09-UT-005

The National Surface Transportation Safety Center for Excellence

Method for Identifying Rural, Urban, and Interstate Driving in Naturalistic Driving Data

Final Report

Brad R. Cannon ● Shane B. McLaughlin ● Jonathan M. Hankey

Submitted: March 12, 2009

Lighting | Technology
Fatigue | Aging

Housed at the Virginia Tech Transportation Institute
3500 Transportation Research Plaza • Blacksburg, Virginia 24061

ACKNOWLEDGMENTS

The authors of this report would like to acknowledge the support of the stakeholders of the National Surface Transportation Safety Center for Excellence (NSTSCE): Tom Dingus from the Virginia Tech Transportation Institute, Richard Deering from General Motors Corporation, Carl Andersen from the Federal Highway Administration (FHWA), and Gary Allen from the Virginia Department of Transportation and the Virginia Transportation Research Council.

The NSTSCE stakeholders have jointly funded this research for the purpose of developing and disseminating advanced transportation safety techniques and innovations.

EXECUTIVE SUMMARY

Most fatal crashes in the United States occur in rural areas despite the fact that only about a third of the distance driven each year is on rural roads (computed from Burgess, 2005).[1] The emergence of naturalistic driving data provides a new opportunity for exploring the rural crash problem. In contrast with crash report-based data sources, naturalistic data provide extremely rich data, from many video views and data streams, describing what occurs instant by instant both in non-crash situations and during actual crashes and near-crashes. Within these data, epochs of driving in rural areas can be located and analyzed further to identify the characteristics of driving in rural areas, to identify what factors contribute to these crashes, to identify countermeasures that will reduce the frequency and severity of these crashes, and in some cases even test the effectiveness of proposed countermeasures.

Naturalistic driving datasets are expected to be large. To prepare for investigations into the rural crash problem using this type and quantity of data, an automated method was developed to determine when participants were driving on rural roads.

A review of previous work on rural road driving was undertaken to find a standard definition of a rural road. While there are some variances in the definitions of a rural road, a definition was selected which is in harmony with the determinations most common in the literature on rural roads.

By employing the functionality of Geographic Information Systems (GIS), code was written which allows for an automated process to compare the Global Positioning System (GPS) data recorded in the naturalistic driving data with geographic map data from the U.S. Census Bureau and road data from various sources, such as state departments of transportation (e.g., Virginia Department of Transportation) or other providers. Points recorded in the naturalistic driving data which fall outside the boundaries of the Census Bureau's urbanized Areas or urban Clusters are determined to be rural. The points are further evaluated to determine whether or not the vehicle was being driven on an interstate highway. Points that are determined to be rural and not on interstate highways are segments of interest in addressing the rural road crash problem.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

AADT	Average Annual Daily Traffic
ESRI	Environmental Systems Research Institute, Inc.
FARS	Fatality Analysis Reporting System
FHWA	Federal Highway Administration
GES	General Estimates System
GIS	Geographic Information System
GPS	Global Positioning System
SAS	Statistical Analysis Software
SQL	Structured Query Language
UA	Urbanized Area
UC	Urban Cluster
VDOT	Virginia Department of Transportation
VMT	Vehicle Miles Traveled

CHAPTER 1. INTRODUCTION

While the ratio of fatal crashes to miles driven in the United States has generally been on a decline since 1994, the number of fatal crashes typically increases each year and remains tragically high. In the United States there were 374,691 fatal crashes during the years 1994 to 2003. Even though only 31 percent of the total miles driven during that time period were on rural roads, 52 percent of the fatal crashes were on rural roads (Burgess, 2005[1]; note: totals recalculated to not include rural interstates with rural roads).

Naturalistic driving data, such as collected in the 100-Car Naturalistic Driving Study (Dingus et al., 2006)[2], provide a new opportunity for exploring the rural crash problem. These data include numeric measures such as vehicle speed, range to other vehicles, geographic location, lateral acceleration, longitudinal acceleration, and yaw, as well as video of the driver and surrounding environment. In contrast with crash report-based data sources, such as the Fatality Analysis Reporting System (FARS) or the General Estimates System (GES), naturalistic data such as collected by Dingus et al. describe what is occurring over time, both in non-crash situations and during actual crashes and near-crashes. Accumulation of naturalistic data continues. Within this type of data, epochs of driving in rural areas can be located and analyzed further to identify how people drive in rural areas, to identify what factors may contribute to crashes, and to perform detailed analyses of what occurred during these crashes.

Naturalistic datasets are expected to be large. The participants in the 100-Car Naturalistic Driving Study drove approximately two million vehicle miles based on odometer readings. Driving time for the study was almost 43,000 hours. Data were collected approximately 10 times per second during each vehicle trip. Current and planned future studies will be even larger. To prepare for investigations into the rural crash problem using this type and quantity of data, an automated method was developed to determine when participants were driving on rural roads. The 100-Car Naturalistic data were used in the development of this method.

CHAPTER 2. PRELIMINARY WORK

REVIEW OF DEFINITION OF RURAL ROAD USED IN PREVIOUS WORK

A review of the literature on rural roads or driving on rural roads found a variety of definitions of a rural road. Often, the qualifier "two-lane" is added to create the term "two-lane rural road" (Ivan et al., 1999; Persaud et al., 2004; Kim et al., 2006; Kim et al., 2007).[3,4,5,6] The two-lane qualifier helps further identify which roads were studied based on an important characteristic of many rural roads. Using this qualifier, limited-access highways (e.g., interstates) in rural areas are excluded from their studies.

O'Neill et al. (1993)[7] wrote extensively about the consideration of selecting a definition for "rural road." They specifically note three commonly used methods. One method is to define a rural road as one that has an average annual daily traffic (AADT) level below a certain amount. They point out that this can be problematic as many roads in rural areas have high AADT levels because they provide access to popular rural recreation areas. Another method they mention is to base the determination on the types of trips made on the road. Because trip type is not recorded in naturalistic data, this method does not lend itself directly to automated processing. In their own work, O'Neill et al. defined a rural road as one outside an urbanized area of 50,000 people or more.

The definition of "rural road" used by O'Neill et al. (1993) is similar to that found in law for Federal-Aid Highways, which uses the Census Bureau's classification of an "urbanized area" and "urban cluster" to classify highways as rural or urban (Title 23, United States Code, §101).

The primary characteristic of an urbanized area (UA) is that it contains 50,000 or more people within "densely settled territory." An urban cluster (UC) is generally smaller, containing at least 2,500 people but fewer than 50,000 people, also within "densely settled territory" (*Federal Register*. Vol. 67, No. 51).

The rural road definition used by a researcher could also adhere to the classification given the road by the government responsible for the road. However, each jurisdiction may use different criteria in making the determination, creating a need for a data collection effort across many government entities, and making it difficult to interpret results from multiple jurisdictions.

SELECTION OF RURAL ROAD DEFINITION

A definition similar to the one used in federal law has been selected for use here. For the purposes of this report, a rural road is any road that is outside of a UA or UC and is not part of the Interstate Highway System. Generally, this includes non-interstate roads outside of areas with 2,500 or more people in "densely settled territory." Using this definition provides consistent criteria for identifying rural roads. In addition to being similar to the definition used by the Federal Government, by excluding interstate highways it is also in harmony with the determinations most common in the literature on rural roads.

Because the definition of a rural road depends on two binary characteristics (i.e., urban versus rural and interstate versus non-interstate), an additional three road types result from the method employed here, creating a total of four road types. In addition to driving on rural roads, the process also finds rural interstate driving (on interstate highways outside of a UA or UC), urban interstate driving (on interstate highways within a UA or UC), and driving on urban surface streets (within a UA or UC but not on interstate highways).

TECHNOLOGY SELECTION

A review of available technology was performed to determine the most efficient method for automating the process of extracting rural driving from the 100-Car dataset using this definition. This review led to a strategy dependent upon geographic information systems (GIS). GIS encompass a number of computer data management systems that specialize in geographic data. A GIS can be used to store, graphically represent, or analyze geographic data.

CHAPTER 3. METHOD

OVERVIEW

Geographic data from various sources are overlaid on pre-processed trip global positioning system (GPS) data to determine whether the vehicle is in a rural or urban area, and also whether the vehicle is on or off an interstate highway. Urban areas are represented in the maps by polygons. Any GPS points that fall within these areas are considered to be urban points. Conversely, any points that are not within these urban boundaries are considered rural. Roads, however, are represented by lines without width. To determine if a vehicle is on a road, the road segment is enlarged to create a polygon with width around the segment. The trip GPS points can then be evaluated to determine if they fall within a polygon representing an interstate highway. A flowchart of the overall process and logic used is shown in figure 1. GPS points locating crashes and near-crashes from the 100-Car data are also processed using the same method in order to apply the same classifications to each of the events.

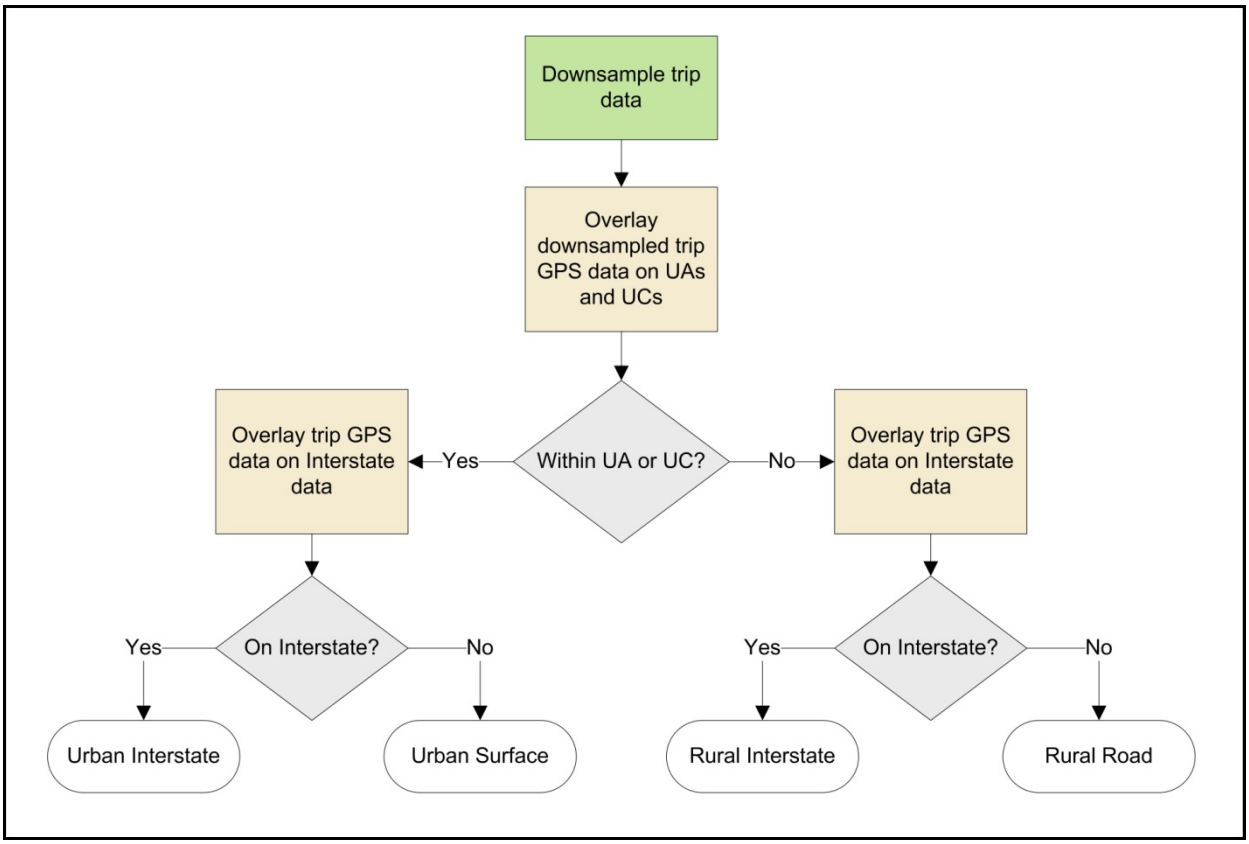

Figure 1. Diagram. Overall method/logic schematic.

MAP DATA PREPARATION

Geographic data describing the boundaries of UAs and UCs were obtained from the Census Bureau in the shapefile format, which stores geographic data for use with commonly available GIS. The UAs and UCs for the 48 contiguous states are shown in figure 2. A close-up view of

the Washington UA and nearby UAs and UCs is shown in figure 3. To improve the efficiency of the code, some manipulation of the file format was performed using ArcGIS® before it was imported into MATLAB®. After some final preparations performed in MATLAB, the result was then saved in MATLAB's native format for quick retrieval by the code used to find rural driving within the 100-Car dataset.

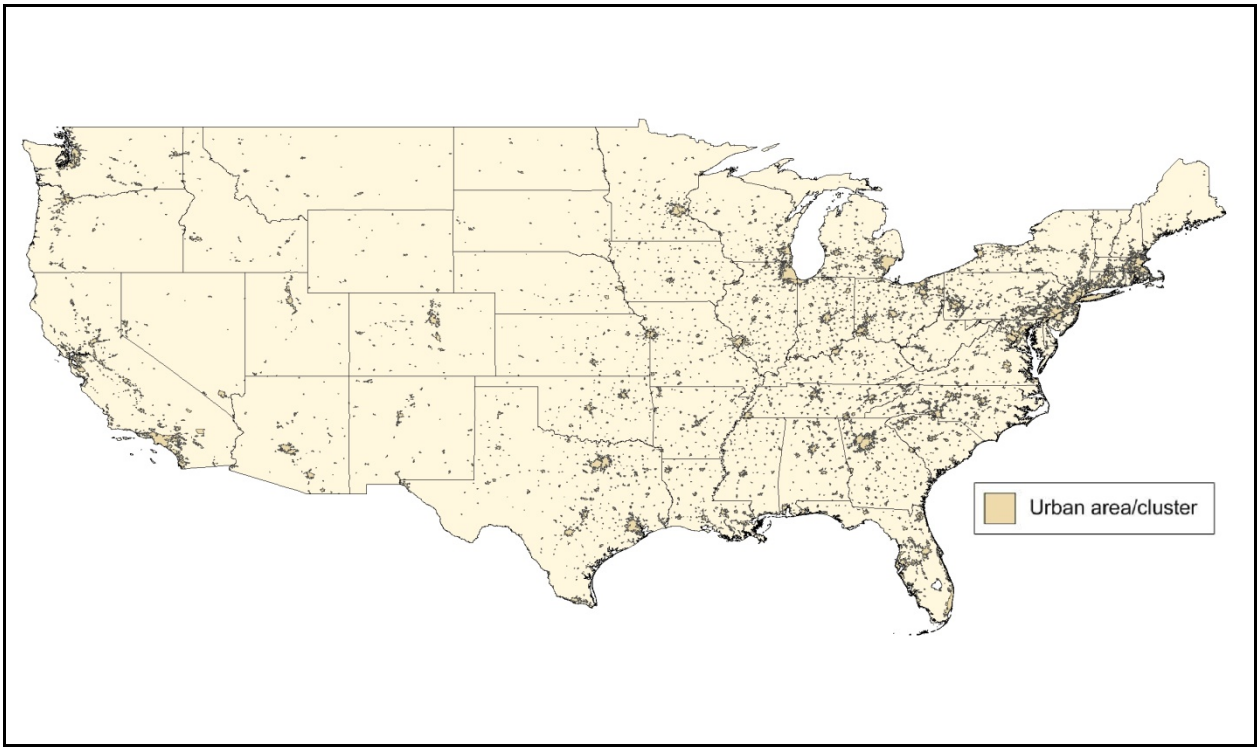

Figure 2. Diagram. Urbanized areas and urban clusters in the 48 contiguous states.

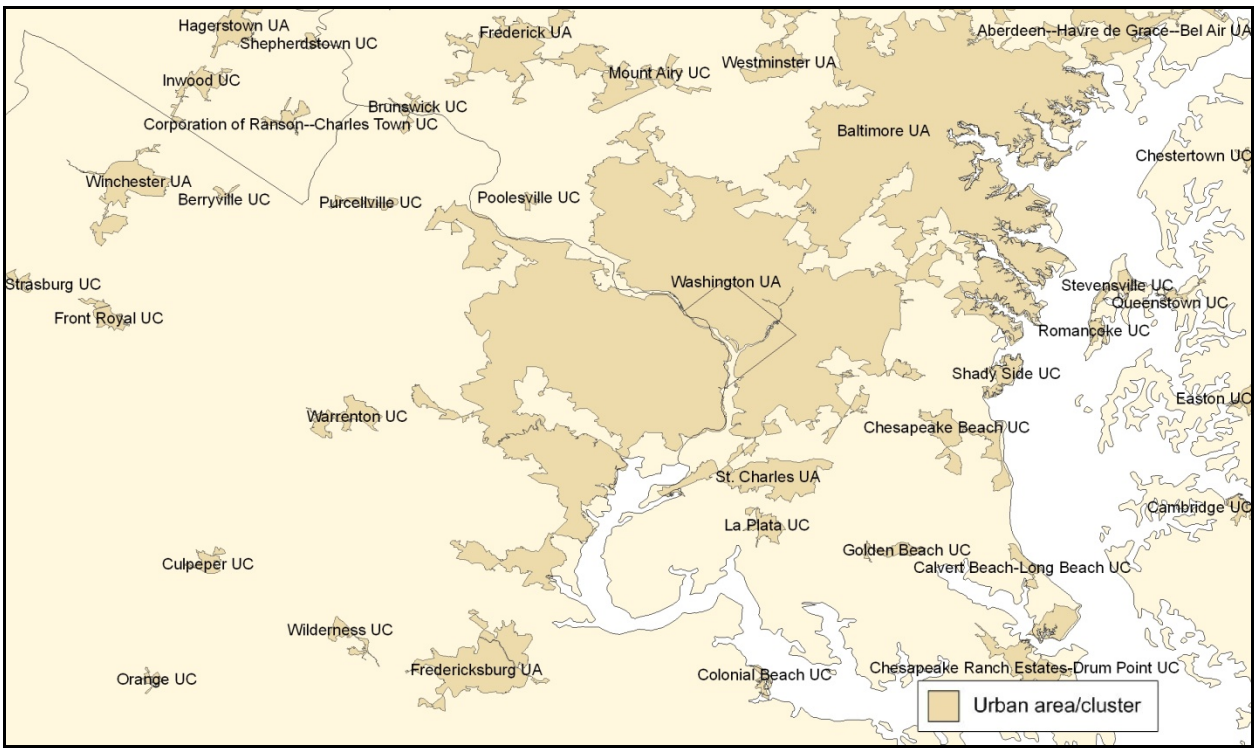

Figure 3. Diagram. The Washington UA and nearby UAs and UCs.

Interstate highway centerline data were obtained in the shapefile format from Environmental Systems Research Institute, Inc. (ESRI). This file includes interstates throughout the country. However, data with a higher accuracy were obtained from the Virginia Department of Transportation (VDOT) for the road network within Virginia. In addition to the higher accuracy, the VDOT interstate data also include a separate centerline for each travel direction. For points within Virginia, the VDOT data were used, and outside Virginia, the ESRI data were used. Figure 4 illustrates the sources of map data used.

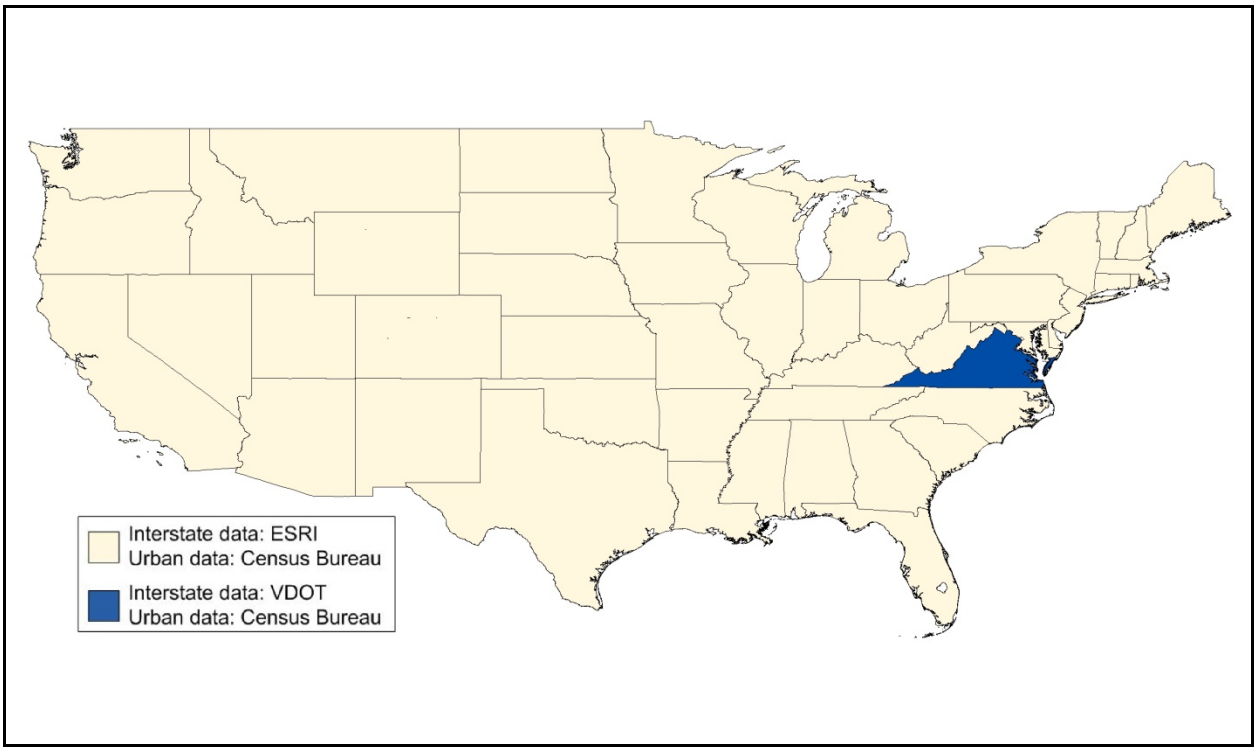

Figure 4. Diagram. Map data sources for Virginia and surrounding states.

As described above, an area around interstate segments was created for use in determining whether or not a GPS point is on or off an interstate. A 200-ft width was identified on both sides of the centerline segments representing interstates in the ESRI and VDOT data. This process, known as buffering, was done in ArcGIS. The size of the buffer was selected to create a balance between conflicting constraints. First, the buffer had to be large enough to accommodate the varying widths of interstate highways and the precision available with GPS. If the buffer was too narrow, interstate driving might be erroneously classified as non-interstate. Second, if the buffer was too large, nearby roads could fall within the buffer. Trips on these roads would mistakenly be identified as occurring on an interstate. This process created a shapefile defining interstates as areas. This shapefile was then imported into MATLAB. After some final preparations, like those done to the data of UAs and UCs, the result was then saved for quick retrieval by the search code.

TRIP DATA PREPARATION

Two characteristics of the 100-Car dataset make it larger than what is required for differentiating rural road driving from non-rural driving using the GIS-based strategy. First, the data include hundreds of variables that are not relevant for isolating rural driving. Second, the data are sampled at a rate which is higher than what is necessary for determining the route of a vehicle. To facilitate the present investigation, a database table was created which represented the path of the vehicles geographically, without including the complete set of vehicle measures. The database also used a reduced sample of the data as follows. For locating rural road driving using the GIS-based method described in this report, the primary variables of interest from the 100-Car

data were latitude and longitude. In the database used for this project, position value storage frequency was translated from the 10-Hz rate found in the 100-Car Study to a distance-based sampling method. This reduction was possible for two reasons. First, the GPS values are available at 1 Hz, but recorded at 10 Hz. So, in the original data, for each GPS position update, there are 10 duplicate rows. Second, the GPS continues to sample, even when the participant-vehicle position is not changing. In situations where the vehicle is stationary, such as when at a light or stopped in traffic, the GPS data can include minutes of redundant position data.

To provide a history of each vehicle's travel path, without including extra points, software was written which used the speed from the vehicle network (i.e., measured from wheel rotation) and time to compute a distance-traveled value for each time sample. This distance measure was then used to collect GPS latitude and longitude values every 50 ft where speed permitted this accuracy. When the vehicle was traveling at approximately 34 mi/h (50 ft/s) or faster, due to the 1-Hz sample rate of the GPS, it was not possible to identify latitude and longitude every 50 ft. At speeds above this 34-mi/h threshold, latitude and longitude were collected at each new GPS reading. In this way, a downsampled dataset was created for the trip files, with approximate minimum spacing between points of 50 ft, but with greater spacing of points as speed increased. Isolating just the variables needed for the present analysis and reducing the frequency and redundancy in position data reduced the network and computational demands during data processing.

ROAD CLASSIFICATION

The downsampled trip data were read and processed by code which was developed to determine if the latitude and longitude points fell within UA, UC, or interstate highway buffers. After processing by the code, the trip data were written to a database table. Two columns in the results table hold Boolean flags. In one column, each point within the downsampled trip data was flagged to indicate whether or not the point is within the boundaries of any UA or UC. A value of 1 in the column marks a point that is urban, and a 0 marks a point that is not urban. The second column has a value of 1 if it is within a 200-ft buffer around an interstate highway centerline, or it has a value of 0 if it is not within 200 ft of an interstate centerline. Using both flags, four types of driving can be identified, as shown in table 1. The driving type of interest for this report is rural-road driving, where both the urban and interstate values are 0.

Table 1. Driving types identifiable from urban/non-urban and interstate/non-interstate classification.

	Urban = 1	**Urban = 0**
Interstate = 1	Urban interstate	Rural interstate
Interstate = 0	Urban surface	**Rural road**

Care is taken throughout this report to only use the term "rural road" as defined above (urban = 0 and interstate = 0). "Rural interstate" (urban = 0 and interstate = 1) and "rural road" are considered mutually exclusive terms in this report. Segments of trips that fall within a UA or UC are classified as "urban surface" (urban = 1 and interstate = 0) or "urban interstate" (urban = 1 and interstate = 1). Figure 5 illustrates the four road types identified in the data. The western segments of the highlighted trip fall outside the urban area/cluster.

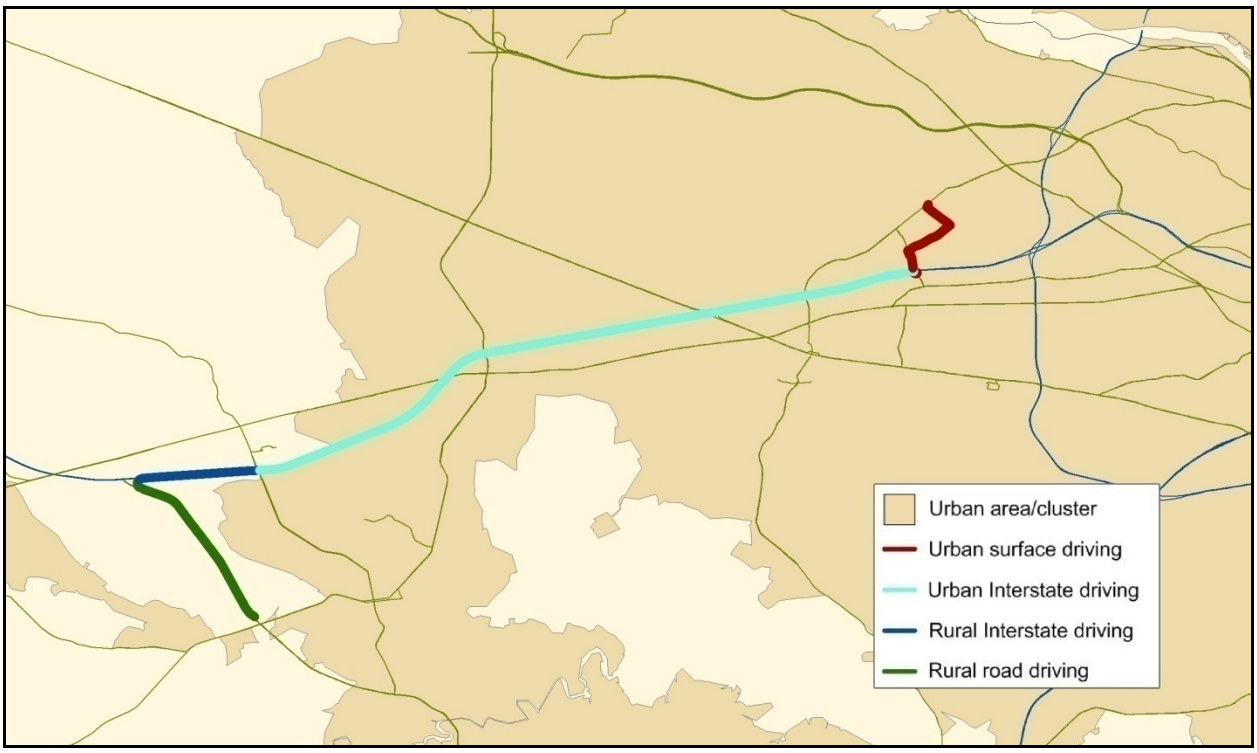

Figure 5. Diagram. Sample trip with driving on the four different road types.

DATA PROCESSING

The data analyzed for this report include 146,580 trips from the 100-Car dataset. This accounts for approximately 35,000 hours of driving. When cast in rows of data, this would create approximately 1,273,464,000 rows of data to be processed. Using a standard desktop, the downsampling task requires roughly 15 s per trip to read in the original trip file, process the file, and write output to a database. The road categorization task involved a combination of reading in downsampled data, overlaying spatial maps, applying GIS functions, and writing output. This task required approximately 30 s per trip. If processed serially for all the trips, one pass through the data would require approximately 76 days of continuous processing. Batch processing techniques were used which permitted parallel processing for the downsampling and road classification tasks. In addition to writing results, data processing functions were used to accumulate trip metadata and task tracking into a database. This permitted process monitoring and controlled restart of processing when necessary.

TRIP TOTALS

After each trip location point was classified by the above process, totals of the distances traveled and time spent on rural roads and non-rural roads were calculated. This was done by taking the difference of the distance traveled and time passed since the previous point and adding that to the respective total. The resulting totals for each trip were written to a database table.

In the 100-Car data, each trip begins shortly after the ignition is started and ends when the ignition is turned off. Therefore, if a driver makes a brief stop, such as at a gas station, while driving somewhere, the driving before and after the stop would be classified as two separate trips.

DATA MARKERS FOR USE IN FUTURE ANALYSIS

One of the results of implementing the method on the 100-Car dataset is the creation of an intermediate dataset for use in subsequent analysis. A table in a relational database was created that identifies points within 100-Car driving as urban or rural and interstate or non-interstate road. Using Structured Query Language (SQL), the table can be queried to find points within each of the four categories. The table is also accessible to any software application which can retrieve data from a database, such as MATLAB or SAS, further increasing the usefulness of the mined dataset. A few example points from a trip in the dataset are shown in table 2. In the first column is a trip ID. In the second column is the sync number which identifies a specific point in time during the trip (each increment of 1 in the sync number represents 0.1 s). The third column indicates whether or not the point is in a UA or UC. The fourth column holds the flag that indicates whether or not the point is within 200 ft of an interstate centerline. The fifth column identifies the road type at that sync number, where 1 = urban surface street, 2 = urban interstate, 3 = rural interstate, and 4 = rural road.

Table 2. Sample results from trip with all four road types.

Trip_ID	Sync	Urban	Interstate	Road type
152815	277	1	0	1
152815	296	1	0	1
152815	311	1	0	1
152815	695	0	0	4
152815	702	0	0	4
152815	716	0	0	4
152815	5576	0	1	3
152815	5582	0	1	3
152815	5593	0	1	3
152815	5603	1	1	2
152815	5614	1	1	2
152815	5625	1	1	2

CHAPTER 4. RESULTS

Three types of results were generated from this effort. The first result, which has been described in the Method section, is the development of a geospatial search procedure that can be used in any geospatial-related research effort. This result will not be described further here. The second result is the availability of a stored tabulated dataset that can be used in future research efforts investigating the rural crash problem or any effort related to the other three roadway classifications established in this effort. The third set of results demonstrates use of the method and provides initial summary measures of interest related to the rural-road crash problem.

TABULATION OF DATA

The output of the method provides a tabulation of data which can be used to address segments of the complete naturalistic dataset and associate additional variables of interest. Using the rural road, which is of interest in the present study, the database table illustrated in table 2 could be used to guide subsequent data mining within only files and sync ranges identified as being on rural non-interstate roads (i.e., road type = 4). In table 3, latitude, longitude, speed, and vehicle heading have been collected for the rural road segments.

Table 3. Driving variables and road type classification data.

Trip_ID	Sync	Latitude (deg)	Longitude (deg)	Speed (mi/h)	Heading (deg)	Urban	Interstate	Road type
152815	695	38.74487686	-77.52091217	48	307.4	0	0	4
152815	702	38.74500275	-77.52111816	49	307.7	0	0	4
152815	716	38.74513245	-77.52131653	49	308.8	0	0	4

Other research efforts might be interested solely in interstate driving. Or, an effort may be interested in comparisons across the classifications (e.g., comparing urban interstate following distances versus rural interstate following distances). Where subsequent research efforts make use of already analyzed geospatial factors, such as the four classifications used here, they can make use of the tabulated data for summary analysis or for guiding additional data mining.

DEMONSTRATION OF METHOD

The following sections provide a selection of descriptive measures collected from the 100-Car dataset to provide a demonstration of the method developed here.

Distance and Time

The 100-Car Study participants were drawn from the northern Virginia/Washington, DC metro area, which is primarily urban and suburban (Dingus et al., 2006, pp. 31–37).[(2)] Therefore, greater amounts of urban and suburban driving were expected. This has been confirmed by employing the method described in this report.

The method was applied to approximately 80 percent of electronically collected data from the 100-Car dataset. This accounts for approximately 1.1 million miles of driving from the 100-Car dataset. The trips that were analyzed were retrieved in a rotational manner across all of the

participating vehicles and selected randomly from across the period of participation. In this way, as data were processed, the results accumulated evenly across vehicles and time of year so that at any point after a reasonable accumulation of miles, the results would likely be representative of the entire dataset. Of the driving analyzed, about 12 percent of the distance traveled was on rural roads. This is contrasted with data from the Federal Highway Administration (FHWA) which show that 31 percent of the miles driven in the United States from 1994 to 2003 were on rural roads (Burgess, 2005 citing FHWA, VMT data 1994–2003).[(1)] Those same rural miles were driven over the course of 3,356 hours, which constitutes about 9 percent of the total driving time analyzed. The proportion of the distance traveled on each of the four road types is shown in figure 6. The proportion of the time spent driving on each road type is shown in figure 7.

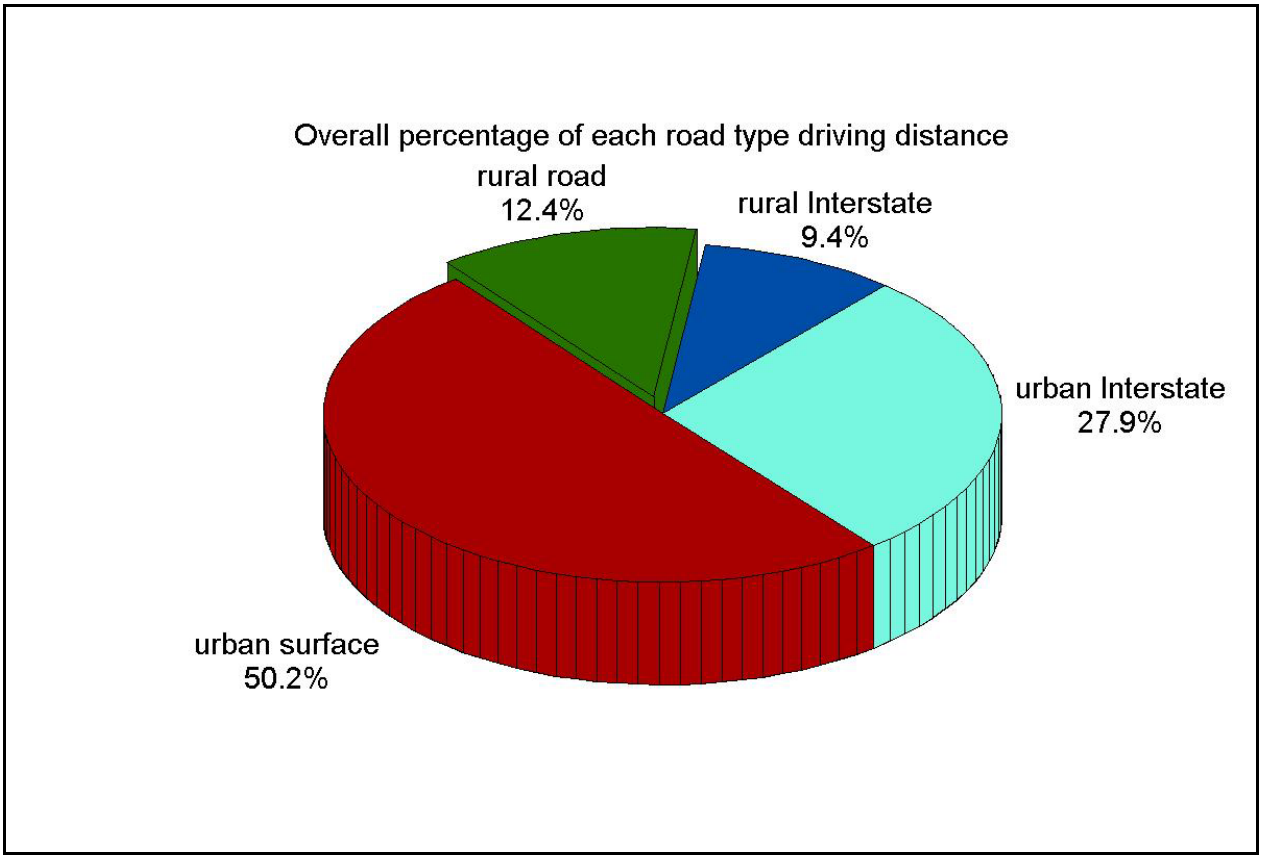

Figure 6. Diagram. Percentage of analyzed driving distance on each road type.

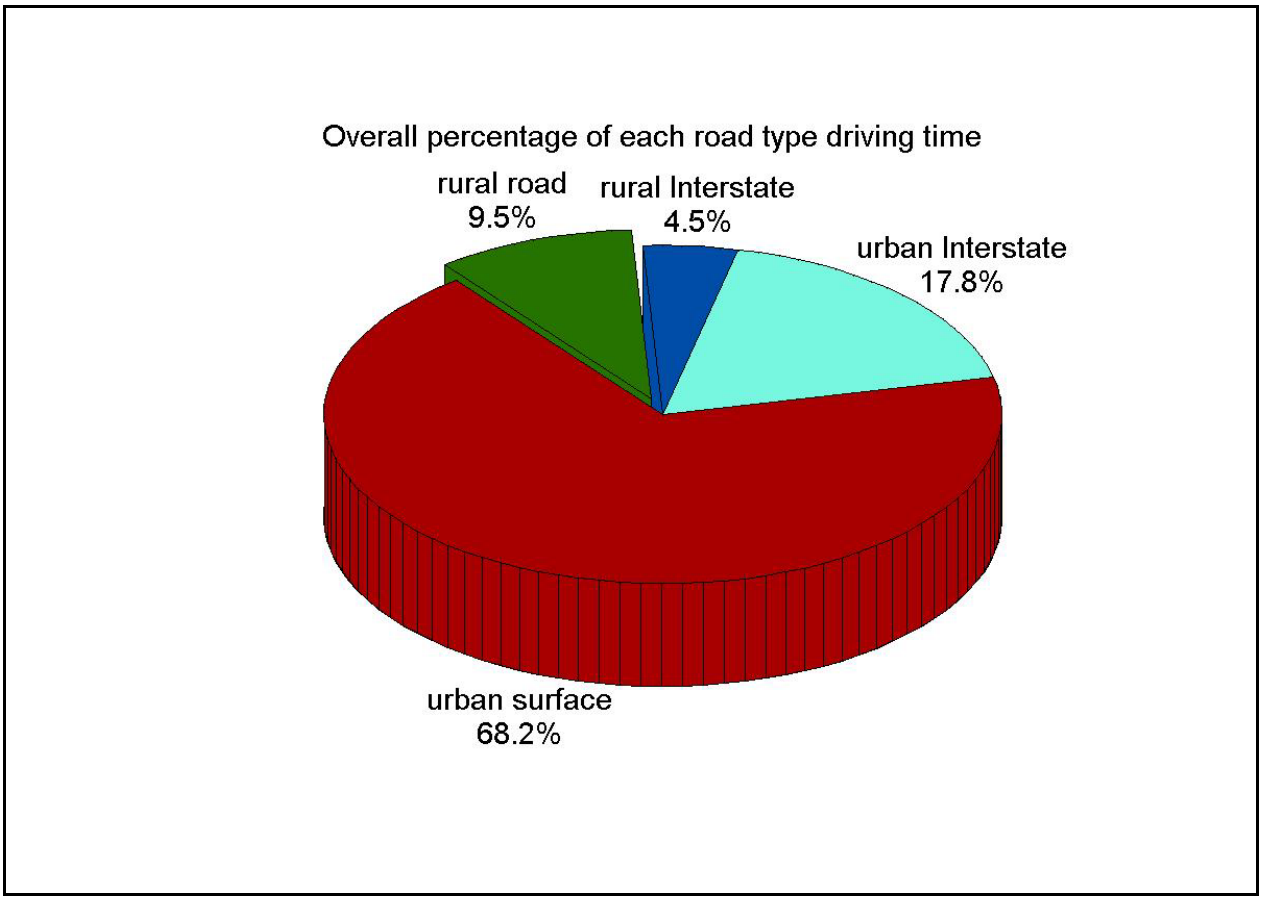

Figure 7. Diagram. Percentage of analyzed driving time on each road type.

Figure 8 shows the distance analyzed for each vehicle. The red represents driving on urban surface streets, the light blue urban interstate highways, the dark blue interstates outside of urban areas, and green rural roads. Figure 8 is sorted by the distance driven on rural roads so that the vehicles with the most rural driving distance are toward the right. Consequently, the vehicle number in the graph does not necessarily correlate with the number assigned to the vehicle during the study. For consistency and to enable comparisons, however, the same vehicle numbers used in Figure 8 are used throughout this report.

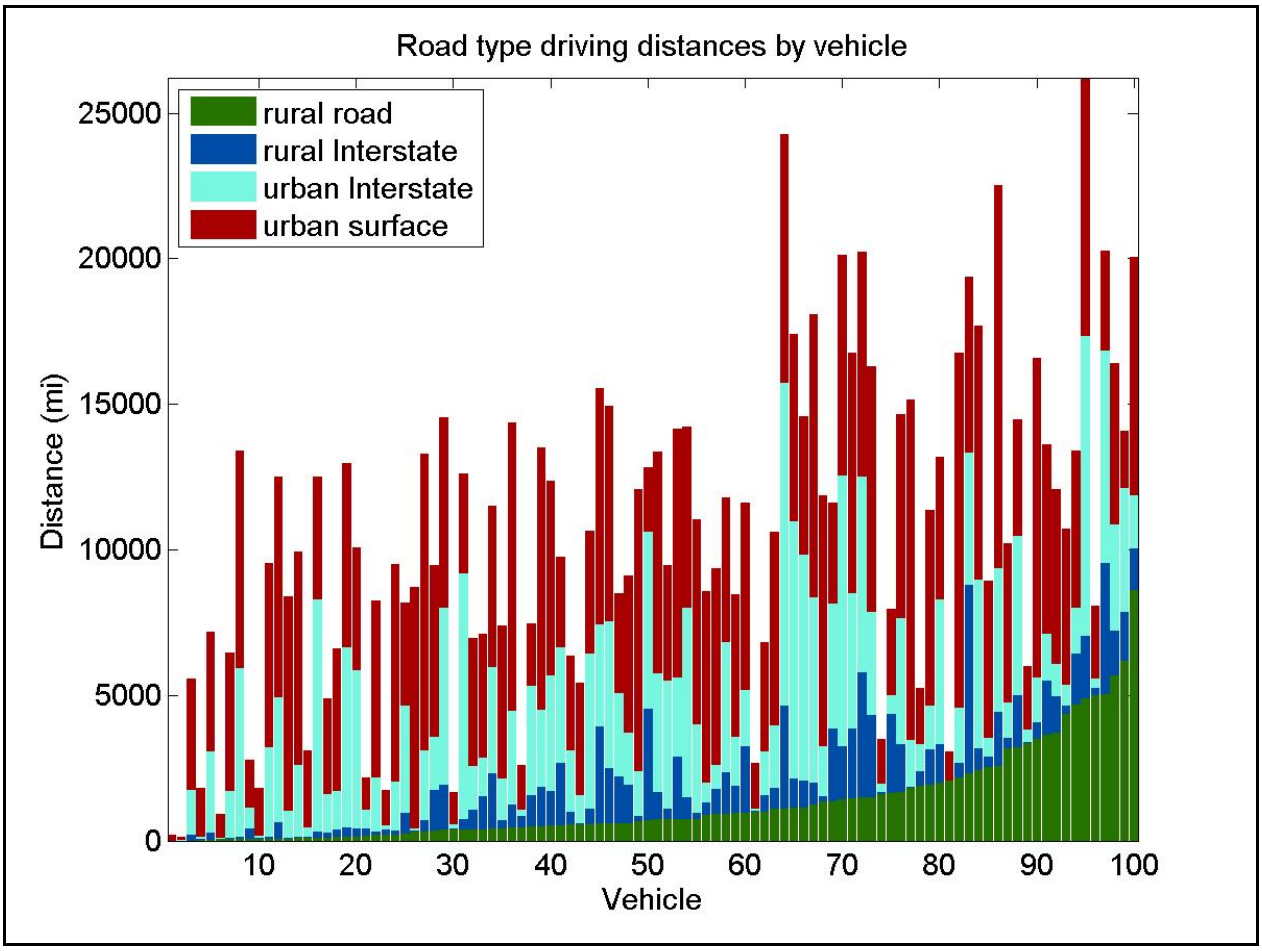

Figure 8. Graph. Driving distances on each road type by vehicle.

There is variety in the proportion of the distance each vehicle was driven on the different road types. The individual percentage of distance driven on each road type is shown in figure 9. A few cars had a high percentage of distance traveled on rural roads, but the larger percentages are typically found for the urban surface streets and urban interstate highways. Within the analyzed data, the distance traveled on rural roads per vehicle ranged from about 7 mi to 8,643 mi. For the vehicle with the shortest distance traveled on rural roads, the rural road driving accounted for about 3.7 percent of the analyzed driving distance for that vehicle. For the vehicle with the longest distance traveled on rural roads, the rural road driving accounted for about 43 percent of the analyzed driving distance for that vehicle.

The percentage of distance driven on rural roads was higher for the vehicles with the greatest distance driven on rural roads. However, the individual rankings do not necessarily correlate between total distance and percentage of distance. For example, the vehicle driven the greatest distance on rural roads was not the vehicle with the highest percentage of rural driving. The vehicle with the highest percentage of rural driving (69 percent) ranked 20th for actual distance of rural driving with 2,093 mi of rural driving.

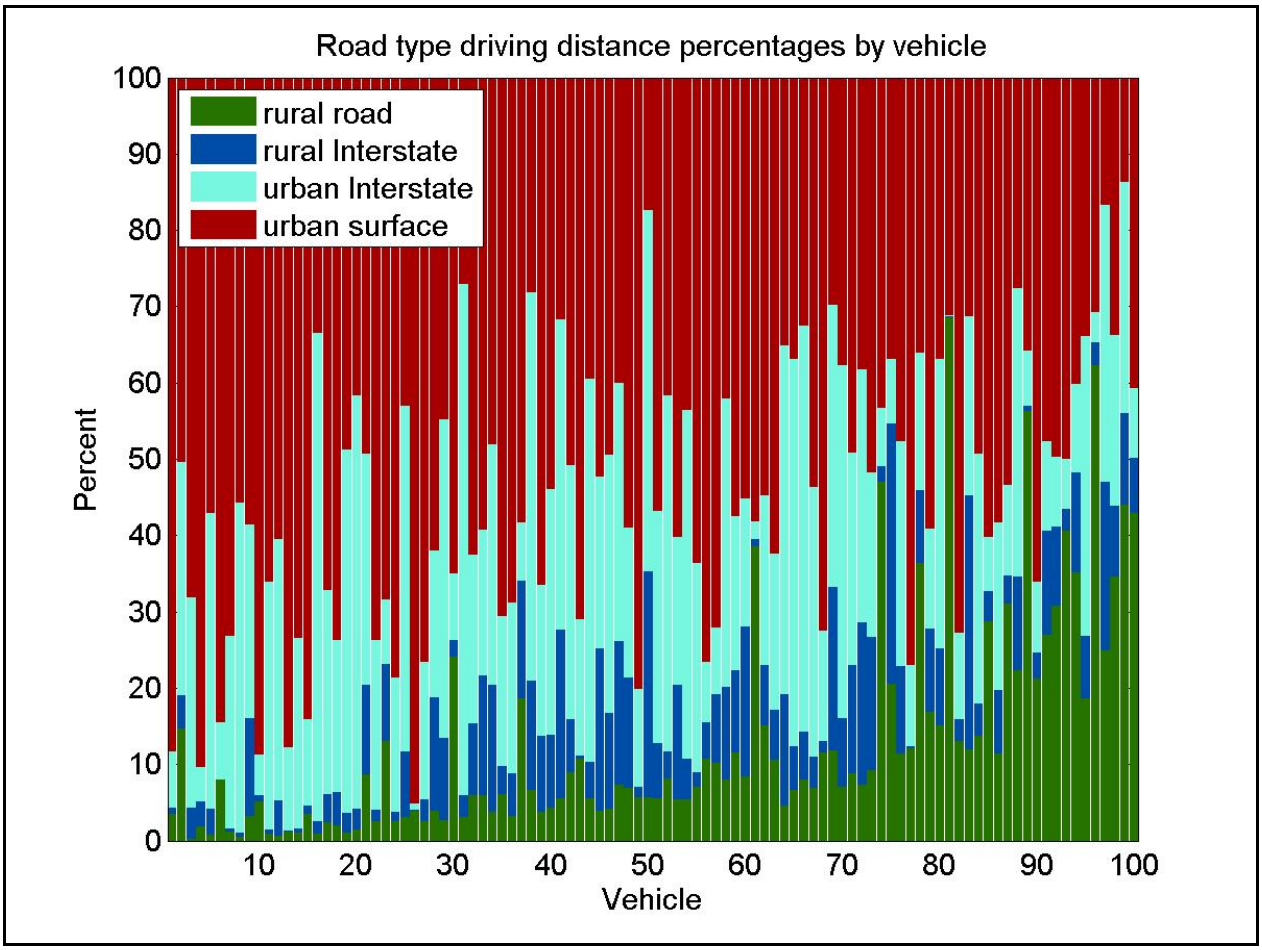

Figure 9. Graph. Percentages of driving distances analyzed for each road type by vehicle.

Figure 10 shows the time analyzed for each vehicle, with the same color coding as used in the previous figures. Note the differences in figure 8 and figure 10 between the distance and the amount of time spent driving on each road type, with the greatest difference between distance and time on urban surface streets. The percentage of the driving time analyzed for each road type is shown in figure 11.

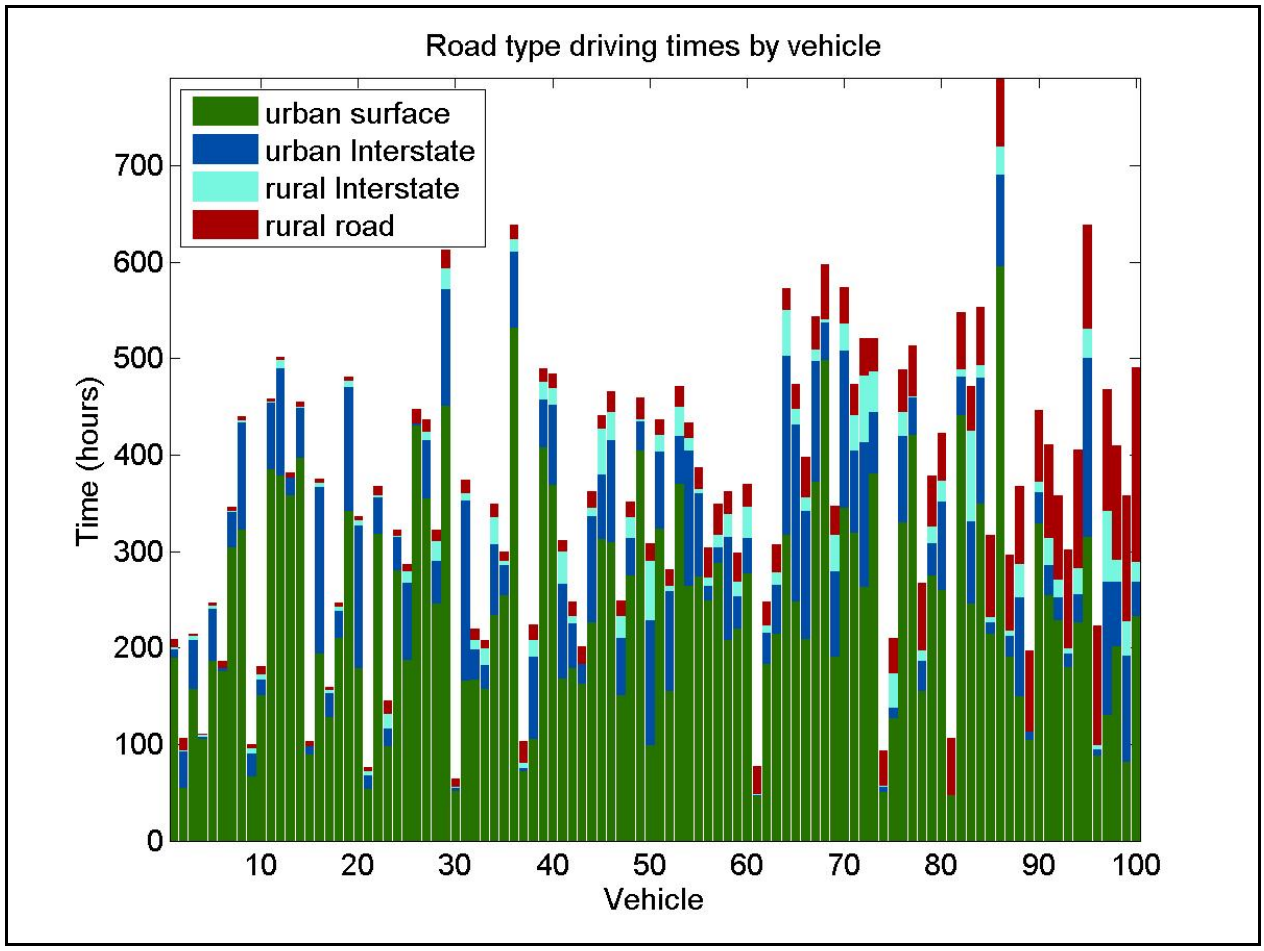

Figure 10. Graph. Time driving on each road type by vehicle.

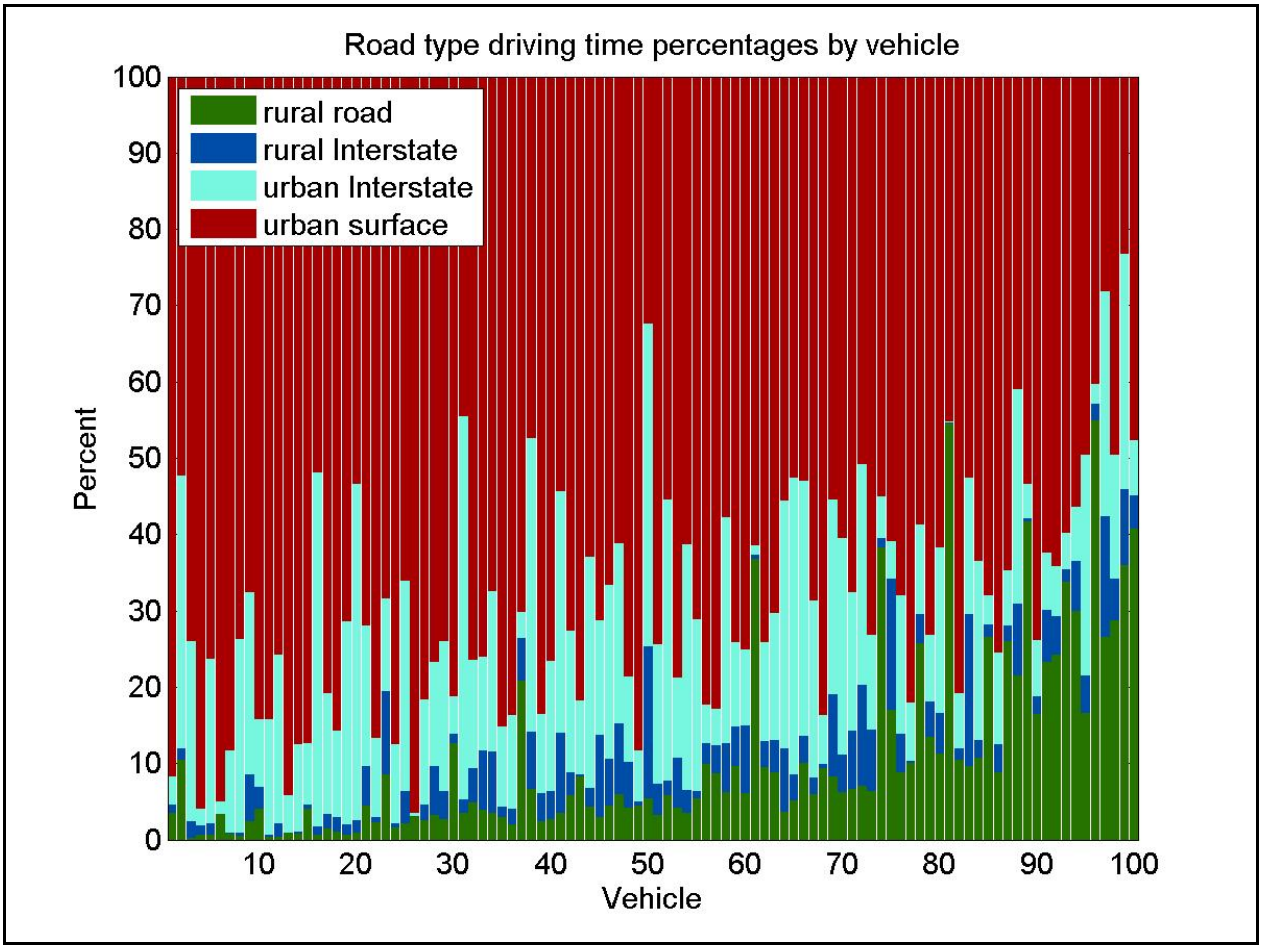

Figure 11. Graph. Percentages of driving time analyzed for each road type by vehicle.

Trip Composition

Most of the trips that include some driving on rural roads also include some driving on other road types. Approximately 21 percent of the trips reviewed from the 100-Car Study included some driving on rural roads. Of the trips with any driving on rural roads, 91 percent included at least one other road type, and thus only 9 percent of them were entirely on rural roads. The numbers of trips with driving on rural roads and other road types are shown in figure 12. The blue portions indicate the number of trips that fit entirely in the respective category, and the dark red portion of the rural bar shows the number of trips with driving on rural roads and at least one other road type.

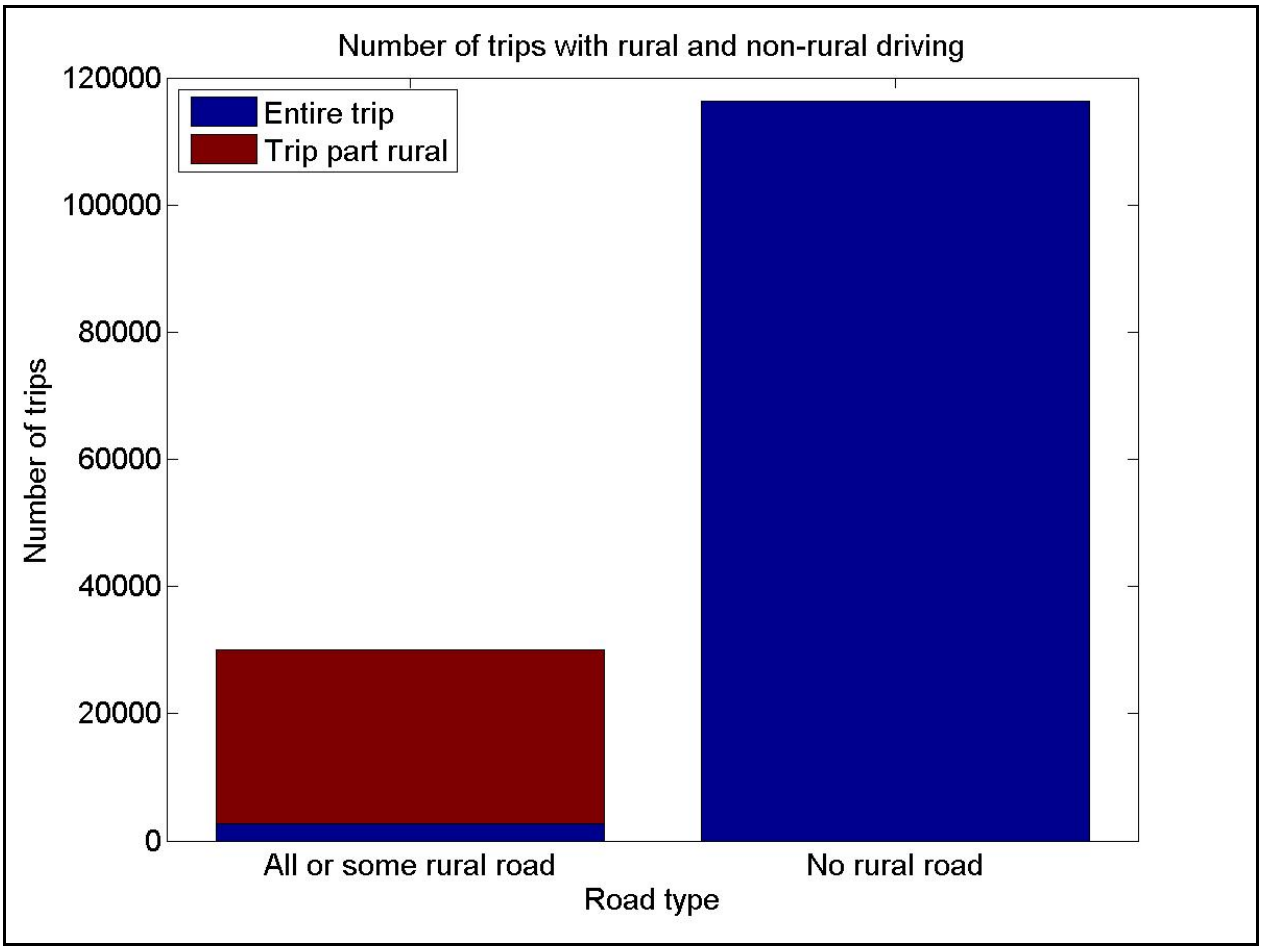

Figure 12. Graph. Number of trips with rural and non-rural driving.

Crashes and Near-Crashes

The crashes and near-crashes collected in the 100-Car dataset were also processed to classify their location according to the four categories. The vehicles in the 100-Car Study were involved in 82 crashes and 761 near-crashes. In order to be used for this analysis, GPS data must have been recorded during the event. GPS data may be unavailable during sensor failure or when the vehicle is not within view of GPS satellites, such as in parking garages or when obscured by buildings and other structures. Of the events in the 100-Car dataset, 55 crashes and 739 near-crashes had usable GPS data collected.

The crashes and near-crashes with available GPS data are shown in figure 13. As would be expected from a dataset made up with drivers from northern Virginia, the bulk of the crashes and near-crashes occurred in that area. One crash and 28 near-crashes happened more than 50 miles outside the Capital Beltway. While the vast majority happened within the Washington UA, the events span an area from Georgia to New Jersey.

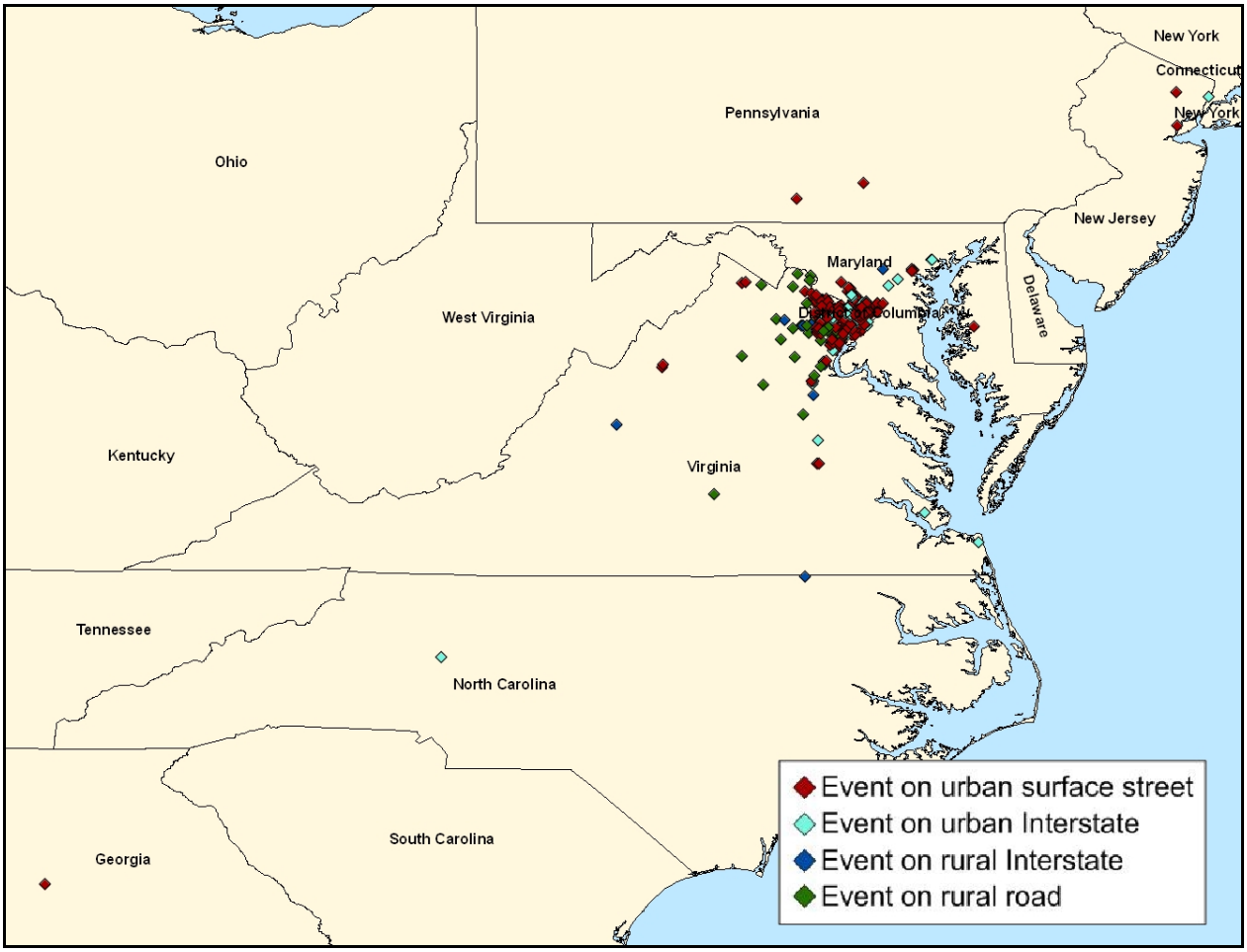

Figure 13. Diagram. All crashes and near-crashes from the 100-Car Study.

Figure 14 shows the 52 crashes that happened in the Washington UA and the two rural crashes from the 100-Car Study. One crash from the 100-Car Study happened outside the limits of the map in figure 14. It occurred just south of the state border separating Virginia and North Carolina on a rural section of Interstate 95.

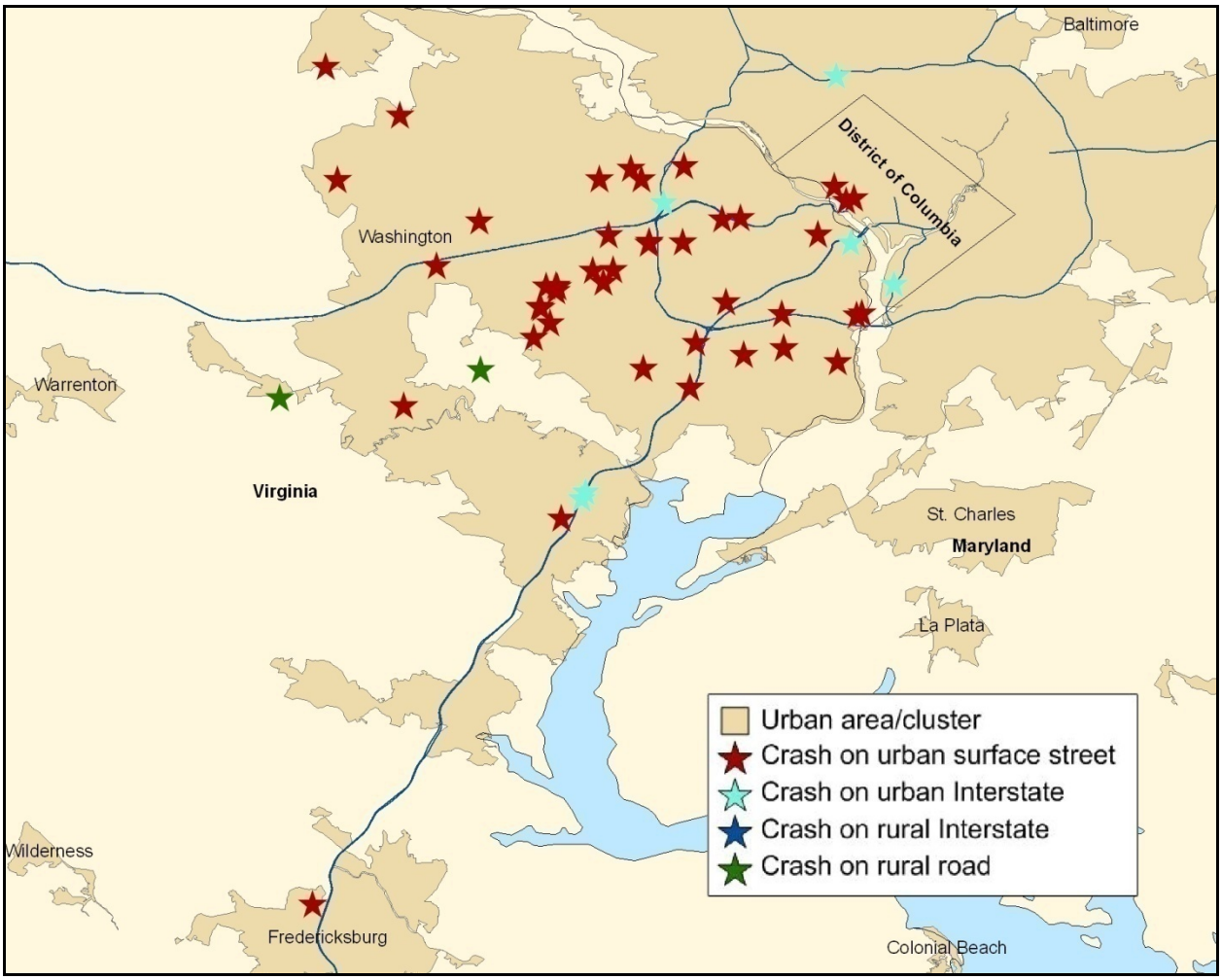

Figure 14. Diagram. Crashes in the 100-Car study.

Figure 15 shows a closer look at both the crashes and near-crashes in and around the Washington UA. As can be seen in figure 15, there are a number of areas where crashes and near-crashes seem to be clustered.

Forty-seven of the near-crashes in the 100-Car dataset were on rural roads. While about 3.6 percent of the analyzed crashes were on rural roads, about 6.2 percent of the analyzed near-crashes were on rural roads.

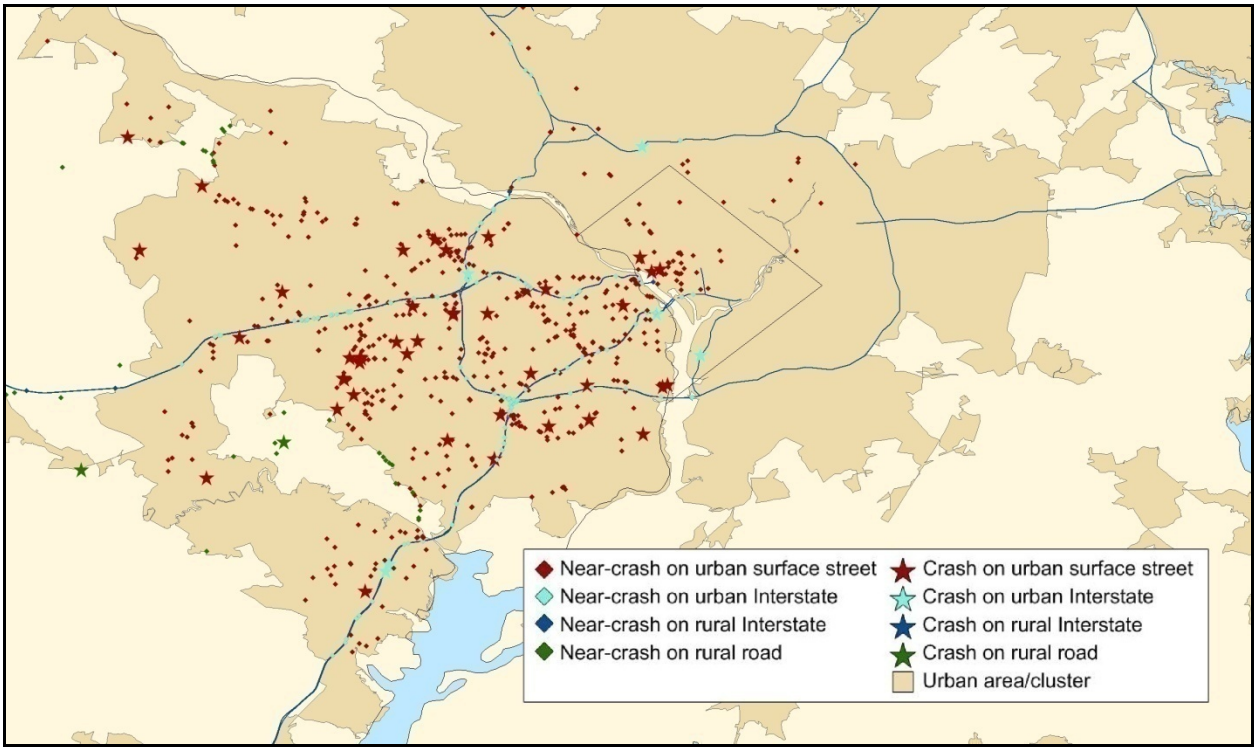

Figure 15. Crashes and near-crashes in Washington, DC vicinity.

CHAPTER 5. CONCLUSIONS AND FUTURE WORK

This method developed here proved to be successful at identifying driving on the four road types defined in this report (rural roads, rural interstates, urban interstates, and urban surface streets). In addition to this specific application, it also provides more general capability using geospatial data-mining approaches for locating driving epochs from within large naturalistic driving datasets. The method can be seen as providing analytical research tools at three levels. At a high (or aggregated) level, the method can be used to support collection and storage of metadata about a dataset. These data might include summaries of driving times, distances, average speeds, or other measures of interest aggregated within geospatial classifications of interest. At an intermediate level, the method can generate reduced datasets that provide methods for researchers to extract, or address, specific geospatial categories within a large dataset. An example of this is querying a dataset for just urban surface street driving for subsequent processing. The third level involves use of the methods demonstrated here to locate cases of driving at specific locations, such as an intersection of interest or a problematic roadway curve.

TECHNOLOGY USE

The preparation of the map data was primarily accomplished using ArcGIS by ESRI. ArcGIS also proved to be very useful for graphically representing the outcome. This was done to both demonstrate and verify results. Code written for The MathWorks, Inc.'s MATLAB with the Mapping Toolbox was used to read in data from the vehicles and to process the data against the map data. The use of these tools proved to be a powerful combination for batch processing large amounts of geospatial data.

LIMITATIONS

Determining a road's urban or rural classification using the Census Bureau's UAs and UCs may not reflect brief changes in road type. For example, a road segment within an urban area may look similar to what a person would subjectively consider to be a rural road. Conversely, a road classified as rural may, in places, be more similar to urban roads. These types of transient changes are not considered a strong threat to initial investigation into the rural road problem, but further work may become necessary to investigate these types of situations. Changes in population density, such as new urban growth, may make UA and UC data out of date.

It was expected that the proportion of rural driving within the 100-Car dataset was smaller than the national proportion of rural driving because the study participants were drawn from the northern Virginia/Washington, DC metro area, which is primarily urban and suburban. Consequently, the data cannot be used to determine national proportions of rural and urban driving. It would seem reasonable to expect, however, that the study participants would drive their cars outside of their home area occasionally over the course of the study's yearlong duration. Consequently, the dataset still provided an opportunity for developing a method to extract naturalistic driving data on rural roads. The rural road driving data mined using this method will also be valuable for future work studying driving on rural roads.

While effort was made to optimize the buffer size to represent the width of an interstate highway, because of the varying width of all road types there are road segments on which cars are likely to

travel outside the area covered by the 200-ft buffer. It is also possible that in other areas, the buffer may be too large and nearby roads, such as service roads, may fall within the buffer. Further refinement of this method could include using GIS tools which are designed to identify roads traveled over a specific route. These tools would be used to eliminate this concern.

It should be noted that due to the powerful abilities of GIS tools and the nature of GPS variables in naturalistic driving data, it is possible to identify the locations that a driver frequently visits or travels. This includes places such as their home or work, making it easy to identify a study participant. Human participants are given a promise of confidentiality when they consent to participate in a naturalistic driving study. This promise is made to encourage the participants to drive as they normally would. Primary to this confidentiality is the promise that identifying information will be separated from other data. Consequently, the study of GPS data requires care to prevent the unintentional release of information that could be used to identify study participants. Various approaches are possible, such as removing location information when it is unnecessary for analyses or ensuring only small segments of location information are available during analysis. Each of these, however, could still expose identifiable information if executed incorrectly (National Research Council, 2007).[8] In summary, privacy concerns of participants may preclude the unlimited availability of location data for secondary analysis.

FUTURE WORK

The process used here shows promise for other geographically based data-mining needs. For example, driving on specific corridors of interest could be extracted from a naturalistic-driving dataset that included latitude and longitude variables. These corridors of interest could be intersections or sections of road with high crash rates, for example. It could also be used to extract driving through areas with abnormal road geometry or any specific road type that a researcher may have interest in studying. The method could be used to pinpoint driving in small areas of interest, such as an intersection, or it could be used to identify larger areas of interest, like neighborhoods, cities, counties, or states.

By combining other variables and data-mining techniques with the process outlined here, summary information could be gathered regarding driving characteristics such as average speed or acceleration on specific roads or road types. As noted above, these summary data could also be collected on driving within a government jurisdiction or other geographic area. These other variables would virtually be constrained only by the limits of the dataset.

REFERENCES

(1) Burgess, M. (2005). *Contrasting Rural and Urban Fatal Crashes 1994–2003* (DOT HS-809-896). Washington, DC: National Center for Statistics and Analysis, National Highway Traffic Safety Administration.

(2) Dingus, T. A., Klauer, S. G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J., Perez, M. A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z. R., Jermeland J., & Knipling, R. R. (2006). *The 100-Car naturalistic driving study, phase II - results of the 100-Car field experiment.* Washington, DC: National Highway Traffic Safety Administration.

(3) Ivan, J., Pasupathy, R. K., & Ossenbruggen, P. J. (1999). Differences in causality factors for single and multi-vehicle crashes on two-lane roads. *Accident Analysis and Prevention*, *31*(6), 695–704.

(4) Persaud, B. N., Retting, R. A., & Lyon, C. A. (2004), Crash reduction following installation of centerline rumble strips on rural two-lane roads. *Accident Analysis and Prevention*, *36*(6), 1073–1079.

(5) Kim, D., Washington, S., & Oh, J. (2006). Modeling crash types: new insights into the effects of covariates on crashes at rural intersections. *Journal of Transportation Engineering*, *132*(4), 282–292.

(6) Kim, D., Lee, Y., Washington, S., & Choi, K. (2007). Modeling crash outcome probabilities at rural intersections: application of hierarchical binomial logistic models. *Accident Analysis and Prevention*, *39*(1), 125–134.

(7) O'Neill, W. A., Ullah, K., & Wang, M. (1993). IVHS and rural road Safety: A Prototype ATIS. *ITE Journal*, *63*(11), 30–37.

(8) National Research Council (2007). *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*. Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data. M.P. Gutmann and P.C. Stern, Eds. Committee on the Human Dimensions of Global Change. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.