation on wages in this sector was felt most strongly between 1980 and 1984 and that wages in this sector declined from 1972 to 1996.

Atreya Chakraborty and Mark Kazarosian do not explicitly address the productivity in the motor carrier industry in their paper, but instead, their analysis assesses the relationship between marketing strategy and information technology (IT). Some firms are likely to use IT to increase the timely delivery of goods, while those that may produce the same level of output (measured in ton-miles) as their IT-using counterparts may be less concerned with timeliness than they are with lower rates. Given these two approaches, which are aggregated into one dataset, analysis of motor carrier productivity is inconclusive without controlling for marketing strategies.

Although the financial condition of railroads has improved significantly since 1980, policymakers are rightly concerned about the future vitality of the sector if costs are not contained. C. Gregory Bereskin's paper focuses on the potential for transcontinental railroad mergers, a controversial issue given the small number of railroads in service and the concern that mergers could result in a monopoly. Mergers would likely increase efficiency and provide a higher quality of service, thus attracting more shippers and raising rail revenues. More specifically, the empirical analysis reveals that unexploited economies of scale could result in monopolistic pricing, a standard theory in market structure research.

In the final paper in this volume, Frank Rusco and W. David Walls examine the effects of deregulation on safety operations. Opponents of freight deregulation often assert that it would create increased competition but at the cost of making freight transportation less safe. The authors apply this hypothesis to the minibus market in Hong Kong, which is both regulated and unregulated. The authors' model shows that minibus drivers in the unregulated market tend to drive faster and

experience higher accident rates than their regulated counterparts.

Deregulation in the transportation sector of the United States has resulted in great savings and increased flexibility. For example, freight transportation costs dropped sharply. Railroad rates fell from 4.2¢ per ton-mile in the 1970s to 2.6¢ per ton-mile in 1998. In addition, the railroad industry became more profitable. The cost of shipping by truck fell by $40 billion from 1980 to 1988, and deregulation has improved flexibility and enabled businesses to provide timelier deliveries, which contributes to a reduction in inventory costs. Despite the multitude of positive benefits from deregulation, there are some interesting policy issues that resulted, and these papers provide further empirical assessment. While readers may choose to peruse papers of topical interest, it would be valuable to read all of the papers because they provide current research on the effects of deregulation on the industry.

Reviewer address: Brian Sloboda, Bureau of Transportation Statistics, U.S. Department of Transportation, 400 Seventh St, SW, Room 3430, Washington, DC 20590.
Email: brian.sloboda@dot.gov.

America's Freight Transportation Gateways

OVERVIEW

The Bureau of Transportation Statistics (BTS) new report, *America's Freight Transportation Gateways,* and the accompanying *Gateways Resource* CD include detailed data on the movement of freight into and out of the United States. Over 400 seaports, airports, and land border crossings currently move international freight in the United States. This new report profiles the top 25 gateways, including the nation's largest gateway by value, the Port of Los Angeles.

Folded into the back cover of the report is the *Gateways Resource* CD, which provides detailed data on over 200 gateways. The CD also contains extensive information on countries of origin and destination for goods passing through U.S. gateways and the firms moving goods across international borders.

The major sources of data for the *Gateways* report are:

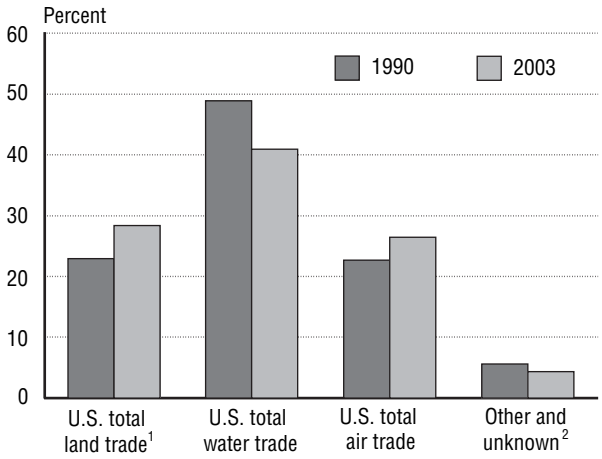
- Transborder Surface Freight Data from the Bureau of Transportation Statistics,
- air cargo data from the Bureau of Transportation Statistics,
- U.S. Merchandise Trade Data from the U.S. Census Bureau,
- border-crossing data from U.S. Customs and Border Protection,
- water data on seaports from the U.S. Army Corps of Engineers, and
- port calls and capacity data from the Maritime Administration

Review by: Jennifer Brady, Analyst, Bureau of Transportation Statistics, Research and Innovative Technology Administration, U.S. Department of Transportation, 400 Seventh St. SW, Room 3430, Washington, DC 20590. Email address: jennifer.brady@dot.gov.

THE REPORT

U.S. international trade increased at an average rate of 6% per year between 1990 and 2003, increasing from $899 billion to about $2 trillion. Water transportation carried the most trade, in terms of both tonnage and value. In 2003, vessels carried 78% of the total weight of trade and 41% of the total value (figure 1). Truck, train, and other land modes (e.g., pipelines) carried 22% of the total weight and 28% of the value. Although air transportation accounted for less than 1% of the total weight of trade, its transport of high-value goods made it responsible for 26% of the value.

FIGURE 1 Land, Water, and Air Gateways' Share of U.S. Merchandise Trade by Value: 1990 and 2003



[1] Includes truck, rail, pipeline, and miscellaneous surface modes.

[2] Includes purchased vehicles such as aircraft or boats moving from manufacturer to customer where the vehicle itself is the shipment, pedestrians carrying freight, and miscellaneous.

Source: U.S. Department of Transportation, Bureau of Transportation Statistics, based on **total trade,** from U.S. International Trade Commission, USITC Interactive Tariff and Trade Dataweb, available at http://dataweb.usitc.gov/ as of Sept. 15, 2004.

The majority of trade is concentrated in a handful of gateways. The top 5 gateways handled more than one-quarter of the nation's trade by value, while the top 15 handled more than 50% and the top 50 handled 80% (table 1).

The top three gateways, by value, represent the three transportation modes (i.e., water, air, and land). The Port of Los Angeles was the leading gateway, with $122 billion in international shipments in 2003. This port grew tremendously from 1999 to 2003, reflecting increased trade with Asia and the Pacific Rim. During this period, imports increased 52% and exports increased 20%. Other gateways experienced an average import/export growth rate of 14%.

Import trade heavily outweighs export trade through the Port of Los Angeles. In 2003, import goods comprised 86% of the value of freight moving through this port. The national average for imports is approximately 66%.

The second largest gateway by value, John F. Kennedy (JFK) International Airport in New York,

TABLE 1 Top 50 U.S. Freight Gateways, Ranked by Value of Shipments: 2003 (Current $, billions)

Rank	Port name	Mode	Total U.S. trade	Exports	Imports	Exports as % of total
1	Port of Los Angeles, CA	Water	122	17	105	13.8
2	JFK International Airport, NY	Air	112	47	65	41.7
3	Port of Detroit, MI	Land	102	55	47	53.5
4	Port of New York and New Jersey	Water	101	24	77	24.0
5	Port of Long Beach, CA	Water	96	17	79	17.9
6	Port of Laredo, TX	Land	79	32	46	41.1
7	Los Angeles International Airport, CA	Air	64	33	31	51.1
8	Port Huron, MI	Land	62	23	40	36.4
9	Port of Buffalo-Niagara Falls, NY	Land	59	27	32	46.1
10	Chicago, IL	Air	54	21	34	37.9
11	Port of Houston, TX	Water	50	21	28	43.0
12	San Francisco International Airport, CA	Air	47	21	26	44.1
13	Port of Charleston, SC	Water	39	13	26	34.0
14	Port of El Paso, TX	Land	39	17	22	42.6
15	Port of Norfolk Harbor, VA	Water	29	11	18	37.4
16	New Orleans, LA	Air	27	14	14	50.0
17	Port of Tacoma, WA	Water	26	5	21	19.8
18	Port of Baltimore, MD	Water	26	6	20	21.9
19	Port of Oakland, CA	Water	25	8	17	30.9
20	Dallas-Fort Worth, TX	Air	24	11	12	48.3
21	Port of Seattle, WA	Water	23	6	17	24.6
22	Miami International Airport, FL	Air	23	14	9	61.5
23	Anchorage, AK	Air	22	6	16	25.5
24	Port of Savannah, GA	Water	21	7	14	34.7
25	Port of Otay Mesa Station, CA	Land	20	8	11	42.0
26	Port of New Orleans, LA	Water	19	11	8	57.9
27	Cleveland, OH	Air	19	10	9	51.3
28	Atlanta, GA	Air	18	8	10	45.6
29	Port of Miami, FL	Water	17	7	10	41.1
30	Port of Champlain-Rouses Point, NY	Land	14	5	9	36.2

TABLE 1 Top 50 U.S. Freight Gateways, Ranked by Value of Shipments: 2003 (Current $, billions) *(continued)*

Rank	Port name	Mode	Total U.S. trade	Exports	Imports	Exports as % of total
31	Port of Hidalgo, TX	Land	14	6	8	43.6
32	Newark, NJ	Air	13	3	10	20.1
33	San Juan International Airport, PR	Air	12	5	7	42.4
34	Port of Blaine, WA	Land	12	5	7	43.6
35	Port of Portland, OR	Water	12	3	9	25.1
36	Port of Jacksonville, FL	Water	11	2	9	20.8
37	Port Everglades, FL	Water	10	4	6	41.4
38	Port of Nogales, AZ	Land	10	4	7	34.2
39	Port of Philadelphia, PA	Water	10	1	10	6.1
40	Port of Morgan City, LA	Water	10	0	10	1.8
41	Port of Brownsville, TX	Land	10	5	5	51.5
42	Port of Alexandria Bay, NY	Land	10	4	6	38.2
43	Port of Corpus Christie, TX	Water	10	2	8	19.8
44	Port of Beaumont, TX	Water	10	1	9	9.9
45	Port of Pembina, ND	Land	9	5	4	53.1
46	Boston Logan Airport, MA	Air	9	6	3	62.0
47	Port of Calexico-East, CA	Land	9	4	5	42.4
48	Philadelphia International Airport, PA	Air	9	5	4	53.8
49	Port of Sweetgrass, MT	Land	7	4	4	48.1
50	Seattle-Tacoma International Airport, WA	Air	7	4	3	56.8
	Total, top 50 gateways		**1,587**	**576**	**1,011**	**36.3**
	Total, U.S. merchandise trade by all modes		**1,983**	**724**	**1,259**	**36.5**
	Top 50 gateways as share of U.S. total (percent)		**80.0**	**79.6**	**80.3**	

Notes: **All data—**Trade levels reflect the mode of transportation as a shipment enters or exits a U.S. Customs port. Flows through individual ports are based on reported data collected from U.S. trade documents. Low-value shipments (imports less than $1,250 and exports less than $2,500) and intransit shipments are not included in trade data. **Air—**Data for all airports are based on U.S. port classifications and include a low level (generally less than 2% to 3% of the total value) of small user-fee airports located in the same region. Air gateways not identified by airport name include major airports in that geographic area in addition to small regional airports. Also due to U.S. Census Bureau confidentiality regulations, data for some of the air gateways include courier operations. For example, data for New Orleans International Airport, include FedEx air cargo activity in Memphis, TN.

Sources: **Air—**U.S. Department of Commerce, U.S. Census Bureau, Foreign Trade Division, special tabulation, August 2004. **Water—**U.S. Department of Transportation, Maritime Administration, Office of Statistical and Economic Analysis, special tabulations from Waterborne Databank, August 2004. **Land—**U.S. Department of Transportation, Bureau of Transportation Statistics, Transborder Surface Freight Data as of August 2004.

handled $112 billion in international trade. The top three origins of goods delivered through JFK are London, Brussels, and Frankfurt. However, further review of the data shows that most merchandise originates in Asia, and Europe is the last link in that supply chain. The top two carriers operating out of JFK, American Airlines and Lufthansa, together transported 21% of the imports and 17% of the exports passing through this gateway.

Detroit, the third largest gateway in the United States and the largest land gateway, moved $102 billion in trade. Of the top 25 gateways, Detroit is one of only three that handled more exports than imports. Trucks carried the majority of merchandise traveling by land through Detroit—83% by value.

The gateway ranking would change if listed by tonnage. Tonnage data are incomplete, however, because for land exports these data are not available

in official records. For example, among seaports, the Port of Los Angeles, which ranks first by value, ranks ninth insofar as waterborne tonnage is concerned. The need for more complete tonnage data is one of the findings of this report.

THE GATEWAYS RESOURCE CD

America's Freight Transportation Gateways profiles the busiest gateways in America. Researchers interested in performing analysis for other gateways may use the detailed data available on the CD, included at the back of the report. The CD contains 11 Microsoft Access files.

Smaller Gateways

All of the data breakouts presented in the *Gateways* report can also be performed for smaller gateways not included in the report. Some of the relevant databases on the CD include: Export Air Cargo, Freight Border Crossing and Entry Data, Import Air Cargo, Maritime Tons by Foreign Ports, Maritime Value by U.S. Ports, TEUs by U.S. Ports, Transborder Surface Freight, U.S. Freight Gateways by Value, and U.S. Seaports by Calls-Capacity.

The CD includes data on trade levels at border crossings, ports, and airports. For example, using the *Maritime Tons by Foreign Ports* database, the trade volume between specific U.S. ports and European cities can be determined. Depending on the data needed, either the port names or the port numbers can be used, as designated by the U.S. Census Bureau. The database can also provide an historical overview of how port traffic changed from 1993 through 2003. Other uses include: comparing the weight of trade moving through ports, airports, and land borders using short tons; and comparing the value of imports for each mode.

Trading Partners

Researchers interested in analyzing trade between the United States and partner countries can turn to a number of relevant databases on the CD: International Freight by Country and Mode, Maritime Tons by Foreign Ports, Import Air Cargo, and Export Air Cargo.

The *International Freight by Country and Mode* database contains data on 232 countries, from 1997 through 2003. This database can be used to compute the total value of imports or exports by mode and the total weight of imported and exported goods by vessel and air.

In 2003, the five countries with the greatest value of imports to the United States (by all modes) were Canada, China (mainland), Mexico, Japan, and Germany. The same five countries supplied the greatest value of imports to the United States in 1997; however, at that time, Japan's trade level was greater than both China and Mexico.

The top five countries for exports by value in 2003 were Canada, Mexico, Japan, the United Kingdom and Germany. Mainland China was the sixth largest recipient of U.S. goods by value. In 1997, Germany was not in the top five and Korea was.

Analysis of the cargo weight (in short tons) by carrier, origin airport, and destination airport can be computed using the *Export Air Cargo* dataset. For example, the weight of trade (in short tons) between Auckland International Airport in New Zealand and all airports in the United States grew 12% from 1999 to 2003. Trade from Auckland to Los Angeles International Airport increased 10%.

Import air cargo is similar to export air cargo, where the country of destination is the United States. Analysis of the cargo weight (in short tons) by carrier, origin airport, and destination airport can be computed using the *Import Air Cargo* dataset.

Industry Profiles

The *Import Air Cargo and Export Air Cargo* databases provide information on trade volume by air. The database provides air carrier codes and names. Use of common air carrier and airport codes allows the user to link the *Gateways* database with other BTS air data products. These datasets can be particularly useful for providing information on where

air carrier firms operate and the level of business conducted by each firm.

Vessel Types

U.S. Seaports by Calls-Capacity includes the number of port calls and capacity for 10 vessel types and 133 ports. Not all ports received all types of vessels. For example, Albany, New York, only received calls from dry bulk vessels and tankers. In contrast, Boston received at least one call from each type of vessel. The largest percentage of port calls in Boston were by tankers and product tankers, and the smallest number of port calls were by crude tankers, general cargo vessels, and combination vessels. The largest number of ports, 97, received dry bulk vessels; the smallest number of ports, 33, received vehicle-carrying vessels.

COPIES AND QUESTIONS

BTS provides this report, with the CD included, free of charge, from:

Customer Service
Bureau of Transportation Statistics
Research and Innovative Technology Administration
U.S. Department of Transportation
400 Seventh St SW, Room 4117
Washington, DC 20590
Email: orders@bts.gov

Online: www.bts.dot.gov, click on Products, then type the name of the report in the Search Products box.

Questions about this report: send an email to answers@bts.gov or call 800-853-1351.

Journal of Transportation and Statistics

call for papers

JTS is broadening its scope and will now include original research using planning, engineering, statistical, and economic analysis to improve public and private mobility and safety in transportation. We are soliciting contributions that broadly support this objective.

Examples of the type of material sought include

- Analyses of transportation planning and operational activities and the performance of transportation systems
- Advancement of the sciences of acquiring, validating, managing, and disseminating transportation information
- Analyses of the interaction of transportation and the economy
- Analyses of the environmental impacts of transportation

See our website at www.bts.dot.gov/jts for further details and Guidelines for Submission.

If you would like to receive a free subscription send an email to the Managing Editor: marsha.fenn@dot.gov

Journal of Transportation and Statistics

CALL FOR PAPERS

Special Issue on Transportation Investment

In conjunction with the national conference on Transportation and Economic Development (TED 2006) to be held in Little Rock, Arkansas, on March 29 and 30, 2006, the editors of the *Journal of Transportation and Statistics* are planning a special issue on transportation investment. We invite submissions relating to any mode of transportation, with an emphasis on economic and statistical models generating policy-relevant results. We are interested in papers dealing with longer term planning, particularly those focusing on safety, infrastructure, environmental, or economic issues, but we will consider other areas.

All papers will be peer reviewed. Please refer to the journal or visit the website (www.bts.gov/jts, and scroll to Guidelines for Manuscript Submission) for editorial requirements. Authors are encouraged to submit an abstract via the registration website for TED 2006 and present your work at the conference. Information on TED 2006 is available at http://www.ted2006-littlerock.org. The cut-off date for submitting completed manuscripts for publication consideration is Apr. 7, 2006. Please send one hardcopy or electronic copy of your paper to Professor Yao, to whom all correspondence should be directed as well (see contact information below). The tentative publication date of this special issue is early 2007.

Editors of the Special Issue

Cletus C. Coughlin
Vice President and Deputy Director of Research
Federal Reserve Bank of St. Louis
St. Louis, MO
Phone: (314) 444-8585
Email: coughlin@stls.frb.org

Randall W. Eberts
Executive Director
W.E. Upjohn Institute for Employment Research
Kalamazoo, MI
Phone: (269) 343-5541
Email: eberts@upjohninstitute.org

Vincent W. Yao (Corresponding Editor)
Senior Research Economist and Director
Transportation and Logistics Program
Institute for Economic Advancement
University of Arkansas
Little Rock, AR 72204
Phone: (501) 569-8453
Email: wxyao@ualr.edu

JOURNAL OF TRANSPORTATION AND STATISTICS

Guidelines for Manuscript Submission

Please note: Submission of a paper indicates the author's intention to publish in the *Journal of Transportation and Statistics* (JTS). Submission of a manuscript to other journals is unacceptable. Previously published manuscripts, whether in an exact or approximate form, cannot be accepted. Check with the Managing Editor if in doubt.

Scope of JTS: JTS publishes original research using planning, engineering, statistical, and economic analysis to improve public and private mobility and safety in all modes of transportation. For more detailed information, see the Call for Papers on page 98.

Manuscripts must be double spaced, including quotations, abstract, reference section, and any notes. All figures and tables should appear at the end of the manuscript with each one on a separate page. Do not embed them in your manuscript.

Because the JTS audience works in diverse fields, **please define** terms that are specific to your area of expertise.

Electronic submissions via email to the Managing Editor are strongly encouraged. We can accept PDF, Word, Excel, and Adobe Illustrator files. If you cannot send your submission via email, you may send a disk or CD by overnight delivery service or send a hardcopy by the U.S. Postal Service (regular mail; see below). Do not send disks or CDs through regular mail.

Hardcopy submissions delivered to BTS by the U.S. Postal service are irradiated. Do not include a disk in your envelope; the high heat will damage it.

The cover page of your manuscript must include the title, author name(s) and affiliations, and the telephone number and surface and email addresses of all authors.

Put the **Abstract** on the second page. It should be about 100 words and briefly describe the contents of the paper including the mode or modes of transportation, the research method, and the key results and/or conclusions. Please include a list of keywords to describe your article.

Graphic elements (figures and tables) must be called out in the text. Graphic elements must be in black ink. We will accept graphics in color only in rare circumstances.

References follow the style outlined in the *Chicago Manual of Style*. All non-original material must be sourced.

International papers are encouraged, but please be sure to have your paper edited by someone whose first language is English and who knows your research area.

Accepted papers must be submitted electronically in addition to a hard copy (see above for information on electronic submissions). Make sure the hard copy corresponds to the electronic version.

Page proofs: As the publication date nears, authors will be required to proofread and return article page proofs to the Managing Editor within 48 hours of receipt.

Acceptable software for text and equations is limited to Word and LaTeX. Data behind all figures, maps, and charts must be provided in Excel, Word, or DeltaGraph. American Standard Code for Information Interchange (ASCII) text will be accepted but is less desirable. Acceptable software for graphic elements is limited to Excel, DeltaGraph, or Adobe Illustrator. If other software is used, the file supplied must have an .eps or .pdf extension. We do not accept PowerPoint.

Maps are accepted in a variety of Geographic Information System (GIS) programs. Files using .eps or .pdf extensions are preferred. If this is not possible, please contact the Managing Editor. Send your files on a CD-ROM via overnight delivery service.

Send all submission materials to:

Marsha Fenn, Managing Editor
Journal of Transportation and Statistics
BTS/RITA/USDOT
400 7th Street, SW, Room 7412
Washington, DC 20590
Email: marsha.fenn@dot.gov

REVIEWERS for 2004

A

Eric Amel	Delta Air Lines, Atlanta, GA
Bill Anderson	Boston University, Boston, MA

B

Sandy Balkin	Pfizer Inc., New York, NY
David Ballard	GRA, Inc., Jenkintown, PA
David Banks	Duke University, Durham, NC
Alfred Brandowski	Gdansk University of Technology, Gdansk Wrzeszcz, Poland
Michael Bronzini	George Mason University, Fairfax, VA

C

Ken Campbell	Oak Ridge National Laboratory, Knoxville, TN
Francis Carr	Charles Stark Draper Laboratory, Inc., Cambridge, MA
Charles Chambers	Inter*VISTAS*-ga^2 Consulting, Washington, DC
Gang-Len Chang	University of Maryland at College Park, College Park, MD
Stephen Clark	Leeds City Council, Leeds, United Kingdom

F

Steven Feinberg	Carnegie Mellon University, Pittsburgh, PA
Philip Hans Franses	Erasmus University, Rotterdam, Netherlands

G

Nicholas Garber	University of Virginia, Charlottesville, VA
Laurie Garrow	Georgia Institute of Technology, Atlanta, GA
Jean-Phillipe Gervais	Université Laval, Québec, Canada
Ralph Gillmann	Federal Highway Administration, USDOT, Washington, DC
Thea Graham	Federal Aviation Administration, USDOT, Washington, DC
David Greene	Oak Ridge National Laboratory, Knoxville, TN
Stephen Greaves	University of Sydney, Sydney, Australia
Mike Greenwald	University of Wisconsin, Milwaukee, WI
Richard Gruberg	Federal Highway Administration, USDOT, Washington, DC

H

Jady Handal	Federal Aviation Administration, USDOT, Washington, DC
Mark Hansen	University of California, Berkeley, CA
Keith Hofseth	Institute for Water Resources, Alexandria, VA
Ho-Ling Hwang	Oak Ridge National Laboratory, Knoxville, TN

I

Towhindul Islam	University of Guelph, Ontario, Canada

K

Anne Koehler	Miami University, Oxford, OH
Young-Jun Kweon	University of Texas, Austin, TX

L

Jung-Taek Lee	University of Illinois, Chicago, IL
Dou Long	Logistics Management Institute, McLean, VA
Dom Lord	Texas A&M University, College Station, TX

M

Fred Mannering	Purdue University, West Lafayette, IN
John Maples	Energy Information Administration, USDOE, Washington, DC
Mike Margreta	Research and Innovative Technology Administration, USDOT, Washington, DC
Tom Maze	Iowa State University, Ames, IA
Sanal Mazvancheryl	State University of New York, Stony Brook, NY
Patrick McCarthy	Georgia Institute of Technology, Atlanta, GA
Mark McCord	Ohio State University, Columbus, OH
Shaw-Pin Miaou	Texas A&M University, College Station, TX
Eric Miller	University of Toronto, Toronto, Canada
Ian Moffat	University of Stirling, Stirling, Scotland

N

Nagaraj Neerchal	University of Maryland, Baltimore County, MD
Debbie Neimeier	University of California, Davis, CA
Bob Noland	Imperial College, London, England

O

Harmen Oppewal	Monash University, Victoria, Australia
Kaan Ozbay	Rutgers University, Piscataway, NJ

P

Brian Park	University of Virginia, Charlottesville, VA
Ram Pendyala	University of South Florida, Tampa, FL

R

Robert Raeside	Napier University, Edinburgh, Scotland
John Rose	University of Sydney, Sydney, Australia

S

Miriam Scaglione	University of Applied Sciences Valais, Sierre, Switzerland
Irwin Silberman	Research and Innovative Technology Administration, USDOT, Washington, DC
Kumares Sinha	Purdue University, West Lafayette, IN
Brian Sloboda	Research and Innovative Technology Administration, USDOT, Washington, DC
Frank Southworth	Oak Ridge National Laboratory, Knoxville, TN
Nikiforos Stamatiadis	University of Kentucky, Lexington, KY
Herman Stekler	George Washington University, Washington, DC
Peter Stopher	University of Sydney, Sydney, Australia

T

Jeremy Tantrum	University of Washington, Seattle, WA
Kent Taylor	North Carolina Department of Transportation, Raleigh, NC
Tomar Toledo	Massachusetts Institute of Technology, Cambridge, MA
Harry Timmermans	Technische Universiteit Eindhoven, Eindhoven, Netherlands

U

Jerry Ullman	Texas A&M University, College Station, TX
R.S. Radin Umar	Universiti Putra Malaysia, Serdang, Malaysia

W

Martin Wachs	University of California, Berkeley, CA
Joan Walker	Caliper Corporation, Newton, MA
Simon Washington	University of Arizona, Tucson, AZ
Steve Welstand	Chevron Products Company, San Ramon, CA
Billy Williams	North Carolina State University, Raleigh, NC
Robert Windle	University of Maryland, College Park, MD
James Winebrake	James Madison University, Harrisonburg, VA
Jeremy Wu	U.S. Census Bureau, Suitland, MD

Y

Melinda Yang	Bureau of Automotive Repair, Sacramento, CA
Chung-Hsing Yeh	Monash University, Victoria, Australia
Ted Younglove	University of California, Riverside, CA

INDEX FOR VOLUME 7*

* A complete index of all volumes of the journal is available online at www.bts.dot.gov.

M

N

O

P

R

S

T

U

W

U.S. Department of Transportation
Research and Innovative Technology Administration
Bureau of Transportation Statistics

JOURNAL OF TRANSPORTATION AND STATISTICS

Volume 7 Numbers 2/3, 2004
ISSN 1094-8848

CONTENTS