

Naturalistic Driving Study: Linking the Study Data to the Roadway Information Database

DETAILS

48 pages | 8.5 x 11 | PAPERBACK
ISBN 978-0-309-31493-0 | DOI 10.17226/22200

BUY THIS BOOK

FIND RELATED TITLES

AUTHORS

Shane B. McLaughlin and Jonathan M. Hankey; Strategic Highway Research Program; Strategic Highway Research Program Safety Focus Area; Transportation Research Board; National Academies of Sciences, Engineering, and Medicine

Visit the National Academies Press at NAP.edu and login or register to get:

- Access to free PDF downloads of thousands of scientific reports
- 10% off the price of print titles
- Email or social media notifications of new titles related to your interests
- Special offers and discounts



Distribution, posting, or copying of this PDF is strictly prohibited without written permission of the National Academies Press. (Request Permission) Unless otherwise indicated, all materials in this PDF are copyrighted by the National Academy of Sciences.

The Second

S T R A T E G I C H I G H W A Y R E S E A R C H P R O G R A M

The logo features a stylized graphic of three white diagonal lines on the left, followed by the text "SHRP 2 REPORT S2-S31-RW-3" in a bold, white, sans-serif font, all set against a dark gray background.

SHRP 2 REPORT S2-S31-RW-3

Naturalistic Driving Study: Linking the Study Data to the Roadway Information Database

SHANE B. McLAUGHLIN AND JONATHAN M. HANKEY
Virginia Tech Transportation Institute
Blacksburg, Virginia

TRANSPORTATION RESEARCH BOARD

WASHINGTON, D.C.
2015
www.TRB.org

Subject Areas

Data and Information Technology

Highways

Safety and Human Factors

Vehicles and Equipment

The Second Strategic Highway Research Program

America's highway system is critical to meeting the mobility and economic needs of local communities, regions, and the nation. Developments in research and technology—such as advanced materials, communications technology, new data collection technologies, and human factors science—offer a new opportunity to improve the safety and reliability of this important national resource. Breakthrough resolution of significant transportation problems, however, requires concentrated resources over a short time frame. Reflecting this need, the second Strategic Highway Research Program (SHRP 2) has an intense, large-scale focus, integrates multiple fields of research and technology, and is fundamentally different from the broad, mission-oriented, discipline-based research programs that have been the mainstay of the highway research industry for half a century.

The need for SHRP 2 was identified in *TRB Special Report 260: Strategic Highway Research: Saving Lives, Reducing Congestion, Improving Quality of Life*, published in 2001 and based on a study sponsored by Congress through the Transportation Equity Act for the 21st Century (TEA-21). SHRP 2, modeled after the first Strategic Highway Research Program, is a focused, time-constrained, management-driven program designed to complement existing highway research programs. SHRP 2 focuses on applied research in four areas: Safety, to prevent or reduce the severity of highway crashes by understanding driver behavior; Renewal, to address the aging infrastructure through rapid design and construction methods that cause minimal disruptions and produce lasting facilities; Reliability, to reduce congestion through incident reduction, management, response, and mitigation; and Capacity, to integrate mobility, economic, environmental, and community needs in the planning and designing of new transportation capacity.

SHRP 2 was authorized in August 2005 as part of the Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (SAFETEA-LU). The program is managed by the Transportation Research Board (TRB) on behalf of the National Research Council (NRC). SHRP 2 is conducted under a memorandum of understanding among the American Association of State Highway and Transportation Officials (AASHTO), the Federal Highway Administration (FHWA), and the National Academy of Sciences, parent organization of TRB and NRC. The program provides for competitive, merit-based selection of research contractors; independent research project oversight; and dissemination of research results.

SHRP 2 Report S2-S31-RW-3

ISBN: 978-0-309-31493-0

© 2015 National Academy of Sciences. All rights reserved.

Copyright Information

Authors herein are responsible for the authenticity of their materials and for obtaining written permissions from publishers or persons who own the copyright to any previously published or copyrighted material used herein.

The second Strategic Highway Research Program grants permission to reproduce material in this publication for classroom and not-for-profit purposes. Permission is given with the understanding that none of the material will be used to imply TRB, AASHTO, or FHWA endorsement of a particular product, method, or practice. It is expected that those reproducing material in this document for educational and not-for-profit purposes will give appropriate acknowledgment of the source of any reprinted or reproduced material. For other uses of the material, request permission from SHRP 2.

Note: SHRP 2 report numbers convey the program, focus area, project number, and publication format. Report numbers ending in “w” are published as web documents only.

Notice

The project that is the subject of this report was a part of the second Strategic Highway Research Program, conducted by the Transportation Research Board with the approval of the Governing Board of the National Research Council.

The members of the technical committee selected to monitor this project and review this report were chosen for their special competencies and with regard for appropriate balance. The report was reviewed by the technical committee and accepted for publication according to procedures established and overseen by the Transportation Research Board and approved by the Governing Board of the National Research Council.

The opinions and conclusions expressed or implied in this report are those of the researchers who performed the research and are not necessarily those of the Transportation Research Board, the National Research Council, or the program sponsors.

The Transportation Research Board of the National Academies, the National Research Council, and the sponsors of the second Strategic Highway Research Program do not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the object of the report.



SHRP 2 Reports

Available by subscription and through the TRB online bookstore:

www.mytrb.org/store

Contact the TRB Business Office:
202-334-3213

More information about SHRP 2:
www.TRB.org/SHRP2

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. On the authority of the charter granted to it by Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. C. D. (Dan) Mote, Jr., is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, on its own initiative, to identify issues of medical care, research, and education. Dr. Victor J. Dzau is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. C. D. (Dan) Mote, Jr., are chair and vice chair, respectively, of the National Research Council.

The **Transportation Research Board** is one of six major divisions of the National Research Council. The mission of the Transportation Research Board is to provide leadership in transportation innovation and progress through research and information exchange, conducted within a setting that is objective, interdisciplinary, and multimodal. The Board's varied activities annually engage about 7,000 engineers, scientists, and other transportation researchers and practitioners from the public and private sectors and academia, all of whom contribute their expertise in the public interest. The program is supported by state transportation departments, federal agencies including the component administrations of the U.S. Department of Transportation, and other organizations and individuals interested in the development of transportation. www.TRB.org

www.national-academies.org

SHRP 2 STAFF

Ann M. Brach, *Director*
Stephen J. Andrie, *Deputy Director*
Cynthia Allen, *Editor*
Kenneth Campbell, *Chief Program Officer, Safety*
Jared Cazel, *Editorial Assistant*
JoAnn Coleman, *Senior Program Assistant, Capacity and Reliability*
Eduardo Cusicanqui, *Financial Officer*
Richard Deering, *Special Consultant, Safety Data Phase 1 Planning*
Shantia Douglas, *Senior Financial Assistant*
Charles Fay, *Senior Program Officer, Safety*
Carol Ford, *Senior Program Assistant, Renewal and Safety*
James Hedlund, *Special Consultant, Safety Coordination*
Alyssa Hernandez, *Reports Coordinator*
Ralph Hessian, *Special Consultant, Capacity and Reliability*
Andy Horosko, *Special Consultant, Safety Field Data Collection*
William Hyman, *Senior Program Officer, Reliability*
Linda Mason, *Communications Officer*
David Plazak, *Senior Program Officer, Capacity and Reliability*
Rachel Taylor, *Senior Editorial Assistant*
Dean Trackman, *Managing Editor*
Connie Woldu, *Administrative Coordinator*

ACKNOWLEDGMENTS

This work was sponsored by the Federal Highway Administration in cooperation with the American Association of State Highway and Transportation Officials. It was conducted in the second Strategic Highway Research Program (SHRP 2), which is administered by the Transportation Research Board of the National Academies. The project was managed by Kenneth Campbell, Chief Program Officer for SHRP 2 Safety. James Hedlund, SHRP 2 Special Consultant for Safety Coordination, also contributed to the project.

This project could not have been conducted without the help of many individuals at the Virginia Tech Transportation Institute. The authors would like to thank Brian Daily for his work tuning the database cluster to maximize throughput in this effort. Special thanks go to Keith Hangland from HERE; this project would never have happened without his efforts to hunt down an earlier version of HERE's digital maps for use in this work. The authors would also like to thank Omar Smadi and Zach Hans of the Center for Transportation Research and Education for their guidance on the structure of the Roadway Information Database and on how they saw it being used.

FOREWORD

Kenneth L. Campbell, *SHRP 2 Chief Program Officer, Safety*,
and James Hedlund, *SHRP 2 Special Consultant, Safety Coordination*

This report details the methodology used to link the second Strategic Highway Research Program (SHRP 2) Naturalistic Driving Study (NDS) data to the SHRP 2 Roadway Information Database (RID), the final critical step in completing the SHRP 2 Safety database. The NDS data set contains extensively detailed data collected continually from more than 5.5 million trips taken by the instrumented vehicles of 3,147 volunteer drivers in six sites. The RID contains extensively detailed data on 25,000 centerline miles of roadways in these six sites, less detailed data on 200,000 centerline miles of roadways in the six states in which the sites were located, and supplemental data on topics such as crash histories, travel volumes, construction, and weather in the six states. The true power of the NDS and the RID comes when they are linked—when each trip is matched to the roadway segments that were traveled and each roadway segment is matched to the trips that traveled on it.

The matching methodology documented in this report uses as input the GPS position data collected once per second by the NDS instrumentation and the NAVTEQ network of road segments of all public roads in the continental United States over which the NDS vehicles could travel. The Matching Algorithm associates each GPS point of an NDS trip with the road segment on which a vehicle traveled. The principal challenges overcome by the algorithm were to accommodate GPS readings that may drift far from the correct roadway and to be operationally efficient in comparing the 3.7 billion GPS readings with the 2.6 million NAVTEQ road segments that were traversed. The algorithm's results are stored in a very large table that associates trip timestamps with road segments.

The SHRP 2 NDS is the first large-scale study focused on collision prevention (as opposed to injury prevention once a collision occurs) since the Indiana Tri-Level Study (*Tri-Level Study of the Causes of Traffic Accidents: Final Report*, DOT HS-805 085, U.S. Department of Transportation, May 1979). Vehicle use was recorded continuously during the SHRP 2 NDS. Information on vehicle travel, or exposure, can be extracted at the same level of detail as for safety-related events, such as crashes and near crashes. Hence, the SHRP 2 NDS is the first large-scale study to support detailed estimates of collision risk. Moreover, crashes are a leading cause of nonrecurring congestion, so collision prevention has added benefits in terms of reduced delay, fuel consumption, and emissions. The NDS provides objective information on the role of driver behavior and performance in traffic collisions and the interrelationship of the driver with vehicle, roadway, and environmental factors.

The SHRP 2 Safety research program was carried out under the guidance of the Safety Technical Coordinating Committee (TCC), which was composed of volunteer experts. The Safety TCC developed and approved all project descriptions and budgets and met semi-annually to review progress and approve any program modifications. The Oversight Committee approved all budget allocations and contract awards. Assistance was provided by expert task groups, which developed requests for proposals, evaluated proposals, recommended contractors, and provided expert guidance on many issues, such as data access policies and procedures. The decisions and recommendations of the governing committees are implemented by the SHRP 2 staff as they carry out day-to-day management of the research projects.

CONTENTS

1	CHAPTER 1 Background
1	Overview
1	Previous Methods for Associating GPS Points with Roadways Using GIS-Based Tools
3	Improving the Methods
4	CHAPTER 2 Research Approach
4	Matching Process Methods
6	File Processing
12	CHAPTER 3 Findings and Applications
12	Linking Table
12	Linking Maps
12	Incident to Roadway Tables and Crash/Near-Crash Count Maps
13	Measures of Amount Matched
17	Sample Code and Instructional Materials
18	CHAPTER 4 Conclusions and Limitations
18	Limitations
19	References
20	Appendix A. SHRP 2 Manual Route Matching Protocol
24	Appendix B. Example of Visualizing File Counts on Links of Interest Using ArcGIS

CHAPTER 1

Background

Overview

Linking the second Strategic Highway Research Program (SHRP 2) Naturalistic Driving Study (NDS) data to the SHRP 2 Roadway Information Database (RID) will support researchers whose work involves relating information from the driving data to information in the roadway data. The SHRP 2 project collected more than 1 million hours of driving data: it measured speed, acceleration, latitude and longitude, distance and speed relative to other vehicles, as well as recorded video of the vehicle's driver and its surroundings. Additionally, the project collected roadway data and made detailed descriptions of a subset of those roads driven, such as the number of lanes, presence of rumble strips, barrier types, and shoulder width. Almost all driver- and vehicle-related safety research questions benefit from considering the surrounding roadway environment, and most roadway safety research questions benefit from the study of drivers and vehicles on comparable sections of roadway. Many believe the highest return on safety research will be achieved through the study of the interactions among the driver, roadway, and vehicle. This project is intended to provide researchers with a means for connecting or associating the data collected from the driver and vehicle subsystems and the roadway subsystem. As described in detail below, that "means" is a massive linking table (approximately 305 million rows with 2.6 million unique links) containing a listing of map-based roadway links traveled by each NDS vehicle on each trip it made. The RID contains the same map-based link identification numbers for each roadway segment for which inventory data have been collected. The number of links for which data are available varies, depending on the type of data. For example, approximately 120,000 links have measures collected by roadway data collection vans.

To deliver this association, the vehicle position, reported by a Global Positioning System (GPS) once per second, had to be associated reliably with the roadway on which it was traveling. Throughout this task report, the processes used to make

this association are referred to as the Matching Process, or the matching of GPS points to the roads on which a vehicle has traveled. Figure 1.1 illustrates accurate GPS data collection from a vehicle overlaid on a digital map.

The reported positions agree well with the location of the roadways shown on the underlying aerial imagery and the digital map. Through visual inspection, one can interpret the GPS points relative to the digital map data and identify the traveled roadway. Figure 1.2 provides an example in which the reported GPS location of the vehicle deviates quite far from the roadway.

The accuracy of the position reported by GPS varies depending on the orientation of the visible satellites to the vehicle. This project applied a data processing algorithm, known as the Matching Algorithm, that interpreted GPS points and digital maps such as those shown in Figure 1.1 and Figure 1.2, identified the actual road traveled, and then recorded the traveled roadway in a database.

Previous Methods for Associating GPS Points with Roadways Using GIS-Based Tools

The digital maps defining roads are made up of many line segments, referred to as links. Links are connected to each other at nodes. Nodes are present at intersections but are also found at other locations along a section of roadway. Later in the report, Figure 1.3 and Figure 2.2 show further details of these digital maps.

Buffers

Previous methods for associating GPS points with roads have used geospatial functions such as buffers or spatial joins (Wu and McLaughlin 2012) within a geographic information system (GIS). Buffers work by defining a buffer area around the

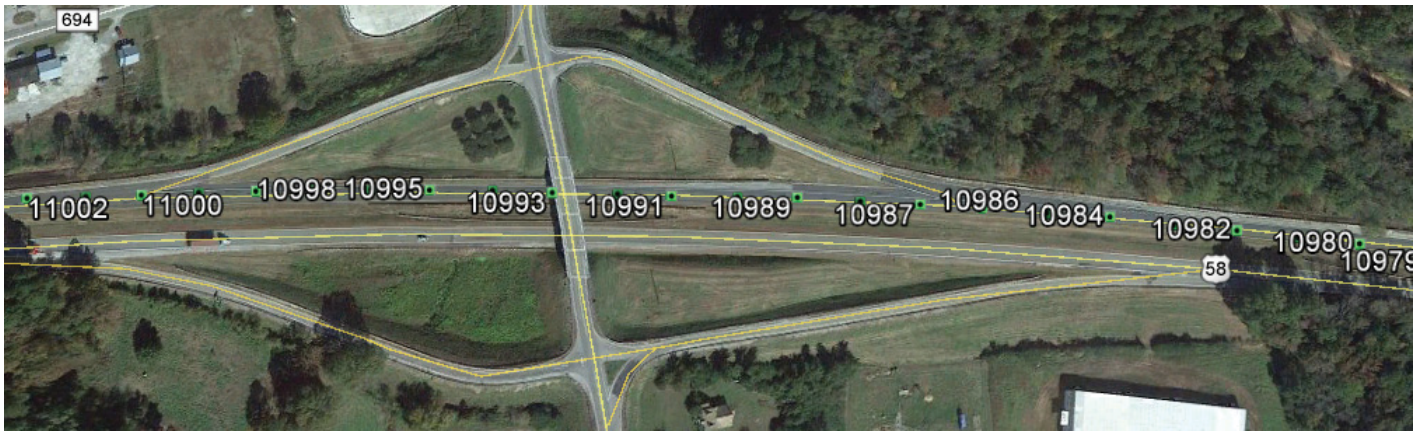


Figure 1.1. Illustration of GPS points that follow a roadway.

link, such as 50 ft on each side of the link, and processing GPS points relative to the buffer. GPS points landing inside the buffer are associated with the link around which the buffer was drawn. Points outside the buffer are either not assigned to a roadway or are assigned to some other roadway if they fall within the buffer around that roadway.

There are several shortcomings of this approach for determining what road a vehicle was on. First, it is difficult to know how large to make the buffer. As we can see in Figure 1.2, the reported position can wander significantly from a road. If the buffer is too small, GPS points will be ignored. If the buffer is too large, GPS points might be assigned to the wrong road. Second, where roads intersect, the buffers will overlap. For example, the left side of Figure 1.2 shows GPS points from a vehicle traveling east to west that land on a road intersecting from the south. A buffer approach would not be able to determine which road to assign these points to.

Figure 1.3 shows an example of the results of this buffer type of method. The left panel of the figure portrays a digital map with lines showing the path of the roads; nodes are marked at intersections with black circles and letters. Note that there is

no node where Link DE crosses Link WX, since Link DE has an overpass over Link WX. This indicates on the digital map that a vehicle cannot travel from Link DE directly to Link WX. GPS latitude and longitude pairs captured by the vehicle are shown as green dots traveling from Point A to Point F. Also, near Node C, a GPS point is recorded on Link CU. Similarly, a GPS point lands on Link WX and Link EZ (see Figure 1.3).

If a buffer is drawn around each of the links and captures the GPS points within the buffer, the solution would indicate the vehicle has traveled on all the correct links but would also incorrectly indicate that the vehicle traveled on Links CU, WX, and EZ. This solution is portrayed in yellow in the right panel of Figure 1.3.

Spatial Joins

An alternative to a buffer approach is to use a spatial join. A spatial join is a geospatial database method that associates objects in space based on their proximity. A spatial join processes roadways as two- or three-dimensional objects through space. The GPS points are also objects in that space.



Figure 1.2. Illustration of GPS points that deviate from a roadway and land on an intersecting roadway.

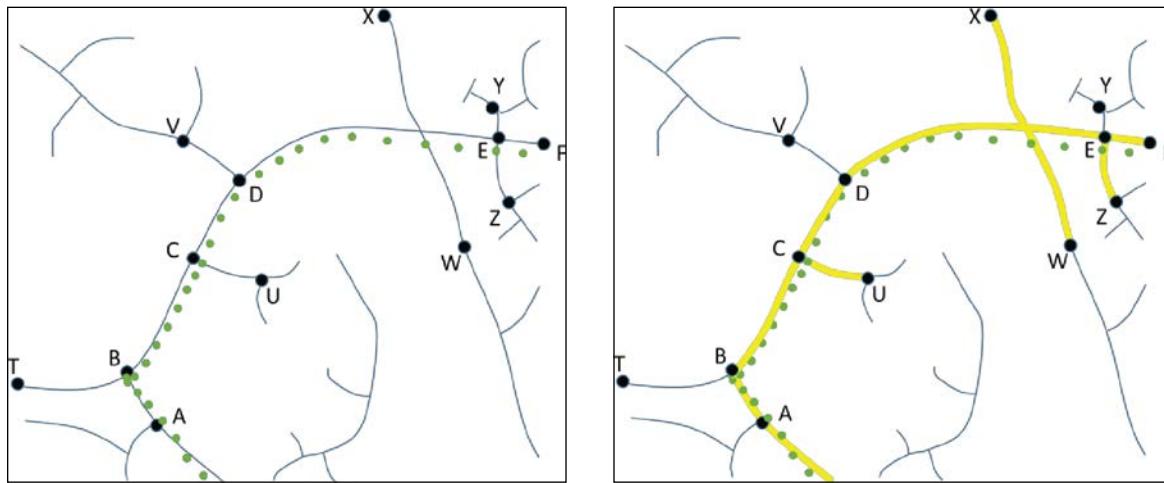


Figure 1.3. Illustration of a route solution based on GPS points relative to roadway spatial join or GPS buffer solutions.

The join computes the distance between the roadways and the GPS points and assigns GPS points to the nearest roadway. Unfortunately, this approach also has limitations. Though this method does not require setting a buffer size ahead of time, it still suffers a similar limitation where roads intersect. If a GPS point lands closer to an intersecting road than to the actual road of travel, it will be assigned to the intersecting road, and the solution will result in errors similar to those shown in the right panel of Figure 1.3. A second limitation of this method is that it is computationally intensive. The roads in the digital map have shape over their length through the space. Computing what part of millions of miles of digital roadway shapes are closest to approximately 3.7 billion GPS points in the SHRP 2 work is a demanding task even for high-performance computing environments. In the end, the solution will still falsely identify roads that are near a GPS point but not actually the road on which the vehicle was traveling.

Improving the Methods

Further enhancements to these fundamental GIS methods are able to improve the accuracy of the output. Li et al. (2014) used a method of joining sequences of GPS points into a “trip line,” and then associating these extended lines to nearby links. In the same work, road names of the roads near trip lines were used within the logic to identify road names that were near at least two of three points defining the line. Those roads that appeared more than twice associated with a trip line were considered the links on which the vehicle had traveled, and the GPS points defining the line were associated with that set of links. These methods increased accuracy beyond the fundamental GIS techniques and are candidates for use with small data sets, but they still have problems with false classifications and cannot be used to handle data sets on the scale of the SHRP 2 NDS.

CHAPTER 2

Research Approach

Matching Process Methods

The method used in the present effort was developed to accurately associate large numbers of GPS latitude/longitude pairs (approximately 3.7 billion) with roadway links using feasible computing demands. At a high level, the approach used for matching was to process vehicle location data relative to digital map data to identify when vehicles were traveling on a specific link. Figure 2.1 shows a schematic of the process.

Position-related data collected in the SHRP 2 NDS were processed through an algorithm that connects the GPS points to the digital map links representing the traveled roadway. Using a digital map database, the Matching Algorithm assembled a network of the road system surrounding a trip, including permitted travel directions and node geospatial locations. This network can be thought of as a swatch of the digital map data, surrounding the extents of the vehicle's GPS data. The algorithm then processed the GPS points from each trip to identify the sequence of traversed intersections (nodes) and, as a result, the links traveled by the vehicle. The results of this algorithm are stored in a table that researchers can use to identify links as well as the timestamps between which the vehicle was on the link. The researcher can also assemble sequences of links into a longer section of road that is of interest. Further detail on this process and the various data sets are described in more detail in the File Processing section of this chapter.

Source Data Sets

The Matching Process relied on data from three sources: vehicle location data, digital map data, and trip metadata.

Vehicle Location Data

Vehicle location data were collected using a GPS receiver integrated into the data acquisition system (DAS). The GPS receiver in the DAS was a Fastrax UP500. This receiver is of a

similar quality to that found in other in-vehicle, consumer-grade navigation devices. It provided location-related measures one time per second (1 Hz). The antenna for the GPS was mounted in the SHRP 2 head unit facing forward out the front windshield.

The GPS receiver provides various position- and trajectory-related measures that were recorded using the WGS84 coordinate system. The variables used in the Matching Process were timestamp, position (recorded as latitude/longitude pairs), speed, number of satellites, and position dilution of precision (PDOP). Both number of satellites and PDOP are indicative of the extent to which the GPS receiver could compute precise location. At least four satellites are needed to compute a position. PDOP is computed on the basis of the angles to the satellites. Broader angles between the satellites and the receiver, and having them lower on the horizon, will improve precision (i.e., make PDOP lower).

The measures were extracted from the DAS into various tables in the SHRP 2 vehicle data database in columnar format using a `file_id`, `variable_id`, `timestamp` recorded, and the data value. For example, Table 2.1 represents a 3-second series of latitude (variable ID -317), longitude (variable ID -318), and GPS speed data (variable ID -330).

The number of satellites and PDOP data are stored in separate relational tables, in the same table organization (i.e., timestamps, variable IDs, and values). These five variables can be joined together using `file_id` and `timestamp`.

Digital Maps

The road network, which for this application can be considered the geospatial location of roads, connectivity of roads, and allowable direction of travel, was provided for use in the Matching Process by HERE North America, LLC. What we could consider a roadway is composed digitally as a set of links and nodes, which when portrayed in a GIS application appears as shown in Figure 2.2.

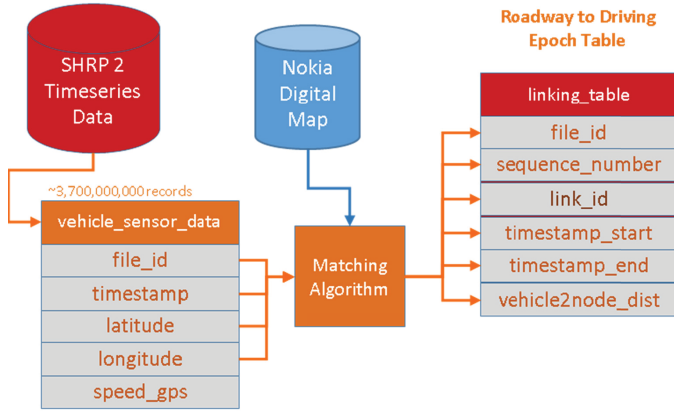


Figure 2.1. Inputs and outputs of the Matching Algorithm.

The digital maps used in the SHRP 2 project are all based on NAVTEQ source data. The RID is based on the ESRI Street-Map Premium 2012Q3 product based on NAVTEQ data. ESRI 2012Q3 documentation defines LINK_ID as, “The NAVTEQ permanent Link ID: Unique IDs for the features” (StreetMap Premium for ArcGIS North America NAVTEQ data dictionary). NAVTEQ’s 2012 data includes more than 40 million links making up the North American road system (newer releases have more links). The Matching Process used all the links in the NAVTEQ data.

The RID includes 1.8 million of the NAVTEQ links stored in a geospatial database format. Within a state, the RID includes all the links in the areas where most of the participant driving occurred. Links in parts of each state that are farther from the central participant area are not included in the RID. This sampling is illustrated in Figure 2.3 for the state of North Carolina.

Table 2.1. Illustration of How GPS Data Is Stored in the SHRP 2 Database (Speed Values in m/s, Latitude/Longitude in Decimal Degrees)

Timestamp (ms)	Variable ID	Data Value
2023327	-330	30.19528
2023327	-318	-75.86013
2023327	-317	37.472706
2024327	-330	30.318735
2024327	-318	-75.86028
2024327	-317	37.472466
2025327	-330	30.102688
2025327	-318	-75.86044
2025327	-317	37.47222

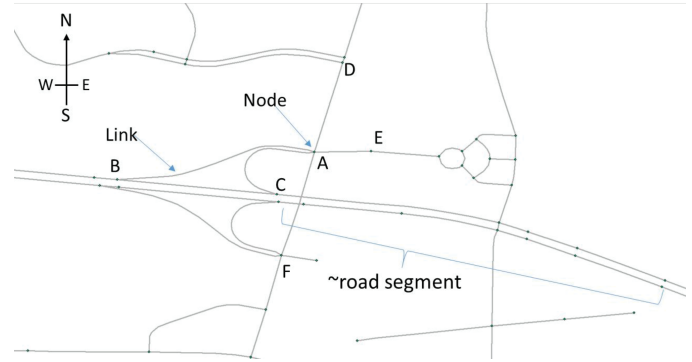


Figure 2.2. Illustration of roads in a digital map made up of links and nodes.

The dense patches of roads in the center of the state are the areas where most of the participant driving occurred, and indicate where the NAVTEQ links are fully included. Beyond these patches, NAVTEQ links representing smaller roads are not included in the RID (see Figure 2.3). Note also that not all NAVTEQ links in the RID will be linkable to detailed roadway inventory data. The extent of the available inventory data is included in a separate report.

The process of matching all of the SHRP 2 NDS GPS records to roads required a different data architecture from the digital map format used for the SHRP 2 RID. The Matching Process used a relational data architecture. NAVTEQ’s relational data format of this type precedes the ESRI product release number by one quarter (ESRI e-mail technical support, 6/24/14). Thus, the NAVTEQ 2012Q2 data was obtained from HERE for use within the Matching Process. These digital map data were stored as relational database tables that could be accessed through standard database queries. The output of the Matching Process was inspected for continuity of routes and for agreement in road names between the two data architectures, further confirming their agreement.

In addition to spatial location information, the digital map data provide other attributes of the links and nodes. Node attributes describing what links are connected at a given node were used in the Matching Process. The digital map data could be queried to determine what other nodes could be reached from a given node. For example, at Node A in Figure 2.2, a query of the digital map data would indicate that five links were connected at that node. The two highway ramps to the west (AB and AC) would be coded as only allowing travel either onto the highway (AB) or off the highway (AC). The other three links (AD, AE, AF) would be coded as allowing travel in both directions. Thus, in this example, four nodes could be reached from the indicated node.

Further, it is helpful to note that a vehicle known to be traveling on the link emanating to the south (AF) from Node A cannot directly access the two parallel east-to-west oriented



Figure 2.3. Illustration of RID link inclusion.

highway sections because the Link AF does not share a node with the east-west highway links.

Trip Metadata

At the same time that the matching processing was being done, participant data associated with each trip were being stored in SHRP 2 database tables as part of the SHRP 2 driver identification activities, and trips by unconsented drivers were being identified. To permit these two activities to be conducted in parallel, the Matching Process was executed on all data, and an intermediate output table was purged of trips by unconsented drivers before generation of final output.

Computing Infrastructure

The computing infrastructure used in the Matching Process was composed generally of the storage architecture, a computational architecture, and the network. The vehicle data, digital map data, and trip metadata were stored in an IBM DB2 database that was configured with one head node and four cluster nodes. The code used in the Matching Process was written in MATLAB language and executed on an IBM Platform LSF cluster with 96 nodes. These computational nodes work simultaneously to read map and vehicle data from the databases, process the data to develop a solution describing the links on which the participant traveled, and then write the solution to another database table that will become the final deliverable for the matching task. These processes are described in more detail in the following section.

File Processing

Creating a File Processing Order

The first step in processing the trip data was to identify file_ids for the trips driven by consented participants and then to randomize these trips before submitting them for processing. The

trip metadata were queried to locate files for the trips driven by the participants. The file_ids for these files (approximately 5.5 million files) were then stored in MATLAB native data format. A random number was assigned to each of the files, and the files were then sorted by this random number. Randomizing file selection during code testing and final processing ensures that adjustments to the code (e.g., correcting logic errors, changing GPS accuracy thresholds) were made without bias from factors that could be indirectly included in the file_id numerical order, such as date of trip collection or data collection site. This file order was then divided up and submitted to the computational cluster in different batch sizes (e.g., 400,000 files per batch in the beginning, 1,200,000 files per batch in the end) until the entire set of files had been submitted for processing.

Processing an Individual File

All processing of trips was done in MATLAB on computational cluster nodes. When a MATLAB computational node was assigned a file to process, it first retrieved the latitude, longitude, timestamp, and PDOP for the trip. These values were stored in the format shown in Table 2.1. An average trip would have approximately 3,600 records describing approximately 720 GPS locations during a file.

After the GPS data were retrieved, the number of satellites and PDOP values were processed to locate the first point in a file at which the number of satellites was greater than five and the PDOP was less than or equal to 17. GPS data recorded before this point in time were omitted from consideration because of their imprecision. Both the number of satellites and the PDOP thresholds were established by reviewing many trips to see where the Matching Process failed and then considering the levels in these two variables coincident with the solution failure. Generally, once the thresholds were met in a file, the solution was reliable. Infrequently, a trip might meet these criteria for a time, and then position quality would decrease for a time and then recover. This might happen, for

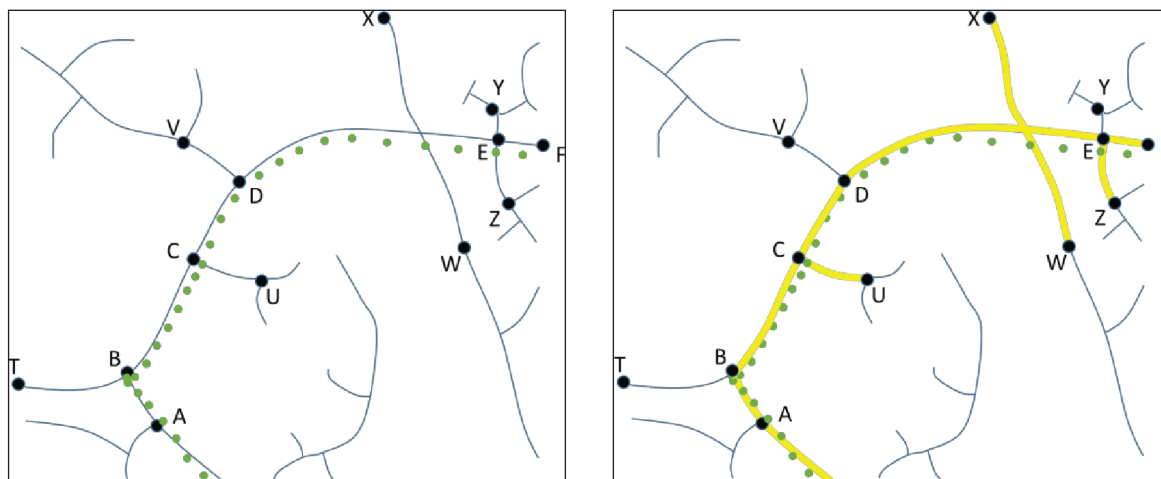


Figure 2.4. Illustration of a road network.

example, when the vehicle was under dense foliage for a period of time. These periods of reduced accuracy later in a file were not omitted from consideration because the previous accuracy would often permit obtaining a good solution even during the period of reduced accuracy.

After determining the reliable GPS points, map data were retrieved for a rectangular area bounding the trip coordinates. The data included the links, the nodes where the links connect, and the allowable direction of travel along the link.

Once the GPS data and the digital map data were available within the computational nodes MATLAB workspace, the distances were computed from each latitude/longitude pair to all of the nodes on the map, and the nodes closest to the start and end of the file were identified. The core Matching Process then was applied. The file was processed sequentially, both from the beginning of the file to the end and from the end of the trip to the beginning.

The sequential processing through the file used a relational database that defines the road system as records in database tables. The relational tables can be queried to determine what

links connect at a node of interest. The simplified road network presented in Figure 1.3 is repeated in Figure 2.4 to further describe the Matching Process.

The connectivity of the road network in Figure 2.4 is cast in table form in Table 2.2. In this example, letters are used to identify nodes. In the digital map database, the nodes are nine- and 10-digit integers. Nodes connected by roads appear on the same row. (See Table 2.2.)

To determine what nodes can be reached from Node D in Figure 1.3, a query such as

```
SELECT node_2, latitude, longitude
FROM table
WHERE node_1=D
```

will return only nodes C, V, and E (see Table 2.3).

Continuing with the example portrayed in Figure 1.3, notice that the results do not return Nodes W or X because they are not accessible from Node D.

A query of this type can be executed, and results set returned, in hundredths of a second using an appropriately designed and tuned database system. With this awareness of the connectivity

Table 2.2. Simplified Example of Table Defining Roadway Network Connectivity

Node_1	Node_2	Node_2 Latitude	Node_2 Longitude
C	D	37.1891515	-80.399365
C	U	37.2198035	-80.418025
C	B	37.1893088	-80.396527
D	C	37.2198035	-80.396527
D	V	37.1893088	-80.418025
D	E	37.1891515	-80.399365

Table 2.3. Simplified Results Set Identifying Nodes Accessible from Node D

Node_1	Node_2	Node_2 Latitude	Node_2 Longitude
D	C	37.2198035	-80.396527
D	V	37.1893088	-80.418025
D	E	37.1891515	-80.399365

of the roadway network, the algorithm logic needs only to follow the feasible sequence of links to which the GPS points track closest. Further, only distances from the GPS points to the nodes need to be evaluated, which greatly reduces the computational demands compared with spatial joins and buffers. In addition to reducing the computational demands, the exposure to errors where GPS points land on side roads (e.g., Links CU and EZ in Figure 1.3) or where roads cross each other (e.g., Links DE and WX in Figure 1.3) are eliminated.

Using this network-aware technique, in its simplest form, the algorithm can iterate through nodes, identifying a small set of accessible nodes ahead (e.g., two to five nodes) and picking the one that the GPS path tracks closest to as the next node in the sequence. The following additions to the logic were used to further improve performance.

In addition to ignoring nodes that could not be reached by a link, the allowable direction of travel on a link was also queried from the database table. In this way, nodes that are only accessible by traveling the wrong way are eliminated from consideration. This assists, for example, in divided highway situations where the GPS is displaced toward an oncoming lane. It also can be used to eliminate consideration of special purpose roads, such as those that only allow emergency vehicle traffic.

Rather than considering only one node forward in time as described in the simplified logic above, the actual logic assembles candidate routes to every node reachable, traveling the correct direction on the link, up to three nodes ahead. It then selects the candidate route with the fewest errors between the three nodes and the nearest GPS points, and then chooses the first node on that candidate route as the next node in the solution. In each iteration, this is essentially building a branching network three steps ahead in all feasible directions, but then backing up two nodes to the best candidate route to identify the next node in the solution. This technique helps avoid routing errors based on transient GPS errors, avoid accidentally selecting a path where GPS points wander at intersections where roads diverge at narrow angles (e.g., interstate deceleration ramps), and avoid cases in which a digital map node is located slightly ahead of or behind the actual location where roads split.

The road network is also processed in this way, both forward and backward through a trip. Generally, the accuracy of the GPS was lower at the beginning of the trip, while the GPS receiver acquired signals from satellites for use in the computation of position. Based on this, the logic located the first nodes in the trip that had vehicle-to-node distances of less than 52.8 ft (1/100th of a mile, selected based on data review during testing). This value was selected by reviewing a number of algorithm solutions for a straightforward cut-off below which matching appeared reliable. Any nodes selected before this were eliminated from the solution because the accuracy was suspect.

Processing the trip in two directions created two separate solutions, one of which needed to be selected as the best. The reverse solution often had better chance of success because it

started at the end of the file, where the GPS had been on and working longer, and so was more accurate. This was not the case 100% of the time, so both solutions were evaluated on two factors to determine which appeared to be best.

The first and most heavily weighted factor in this evaluation was the amount of time of the trip that was matched to the digital map. This value was computed as a percentage (in time) for both the forward and reverse solutions. If the forward solution could not obtain a good solution based on vehicle-to-node distances until a timestamp 25% into the trip and then solved all the way to the end of a trip, but the reverse solution solved from the end of the trip all the way back to a timestamp 10% from the beginning, then the reverse solution would have solved for 90% and the forward solution for only 75%.

The second factor was a percentage of vehicle-to-node outliers found internal to a solution. If a file solution located 100 links, but two of them had vehicle-to-node distances greater than 105.6 ft (1/50th of a mile, selected on the basis of data review during testing), then 2% of its points were considered outliers. An example of this is a solution that got off on the incorrect road, possibly a parallel service road, and then recovered. The points where it was off-track would have a greater vehicle-to-node distance. This was a less common failure mode and also considered a less serious failure, in that it was for a short portion of the trip.

The percentage of time solved was weighted as 80% and the percentage of outliers was weighted as 20%. The two solutions were then compared and the one with the higher total score was retained as the final route solution and written to the table collecting the results.

Process Monitoring

As files were processed, a separate set of database tables was used to track the processing of the work. These tables tracked whether or not a file was successfully processed, and if it was processed, the number of links that were identified in the solution. These tracking tables were used to identify the number of files that were not processed completely and the reason they were not processed. In some cases, files might have sufficiently poor GPS accuracy that a solution cannot be established in either the forward or reverse directions. A file might not be processed fully if it was an exceptionally long trip that exceeded the memory space of a computational node. When large numbers of failures of the same kind were found, the tables were used to guide revisions to handle special cases. Additional details on file processing outcomes are provided later in this document.

Quality Assurance Accuracy and Error Estimates

To evaluate the quality of the results, a three-step process was used to establish a “correct solution” against which the Matching Algorithm could be evaluated.

Manual Route Reduction

First, data reductionists visually translated GPS point paths from files into a list of road names, similar to a set of turn-by-turn directions defining the route. Reductionists were provided a list of files by consented drivers in a random order. In this task, reductionists opened a file in the Virginia Tech Transportation Institute's (VTTI) data reduction software (Hawkeye) and selected the Map Dashboard. This function overlays the GPS points (shown in yellow) on aerial imagery (see Figure 2.5).

Reductionists visually inspected the map and GPS points to determine the road the vehicle was on. Reductionists moved the map using mouse click-and-drag and/or referenced the vehicle yaw values to slew the map to where turns were made.

In addition to identifying the street names, the reductionists approximated the timestamps at which transitions occurred between one road and another. While identifying the route of travel, reductionists also coded the accuracy of GPS. Levels of accuracy were approximated into five levels: no displacement, slightly displaced, moderately displaced, extremely displaced, and poor geometry (the example in Figure 2.5 was considered moderately displaced). Each level was operationally defined using visual examples provided in the protocol. Additional detail on this method can be found in Appendix A. While not the same trip shown in Figure 2.5, Table 2.4 provides an example output of the reductionists' work.



Figure 2.5. Portrayal of GPS points from a trip overlaid on aerial imagery as seen by video reductionists.

Table 2.4. Example Results from Visual Inspection of GPS Points on Aerial Imagery

Manual				
Road Order	Road Name Read from Map	Start (ms)	End (ms)	GPS Rating
1	None	0	334943	Poor geometry
2	Shadagee Rd	334944	391149	None
3	Southwestern Blvd	391150	1541418	None
4	Angle Rd	1541419	1630939	None
5	East and West Rd	1630939	1670095	None
6	Main Dr	1670096	1784703	None
7	None	1796368	1841680	None

Comparison of Manual Solution to Matching Algorithm Solution

Once the data reductionist completed review of a file, another researcher checked the manual solution against the output of the Matching Algorithm. This task was a check of both the manual results and the algorithm. To accomplish this check, code was used to create an Excel file in which each tab described a trip, and it included the manual solution (Table 2.4) and the algorithm solution (Table 2.5).

The results of the manual solution are provided in Table 2.4. In this example, the reductionist identified five roads by name (road order 2–6). Travel on these roads was preceded and followed by some time not on a named road. In Table 2.5, the output of the algorithm is presented. The algorithm reports every link traveled on, whereas the manual reductionist only identifies where road names change, such as at a turn. To make the algorithm output comparable to the manual output, code was used to reduce repetition of a road name over many LINK_IDS into a single time span. For example, Southwestern Blvd read manually off the map is identified as one road but is actually made up of 117 LINK_IDS (the 5th link in the file to the 121st link in the file). It is also stored in the database as “Southwestern Blvd” and “US-20.”

The researcher manually compared the road names, sequence, and timestamps between the reductionist's version and the algorithm version. When disagreements were identified, the researcher opened the file and confirmed whether the reductionist had made a mistake or the algorithm had made a mistake. Corrections were made as necessary to the reductionist's manual output (Table 2.4).

To validate the algorithm results, the second researcher generated a random number from one to the number of links reported by the reductionist. In the example in Table 2.4, this

Table 2.5. Example Results from Algorithm

Algorithm					
Link Order (n)	Start (ms)	End (ms)	Road Name 1 from Database	Road Name 2 from Database	Road Name 3 from Database
1	338922	395923	S Creek Rd	Shadagee Rd	CR-476
5	395923	1543942	Southwestern Blvd	US-20	
122	1543942	1631943	Angle Rd		
130	1631943	1671944	East and West Rd	CR-363	East and West Rd
134	1671944	1787946	Main Dr		
137	1787946	1817946	Honeysuckle Dr		

would be a random number from 1 to 7. The researcher then checked this row for algorithm accuracy. For example, if the random number 2 was generated, the researcher looked at the road name provided by the reductionist and saw “Shadagee Rd.” In the algorithm solution (Table 2.5), “Shadagee Rd” is placed there through an automated process. The researcher then checked the timestamps at the start and the end of the road to make sure they were in agreement. In this example, the manually identified start of “Shadagee Rd” is about 4 seconds earlier than when the algorithm identified the start. The end of the road has a similar difference. This agreement indicates that the algorithm agreed with the manual method for links 1 through 4, which accounts for the time between timestamps 338922 and 395923, or 57 seconds. This time span would be classified as a true positive.

Only one row was checked in each trip. However, reusing the trip shown in Table 2.4 and Table 2.5 for illustration, if the random number picked had been 7, the manual method indicated the participant was not on a road, but the algorithm incorrectly indicated the vehicle was on Honeysuckle Dr after timestamp 1787946. This 30 seconds would be

classified as a false positive. If the random number picked had been 1, the manual method indicated the vehicle was not on a road or not on a named road. The algorithm (Table 2.5) did not assign a link to this timespan, so this is classified as a true negative.

This process was conducted for 100 files, providing a check of 100 manually identified roads. Recall, because a manually identified road is generally made up of many links in the database, this represents checking of many links. See Figure 2.6 for the results of this analysis presented in a confusion matrix. In this matrix, the actual (i.e., manually identified) situation is presented in the rows, and how the algorithm defined the situation is presented in the columns. If the Matching Algorithm agrees with what was found through the manual process, then it is correct.

As shown in Figure 2.6, the algorithm correctly assigns the driving data to the correct link 91% of the time. In 86% of the time the algorithm does not assign the data to a link, the GPS data is actually not on a link. In 2% of the time when the vehicle is not on a link, the algorithm assigns the data to a link.

Driving Seconds		Algorithm				
		Link N	Not on Link			
Actual	Link N	6,470,066 True Positive	620,613 False Negative	91%	Sensitivity	Method finds x% of link time correctly.
	Not on Link	131,012 False Positive	774,447 True Negative	86%	Specificity	x% correct ignoring time that is not on a link.
		98%	56%			
		PPV	NPV			

Figure 2.6. Algorithm validation confusion matrix.

Masking Personally Identifying Information

The Matching Algorithm attempted to identify every link on which the vehicle traveled. Various approaches have been explored for removing or masking locations, such as where the participant lived or worked. The current best approach is considered to be removing links leading up to an intersection that

has many (e.g., three or more) approach options. In this way, from the data included in the file, many possible travel paths might have been taken by the vehicle and thus the actual destination would not be apparent. However, further work and review of this approach is required. The Linking Table will continue to be treated as personally identifying information (PII) until a masking solution can be identified and approved.

CHAPTER 3

Findings and Applications

As each file was processed, the results of the matching solution were written to a database. This database is referred to as the Linking Table. It lists each link, identified with a unique LINK_ID shared with the RID, the timestamp within each file when the vehicle started on the link, the timestamp when the vehicle left the link, and the vehicle-to-node measure at the start of each link. The timestamp in the table is the same timestamp as the timestamp of the closest GPS point to the start (and end) of a link. The following sections describe the findings from the Matching Process and data developed through the process.

Linking Table

The primary output of the matching work is the Linking Table. It lists the LINK_IDs on which participants drove, as well as timestamps at the beginning and end of the links and the file that includes the rest of the vehicle data describing these epochs. Researchers can use this table to index into the vehicle data for analysis of driving data. The Linking Table contains approximately 305 million rows. See Table 3.1 for a small sample of the Linking Table.

Fields in the Linking Table

The fields in the Linking Table are defined as follows:

- **FILE_ID.** The unique number identifying the trip file within the SHRP 2 NDS database. One trip file is typically one key-on to key-off ignition cycle, but sometimes trips are divided into multiple files.
- **SEQUENCE_NO.** A within-file sequence number identifying the order in which the LINK_IDs were traversed in the file.
- **LINK_ID.** A permanently unique ID for a roadway link, within the digital map databases. A road or road segment is made up of multiple links.
- **TIMESTAMP_START.** The time (ms) into the file at which the vehicle began traversing the link.

- **TIMESTAMP_END.** The time (ms) into the file at which the vehicle finished traversing the link.
- **VEHICLE_TO_NODE_DIST.** The estimated straight-line distance (ft) between the latitude/longitude location reported by the vehicle GPS when passing closest to the start of the link, and the latitude/longitude location in the digital map. This value provides a measure of how closely the vehicle GPS output is following the digital map.

The example, Table 3.1, shows six links traversed within FILE_ID 1000019. Over the entire file, these are the 13th through 18th links traversed. The vehicle entered the 17th link, with LINK_ID 41839022, 213.5 seconds into the file. The vehicle left the link 33 seconds later. The straight-line distance between the start of the digital map link and the position measured by the vehicle GPS was 19.8 ft.

Linking Maps

Maps provide a helpful tool for visualizing the data stored in the Linking Table. In Figure 3.1 through Figure 3.6, the count of the number of participant traversals over a link is color coded on a map. Links shown in green have fewer traversals, and links in red have the most traversals. These counts were determined by querying the 305 million rows in the Linking Table to count the number of traversals on each unique link and the number of participants, restricting the result to links with more than 10 participants and 30 traversals, joining these counts to the geospatial data for a link, and portraying it in a GIS application, in this case ArcGIS.

Incident to Roadway Tables and Crash/Near-Crash Count Maps

Through the roadway LINK_IDs, connections between incidents and many roadway-related descriptors can be created to explore the relationships between roadway attributes and

Table 3.1. Linking Table Sample

FILE_ID	SEQUENCE_NO	LINK_ID	TIMESTAMP_START	TIMESTAMP_END	VEHICLE_TO_NODE_DIST
1000019	13	771927354	181461	191461	29.91060377
1000019	14	771927355	191461	192461	31.48441077
1000019	15	122666697	192461	202461	25.67103791
1000019	16	771927351	202461	213462	43.94671698
1000019	17	41839022	213462	246462	19.82166502
1000019	18	41835835	246462	248462	6.555630753

different incident types. Attributes that can be joined include Highway Safety Information System (HSIS) category, number of lanes, functional class, speed limit, and lane width. This incident data can also be visualized on maps. In Figure 3.7, links with different numbers of incidents (crashes or near crashes) are color coded.

Roadway-related attributes for the links are listed in the boxes in the figure. For example, the red link has two lanes, is Functional Class 1, and is classified as a ramp.

and the roadways on which the participants drove were cataloged. This catalog is the main deliverable of this work, and constitutes approximately 305 million rows in a database table (one row per link traversal). In total, the database identifies the roadway for approximately 1,026,000 hours of driving. The following sections describe the results of the data processing in more detail, with emphasis on quantifying what has not been matched.

Measures of Amount Matched

As of December 2014, the main data processing effort has been completed. Approximately 3.7 billion latitude/longitude pairs from 5,512,844 trips by consented drivers were analyzed

Percentage of All Files

The number of files not processed constitutes approximately 1.8% of the total files to which algorithm code was applied. The 1.8% difference is made up of cases in which *(text continues on page 17)*

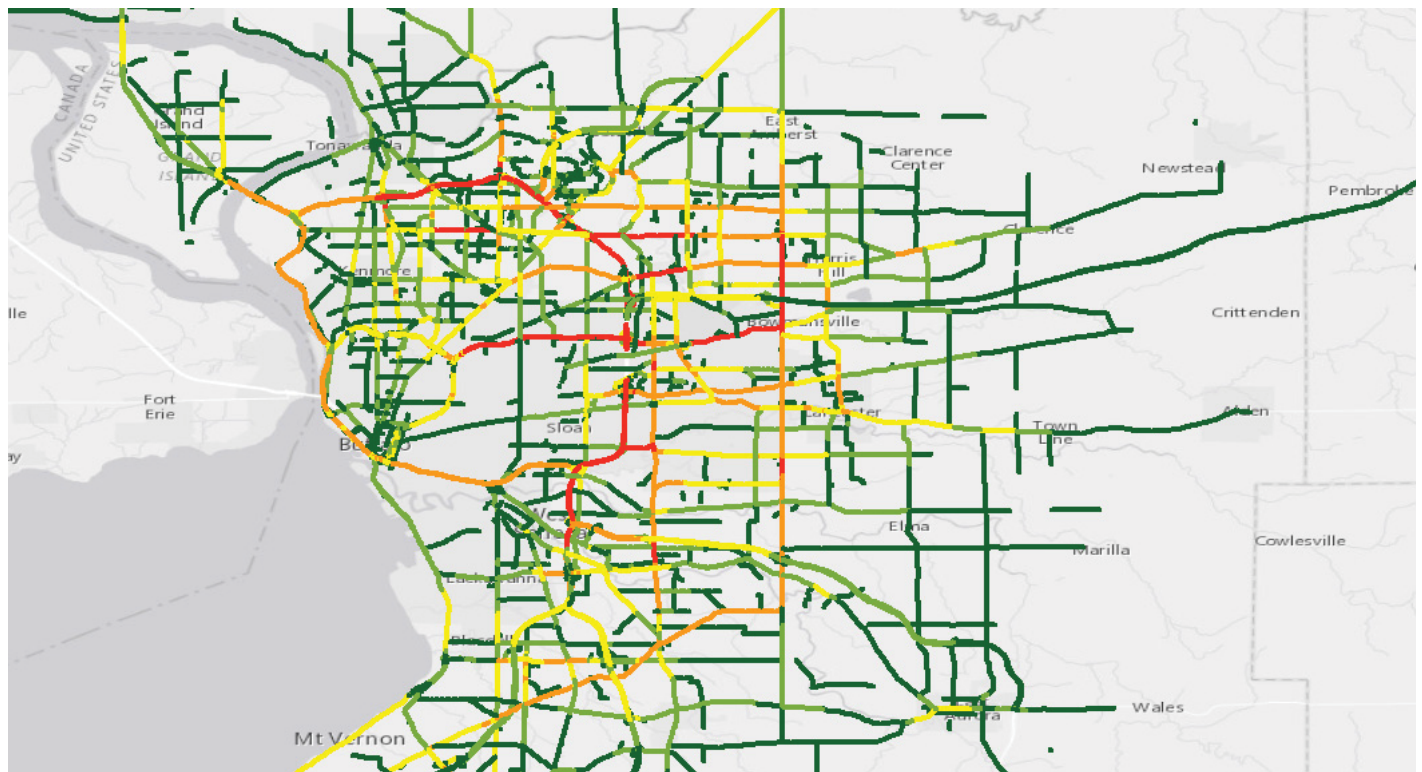


Figure 3.1. Buffalo, NY-area map indicating links and file count (green = low, red = high).

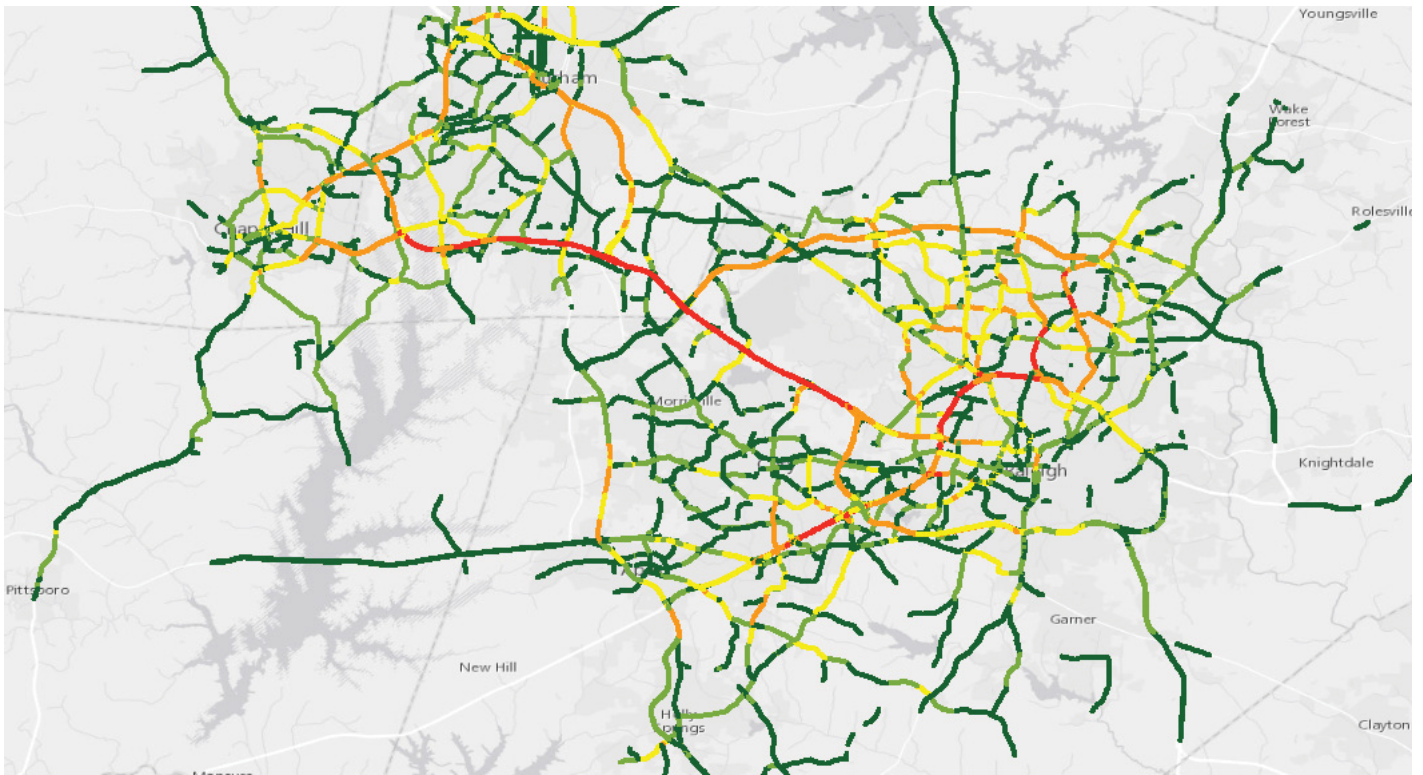


Figure 3.2. Raleigh/Durham, NC-area map indicating links and file count (green = low, red = high).

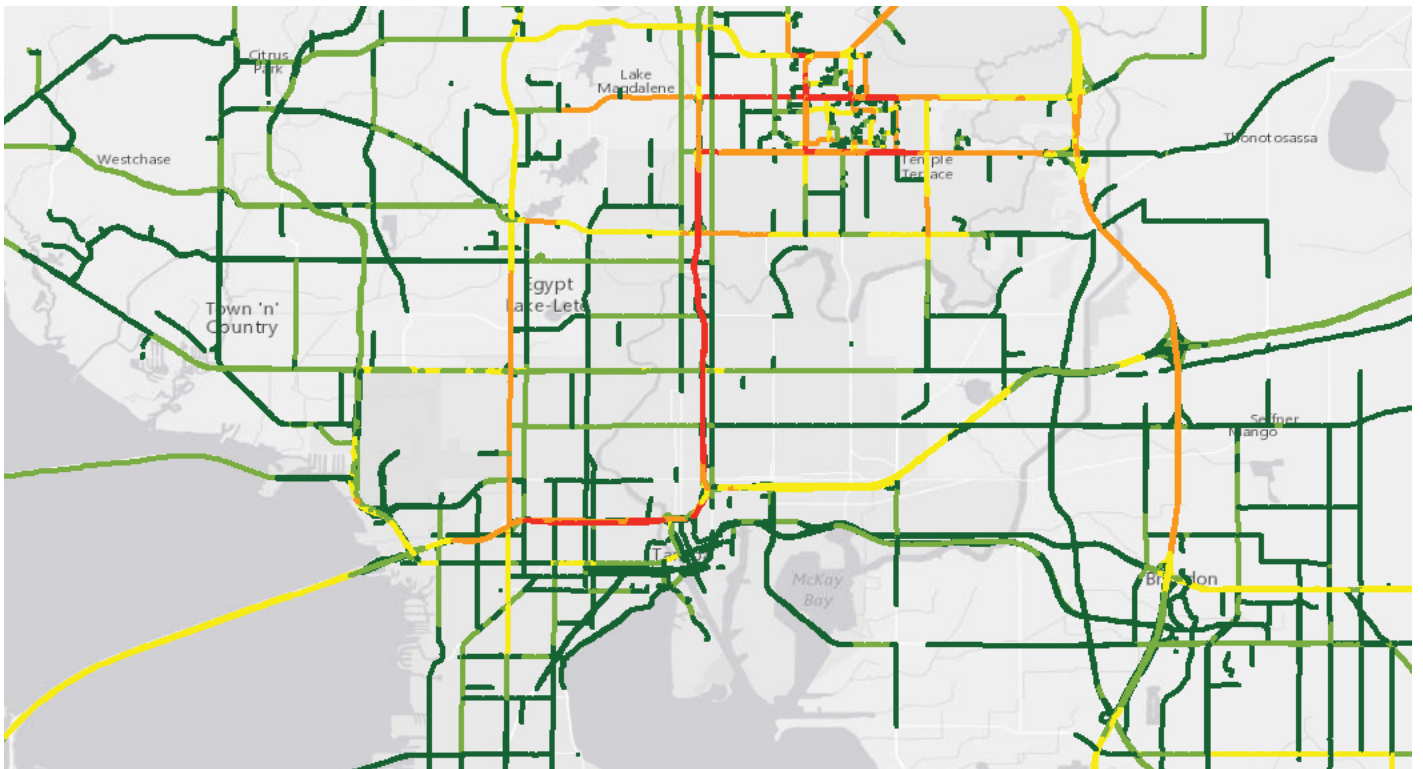


Figure 3.3. Tampa, FL-area map indicating links and file count (green = low, red = high).

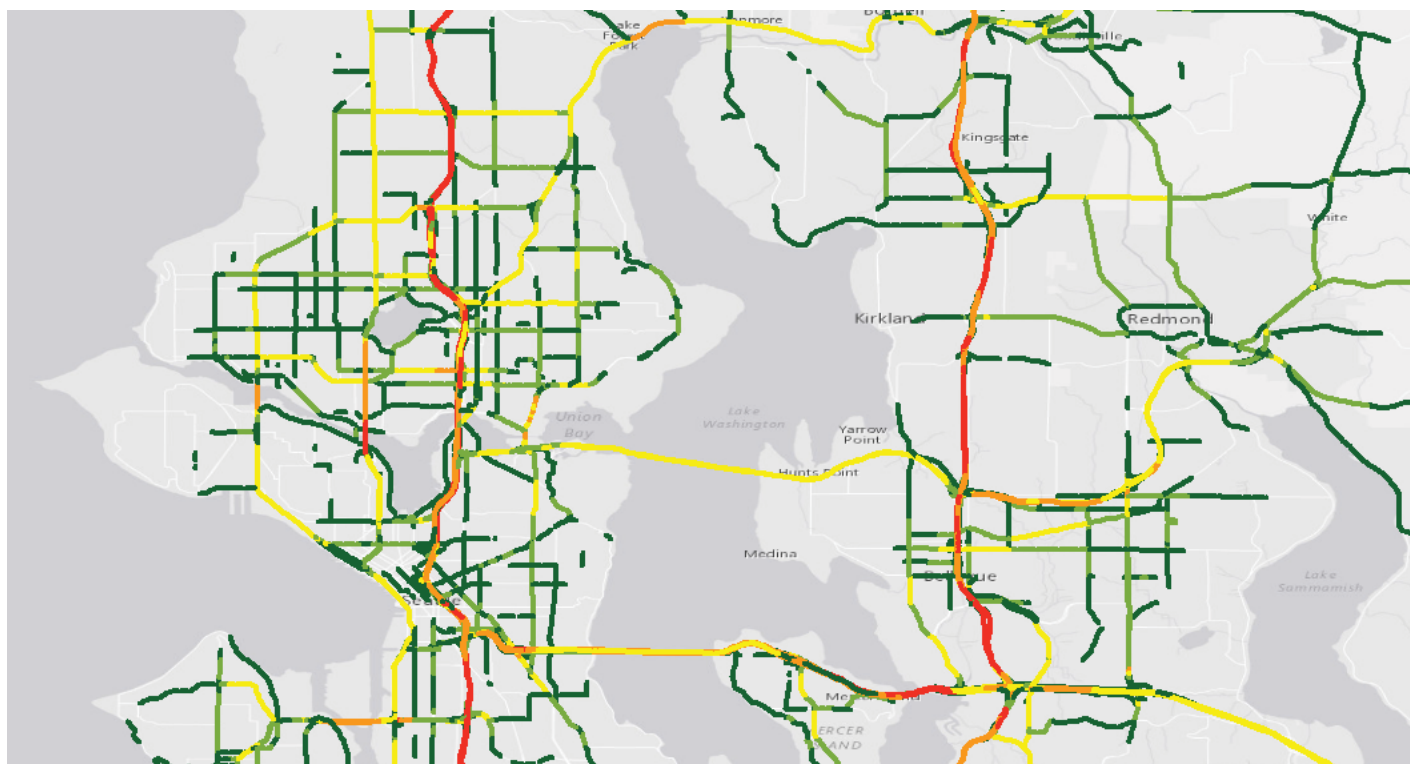


Figure 3.4. Seattle, WA-area map indicating links and file count (green = low, red = high).

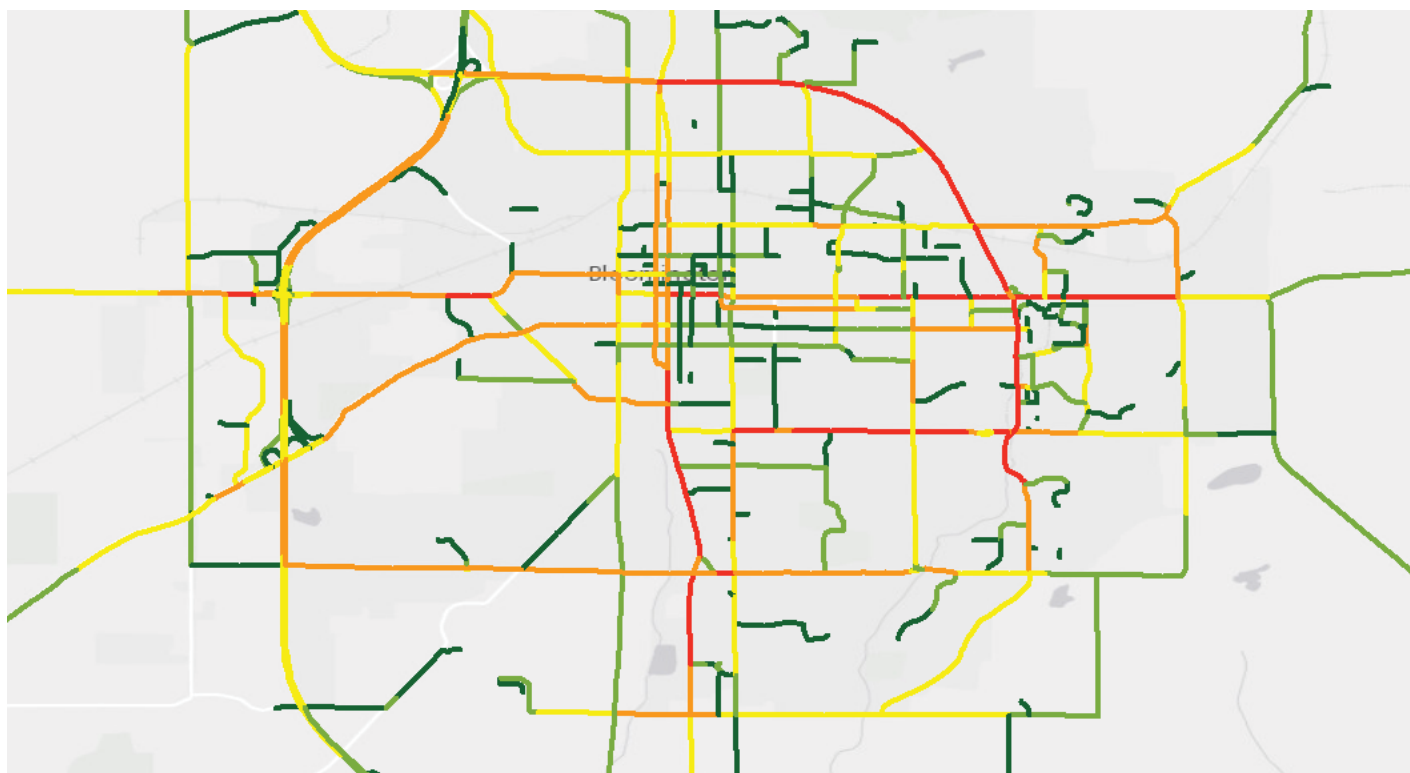


Figure 3.5. Bloomington, IN-area map indicating links and file count (green = low, red = high).

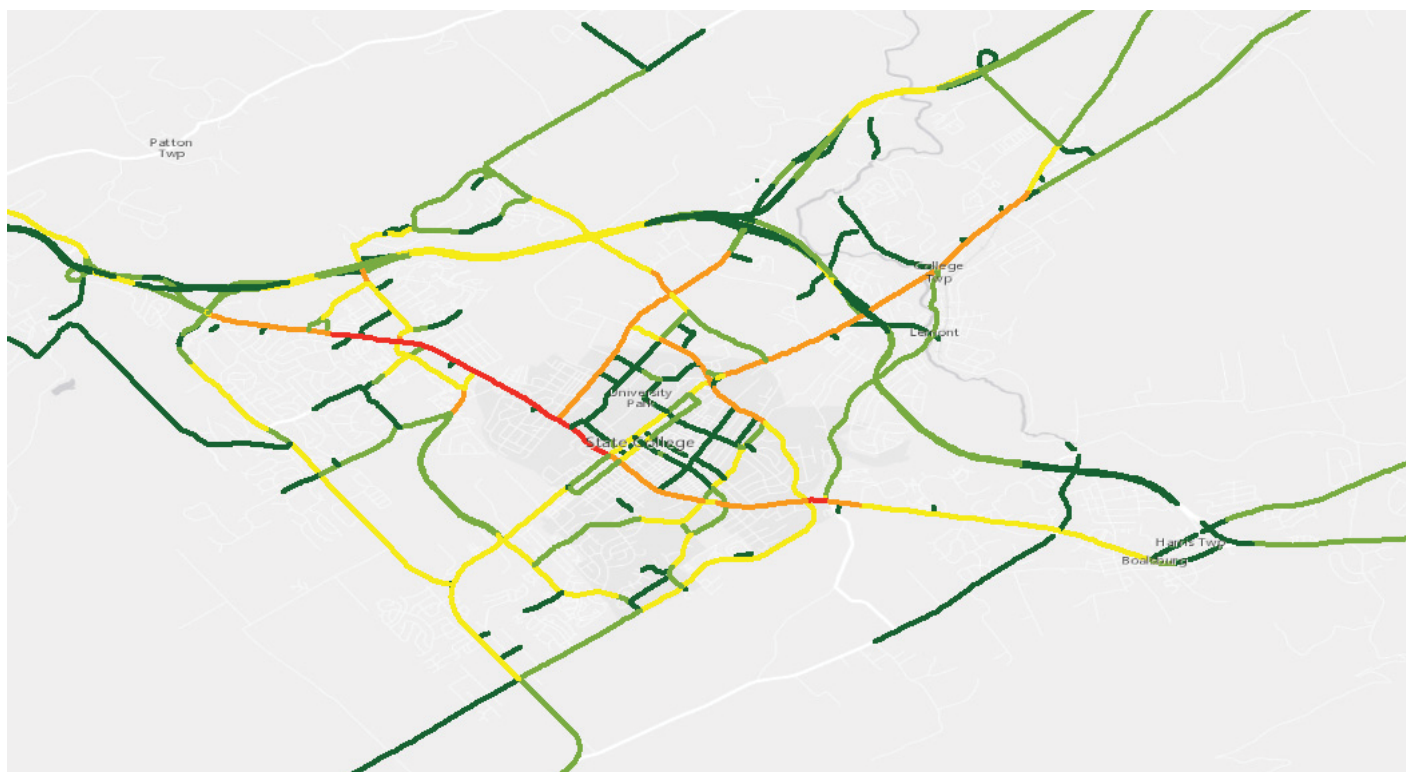


Figure 3.6. State College, PA-area map indicating links and file count (green = low, red = high).

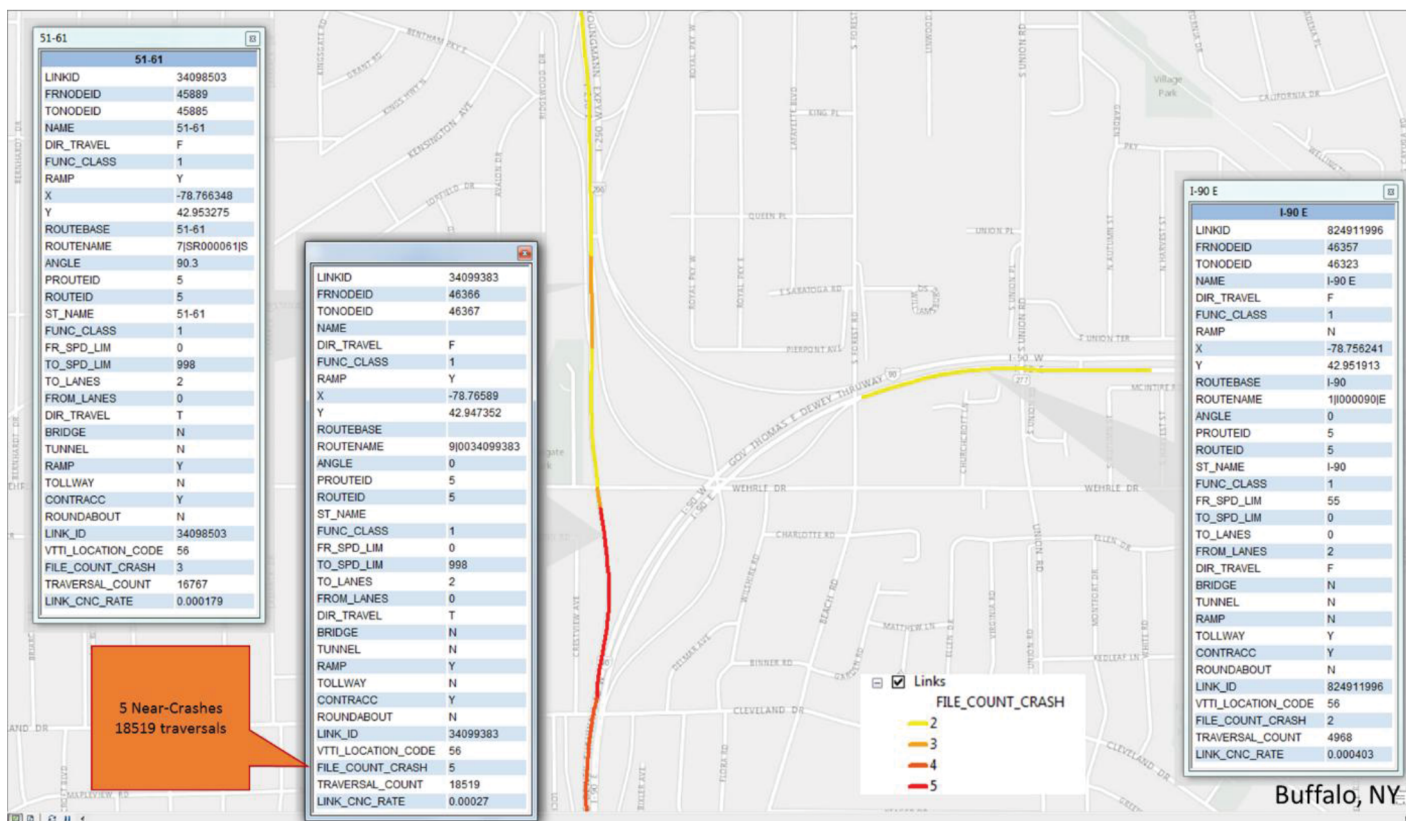


Figure 3.7. Illustration of links coded with the number of crashes or near crashes identified.

(continued from page 13)

GPS data were never present or never reached the accuracy criteria, the file was too short for GPS to begin reporting, the vehicle was started but never moved, out-of-memory errors occurred for some very long trips, for some reason a succession of road links could not be established for the file, and so on.

Percentage of Time Within Files Matched to Roads

For the files that were processed, an analysis was conducted to determine what percentage of the driving time in each file was matched to roadways by the algorithm. Twenty files that were manually reviewed and assigned to roads (see Chapter 2) were checked to determine the amount of time that the participant was on roads. That is to say, time spent on a driveway or in a parking lot was omitted. The amount of time that the Matching Algorithm allocated to links was also computed. The percentage of time assigned by the Matching Algorithm was divided by the time on roads determined manually. The median percentage allocated by the algorithm was 87% of the time spent driving on roads within a file. Time was generally lost at the beginning of the trip as the GPS came online and established reception with a sufficient number of satellites to provide good positioning.

Sample Code and Instructional Materials

Bulk Inserting CSV Files into a Database

Once methods are developed for masking PII, the Linking Table will be made available in multiple CSV files, which in total will be 305 million rows of data in six columns. A sample of this table is shown in Table 3.1. Though details of a procedure to import CSVs into a database depend on the operating system and database platform, importing CSVs into a database is a standard method. In general, the technique will loop over files that meet a specification, executing the database command line utility for bulk inserting for each file. For example, if the map link data are written to CSV files with file names MAPLINKxxxx.csv, where xxxx is an integer from 0000 to 9999, then the following is an outline of how such a script would look:

```
for filename in MAPLINK*.csv
bulkload filename
```

If the operating system is Windows and the target database engine is MS SQL Server, then the following is a more specific example of a script to load the CSV files:

```
for %f in ( MAPLINK*.csv ) do
bcp database.dbo.map_linking_table in %f -T -c
```

where database is the name of the target database on the local machine and map_linking_table is the name of the target table in that database. If the database is on a remote machine, the server can be specified using the -S server_name option.

Counting Number of Files Traversing a Link

Once imported into a database, use of the table to, for example, determine how many files include traversals of a link, is straightforward. The SQL coded below orders a list of links with the most frequently traversed at the top and provides a count of the files that include traversals on the link.

```
select link_id, count(unique(file_id)) as file_count
from map_linking
group by link_id
order by file_count desc
```

Six lines of output from the query above are shown in Table 3.2.

Visualizing the Link Counts in ArcGIS

It is also sometimes helpful to portray the number of traversals on roadways. A query such as the one shown above, restricted to a specific location, could be portrayed in GIS similarly to the map in Figures 3.1 at the beginning of this chapter. Detailed instructional materials for this type of task are provided in Appendix B.

Table 3.2. Sample Output from Link Count Query

LINK_ID	FILE_COUNT
126684274	22375
126684273	22360
34129178	22356
779335195	22295
824914307	22292
824914306	22286

CHAPTER 4

Conclusions and Limitations

The SHRP 2 NDS contains a record of personal vehicle routes of travel on a scale well beyond what has been collected in the past. The data contain more than 5.5 million trip records from consented drivers. The time totals more than 1 million hours of driving, generating approximately 3.7 billion latitude/longitude pairs. The separation between each pair of points was different depending on vehicle speed, and each pair included different levels of precision depending on the reception of GPS and geometry of the visible satellite constellation as the vehicle moved through the network. The roadway database making up the RID includes 1,820,954 links. The objective of this project was to create a catalog of driving epochs and identify on which of the 1.8 million links they occurred.

The catalog, exported to the Linking Table, contains 305,000,000 link traversals and the time at which the participant vehicle entered and exited the link. Through the Linking Table, in seconds, researchers can locate as many as 27,000 traversals on a single link out of more than 1 million hours of data.

Limitations

The output of any large-scale data mining effort should be used cautiously. When processing large amounts of data, it is often difficult to detect how the process might be affecting the output. The Linking Table itself should not be considered completely representative of all the SHRP 2 participants and all their trips. As previous sections have indicated, approximately 1.8% of the trips were not processed. Within trips, the amount that could be processed varied. Some vehicles may have had a faulty GPS for a period of time. Some participants may live in locations that have exceptionally good or exceptionally poor GPS coverage compared with other participants. Researchers are encouraged to check their sample for biases before conducting analyses. This can be done by checking trip counts for different participants against each other, or spot-checking each participant's data to establish a rate of files that are missed and determine if it is more or less than other participants.

References

Li, Y., R. Gibbons, and A. Medina. 2014. *Feasibility for Linking an Adaptive Lighting Database with SHRP 2 Naturalistic Driving Data*. Draft. National Surface Transportation Safety Center for Excellence, VTTI, Blacksburg, Va.

Wu, S., and S. McLaughlin. 2012. Creating a Heatmap Visualization of 150 Million GPS Points on Roadway Maps via SAS. Presented at 20th Annual SouthEast SAS Users Group Conference, Durham, N.C.

APPENDIX A

SHRP 2 Manual Route Matching Protocol

Software needed: Data/Video/Map viewer, spreadsheet (VTTI's Hawkeye data viewer and Microsoft Excel were used for this work).

Collections: NEW SHRP 2.

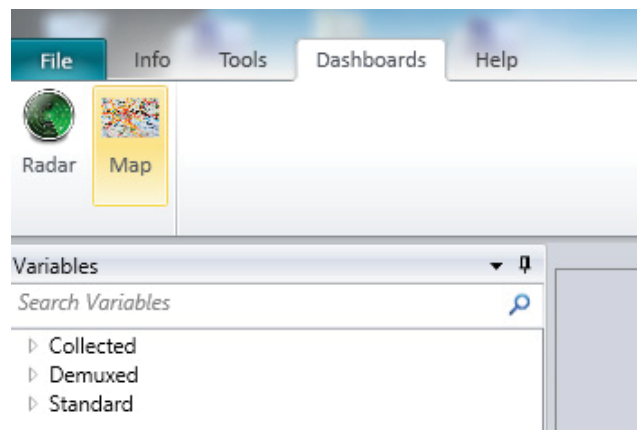
The goal of this task is to review files in Hawkeye from the NEW SHRP 2 data collection to provide sequential names of roads traveled during a trip and also to categorize problems if any part of the trip is not clear on the Hawkeye map. This document provides guidance and definitions required for completing these analyses consistently. Please bring any questions to the data reduction coordinator as soon as possible so that consistency and accuracy can be maintained.

Reduction Log

1. Locate SHRP 2 Linking QA reduction.xls.
2. In the first worksheet (named "1000 files"), find the first row that has not been marked by a reviewer. Enter your name in that row under Reviewer, and proceed to Hawkeye to open the file.

Setting Up Hawkeye

1. Load Hawkeye.
2. Load the NEW SHRP 2 collection.
3. Copy and paste the File_ID from the Excel log for the file you have signed out into the Trip field in Hawkeye. Click on Enter or Load Trip.
4. Open the Map (found in the Dashboards tab).



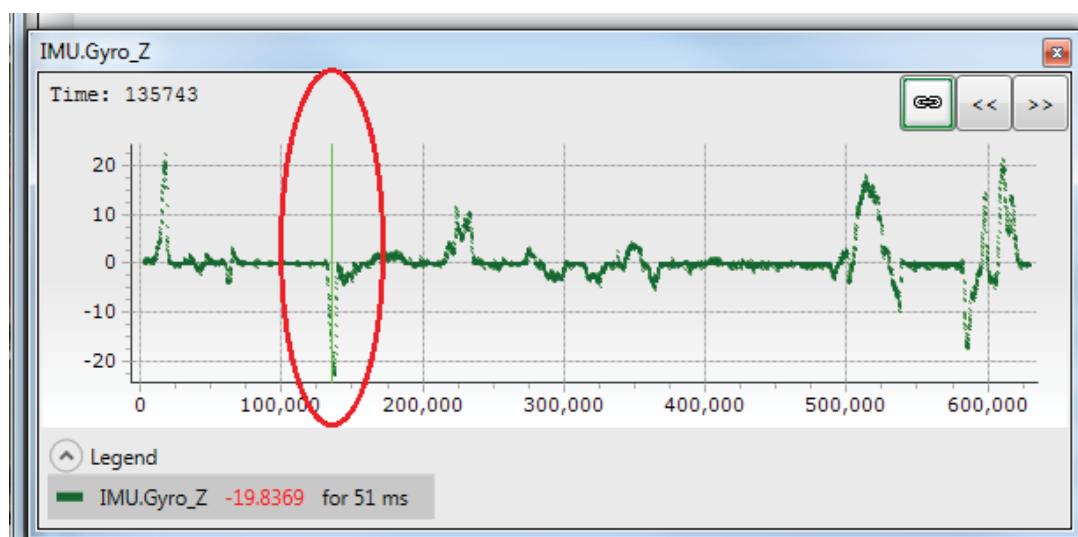
- a. Enlarge the map if necessary.
 - b. Zoom in and out by using the scroller on the mouse and move the map by dragging.
5. Open the Gyro_Z variable graph (found in the Variables menu, in the IMU module).

Reduction Spreadsheet

1. Open the SHRP 2 Linking QA Reduction spreadsheet.
2. Create a new sheet for the current file by copying the worksheet named Template. Rename this worksheet with the File_ID.

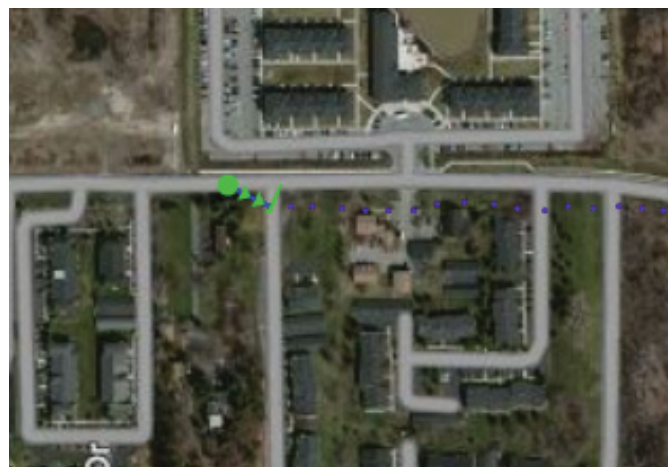
Begin Data File Route Assessment

- When the play button is clicked and the trip data collection begins, a green dot will show up on the map, indicating what should be the vehicle location. There should also be a dotted path (usually blue or purple), showing the vehicle's entire travel path. At times the green dot will not be on the actual (dotted) travel path when the file starts playing. *The data reduction will begin when the green dot begins to move along the dotted path.*
- The goal of the assessment is to record each road that the vehicle travels on, as indicated by the map travel path (green dot moving along the dotted path), using approximate timestamps for time on each road. If the travel path is not actually on a road, this should also be noted by timestamp duration, along with a categorization of the nature of path deviation (described below). To save time and advance the green vehicle indicator to the next road, double-click on the next location on the Gyro_Z graph where there is a substantial jump in value (this should indicate a notable change in direction).



- Record the File_ID in each completed row.
- Record the sequence order of each road in the Road Sequence column and the corresponding name of each road in the column Road Name.
 - Add more values or remove excess in the Road Sequence column as necessary.
 - Record road name exactly as it is written on the map. For special unnamed road configurations, code as applicable (e.g., On ramp, Off ramp, Cloverleaf, Parking lot, Driveway). Otherwise, if the road name is not indicated on the map, code as "No name."
 - If there is no path or you cannot tell what road the path corresponds to, enter "none" for Road Name.
 - If there are questions about the path, videos may be opened for clarification.
- To indicate whether there are cases in which the path is not on a road segment for any period, code each segment using the following categories (provided in a dropdown menu in the GPS Problems column):
 - *None*. There are no problems with correspondence between the indicated travel path and the roads presented on the map. Points indicated on dotted travel path might not be completely on the road segment, but are less than a road width from the actual roadway (Road Name will be entered).
 - *Slightly displaced*. Points indicated on dotted travel path are not on the road segment, and are generally between one and two road widths from the actual roadway (points follow the general travel direction of the road but are slightly shifted as

a unit) (Road Name will be entered if it is known; “None” if it is not). The following figure shows an example of a portion of a travel path that is slightly displaced.



- *Moderately displaced.* Points indicated on dotted travel path are not on the road segment and are generally more than two but less than five road widths from the actual roadway (points follow the general travel direction of the road but are moderately shifted as a unit) (Road Name will be entered if it is known; “None” if it is not). The following figure shows an example of a portion of a travel path that is moderately displaced.



- *Extremely displaced.* Points indicated on dotted travel path are not on the road segment and are generally five or more road widths from the actual roadway (points follow the general travel direction of the road but are extremely shifted as a unit) (Road Name will be entered if it is known; “None” if it is not). In the following figure, the portion between points 2089 and 2104 illustrates a travel path that is extremely displaced.



- *Missing.* There are no points on the map indicating travel path; there is a gap in the actual travel path (“None” will be entered for Road Name).
- *Poor geometry.* Either the travel path points are scattered (they do not follow a route and look haphazard or clustered), or they indicate a travel path not related to the actual roadway (e.g., a diagonal path cutting across horizontal streets), or the green dot indicating vehicle position has become stationary (but vehicle is actually still moving, seen in video) (“None” will be entered for Road Name).
- If a particular road segment contains a travel path with poor points, but not all for the same reason, add another row to the spreadsheet with a duplicate Road Sequence number and Road Name to indicate each problem and the timestamps for which that problem occurs. Also enter duplicate rows if part of the segment has no problems and part does have problems. See the following example.

Road Sequence	Road Name	Timestamp Start	Timestamp End	GPS Problems	Comments
1	Main Street	10000	50000	None	
2	Elm Avenue	50001	70000	Slightly displaced	
2	Elm Avenue	70001	100000	Missing	
2	Elm Avenue	100001	400000	None	
3	Second Street	400001	500000	Poor geometry	
3	Second Street	500001	700000	None	
4	Third Street	700001	900000	None	

- If it appears that there is no path to record for a trip (no GPS points to indicate road segments), open the graphs for Latitude and Longitude (both in the Collected>Head_Unit module) to be sure that GPS points do not appear at a point later in the file. If there are any points at all from which road segments can be determined, record those road segments with applicable road names.
- If there are any situations that make reduction difficult, provide explanation in the Comments column.

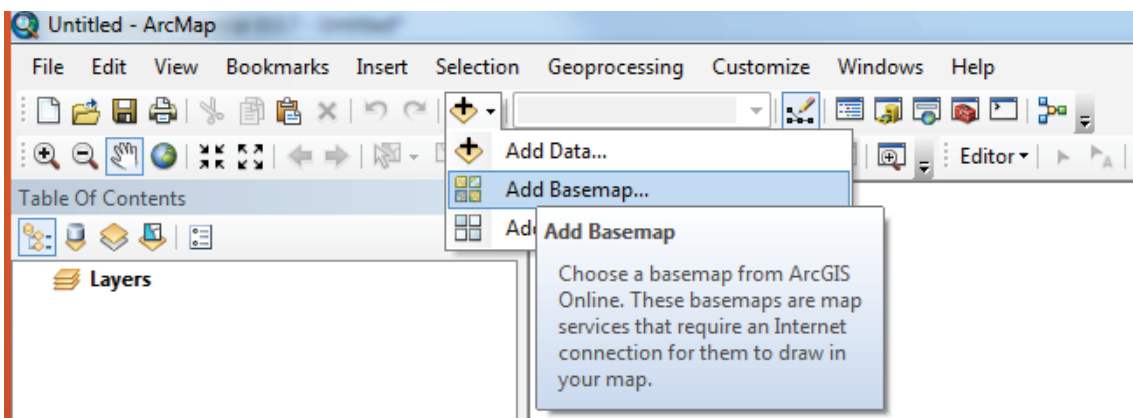
APPENDIX B

Example of Visualizing File Counts on Links of Interest Using ArcGIS

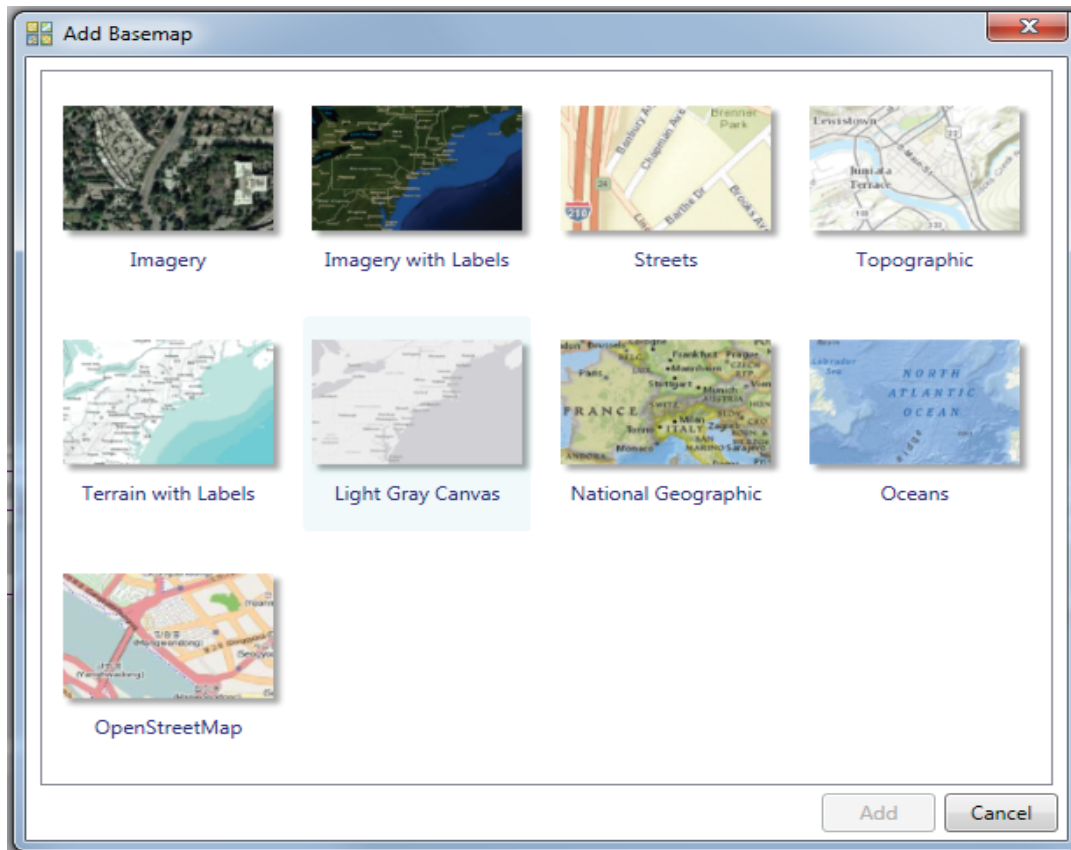
The following steps illustrate how to visualize the counts of traversals on links using ArcGIS.

Setting Up the Initial Map

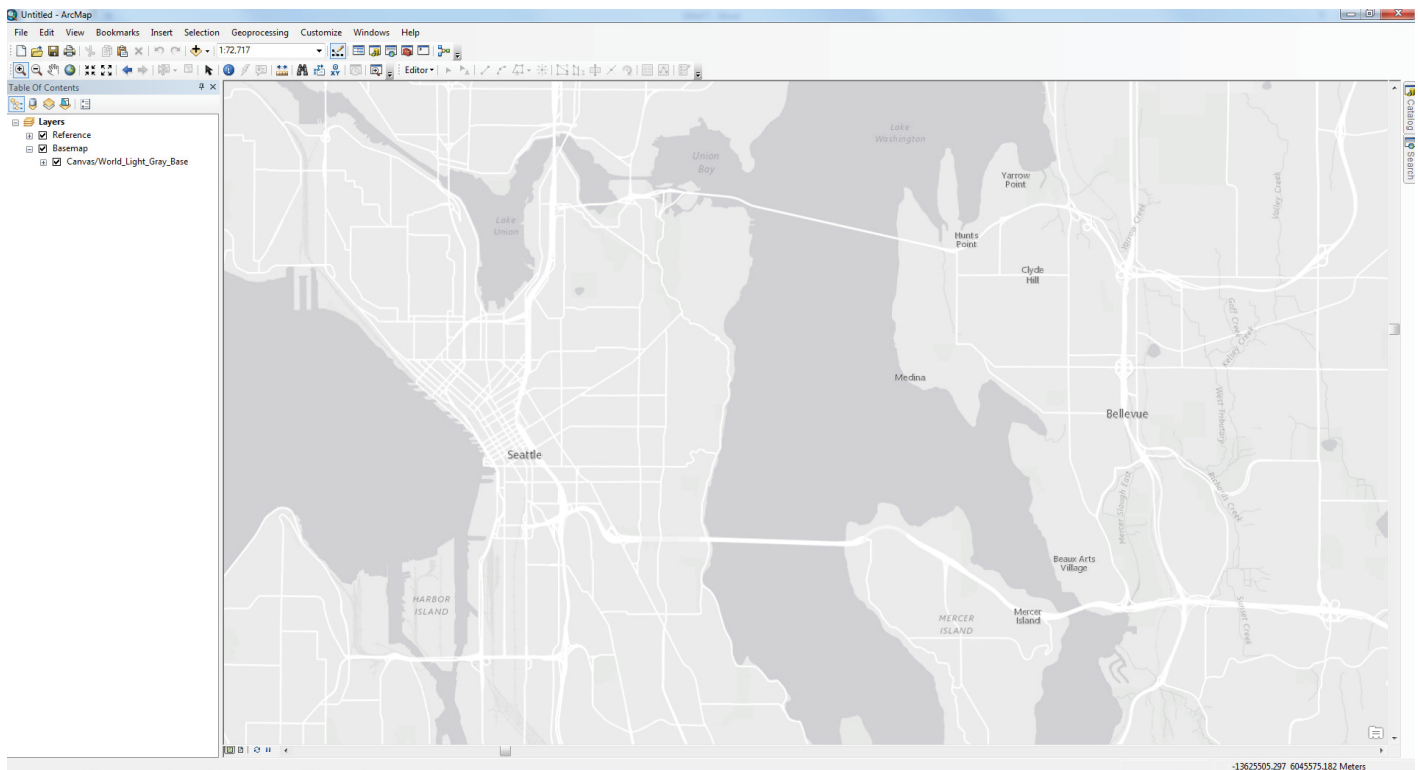
1. Adding a basemap.
 - a. From the icons along the top, select the Add Data icon and then select Add Basemap.



- b. Then select the basemap you would like to use. The examples used in this document use the Light Gray Canvas basemap.

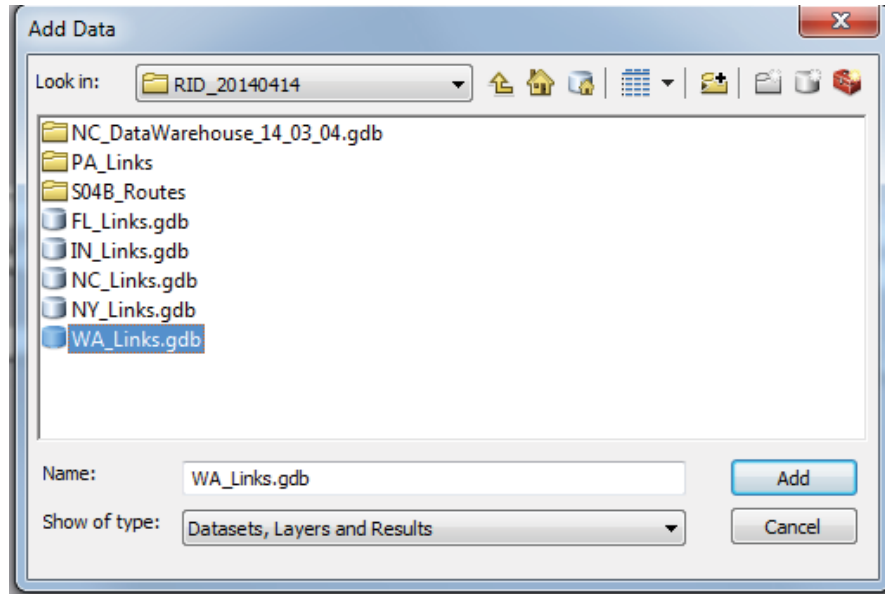


- c. Once you have selected the basemap, it should look like the following figure.

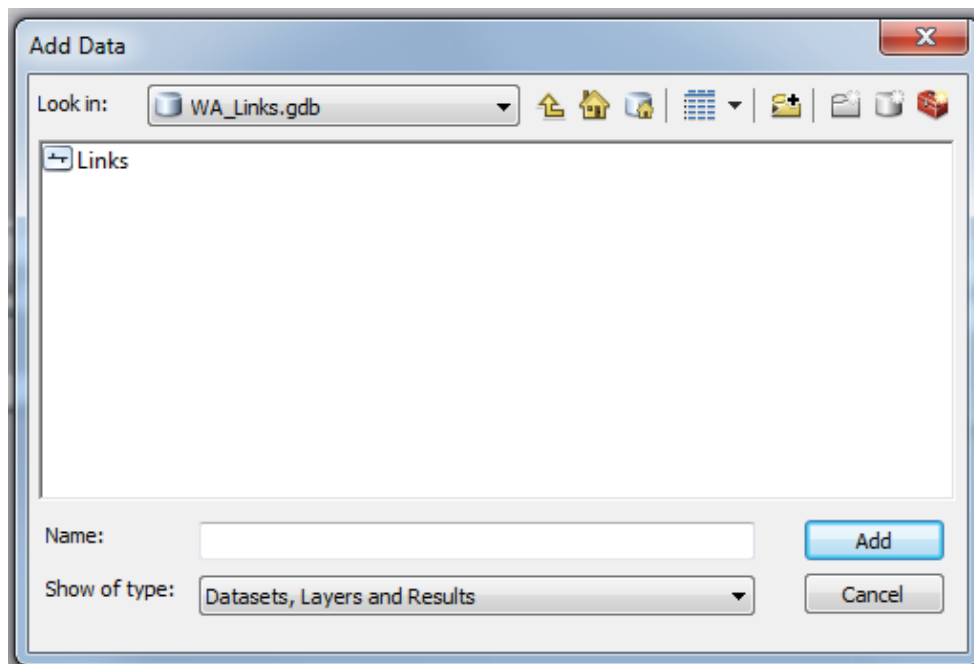


2. Adding links from the Roadway Information Database (RID) to the map.

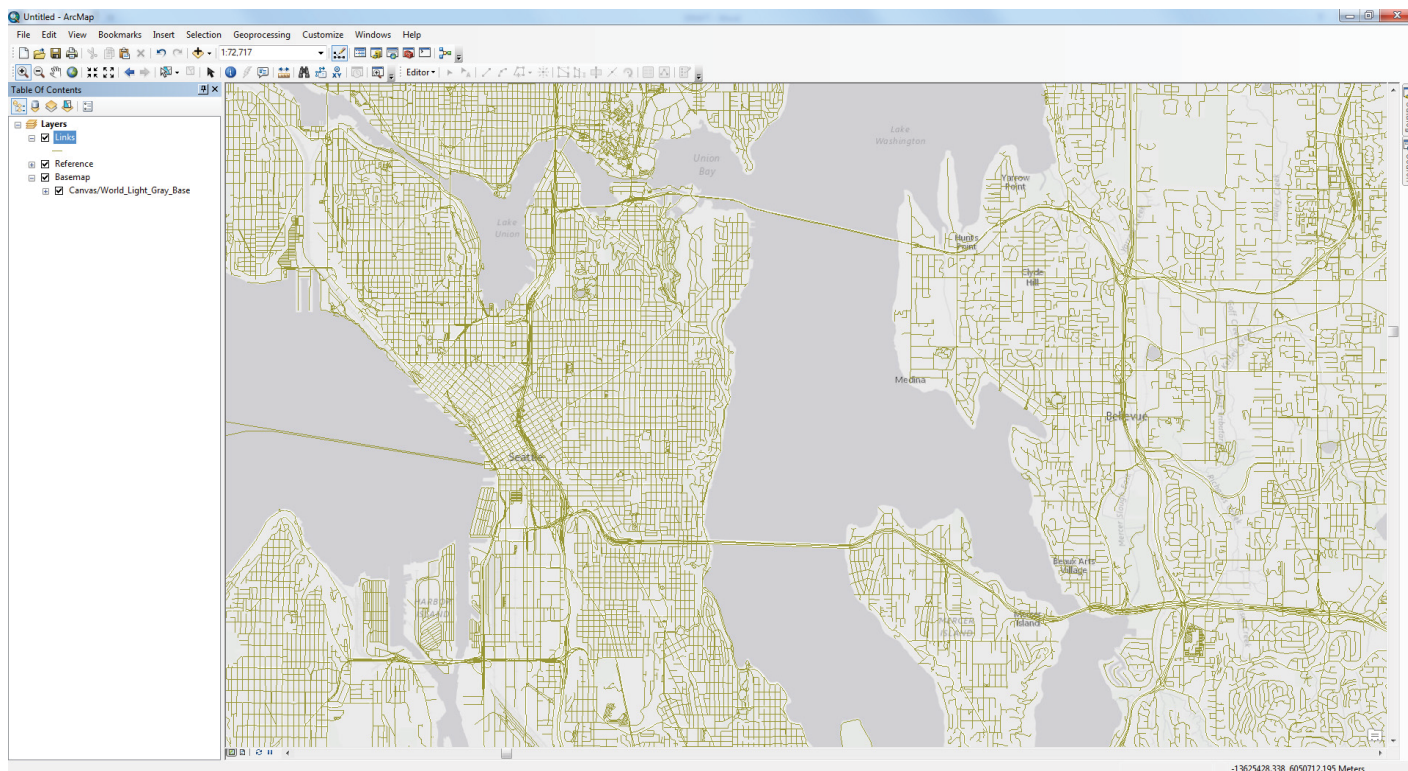
- a. The next step is to add the links from the RID to the basemap. Similar to how you added the basemap above, again use the Add Data icon, but this time select Add Data. Having selected Add Data, browse on your hard drive or network to where you have stored the RID. Once you have located it, select the links of interest. In the following example, links from the state of Washington are being added to the map from a geodatabase named WA_Links.gdb. This information may be slightly different depending on the final organization of the RID.



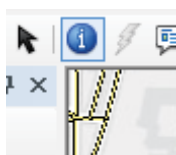
- b. Having selected WA_Links.gdb, the last step for adding links is to select the Links layer from within that geodatabase.



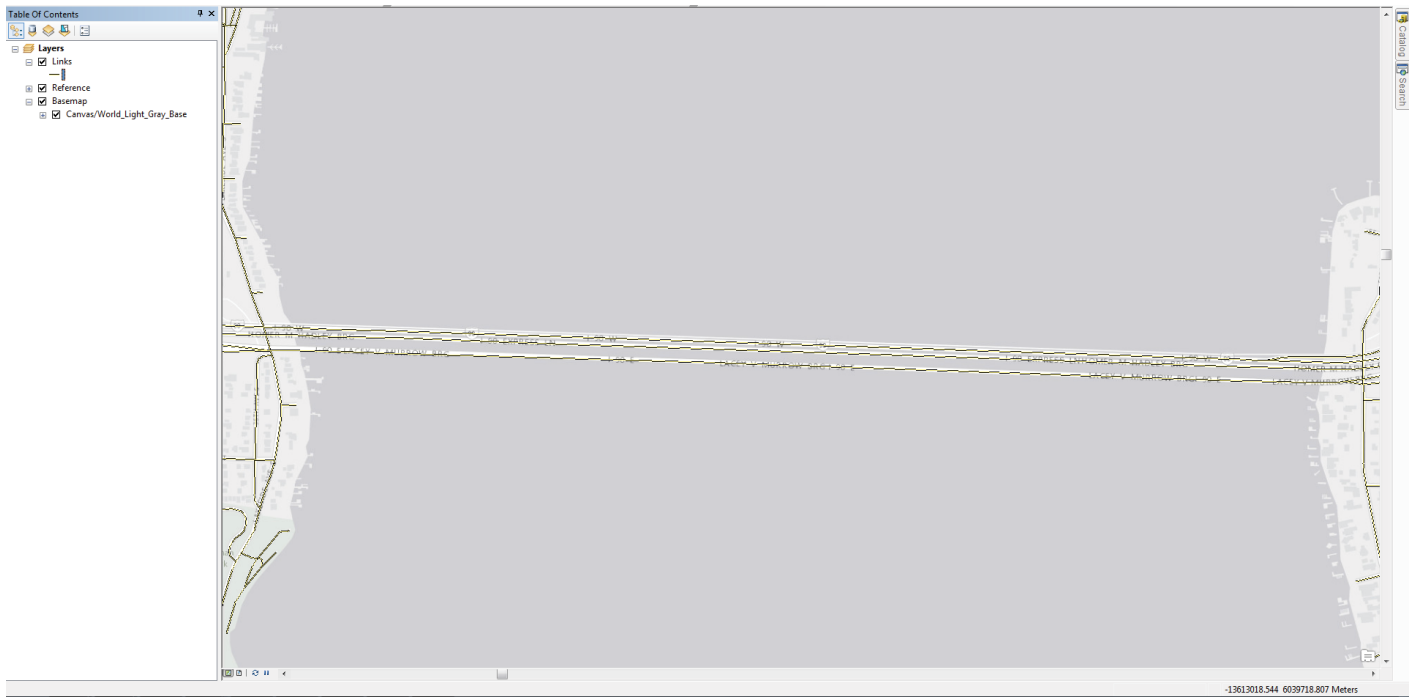
- c. Once this step is completed, you will see the links drawn on the map, similar to what is shown in the following figure. The digital map is shown in the map portion of the display, and a Table of Contents window to the left (in default setup) shows the Link layer you have just added.



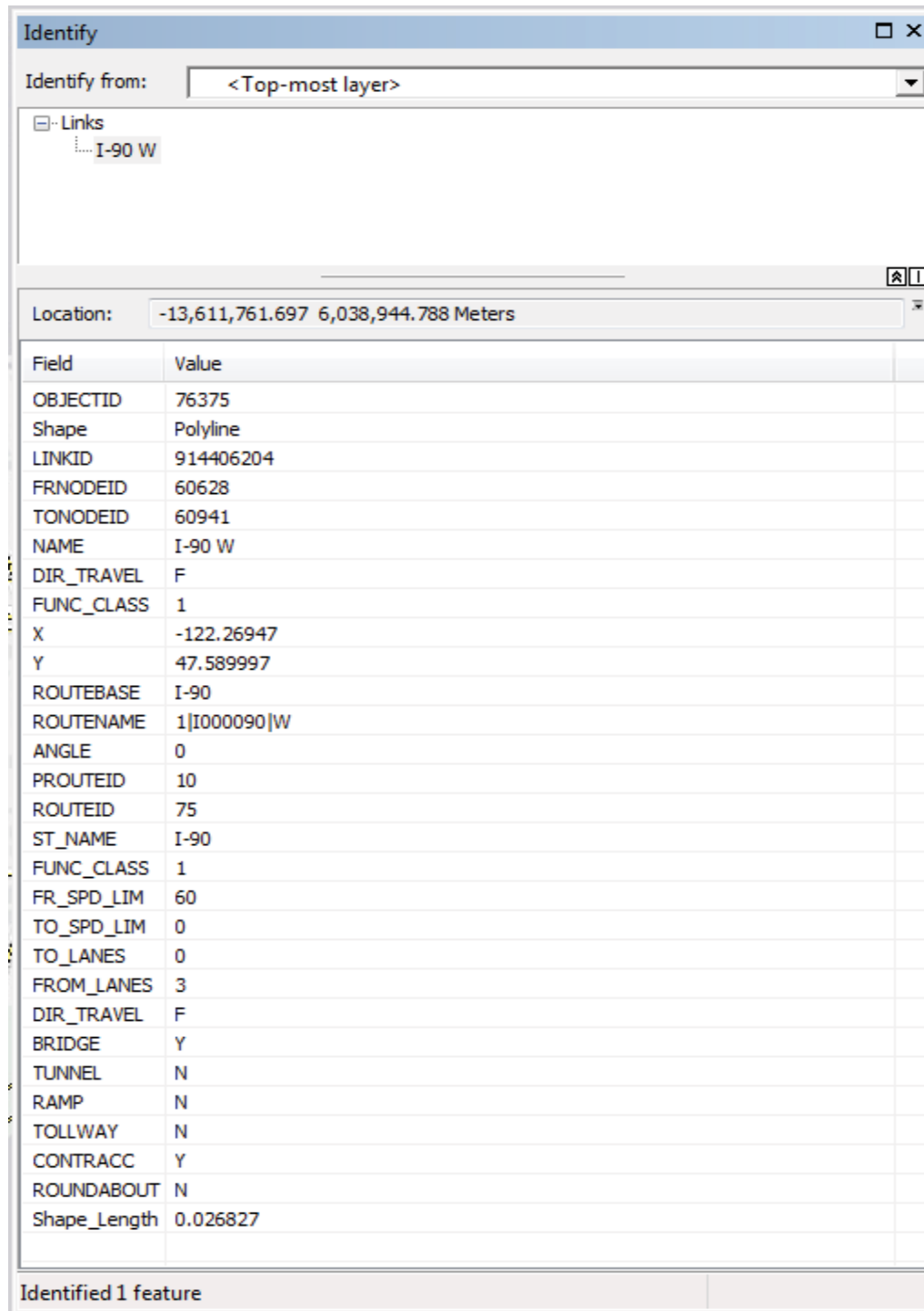
- d. If you want to identify the RID attributes associated with a link, you can use the Identify icon (blue circle with “i”) in the center, as shown in the following figure.



- e. Click on the Identify icon and then click on a link of interest. For example, if you click on one of the links shown in this figure ...



- f. ... then the following dialog box will open listing the attributes associated with the link. Note the LINKID near the top of the list.



Portraying Links of Interest on a Map with Counts

1. Exporting links and counts.

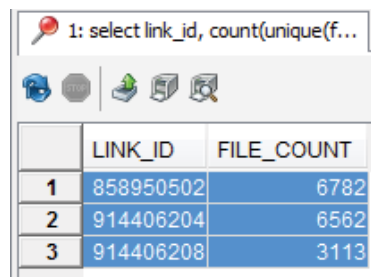
In this example, we will count the number of traversals for three links of interest and then portray the links on a map with link counts.

a. In your database client (e.g., DbVisualizer), run the following query:

```
select link_id, count(unique(file_id)) as file_count
from map_linking
where link_id in (914406208, 914406204, 858950502)
group by link_id
order by file_count desc
```

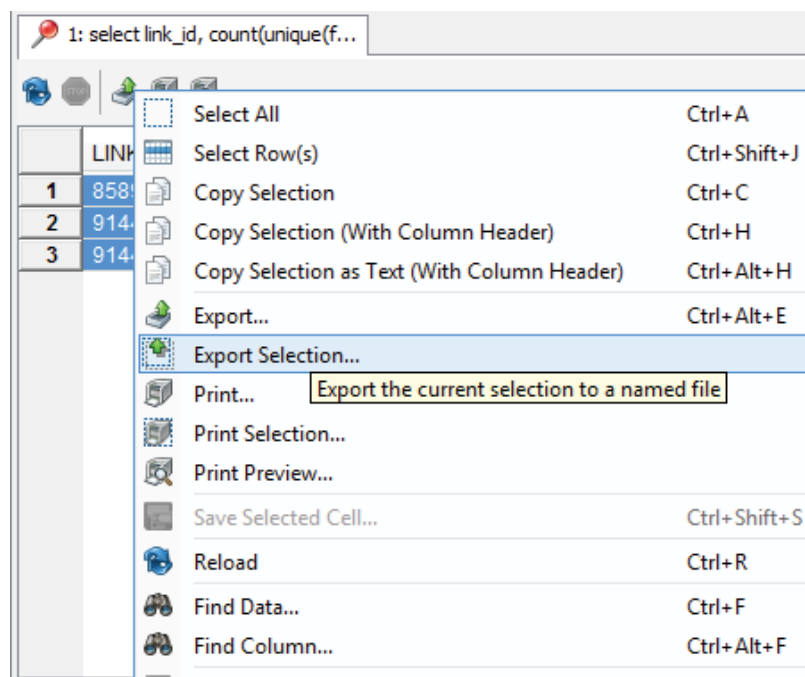
Note that this example uses a small number of link IDs for demonstration. Eliminating the “where link_id in (...)” would do this for all LINK_IDS.

b. The results will look something like this:

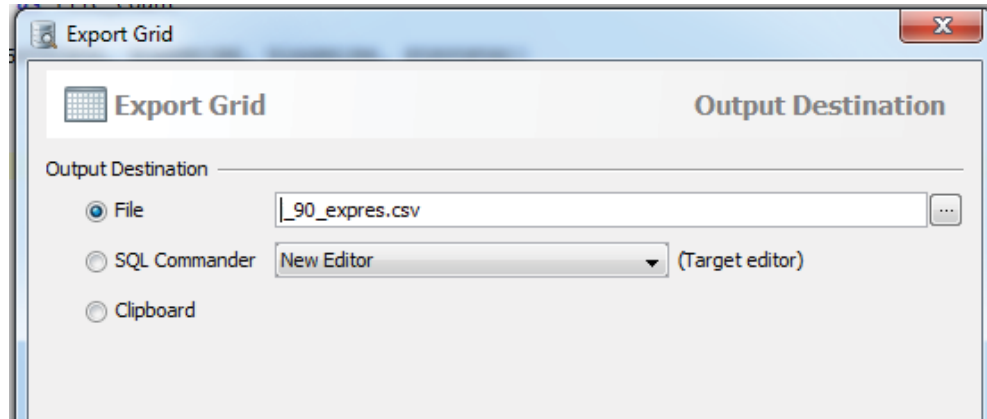


	LINK_ID	FILE_COUNT
1	858950502	6782
2	914406204	6562
3	914406208	3113

c. Next, export the LINK_IDS and counts to a CSV file. The following figure is from DbVisualizer. Select the values of interest and then right-click on the selected area to access the menu shown.



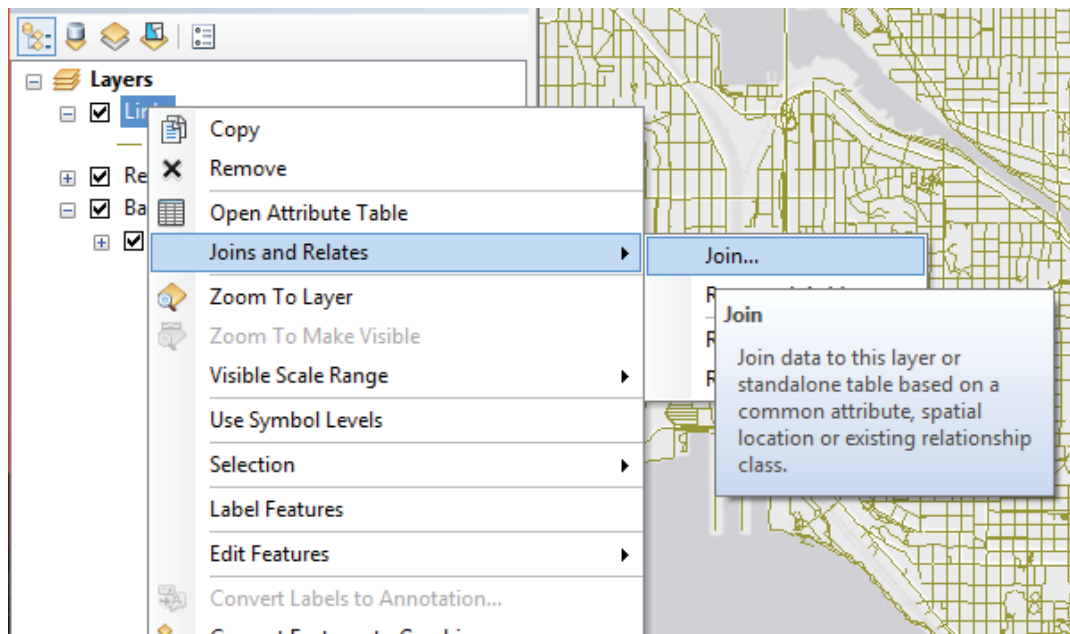
- d. Next, provide a name for the export file.



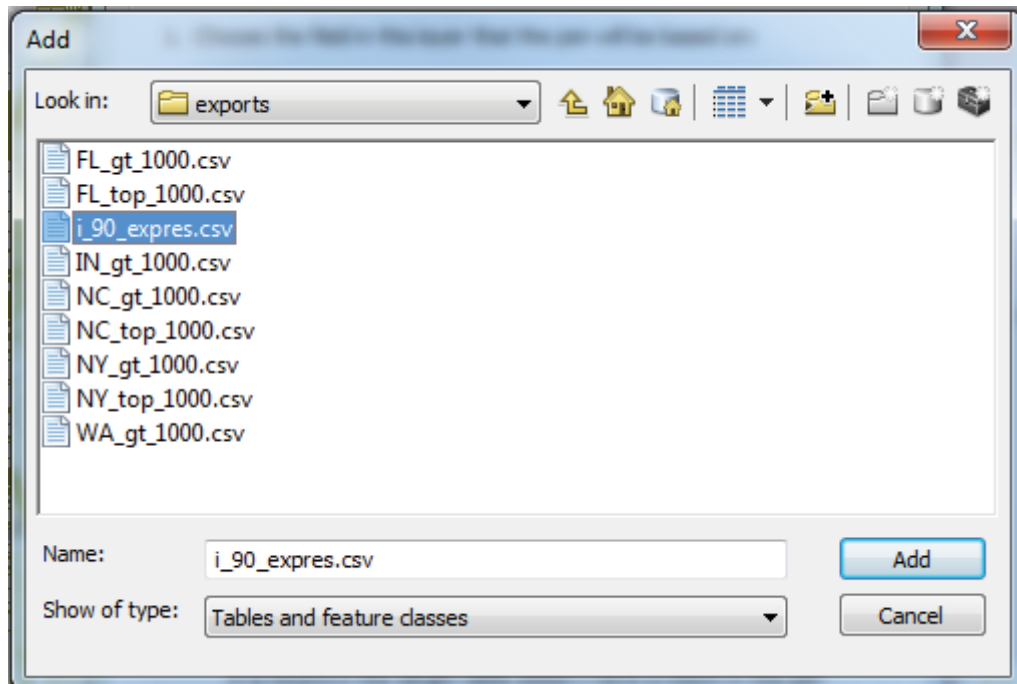
2. Joining the exported data to the RID map.

Before starting this task, you will need the link layer from the RID that includes the LINK_IDs that are in your export. (For example, the Washington State link layer will be needed if the links are in Washington State). See Step 2 from the previous demonstration, Setting Up the Initial Map. Once you have the link layer available, you can join exported LINK_ID counts with the links in the digital map as follows:

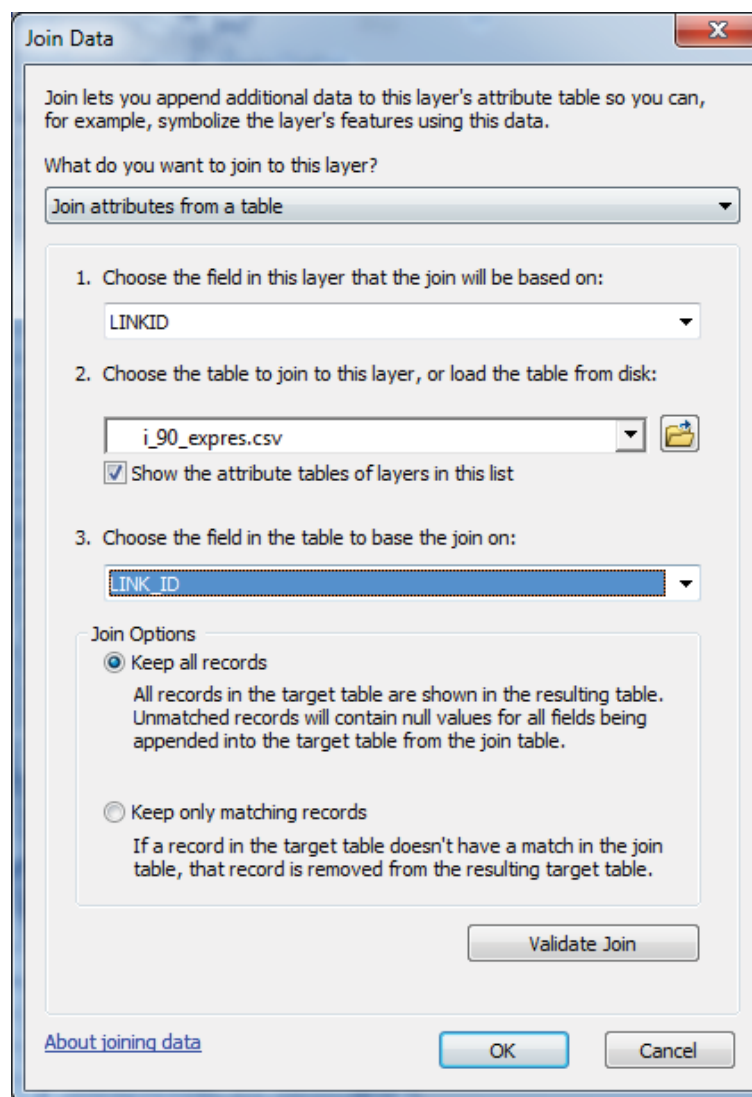
- a. In the Table of Contents window, to the left of the display in the default arrangement, right-click on the appropriate Link layer. You will see a menu like the one shown here. Select Joins and Relates and then Join.



- b. Then select the CSV file you have exported. In this example, a file named `i_90_express.csv` is selected from the exports folder. Click Add when selected.



- c. Next, you must make a series of selections indicating what fields in the digital map (RID) are going to be used to relate to the values in the CSV. The LINK_ID is the common field. As shown in the following dialog box, in Dropdown 1, the LINKID field is selected in the layer. For Dropdown 2, if it does not already show it, browse to where you have stored your CSV, and select it. In Dropdown 3, select LINK_ID, which is the common field from the CSV. Next, click OK.

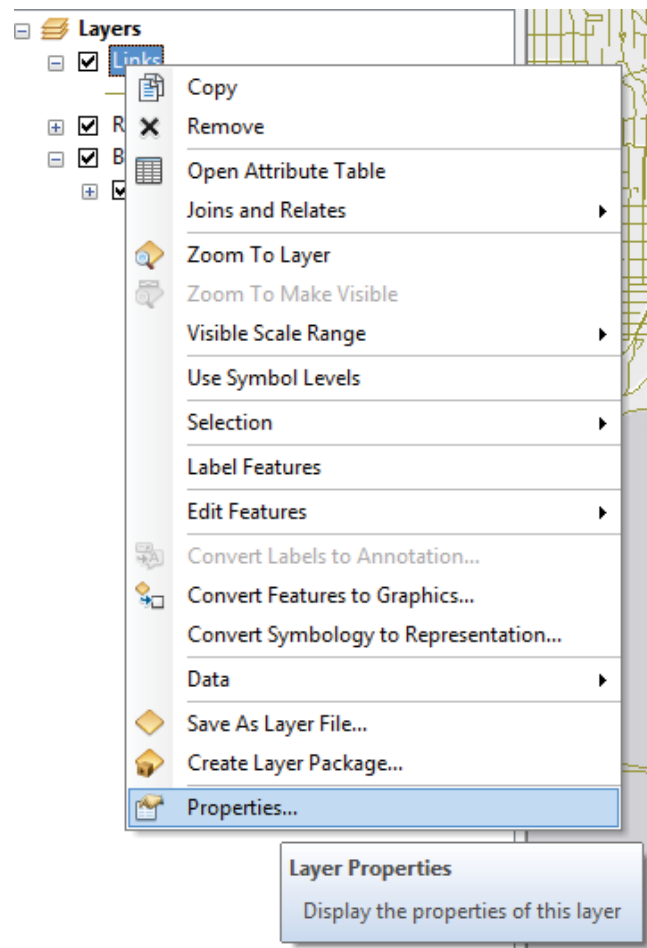


- d. After the join is complete, you will be returned to the map screen.

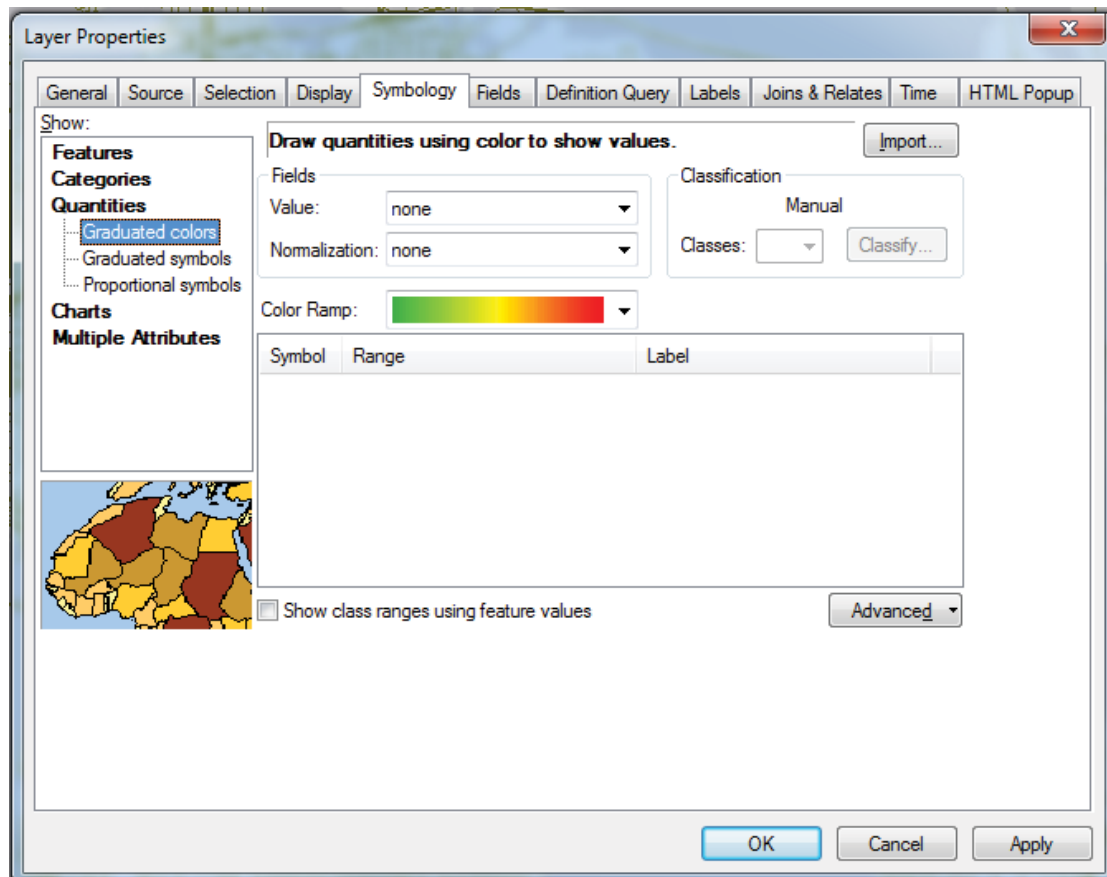
3. Setting the color coding.

It is often useful to color code the links according to counts—of traversals, for example. Having joined the link counts to the link layer in the previous step, the link count is now an attribute of the links.

- a. To color code the links according to the count on the link, first right-click on the link layer and select the Properties option, as shown in the following figure.

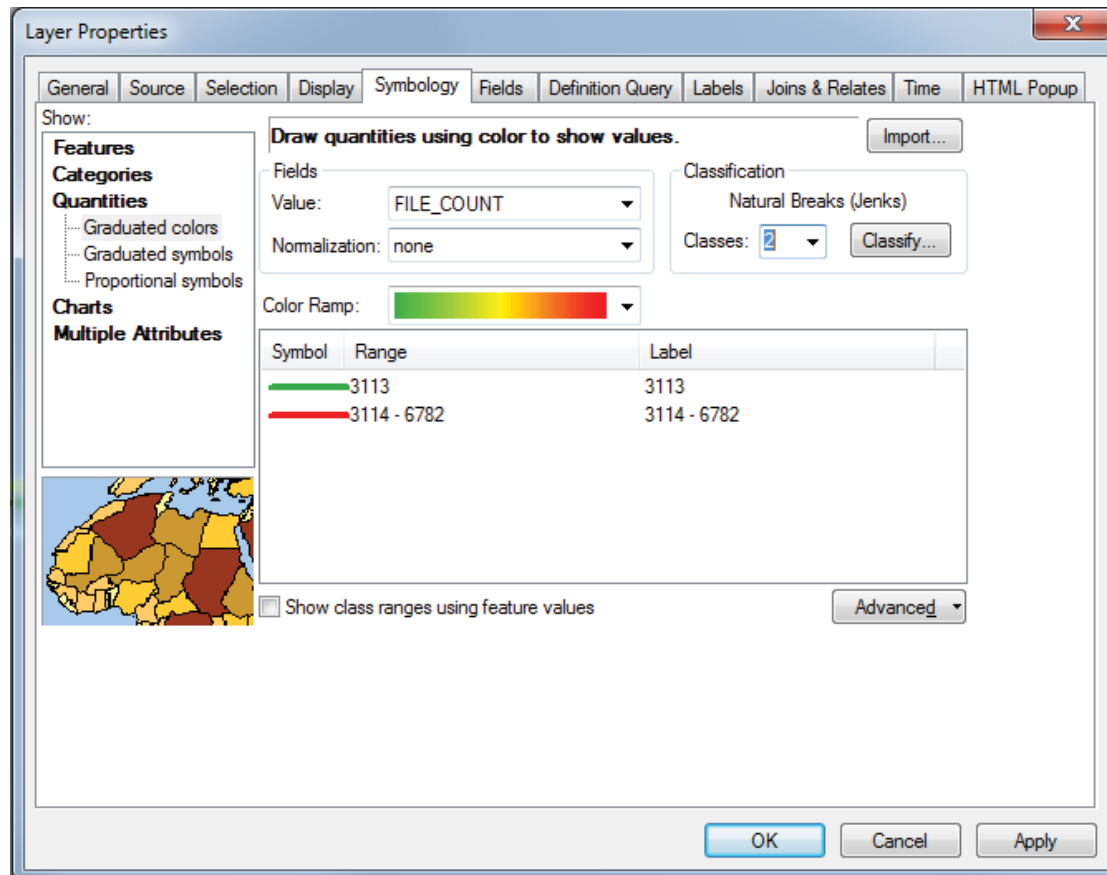


- b. The Layer Properties dialog box will open.



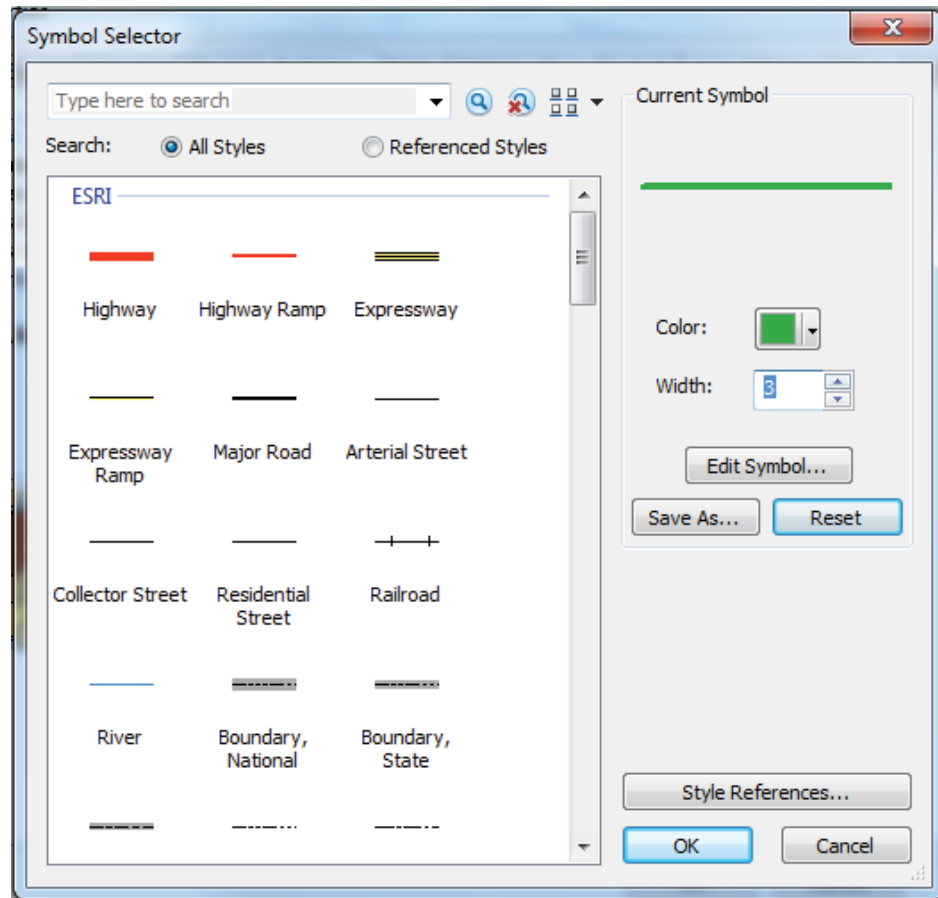
- c. Next select the Symbology tab.
d. Next select the Quantities, Graduated colors from the left side (as shown in the previous figure).
e. Select the dropdown list in the Value field.

- f. Select FILE_COUNT.



- g. Set Classes to 2, indicating the software will use two color classes. If you are doing this with a larger list, you can go with the default number of classes.

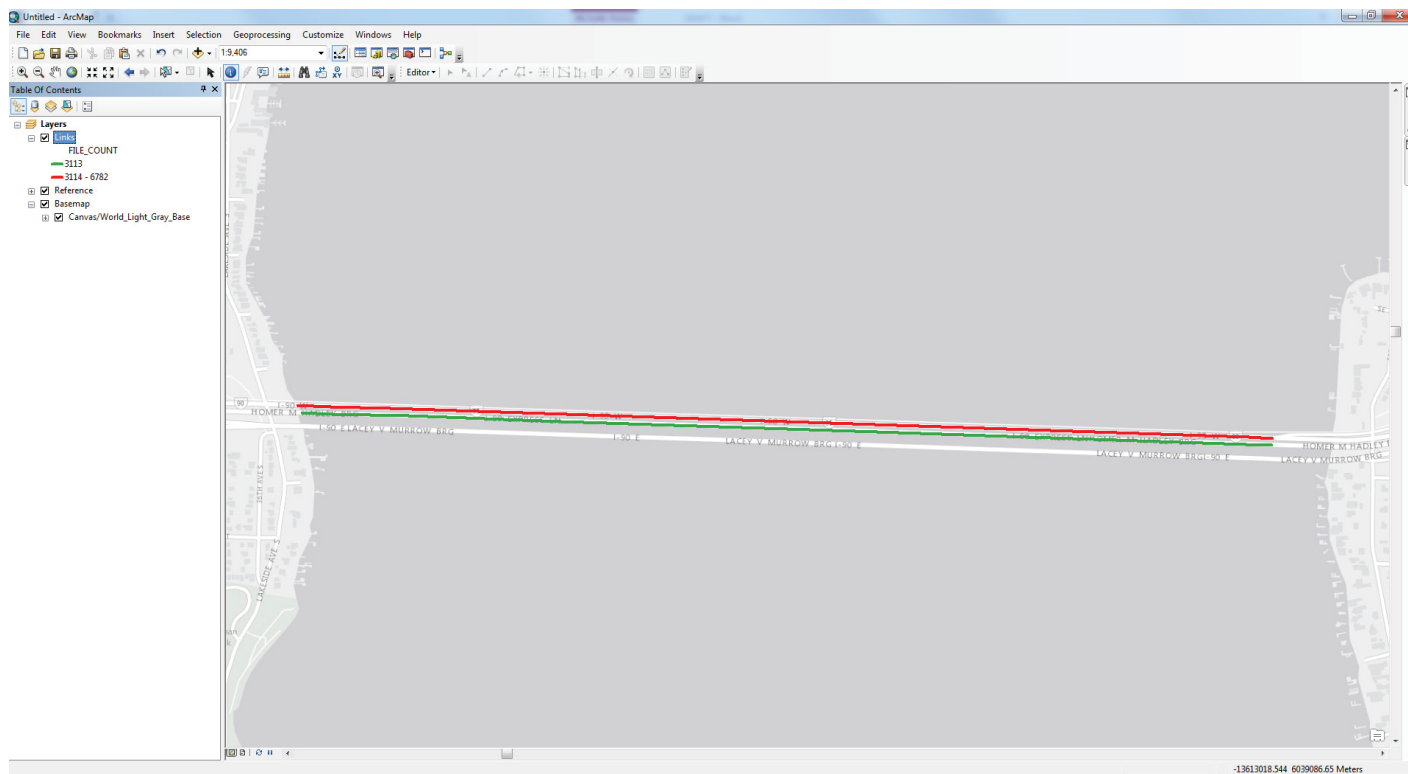
- h. Double-click on the two color codes and increase them to 3 pt, as shown in the following dialog box.



- i. Click OK.

4. Viewing the result.

At this time, three color-coded links should appear, demonstrated in the following figure.



TRB OVERSIGHT COMMITTEE FOR THE STRATEGIC HIGHWAY RESEARCH PROGRAM 2*

CHAIR: Kirk T. Steudle, *Director, Michigan Department of Transportation*

MEMBERS

H. Norman Abramson, *Executive Vice President (retired), Southwest Research Institute*
Alan C. Clark, *MPO Director, Houston–Galveston Area Council*
Frank L. Danchetz, *Vice President, ARCADIS-US, Inc. (deceased January 2015)*
Malcolm Dougherty, *Director, California Department of Transportation*
Stanley Gee, *Executive Deputy Commissioner, New York State Department of Transportation*
Mary L. Klein, *President and CEO, NatureServe*
Michael P. Lewis, *Director, Rhode Island Department of Transportation*
John R. Njord, *Executive Director (retired), Utah Department of Transportation*
Charles F. Potts, *Chief Executive Officer, Heritage Construction and Materials*
Ananth K. Prasad, *Secretary, Florida Department of Transportation*
Gerald M. Ross, *Chief Engineer (retired), Georgia Department of Transportation*
George E. Schoener, *Executive Director, I-95 Corridor Coalition*
Kumares C. Sinha, *Olson Distinguished Professor of Civil Engineering, Purdue University*
Paul Trombino III, *Director, Iowa Department of Transportation*

EX OFFICIO MEMBERS

Victor M. Mendez, *Administrator, Federal Highway Administration*
David L. Strickland, *Administrator, National Highway Transportation Safety Administration*
Frederick “Bud” Wright, *Executive Director, American Association of State Highway and Transportation Officials*

LIAISONS

Ken Jacoby, *Communications and Outreach Team Director, Office of Corporate Research, Technology, and Innovation Management, Federal Highway Administration*
Tony Kane, *Director, Engineering and Technical Services, American Association of State Highway and Transportation Officials*
Jeffrey F. Paniati, *Executive Director, Federal Highway Administration*
John Pearson, *Program Director, Council of Deputy Ministers Responsible for Transportation and Highway Safety, Canada*
Michael F. Trentacoste, *Associate Administrator, Research, Development, and Technology, Federal Highway Administration*

*Membership as of January 2015.

SAFETY TECHNICAL COORDINATING COMMITTEE*

CHAIR: Forrest M. Council, *Senior Research Scientist, Highway Safety Research Center, University of North Carolina*

MEMBERS

Timothy E. Barnett, *State Safety Engineer, Alabama Department of Transportation*
James A. Bonneson, *Senior Principal Engineer, Kittelson and Associates*
Leanna Depue, *Director, Traffic and Highway Safety Division, Missouri Department of Transportation*
Bradley M. Estothen, *State Traffic Safety Engineer, Minnesota Department of Transportation*
Jurek Grabowski, *Research Director, AAA Foundation for Traffic Safety*
Jeffrey Greenberg, *Senior Technical Leader, Ford Motor Company*
Joanne Harbluk, *Human Factors Specialist, Transport Canada*
Brent Jennings, *Highway Safety Manager, Idaho Transportation Department*
Alan F. Karr, *Director, National Institute of Statistical Sciences*
Bani K. Mallick, *Distinguished Professor, Department of Statistics, Texas A&M University*
John C. Milton, *Director, Enterprise Risk & Safety Management, Washington State Department of Transportation*
Harlan J. Onsrud, *Professor, School of Computing & Information Science*
Michael Perel, *Safety Knowledge Engineer*
Charles W. Reider, *Chief Safety Engineer, Nevada Department of Transportation*
David Shinar, *Professor, Department of Industrial Engineering and Management, Ben Gurion University of the Negev*
Alison Smiley, *President, Human Factors North, Inc.*
Thomas M. Welch, *State Transportation Safety Engineer (retired), Office of Traffic and Safety, Iowa Department of Transportation*

AASHTO LIAISONS

Kelly Hardy, *Safety Program Manager, American Association of State Highway and Transportation Officials*
Pam Hutton, *SHRP 2 Implementation Manager, American Association of State Highway and Transportation Officials*
Jim McDonnell, *Program Director for Engineering, American Association of State Highway and Transportation Officials*

FHWA LIAISONS

Monique Evans, *Director, Office of Safety Technologies, Federal Highway Administration*
Michael Griffith, *Director, Office of Safety Integration, Federal Highway Administration*

AUTO INDUSTRY LIAISONS

Michael Cammisa, *Director, Safety, Association of Global Automakers*
Scott Schmidt, *Director, Safety and Regulatory Affairs, Alliance of Automobile Manufacturers*

EUROPEAN SAFETY LIAISON

Fred Wegman, *Managing Director, SWOV Institute for Road Safety Research, Netherlands*

FMCSA LIAISON

Martin Walker, *Chief, Research Division, Federal Motor Carrier Safety Administration*

NHTSA LIAISONS

Richard Compton, *Director, Office of Behavioral Safety Research, National Highway Traffic Safety Administration*
Tim Johnson, *Director, Office of Human-Vehicle Performance Research, National Highway Traffic Safety Administration*
Seymour Stern, *Team Leader, State Based Systems, National Highway Traffic Safety Administration*

*Membership as of July 2014.

Related SHRP 2 Research

Naturalistic Driving Study: Development of the Roadway Information Database (S04A)

Design of the In-Vehicle Driving Behavior and Crash Risk Study (S05)

Naturalistic Driving Study: Technical Coordination and Quality Control (S06)

Naturalistic Driving Study: Collecting Data on Cell Phone Use (S06)

Naturalistic Driving Study: Field Data Collection (S07)

Analysis of Naturalistic Driving Study Data: Safer Glances, Driver Inattention, and Crash Risk (S08A)

Analysis of Naturalistic Driving Study Data: Offset Left-Turn Lanes (S08B)

Analysis of Naturalistic Driving Study Data: Roadway Departures on Rural Two-Lane Curves (S08D)

Naturalistic Driving Study: Descriptive Comparison of the Study Sample with National Data (S31)

Naturalistic Driving Study: Alcohol Sensor Performance (S31)