SPATIAL ANALYSIS OF NORTH CENTRAL TEXAS

TRAFFIC FATALITIES 2001-2006

Paula S. Rafferty

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

December 2010

APPROVED:

Pinliang Dong, Major Professor
Donald Lyons, Minor Professor and
    Program Coordinator
Chetan Tiwari, Committee Member
Paul Hudak, Chair of the Department of
    Geography
James D. Meernik, Acting Dean of the
    Robert B. Toulouse School of
    Graduate Studies

Rafferty, Paula S. Spatial Analysis of North Central Texas Traffic Fatalities 2001-2006. Master of Science (Applied Geography), December 2010, 80 pp., 7 tables, 25 illustrations, references, 34 titles.

A traditional two dimensional (planar) statistical analysis was used to identify the clustering types of North Central Texas traffic fatalities occurring in 2001-2006. Over 3,700 crash locations clustered in ways that were unlike other researched regions. A two dimensional (*x* and *y* coordinates) space was manipulated to mimic a one dimensional network to identify the tightest clustering of fatalities in the nearly 400,000 crashes reported from state agencies from 2003-2006. The roadway design was found to significantly affect crash location. A one dimensional (linear) network analysis was then used to measure the statistically significant clustering of flow variables of after dark crashes and daylight crashes. Flow variables were determined to significantly affect crash location after dark. The linear and planar results were compared and the one dimensional, linear analysis was found to be more accurate because it did not over detect the clustering of events on a network.

ACKNOWLEDGEMENTS

I would like to thank the three professors, Drs. Pinliang Dong, Donald Lyons, and Chetan Tiwari, who helped me throughout by giving direction to this work, Donna Plummer from the TxDoT for the crash statistics for the region, Dr. Bruce Hunter from the Center for Spatial Analysis and Mapping for keeping the GIS up and running, and Prof. Atsuyuki Okabe of University of Tokyo for providing the SANET software.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Statement of the Problem

Traffic fatalities are considered a global epidemic and an increasingly important issue. The World Health Organization estimates 1.3 million people worldwide die every year as a result of car crashes (World Health Organization [WHO], 2009). Research suggests that crashes and the resulting fatalities can be avoided (Baker and Haddon, 1974; Cook and Tauchen, 1984; Crandall and Graham, 1984; Ross, 1973). GIS (geographical information system) is a tool that can be used to spatially analyze crash patterns so that crash reduction measures may be fed into the existing roadway structure.

Spatial statistics are often used to analyze spatial phenomenon and can also be used to identify crash patterns. Mapping crash points allows specific patterns to be isolated while analyzing the associated crash variables that highlight the dynamics associated with each crash pattern. Intentional driver behaviors that result in crashes and that can be distinctly isolated from crashes that are a result of unintentional driver behaviors are of importance when identifying the causes of crashes. Analyzing existing crash databases alongside the existing policies in law enforcement can effectively pinpoint successful management practices. Finally, comparing the traffic management outcome in a GIS can be beneficial in visually relaying information to decision makers in local areas.

Traffic accidents have been a leading cause of death in the United States (National Highway Traffic Safety Administration [NTHSA], 2008) as is the case with most other industrialized countries (WHO, 2008). The United States can boast of a decline in traffic fatalities in recent years, but without a vigilant review of reduction practices, increases are inevitable (NHTSA, 2004). In the year 2000 alone, the national economic costs of traffic crashes exceeded $230 billion (NHTSA, 2002). Currently, Texas experiences a traffic fatality on the average of every 2 hours and 26 minutes (Texas Department of Transportation [TxDOT], 2009).

The 16 county region of North Central Texas (NCT) expects about a 45% population increase by the year 2030 (North Central Texas Council of Governments [NCTCOG], 2008). This is an important consideration for city planners and transportation engineers of the region because the present road capacities will not accommodate such an increase. Crashes occur every 70 seconds on the roadways of Texas (TxDOT, 2009) and with levels of congestion on the rise, crashes are likely to increase. The subject of this research concerns the traffic crashes within the 16 counties of NCT and the associated dynamics of their patterns of location.

Literature Review

Accidents

Historically, traffic studies were grounded in the economic and political sciences (Kendall, 1950; Lewis 1927). The transportation system was used as an indicator of economic development, while traffic congestion was used to analyze the characteristics of road users. Network planning and public safety regulations became necessary in order to deal with the emerging problems associated with traffic congestion.

Network planning came in the form of throughways that enabled swift roadway passage between some starting point and an end destination. Thoroughfares with constant speed and without the competition between through and local traffic provided a solution to congestion in the form of a smoother flowing, complex, and interdependent system. Inbound vehicles from numerous directions and vicinities could merge onto a broad roadway that in turn provided exits that joined other exterior local routes. Coexistence of local and through traffic was achieved because there were not intersections and stops for through traffic. Once moving, a vehicle could continue moving until the local route was reached.

Regulation was necessary to provide a safe environment for the road user and was enacted early on. There were three main courses of action through which public safety legislation was implemented: infrastructure planning, vehicle safety measures, and driver monitoring. The first dealt with land use specifications and included engineering requirements for the road in order to incorporate the different modes of transportation along with their specific pattern needs (Pruitt, 1979). Land use designations included rural and urban classifications. Today, patterns of traffic movement are a part of what defines the roadway system and are a reflection of the everyday interactions in American society. The second method of regulation was to ensure safe vehicle design. Safety features like shatterproof windshields, seat belts, and structural elements, prevented injury and reduced the trauma associated with traffic accidents (Huelke, 1968). The third course of action was to enact driving rules aimed at the road user. Speed limits and licensing are two ways in which driver behavior is monitored.

Within the last 50 years, the assumption has been that drivers can avoid fatal crashes if the right precautions are taken (Baker and Haddon, 1974). Restrictive laws, such as restraint use or drinking and driving laws, are based on this perspective that the cause of death may be attributed to driver behaviors. Statistics show that more than half of the people killed in car crashes were not using seat belts (Fatality Analysis Reporting System [FARS], 2009) and it is estimated that over 75,000 lives have been saved due to the effectiveness of the restraint law over the years. Overall, alcohol legislation such as minimum age requirements and license revocation has also been effective in reducing alcohol related crash fatalities. The 2008 statistics for fatalities involving alcohol were at 32% of all fatal accidents, down from a high of 50% in 1980 (Mothers Against Drunk Driving [MADD], 2010). However, driver behavior is not always a result of actions under the driver’s control. Visibility, road geometries, and the weather often create conditions where the driver’s response is uncontrolled, and so also play a role in motor vehicle accidents.

Once a vehicle is involved in a crash, police gather information at the site and record it. Police records of the accident include information about the driver, the people involved, all of the vehicles involved, and information about the crash itself. This information is used to analyze traffic accident patterns. The analyses results lead to new measures that aim to reduce traffic accidents. For instance, if it can be shown that many crashes involve drivers without licenses; local jurisdictions may need to enforce stricter compliance methods to promote responsible driver behavior.

Driver Behaviors-Intentional

The research into traffic accident patterns often considers the crash location a consequence of measureable driver behaviors: that is, an accident is a result of predictable intentional or unintentional driver decisions. Most accidents are caused by intentional driver behaviors. Reckless driving is considered a conscious driver decision because the actions are based on personality traits such as thrill seeking or driver aggression (Taubman-Ben-Ari et al., 2002). Thrill seeking behaviors include drinking and driving, speeding, and lane weaving. Driver aggression behaviors include following too close behind cars ahead of the driver and cutting off other drivers. Other intentional driver behaviors consist of negligent behaviors such as careless driving (lack of restraint and disregard of roadway protocols like blinker use). Intentional driver decisions often manifest themselves through actions that are monitored through strategies within enforcement agencies such as the courts, police departments, and through law making bodies. The strategies may consist of regulating speed limits, enforcing vehicle ownership responsibilities, and penalties for not abiding by the rules.

Driver Behaviors-Unintentional

Unintentional driver behavior is largely spontaneous on the driver's part and the behaviors are usually quick reflex responses to spatial aspects within the driver's environment. Unintentional driver behavioral crashes are the result of a driver's inability to avoid a collision. The unintentional crash may be the most difficult behavior to monitor or change. The unintentional behavior is important in traffic accident pattern analysis because there may be issues involved that are common, and relevant, to other accidents. For instance, a rear end collision is common in the outer lanes of congested roadways (Golob et al., 2003). The stressors of congestion on the driver lead to behavior that causes a misjudgment in the rate of traffic flow movement. The driver usually has initiated a defensive tactic for control of the vehicle, but was incapable of full avoidance based upon the time and distance required for adjustment to the traffic flow. Research shows that the crash site will not only be an indication of the speed of the vehicle at the time of the crash, but also the amount of traffic at the time, and may also be an indicator of the severity of the accident (Golob et al., 2004).

Traffic Flow Accidents

The patterns associated with the location of traffic accident locations may be divided into two categories: accidents where the cause is linked to intentional driver behaviors and accidents that are linked to unintentional driver behaviors. Accidents where unintentional driver behaviors are at fault and are largely a reaction to the movements within the flow of traffic are called traffic flow accidents. Traffic flow accidents are important elements to study in traffic systems. The traffic flow system has its own set of variables and safety issues, and research has shown that traffic accidents do not have a linear relationship with traffic flow (Mountain, et al., 1996). Traffic flow is dependent upon the roadway type (function) on which the vehicle travels from one place to another, not upon the vehicles within the flow. The numbers of lanes, whether it is rural or urban, or divided, are some of the elements that constitute the roadway upon which traffic flows. Traffic roadways take into account the uniqueness of location and include entrance and exit ramps and features that are inherently necessary for the network structure. A smooth, gradual transition of incoming and outgoing traffic produces an efficient traffic flow. An efficient, smooth traffic flow reduces travel times, vehicle emissions, and stressors on the driver. Egress methods, as well as the numbers of lanes, should facilitate traffic flow. This environment, although designed as a system for movement, often clogs with congestion. Vehicular volume, density, and speed are some elements that describe the traffic flow facility, or congestion level, of a roadway (Golob et al., 2003). Preventive measures are put into place to prevent accidents that are a result of traffic flow driver behavior, but unintentional driver behavior is sometimes evident on the roadway.

How to Identify a Traffic Accident Pattern

The most basic element of a traffic accident pattern is the designation of a single crash point in a geographic location. Once the basic elements are plotted on a map, the crash points can be tested for clustering and significance. Crash patterns may also be recognized through interpretations of the variables associated with each crash location. The variables define the crash in terms of road construction, speed limits, or alcohol use. Statistical analyses are usually performed on one or more of these factors so that driver behaviors may be interpreted and a pattern of crash causes may be recognized.

There are some major difficulties when trying to isolate accident patterns. Most studies of traffic accident patterns discuss the weaknesses of common pattern analysis (Xie and Yan, 2008). Many statistical tests measure distances between a set of crash locations and compare them to a normal or random population distribution. The population of a normal or random distribution is usually spread throughout a two-dimensional area, i.e. a square grid, set upon an *x* and *y* axis, but not along a pre determined, one dimensional stretch, such as a roadway. Figure 1 shows the biases of accident locations. The data includes random point placement (triangles) as well as the placement of real recorded accidents (dots). Rather than a randomly spaced collection of accident locations, the crashes are clustered in a linear fashion. The roadway structure is missing, yet the accidents clearly follow the roadway boundaries. Two-dimensional methods are flawed when measuring traffic accident distances because, by default, accidents are restricted within roadway boundaries. Traditional two-dimensional pattern analysis will always over-detect traffic accident cluster patterns when the variables are associated with geographic location.

Crash patterns that are not evident at the network scale, i.e. analysis of traffic crashes at a county or regional level, may be recognizable on a larger scaled map of a neighborhood or city block (Thomas, 1996). Adjusting the study area by modifying the areal unit may often times uncover patterns that highlight the most important elements of the accident.

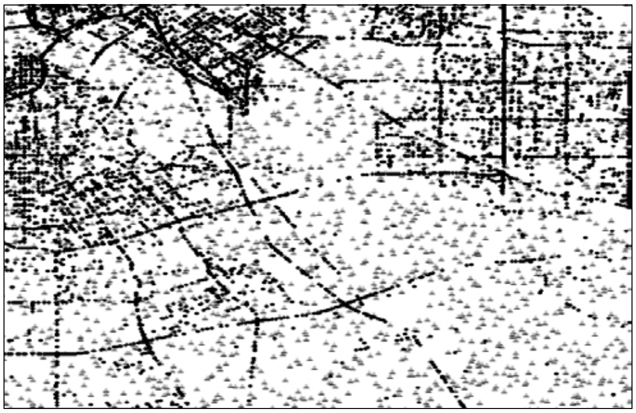


Figure 1. NCT traffic crashes without the underlying roadway structure. Triangles - random points generated with Hawth’s Tools for ArcGIS; Dots - real traffic accidents.

According to Xie and Yan (2008),segment length will significantly affect the local variations of the crash point data pattern. Longer segments will group more crashes and obscure smaller clusters. Analyses of crash points on a large scale map help determine if areas are hazardous or crash prone. The hazardous areas are often called “hot spots” or “black zones.” These high risk road sections may not be apparent at a network scale because the scale contains so much area that is unused by the roadway system. A large scale areal map provides less area coverage and more detail (1 centimeter = 1 meter) than a small scale areal map that provides more area coverage with less detail (1 centimeter = 100 meters). Large scale analysis is favorable in traffic studies because the crash location takes up more of the study area when unused, off-road space is eliminated. The larger scale makes each point and its associated variables more important, thus the data clusters that are linked specifically to the geography of the region may be highlighted and uncovered (Eksler et al., 2008). Scale changes may be made by any method that narrows the scope of the data. The difference in scale alters the pattern because each crash location is analyzed and compared within a different framework. To fully understand why accidents are occurring, it is important to control variables. Isolating similarities among many traffic accidents, whether through driver behaviors or geographically, lead to pinpointing traffic accident patterns and enable a more precise look at the specific causes.

Traffic Accident Patterns are a Function of Traffic Flow

Traffic flow can be looked at as the geographic location of a vehicle coupled with the linear momentum associated with its change in geographic location. When traffic flows smoothly, traffic flow and congestion are predictable. Rush hours are predictably more congested than non-rush hours if traffic is managed in a manner that works with the local environment. Traffic flow accidents can also be predictable. Not predictable in a linear regression model, but because of the correlations with the distinct roadway phenomenon of traffic flow and congestion. When preparing traffic flow accident cases for analysis, the traffic flow existing at the time of the crash is considered in order that accidents may be profiled with other accidents exhibiting similarities in traffic flow. By organizing crashes in this way, traffic flow accidents may be characterized and grouped into categories.

Traffic flows at different rates according to the lane being observed. The right hand lane behaves differently than the other lanes because it is used intermittently as an on/off roadway. The right hand lane typically involves off-the-road accidents (Golob and Recker, 2004). Lanes that are more indicative of flow rates are the interior and left hand lanes. Generally, the interior lanes experience lower traffic flow than right and left hand lanes. This type of traffic flow leads to side swiping accidents as vehicles weave in and out of the interior lane. The higher flow lanes (exterior left hand lanes) usually have the rear-end collisions. Research has shown that the collision lane is more a function of traffic flow than density levels (Golob et al., 2003).

There is a general consensus that traffic congestion is a problem. This is not always true. Local planning boards plan for a certain amount of congestion; for instance, NCT plans for a 30% increase in congestion levels (a moderate level of congestion) for the year 2030 (NCTCOG, 2009). Some level of congestion actually reduces traffic accidents (Golob and Recker, 2002). Managing congestion within some predetermined level helps ensure precautions are taken to reduce the occurrence of the predictable crashes.   Without some level of congestion, free flow conditions set the stage for speeding and other risk taking behaviors.

The density level affects many variables associated with crashes; for example, crashes involving a single car decrease with increasing congestion, and crashes involving multiple cars, increase. Levels of congestion range from complete free flow conditions to full roadway capacity to severe congestion. Free flow conditions promote excessive speeds, while roads that are past nominal capacity have low flow, lower speeds, and very little variability in both. As congestion levels increase from free flow, the numbers of vehicles traveling at above average speeds also increase with the demand for road space.   As capacity is approached, speed and speed variability drops. At this point, the variability in speed and flow begins to increase once again, due perhaps to the lane jumping in the outer lanes that are necessary for right hand lane exits or in a last ditch effort to beat the traffic of heavier congestion. The final stages of congestion have very little speed and flow, only a variation of speeds and movements between vehicles.

In dark, low flow conditions, two vehicle crashes predominate, until at lowest flow conditions, rear-end and weaving crashes dominate. In the first stages of darkness, the speed and the flow of the vehicles increase with increasing congestion, but then speed and flow decrease. It is possible that diminished visibility of the environment encourages a driver behavioral response to follow traffic movements but then the drivers tend to slow down in efforts to adjust to light conditions of the surroundings. Nighttime congestion does not usually reach the levels of daytime congestion, but it is also possible that congestion adds a dimension to night time traffic that is not apparent in the daytime.

Traffic counts measure the rate of traffic and are used as a way to measure the level of congestion on a roadway. Traffic counts for crash analysis come in two forms: aggregated and disaggregated. Aggregated counts are data compiled in respect to their geographic location, over some time period; for example, traffic fatalities in North Central Texas between 2001 and 2006. These counts are readily available in many databases and are usually equipped with query methods that enable the user to define their search. Disaggregated counts emphasize the individual crash. The data used to support disaggregated analysis are the physical characteristics of the crash, the environmental conditions (including roadway geometry), but most importantly, the description of the traffic flow facility at the time of the crash (Golob et al., 2003). Golob et al. (2003) uses disaggregated counts to avoid the comparisons of an individual crash to crashes in general (ecological fallacy), often associated with aggregated counts. Each crash is analyzed in relation to its specific traffic flow state. Disaggregated counts are derived from single (occasionally double) inductor loops placed along traffic corridors to measure real time traffic volume. Hourly volume rates allow analysis of the traffic flow at the time of the accident, but are difficult to obtain and not a viable consideration for this study. Although street level counts are available through the Texas Department of Transportation, the availability is without an interactive query method and time constraints prohibit data collection. Consideration of the other factors involved in the makeup of a traffic accident is therefore necessary. Using crash data that is readily available, combined with known driving behaviors and their consequences, an inferential study that is uniquely regional is possible and provides a reliable interpretation of NCT crash statistics.

Research Questions and Objectives

Relating the NCT traffic crashes to crash lane research can be done by first dividing the NCT crash dataset into the number of traffic lanes the crash site contains. Test results will show if the NCT region experiences accidents that are as predictable and common as the results of other traffic flow research suggest. When traffic crashes occur as a result of traffic flow there are identifiable characteristics. According to traffic flow studies (Golob and Recker, 2002) moderately congested traffic flow favors a reduction in traffic accidents. If this is true, then any traffic flow that is not moderately congested is likely to contain more crashes. Traffic flow changes as congestion increases and decreases, making density the most explanatory factor for the cause of traffic accidents. As the density levels change, the traffic flow compensates and the crash locations change.

Traffic flow accidents can also be identified with a look at the collision type. There are lane specific maneuvers that result in lane specific crash type patterns. Even though these accidents are regarded as a function of traffic flow, density plays a role in where the crash occurs, therefore, the number of vehicles involved can help determine if congestion within the traffic flow is a factor in NCT crashes. The goal of this first research question is to identify the types of traffic accidents common in North Central Texas by analyzing the number of lanes a roadway has, the collision type, and the number of vehicles involved. These variables should point to the specific lane a crash occurs in and if the crash locations are related to congestion issues.

Congestion levels may vary significantly throughout the normal course of a day. Morning and evening rush hours are predicted to be, in most cases, more congested than at other hours. Traffic flows according to three important elements: the speed of the vehicle, the density of the traffic, and the roadway function. As congestion decreases after rush hours, speed variances across all lanes increase while the traffic flow variances decrease. Golob et al. (2003) showed that the collision lane was a function of the traffic flow. If traffic flow is dependent upon the roadway design, and if the roadway design is based on the number of lanes necessary for transport from point A to B, the important question to ask is whether the roadway design (or roadway geometry) is a contributor to serious roadway accidents.

The third research question deals with the manner in which traffic flows during dark hours. The NCT crash dataset will be regrouped by light conditions and analyzed. Driving after dark creates conditions for traffic that are different from day time traffic conditions. Whereas density is the most explanatory factor for why crashes occur during the daytime, traffic flow (how the vehicles move through space) most fully explains why the crashes happen after dark (Golob et al., 2004). With the contributing factors for crashes reversing for nighttime conditions, it is possible that visibility or other variables are an important issue. If NCT nighttime fatalities can be attributed to low congestion levels and high flow, isolating the regions that experience the most crashes should reveal the types of contributing influences in the form of the numbers of vehicles involved and specific geographic variables that point to a sudden loss of driver control, if they are present in that area. The factors that explain nighttime traffic crashes should also be evident in the types of crashes that can be attributed to nighttime conditions. Single vehicle crashes with an object should be the predominant type of crash in dark, high flow traffic conditions. The results of testing will determine if traffic density or traffic flow is the most explanatory reason to significantly affect traffic crashes.

CHAPTER 2

STUDY AREA AND DATA

Study Area

The geographic boundaries of the study area begin slightly west of the 95th longitudinal and extend west to the 99th longitudinal. The upper and lower extents are the 31st and 34th parallel. North Central Texas (NCT) is a region comprised of 16 counties: Collin, Dallas, Denton, Ellis, Erath, Hood, Hunt, Johnson, Kaufman, Navarro, Palo Pinto, Parker, Rockwall, Somervell, Tarrant, and Wise counties. The study area encompasses approximately 20,592 km^2 (close to 8.2 million acres), and although a rather large area, the combined length of roadway accounts for 100,980 kilometers (about 923 km^2 of impervious surface, with approximately 9 meters width), or just under 230,000 acres (Figure 2).

The roadway geometry is divided into groups that are based on road type. The roadway types consist of public, paved, urban and rural roads that data wise are grouped into functional classes such as rural and urban principal arterials (including interstates and freeways), minor arterials, major and minor collectors, and local streets. Single, undivided roadways are coded differently than two lane divided highways and interstates with three or more lanes. Roadway crashes will be analyzed by the number of lanes in the roadway on which the crash occurred, the number of vehicles involved in the crash, the roadway design, and the collision type. The geographical data was downloaded from the North Central Texas Council of Governments (NCTCOG) website in shapefile format.

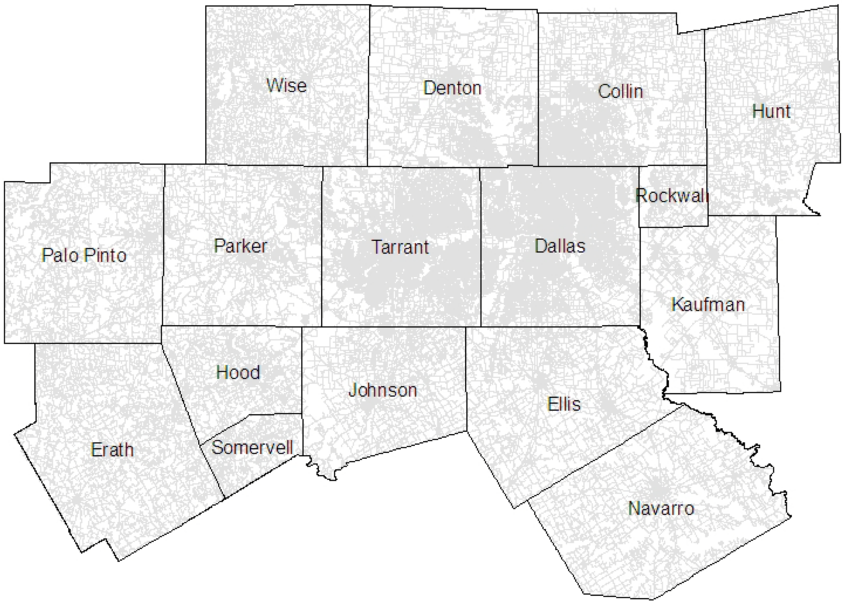


Figure 2. The study area of NCT contains the roadways of sixteen counties.

Data Sources

The crash data used to analyze the traffic flow were obtained from the Fatality Analysis Reporting System (FARS), part of the National Highway Traffic Safety Administration (NHTSA) and spanned the years 2001-2006. The FARS data consists of 10,000 records, of which, only 3,706 were actual crash locations. Each of the remaining records are associated with the fatality location but the records describe other aspects of the crash; for instance, information about other vehicle occupants, other vehicles involved, and the drivers of the vehicle(s).

The dataset used to test the road geometry of the region was gathered from the Texas Department of Transportation (TxDOT). The crash data contained any crash where a police record was filed from the years spanning 2003-2006. The severity of the crashes was listed as fatal, incapacitating, non-incapacitating, and unknown.

Crash data at the state level are not as easily accessible and do not contain as many variables associated with the crash. The data were collected from GPS units at the time of the crash in WGS84 and had to be projected in the GIS to State Plane 83, North Central Texas 4202, to match the road segment projection. The state data contain *x* and *y* coordinates, the county in which the crash occurred, the date of the crash, and crash severity but do not include the roadway type.

Data Preparation

Crash Data in ArcGIS

Crash data can be downloaded from public databases into an excel table and imported into a GIS as a data base file (.dbf). At the federal level (NHTSA) statistics are generally available for the last thirty years. Data with *x* and *y* coordinates is available from about 2001 to the present. The *x* and *y* coordinate data facilitate crash point placement because crash points lay accurately upon the roadway segments if the data are projected on the same coordinate system. FARS data contain enough variables to compare fatalities to ongoing crash research in other regions and only need to be organized into workable subsets of county crashes separated into years.

If the *x* and *y* coordinates are not included in the database, the crash database table must include a measurement field so that the computer system can match and "attach" the point data to the roadway. A point that designates each crash location will be visible as a separate layer on top of the roadway. Older traffic accident reports include a mile marker field as a location reference in addition to the roadway name. The shortcomings of this system of location identification are that placements of the crash points are dependent upon the measurement parameters set forth by the GIS user and the initial data recordings are relative, done by estimating distances between mile markers. This method results in approximations of crash locations. Additionally, not all crash records contain information in comparable detail.

The biggest challenge to displaying crash database tables geographically is the reconciliation of the crash descriptions in the crash database to the road database. Reporting agencies seldom use the same nomenclature. Although “Loop 12”, “Lp 12”, and “12” designate the same roadway, they are not the same to the computer. Standardizing the data is therefore a necessary part of preparation. If the traffic crash database includes an *x* and *y* coordinate field the process is simplified, not only because road name conciliation is not necessary, but also because it is more accurate. The *x* and *y* coordinate fields have become standard in traffic crash reports over the past five years. If the road segment map is projected in a coordinate system, the traffic crash locations can be layered on top of the road segments without having to perform any adjustment of the roadway file structure.

Roadway Layer in ArcGIS

The roadways are a basic element in this study and can be the most unimportant, yet most useful part with which to construct the study framework. Over the past ten years, the preparation of the roadway network has changed. A roadway file consists of numerous line segments; each with a name field. Each segment connects to a node that represents an intersecting street. The name field provides a means to combine records that share a common road name designation. Multiple segments with the same name designation can be combined into one record.

To reduce the number of records associated with the roadway database (the largest counties have over 85,000 entries), the road segments were prepared for analysis in ArcGIS by transforming the road segments into “routes”. ArcGIS has a linear referencing toolbox with a “Create Routes” tool that combines road segments of the same name. This step reduces the amount of street name reconciliation to crash data. An identification field, in this case the road name field, is required as input for the tool to work. The road map is important because it provides regional orientation but is also very helpful in recognizing one of the main points of this thesis: the difference between a two dimensional planar space and a one dimensional linear space. The roadways are also an important visual tool for isolating the road geometry changes that are responsible for some traffic crashes. In this respect, the roadways are very useful, but they are not part of all of the analyses because the tabular crash data contain the *x* and *y* coordinates, allowing for the measurements for analysis by the GIS.

The definition for traffic flow varies according to the agency analyzing the data. Traffic flow classifications that are used by FARS for fatal crash locations include urban, rural, grade, curvature, and roadway design (divided or undivided). The classes that state reporting agencies, such as the NCTCOG, may use to describe traffic flow classes are: Access Ramp, Major and Minor Arterials, and Primary and Secondary Highways. Analyses that are performed on crash data using the traffic flow variable in this study will use the term “roadway function” to prevent any confusion with the traffic flow research discussed earlier. The roadway function terms have been re-termed in order to clarify the difference between roadway structure, sometimes referred to as function, and traffic movements that have been previously researched and are referred to as "traffic flow."

Crash Data in SaTScan

State level databases are not as attribute rich and are not as easily assessable as federal level databases. The TxDOT crash data consist of all police reported traffic crashes from 2003-2006. SaTScan (Software for the Spatial, Temporal, and Space-Time Scan Statistics, v 8.1, 2009) is software that analyzes spatial and temporal clusters over space or time. It uses a Poisson or Bernoulli statistical model and measure distances in a scanning fashion. Preparing the TxDOT data for the SaTScan analysis was more difficult than for ArcGIS because the data was manipulated to mimic a one dimensional testing environment. The following flowchart (Figure 3) maps out the course of data preparation for the crash data.

The TxDOT crashes were too data intensive to process. The 393,000+ crash locations had to be aggregated into groupings before the Monte Carlo simulation could replicate the dataset. The dataset has to be replicated 999 times for the Monte Carlo simulation. A linear grid of .005 degree squares was placed upon the crash location extent, and a crash point per grid square count was taken. The .005 degree grid square in zone 14N equates to approximately 457 meters. Grid squares that partition the crash data facilitate analyses by decreasing the crash point input to a fraction of the original dataset. Each grid square was given a unique numerical designation. The crashes located within each grid square were counted and the total count per grid square was appended to the grid table as an additional field. The counts per grid square ranged from no crashes to 455.

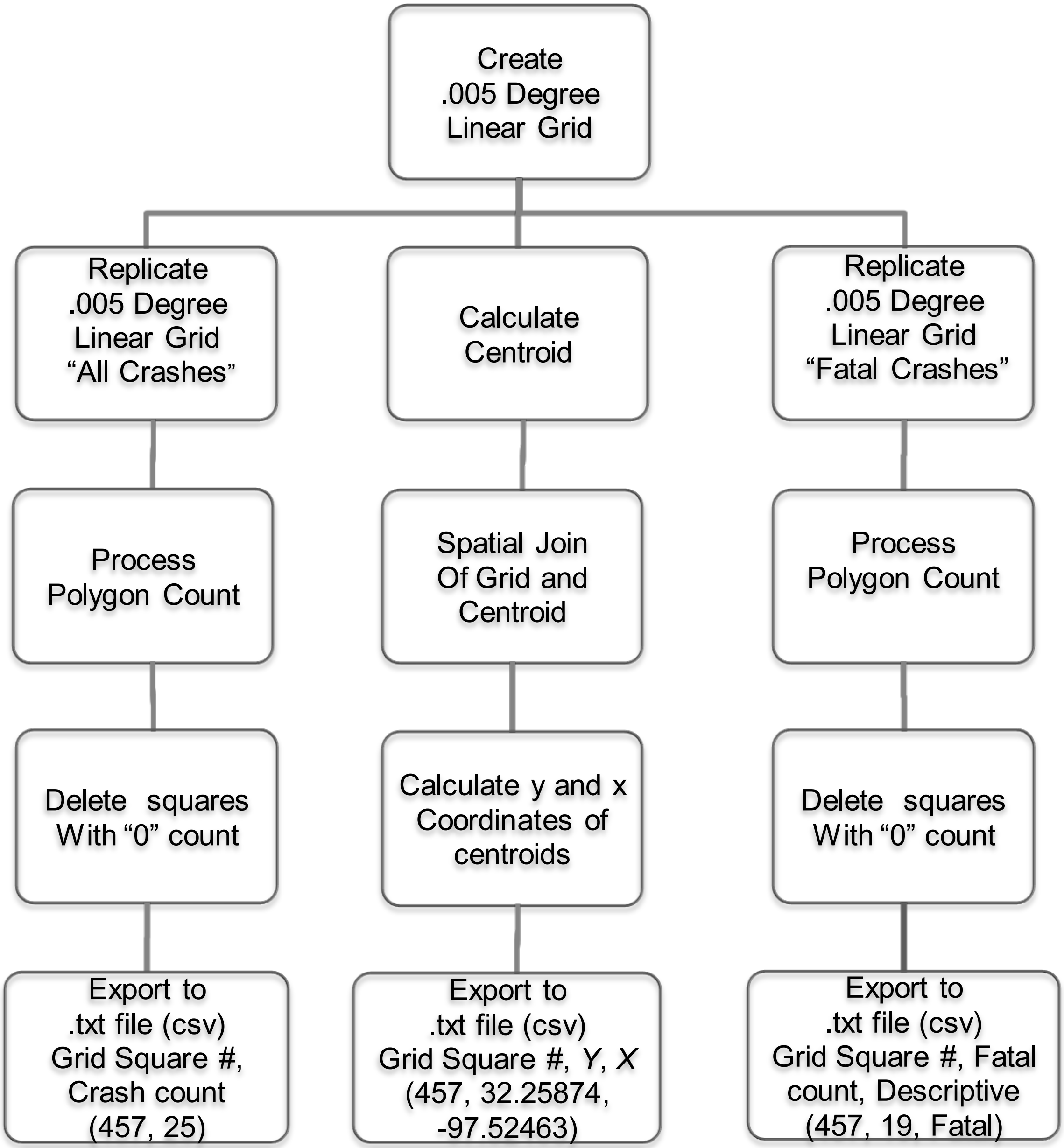


Figure 3. SaTScan data preparation included three identical .005 degree grids. Each grid square contained a unique identifier. The first grid was for the control crash group, the second for a coordinate file, and the third for the case, or fatality, group.

An identical .005 degree grid pattern was used for the fatality database. The grid squares were identical in number sequence as well as coordinate location. All grid squares in the crash and fatality databases that were without a crash point were deleted, leaving a patchwork of grid squares across the region (Figure 4). Each grid square's center location was calculated and designated the centroid. The centroid coordinate location was used to represent the location of all of the crashes per that grid square. Each dataset (crash, fatality, and grid centroid location) was converted to a comma separated value (.csv) text file.

The TxDOT crash file is the control file and includes a unique identifier - the grid square number before the deletion of "0" crash point polygons - and the crash point count of each polygon (457, 25). The fatal crashes are considered the case file and include the corresponding unique identifier, a polygon point count, and an optional description, in this case – fatality (457, 19, fatal). The case files are always a subset of the control file. The coordinate file - the grid centroid locations file - is identical to the control file except that the latitude and longitude location of the centroid replaces the polygon crash count (457, 32.25874, -97.52463).

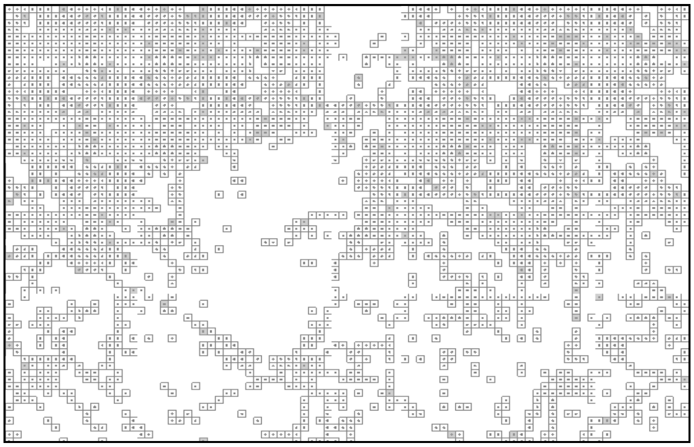


Figure 4. Grid square patchwork containing all crashes (□), fatal crashes (▨), and centroids (⊡).

Roadway Layer in SaTScan

The NCT roadways were buffered about 15 meters on each side creating a 30.5 meter wide roadway polygon layer that will replace the polyline road layer. A regularly spaced non random point array was created over the region at one 30.5 meter intervals (Figure 5). The point distance of the non random point array is based on the roadway buffer width, ensuring non random point placement within the buffer. Any point intersecting the roadway buffer was extracted to create a new layer, a grid file, to be used as a starting point for the distance measurement of the scan statistic.

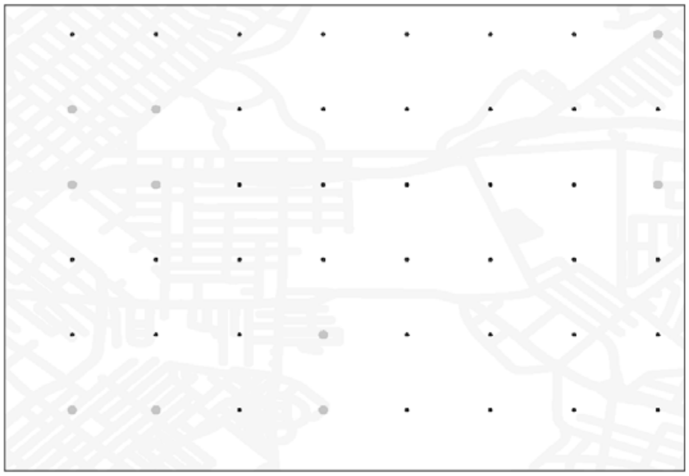


Figure 5. Non-random point placements (•), 30.5 meter roadway buffer, and highlighted (●) intersecting points.

The decimal degree coordinate location was appended to each point that intersected a roadway buffer and exported to .csv text file for a grid file that will be used in conjunction with the coordinates file obtained from the centroid location of each grid square. This preparation attempts to exclude unused space on the grid plane so that analysis emphasis is on the roadway. The roadway buffers more closely represent the actual study area than the open 20,600 km^2 of the NCT region. In theory, the crash groupings appended to the grid centroid are measured in reference to the non random point locations which will be used as the starting point of measure for each scan. Using a separate grid coordinate text file in addition to the crash coordinate text file compares the groupings of accidents to a regularly spaced, non random location.

The distance measurements were then averaged and compared. The scan statistic is programmed to use the Bernouli statistical model and identifies the three most clustered groupings of fatal crashes. It is a purely spatial statistic, and compares the clustering of fatal accidents, the case group, to the population of all traffic crashes, which is the control group. Additional inputs of the grid centroid coordinate file and a regularly spaced, non random grid file accompanies the fatal cases and the control group. This is, in reality, a two dimensional procedure altered to compensate for a one dimensional dataset.

Crash Data in SANET

Okabe's SANET (Software for the Spatial Analysis on a Network, 2006) is a toolbox program from the University of Tokyo that converts two dimensional spaces to a one dimensional space. Before the NCT fatal crashes from FARS can be used in SANET the study area has to be redesigned. It is a one dimensional statistical testing program that was developed to measure clustering of data that set upon a roadway grid pattern, such as roadway crashes or datasets that contain street addresses. The NCT fatality dataset was queried for light conditions to narrow the scope of study to a temporal level that will give a better understanding of congestion issues. Light conditions were specified in the dataset as light, dark, dark but lighted, dawn, and dusk. Organizing and controlling the data through a diurnal context allow the variability of flow conditions that influence driver behavior to be minimized and more carefully observed. The scale change is accomplished by extracting the relevant data; in this case fatalities grouped by light conditions, from the FARS NCT fatality dataset to determine the area where the most crashes are occurring. Both datasets are displayed in State Plane 1983, with the foot as the unit of measure.

The density outputs of Spatial Analyst were, by default, classed into 9 Jenks Natural Breaks interval classes. Those intervals were reclassified into 4 Jenks Natural Break interval classes, the tightest groups reserved for analysis. A polygon shapefile was created to cover the extents of the different dark related overlapping crashes. The shapefile was used as a clip pattern to isolate those fatalities from the rest of the NCT dataset. The new clipped fatality dataset contains 82 fatalities and will be manipulated by SANET tools in a transition from a separate dataset to one that is infused within the linear network.

Roadways Layer in SANET

SANET may be best understood in terms of a time line. Rather than assuming that location points "A" and "B" are static and independent of time, as in planar space, location differences in one dimensional space should be viewed as location changes that are consistent with time progression. For instance, two points A and B exist simultaneously on a grid plane and carry static coordinate variables that describe each location (96.4946, 33.53678) and (97.76324, 32.67540). In one dimension, the locations A and B are placed on a single line with the distance between them used as a description, easily defined in terms of a time vector that moves from one point to the next. The locations are separated only by the shortest distances that divide them. This is a network; progression from a starting point to an end destination following a pre-determined, linear sequence.

The roadway layer is an integral part of the analysis with SANET. The roadway is split into segments and later rejoined with the connectivity information attached. The study data is also attached after the roadway is assembled. Distance measurements are then made solely within the new network. The target area of interest is redesigned for use in SANET by creating a polygon shapefile in the shape of the most dense crash area to become a clip pattern for the NCT roadways (Figure 6a and Figure 6b). It is the same polygon shape that was used for isolating the fatalities within that area. The new clipped roadway will be manipulated into a one dimensional network space.

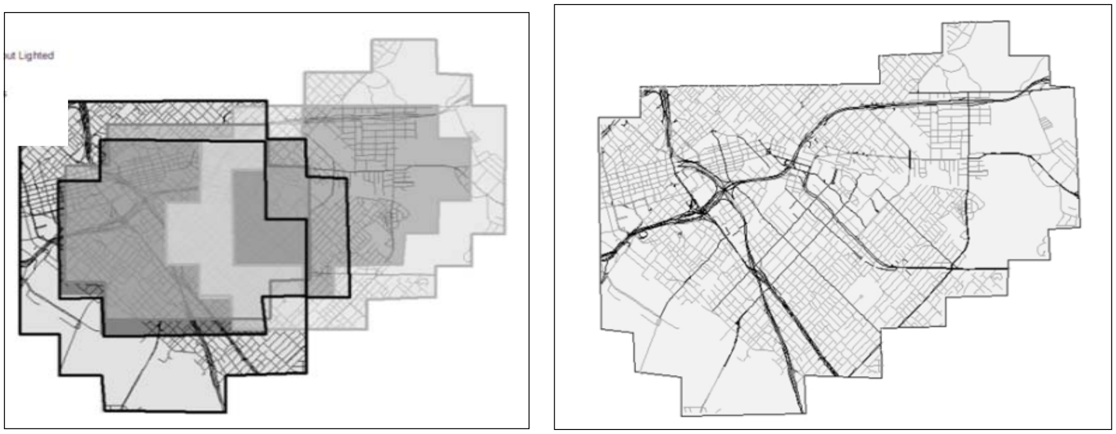


Figure 6a and 6b. The area of interest clip that isolates the roadways to be manipulated in a SANET prepared roadway network.

CHAPTER 3

METHODOLOGY

Standard Geographic Analysis

Because traffic flow has been shown to have correlations to traffic crashes, the crashes that experience similar traffic flow conditions should be tested for clustering. It is important to know that arterials with differing roadway functions have differing density capabilities. Also, as the road density diminishes in the outer sections of busy metropolitan areas, so too do congestion levels. Although disaggregated traffic counts are not available, the crash variables attached to the fatality crash database should provide clues to the congestion level and flow conditions at the time of the crash. The severity of the accident is a function of traffic density; the more severe the accident is, the more it is a function of density. Using a fatality database helps to narrow the scope of the study considerably: property damage and injury accidents are categorized separately because each group displays qualities characteristic of their placement.

Collision location is a function of traffic flow conditions, and traffic flow is dependent upon the roadway type in terms of its movement. Using the number of cars involved and the manner of collision behavioral variables will help verify other traffic flow research results, and those results will help determine if North Central Texas (NCT) exhibits the same level of predictability in traffic occurrences (Figure 7).

The test type must lead to the most accurate interpretation of the data. Changing the scale of the study area will influence the interpretation of statistical results. The traffic fatalities in NCT were divided into number of lane datasets to test for clustering of the known traffic variables that contribute to crashes.

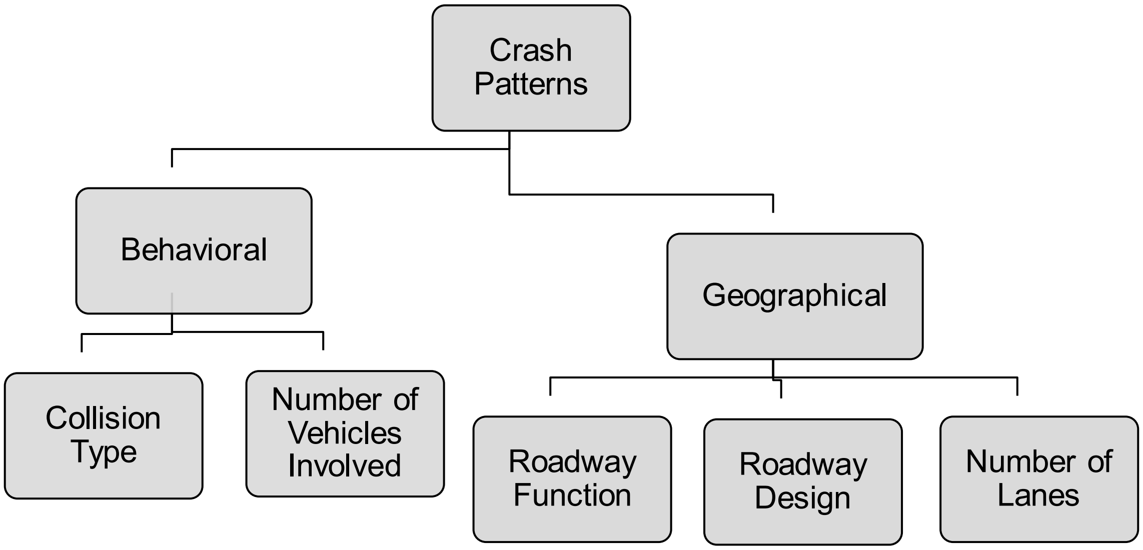


Figure 7. Crash pattern variables can be classified into two categories; a behavioral group and a geographical one.

A spatial statistics test, such as the Moran's I autocorrelation, will analyze the data for the correlation of specific variables that influence the crashes over an entire region. This spatial statistic considers both the spatial relationships of features and the values of attributes associated with the features to measure the clustering of the data. In a sense, this spatial statistics test looks for relationships that have been weighted by the attribute values, thereby adding importance to that variable, and will be used to find out if crashes that occur in the same lane are forming clusters. The test does not analyze individual crashes but all crashes in a group, in this case the number of lanes a road contains, to determine the level of spatial dependency that the crashes have to this traffic flow variable. The Moran's I spatial autocorrelation will later be used to analyze the correlations between the collision types, the roadway design, the roadway type, and the number of vehicles in a crash so that a comparison may be made to results with other vehicles experiencing the same conditions.

Dividing the North Central Texas (NCT) dataset by traffic variables changes the scale of the data and provides additional clues to the crash in terms of the traffic flow. The Moran's I autocorrelation will allow a literature based interpretation that compares NCT crashes to other regional studies. For a large study area with many crash points, this analysis is sufficient for identifying the trends present in NCT vehicle crashes. The Moran's I test is beneficial, even though it is Euclidean; but it only indicates if clustering is present. The test does not identify the location of the clusters.

To pinpoint the location of crash clusters, the Getis-Ord "Hot Spot" analysis, or Gi statistic, was used. The crash lane location should be identifiable with a look at what type of crash occurred, and the unintentional driver behavior traffic flow variables. The Getis-Ord "Hot Spot" analysis normalizes the crash data and scores the data based upon how prevalent a feature variable is. It is a measure of the intensity of the feature variable. If the resulting *z*-score is statistically significant, the most clustered groupings will have the highest *z*-score values. The Getis-Ord "Hot Spot" analysis visually displays where the normal distributions are as well as indicating the dispersed and clustered groupings by analyzing and measuring distances in terms of each of the feature's neighbors.

Modified Two Dimensional Planar Analysis

Population clustering is often measured with Euclidean distances from one data point to another, then deriving the square root of the average sum of the mean distances of each data point squared from the mean. The geographical area of study, or plane, is then divided into a grid and the data plotted onto the *x* and *y* axis of a

Cartesian coordinate system. Traditional point pattern analyses are designed to compare the data point densities to a normal distribution over planar space (an *x* and *y* axis). A fundamental problem involved when using planar statistical tests on traffic studies is that the vehicle movements, and the ensuing crashes, are constrained within the boundaries of roadway systems (Goodchild, 2000). Tests that analyze crashes in geographical space will identify clustering, but disregard the presence of the roadway structure. Traffic crashes cannot be compared to a normal population in the sense of a distribution spread across an *x* and *y* axis of geographical space, because crash locations are bound to this linear partitioning within the planar space. Standard statistical tests require adjustments to compensate for the boundary restrictions.

Golob et al. (2004) showed that crashes were a function of speed in all levels of traffic flow. While congestion is necessary to inhibit the speed associated with free flow, traffic flow predictions for one roadway type may not be suitable for different road geometries. Road geometry consists of urban routes and rural routes in addition to when the road curves or contains an intersection. In order to determine if specific road geometry promotes roadway crashes, the different road geometries are analyzed in terms of the number of crashes and their variables. For this phase of the research, a downloadable statistical program called SaTScan (Software for the Spatial, Temporal, and Space-Time Scan Statistics, v 8.1, 2009) is used to compare the number of traffic fatalities in an area to the total number of crashes. The scan statistic will be used to verify whether or not the different road geometries of NCT roadways exhibit significant differences in crash counts.

To analyze a crash distribution for patterns, a random distribution comparison must be made. One method used to overcome the difficulties associated with crash distribution is to adjust the sample population assumptions. The Monte Carlo simulation compensates the loss of a planar distribution by comparing the network distribution to a massive repetition of randomly distributed populations. The Monte Carlo simulation uses this repetition of randomizations to act as a comparison model to the observed dataset, often 999 repetitions for a .01 percent chance of error. The Monte Carlo simulation will provide a random distribution for this study in order to test crash locations for clustering. The simulation is used frequently in traffic analysis as well as other network constrained analyses. In effect, the sheer number of data points that will compare with the network distribution makes up for the spatial variations common in network densities.

A scan statistic often times uses a bounding circle or rectangle, the size determined by the user, to position over the study area in an overlapping, or scanning, fashion. The data within each bounding shape is measured and compared to a control group. The statistical test used for this dataset is a probability model, the Bernoulli statistic, and offers a "cluster" or "no cluster" (q = 1-p) result, sometimes referred to as a Boolean, or yes/no statistic. The result will indicate the three most significant cluster locations (p < .001) where traffic fatality clustering has occurred most often amongst all recorded police crashes.

Distance measurements from the Monte Carlo randomizations are made from point to point and averaged with the other 998 randomizations (plus the observed dataset). Once all of the bounding shapes are analyzed within the Bernoulli model, the data that form the tightest clusters are identified. The dataset is clustered if the observed data points are more numerous than the expected data points. Expected data points are data that are predicted to be at a level that is derived from the rest of the data scans.

One Dimensional Linear Analysis

Preliminary cluster tests show that the number of single vehicle crashes after dark is significant. Locating the most clustered areas of night time accidents should be done to determine the reasons why. Density tests for traffic crash locations will of course coincide with the roadway densities, but once the area is isolated, the density clusters within the structural framework can be tested for factors other than what is expected from daytime traffic flow crashes. The roadway infrastructure consists of adjoining roads that increase and decrease in density as demand for road space increases and decreases; usually road networks become less dense as the area expands outward from busy business centers. The roadway network consists of interstates, highways, roads, and ramps; both urban and rural. Technically, the linear roadway system is a one-dimensional system; in order to accurately compare the locations of crash points on a linear system to randomness, the comparison model must also be one dimensional. Planar analysis assumes homogeneity of objects in two dimensional spaces (Xie and Yan, 2008); hence the perceived clustering of accidents when compared to a random distribution of objects. The formulae distance measurements are in the form of Euclidean distance measurements and have to be reworked in order to derive a more accurate measurement of data points on a linear surface. Once a data point is positioned in a planar field, the number of locations available for distance measurements is greater than on a network. A planar space may encompass additional data points that are beyond the same distance as measured on a network space, making for a higher mean data point count. The same number of data points in an unbounded area leads to a perceived clustering because of chance placements that fall within Euclidean distance, but beyond the network distance. Crash points on a roadway or other datasets that are set atop street grid patterns require a different type of analysis other than the traditional statistical analyses.

A network is an interlocking system of pathways and a very important concept in traffic studies. Analyses associated with networks adhere to the constraints of the roadway, but the roadway system is not used in a dynamic way. The primary building block of networks is the use of decision making choices and processes that are programmed at the nodes; speed limits, one ways, and traffic light information are some examples. Datasets that are constrained by this underlying structure must use the structure as a means of travel from one point to another. The restriction of following the underlying structure from one point to another is called the shortest pathway. Network procedures are based on topology and the standard uses for it consist of routing, allocation, and finding the shortest pathways. These procedures allow course plotting from some starting point to an end point, usually by following the shortest path along the line segments. The shortest pathway is in contrast to Euclidean distances in which distances are measured by a straight line between points A and B. In the best scenario, linear analysis should be applied to road networks if crash patterns are to be picked up and recognized as a pattern that is bound by the roadway structure and not a distribution that is independent of this underlying structure. The network tests available in ArcGIS are really not suitable for this type of static pattern analysis either.

Currently, many spatial cluster tests seem to suffice for populations that are attached to an underlying framework, as is evidenced by the proliferation of analyses performed with them (Aguero-Valverde et al., 2006; Eksler et al., 2008; Flahaut et al., 2003). Basically, the traditional planar tests have been used to measure a different type of scale problem, one consisting of a dynamic system set atop a static one (Goodchild, 2000). This problem is exemplified when roadway geometry is an issue (Figure 8). Overpasses pose a particularly interesting situation in which many crash points may appear to occupy the same space. Most two-dimensional planar tests will not separate these two populations. In reality, the data points consist of multiple, separate populations. Roadway systems' grid and movement patterns should necessarily be considered when isolating traffic crash patterns. The system is a dynamic one, only the network is dynamic and the crash points, static. A planar test will measure distances from one crash point to another, but will not produce an accurate reflection of the traffic flow.

One attempt to achieve linear results with a planar test is to rework the distribution of planar space so that the data is organized into a format that facilitates measurement of crash points. Linear referencing and segmentation will supply a metric that standardizes the data. Essentially, line segments are cut into equal lengths, each with two end nodes containing relevant information about location and connectivity. After the new lengths are created, the segments are rejoined into pathways. Next, data are attached to the roadway in relation to the new roadway metric. This allows a point location reference on a linear surface, but the analysis associated with this system does not measure the distances along the two directions in which the measurement is needed, to the right and left of the data point.

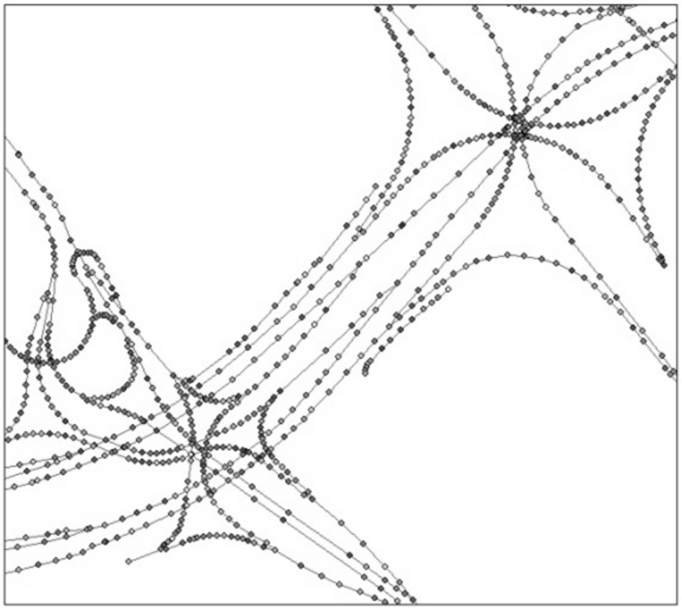
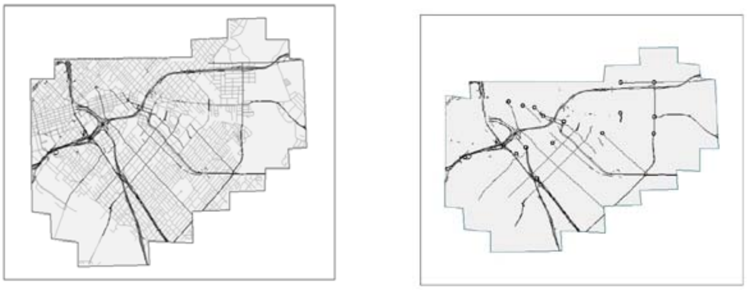


Figure 8. Overpass line segments in Dallas and the nodes that direct the many separate populations. The populations travel in differing directions and on different levels.

One dimensional network tests are planar statistical tests reworked to be a variation of their counterparts. The one dimensional tests require that the test site be preprocessed. This will include some type of linear segmentation, inserting nodes at the endpoints, and rejoining the segments with connectivity features assigned to the nodes. The test area is then called a network. The distances used for one dimensional network analysis are measured along the linear distances, following the roadway path, rather than a standard Euclidean distance. The analysis is a process where point distances are measured along the roadway and compared in two directions; on either side of the initial data point, measuring only along the roadways within the connected structure. Yamada and Thill (2004) demonstrated how the planar *k*-function over detected the occurrences of cluster patterns in a distribution. The network version of the *k*-function test was more effective in detecting cluster patterns because it did not over detect the patterns as did the standard *k*-function. Xie and Yan (2008) described a transition from the planar kernel density estimation (KDE) to a network version using the linear reference system (LRS) process called segmentation. The planar KDE definition uses area references such as r^2 and π. The network version of the KDE was defined as the point density at a location being equal to the sum of the observed point distances multiplied to one divided by the radius times the weight of the point that is usually a ratio of the distance of the point over the radius. The ratio provides recursive "distance decay." As the ratio changes, so does the weight of each data point. Although one dimensional network versions of standard statistical tests prove to be more reliable than their planar counterparts, the tests are not readily available.

The network *k*-function is similar to the standard *k*-function tests in that it recursively measures a distributed set of points, but the network version is an extension of the planar version in that the distances between points that are distributed along a network are measured within that network. The network *k*-function definition uses division by "point density per line segment unit" rather than "point spread intensity" to measure the mean number of points within a user specified distance. The network *k*-function distance for the observed set of crash data points is equal to the total length of the network divided by the number of crash points multiplied by the number of crash points minus one (for the one degree of freedom lost due to estimation). The quotient is then multiplied by the sum of the point distances times the sum of the point distances that have been multiplied to the indicator function, which is also recursive. If the distances of the points are less than the distance of the network, the indicator function multiplies the point distances by one. If the distance is greater than the network distance, the indicator function multiplies the point distances by zero. The *k*-function is determined, or is returned as “0.” It is a recursive binomial process and plugs results of the function back into the equation until a result is achieved.

SANET requires implementing a series of processes in order to set up the testing environment. Each process produces an output that is used for the next process. First, the roadway segments have to be cleared of any pseudo lines and overhangs. Once cleared, the remaining line segments are split at intersections, turns, and curves, and then capped on either end with nodes (very similar to “segmentation”). The separate segments are then recombined into a continuous graph, the pathway connectivity checked, and each flanking node numbered and provided with identifying information associated with the segment and adjacent nodes. The information is then set into an output table. An adjacent nodes table contains the identification along with the length of the segments between the nodes. A numerical designation is added to the end of that table to represents the direction in which the segment lies in relation to the node; labeled as either forward (1) or backwards (-1) of the pathway node. The binary directional component within this dynamic system provides a route for the static dataset, and enables placing of the random dataset upon the roadway boundaries during analysis. An additional text file records the input and output file names. To minimize the amount of input segments and nodes, the roadways that did not have a fatality data point were deleted (Figures 9a and 9b).



Figures 9a and 9b. The clipped roadways contain the Dark, Dark but Lit, and Dusk fatalities before and after the deletion of roads without crash data.

The fatality data is now added to the segment lengths. The resulting output attribute table includes a new identification number for the fatality, an *x* and *y* coordinate, a pointer to the adjacent node table, and a pointer to the reference table. The attribute table also assigns a “1” if the crash data point is inserted as a new point or a “0” if the crash point is inserted directly on top of an existing node. SANET determines the closest reference node and assigns the crash variables data to the node location on the roadway (Figure 10).

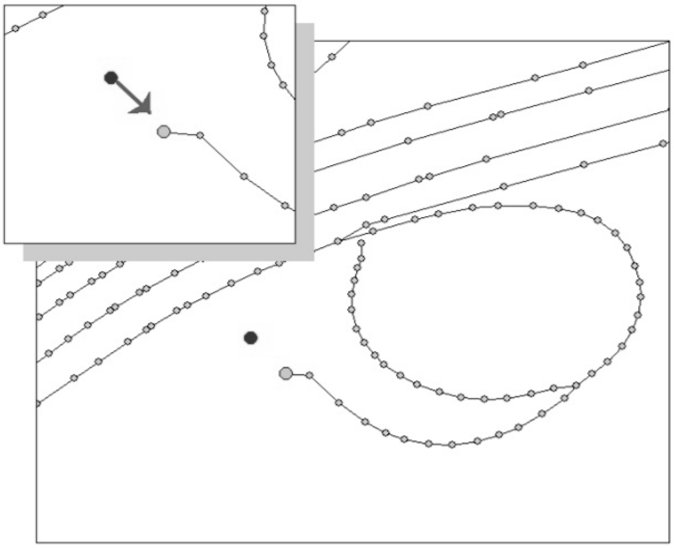


Figure 10. Fatality (black) and roadway segment nodes (gray). The large gray marker is the new location of the fatality after it has been attached as a “new” point, as opposed to an “existing” point.

There are three new associated tables created as output. The point reference index contains the original file identifications for points and nodes. The network index attribute table contains the adjacent node information, segment lengths, and segment direction to the node. The reference database file contains the identifications of the original crash points and the inserted crash point direction to the node: “1” is on the left, “2” is the right, and a “0” means that the inserted point was directly on the node. The final output file (Light_Condition_v) contains 4,731 nodes, which includes 68 inserted fatality points (Figure 11). The dark, dusk, dark but lighted fatality dataset contained a total of 82 fatalities, with 14 fatalities inserted on top of existing nodes. Note that the resolution of the dataset prohibits viewing the 82 fatalities in one area of interest map. Many crashes occupy an area that is similar in location to other crashes, istinguishable only at larger scales.

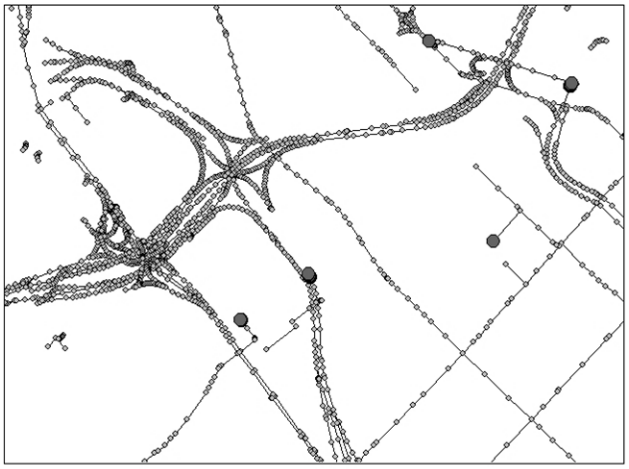


Figure 11. Final _v file with the 82 original fatality points (large, gray dots) added for visual clarity.

Two datasets were analyzed in SANET. The Darks dataset, which includes the dark but lit and dusk fatalities, and the daylight dataset were compared for clues to the underlying factors of the crash. The extracted area of interest is different for the two datasets, even though the geographical coordinates are the same. The difference is in the data and how the location conforms to the data. Network analysis enables a tailoring of the geography to circumstances. In this case, there are aspects of the geography that are taken out because the network does not connect and is unnecessary. The resulting environment looks and behaves differently (Figures 12a and 12b).

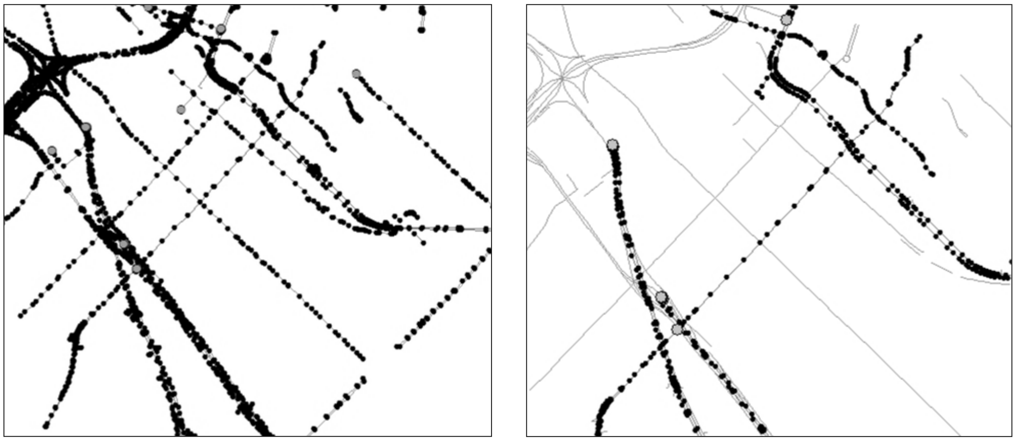


Figure 12. (a) Darks dataset with roadways that contain a fatality. (b) Daylight dataset with the Darks streets in light gray added for emphasis. The two areas are identical in location, but provide different study areas for analysis.

CHAPTER 4

RESULTS OF ANALYSIS WITH DISCUSSION

Predictability of NCT Crashes

Analysis of the North Central Texas (NCT) crash sites attempts to align itself with other research analyses that show that traffic flow significantly affects the location of crashes. The NCT fatalities were initially tested for clustering using weighted variables that are known to contribute to crash statistics. The results are provided in Table 1.

Each of the geographical variables (road design, road function, and traffic lanes) is clustered, each exhibiting a probability of error value of $p < 0.001$. These geographical variables should automatically be dismissed because of the natural tendency of two-dimensional pattern analysis to over-detect cluster patterns of traffic variables that are specifically location based. For example: road function is clustered because the roads are divided into urban and rural designations. Urban roadways are much closer together and are designed to transport a greater number of vehicles. The crashes that occur on urban roadways will therefore have less distance between each crash location. It nonetheless makes sense to analyze the geographical variables in order to follow the effects it has on any dependant variable.

The behavioral category consists mostly of traffic flow variables that are associated with congestion, while the geographical category consists of location based variables mostly out of the drivers' control. In order to accept or reject the null hypothesis that NCT crashes exhibit predictable characteristics that are specifically related to traffic flow, the collision type and the congestion level should meet the criteria shown previously to be indicative of crashes known to occur during certain types of traffic situations.

Table 1

Moran's I Geographic and Behavioral Variables

	Collision Type	Number of Vehicle	Day of Week	Hours/ Rush Hours	Road Design	Road Function	Traffic Lanes
State	Cluster	Cluster	Rand	Cluster/ Rand	Cluster	Cluster	Cluster
Moran's Index	0.080	0.074	0.003	0.035/ 0.056	0.350	0.327	0.244
Z-score	8.957	7.922	0.235	3.728/ 1.404	37.131	34.696	25.972
p-value	0.001	0.001	0.814	0.001/ 0.160	0.001	0.001	0.001

Traffic flow (the manner in which traffic moves from one location to the next) has been shown to influence the crash lane, while congestion has not only been known to affect the number of vehicles involved in a crash, but also to affect the severity of the crash. If collision type and density characteristics can be shown to match the criteria of other traffic flow crash research, a reasonable interpretation is that NCT crashes are most likely the result of traffic flow.

To prove this, the geographical variables went through a scale change so that a better observation may be made about the crash distribution. The fatality dataset was divided into traffic lane subsets: two lane, four lane, and six lane subsets. These subsets cover a wide range of NCT roadway types. The lane datasets were then tested for clustering by utilizing the known behavioral variables in this new geographical scale.

A Global Moran's I autocorrelation was run on the three individual datasets using, once more, the manner of collision as the weighted cluster variable. Results from the Global Moran's I autocorrelation indicate that there are similarities within lane groups and that there is clustering depending upon the number of traffic lanes involved. These results were expected and the tests were based upon a Golob and Recker's 2002 traffic flow study. The global Moran's I autocorrelation results also indicate that the manner of collision clustering are statistically significant (Table 2).

Table 2

Moran's I Manner of Collision

	2 Lanes	4 Lanes	6 Lanes
Moran's Index	0.039	0.338	0.227
z-score	2.039	13.172	.940
p-value	0.041	0.001	0.347

The Moran's Index is a score that rates the clustering of the dataset. The score ranges from negative one to a positive one. A score on the positive side of zero indicates clustering with a "+1" indicating complete clustering. A score on the negative side of zero indicates dispersion with a "-1" being completely dispersed. A "0" is a dataset with neither clustering nor dispersion. The value of each of the *z*-scores indicates how far the normalization of the collision type dataset distribution falls outside of the standard deviation of a normal collision-type dataset. The *p*-value is the probability of error that the dataset clustering or dispersion is randomly distributed.

Roadways that contain two lanes, each lane intended for travel in the opposite direction, displays some very slight clustering (Moran's I = 0.039) with a normalized *z*-

score of 2.04, within an accepted 2.59 standard deviation. The *p*-value of 0.04 is acceptable under most situations that require a 5% probability of error. The four lane roadway dataset contains a more credible clustering score of 0.338; the normalized *z*-score of 13.17 is clearly demarcated as outside of the normal standard deviation, and the probability of the sample being randomly distributed is less than .001 ($p < 0.001$). The four lane roadway dataset was set aside for further testing because of its better indication of clustering than the two and six lane roadway datasets. The six lane roadways test results did not show any significant clustering.

If the traffic crash patterns are a function of traffic flow and traffic flow is dependent upon the roadway function, the crash location should be predictable based upon the state of the traffic flow at the time of the crash. Variables that are associated with congestion often include the number of vehicles that were involved in each crash (because it can be an indication of how dense the traffic is). We know that the level of congestion has an effect on the crash statistics: crashes involving a single car decrease with increasing congestion, and crashes involving multiple cars, increase. The second test of the global Moran's I autocorrelation was to analyze the correlations between the number of lanes a roadway contains and the numbers of vehicles involved in the crash (Table 3).

Table 3

Moran's I Number of Vehicles Involved

	2 Lanes	4 Lanes	6 Lanes
Moran's Index	0.231	0.068	0.292
z-score	10.664	2.167	1.245
p-value	0.001	0.030	0.213

The two lane roadways and the four lane roadways datasets show that some clustering was present when using the number of vehicles involved in a crash as the weighted variable. The two lane dataset indicated light clustering with a Moran's I score of 0.231.

A high normalized *z*-score of 10. 66 with the probability of error that the sample is randomly distributed was less than .001 ($p < 0.001$). The four lane dataset had a lower Moran's I score of 0.068. The normalized *z*-score of 2.17 was within a normal standard deviation of the population, and an acceptable probability of error value of $p = 0.03$ chance that the sample was randomly distributed. The six lane dataset did not have any statistically significant clusters.

Test results of lane datasets using the collision type as a weighted variable were different from the test results of the number of vehicles involved when it was used as a weighted variable. Specifically, the *z*-score for two lane roadways with the number of vehicles involved were well outside the accepted standard deviation, whereas the two lane roadways with the collision type as a weighted variable were not. The *p*-values for each were different as well. The two lane roadways with the number of vehicles involved were statistically significant with a *p*-value of less than 0.001 while the collision type weighted dataset probability of error results were higher at $p = 0.03$. The six lane roadways datasets did not indicate any significant clustering. These results indicate that the roadway function (and more specifically, the number of lanes that a roadway has) affects the two variables differently.

Because the traffic variability in speed may be measured across the lanes and contributes in lane specific crash patterns (Golob et al., 2004) the rush hour/non rush hour variable was tested in order to compare which collision types were prevalent. Variability changes with congestion, therefore, the collision type should indicate the lane that the crash occurred in. The number of crashes occurring during rush hours was used as a weighted variable in the lane datasets (Table 4). Unfortunately, while the fatality dataset using crash hours as a weighted variable tested as clustered, the dataset tested random when divided into rush hour and non rush hour subsets. A clustered result would have meant that collision type was indeed lane oriented.

Table 4

Moran's I Rush Hours

	2 Lanes	4 Lanes	6 Lanes
Moran's Index	0.004	0.013	0.106
z-score	1.580	0.442	0.478
p-value	0.114	0.659	0.633

The roadway design is the most generalized variable because the roadway structure lends itself to urban layouts in much the same way that it does with rural ones. An undivided roadway is undivided wherever it is. There are roadway designs, however, that are more common in rural areas than urban ones. For this reason, the lane datasets with a roadway design weight is the most revealing. The two lane dataset had a Moran's Index of 0.653; the most clustered so far (Table 5). The *z*-score of 30.19 was far outside of the standard deviation, with a p-value of less than .001 (p<0.001). Also of note is that the four lane dataset contained a higher statistically significant secondary score than the second cluster group of the other lane datasets. A slight Moran's Index of 0.141949 and a *z*-score of 4.374 standard deviations with a probability of error value of less than .001 chance of being randomly distributed (*p*<0.001) qualifies as an additional clustered grouping. What this means is that the geographical variables weigh heavily in crash pattern analysis and scale changes contribute in uncovering additional information.

Table 5

Moran's I Road Design

	2 Lanes	4 Lanes	6 Lanes
Moran's Index	0.653	0.142	0.236
z-score	30.109	4.374	0.974
p-value	0.001	0.001	0.330

The Getis-Ord Gi "hot spot" analysis shows geographically where the significant crashes are. The analysis contains two additional fields appended to the end of the dataset table with the normalized *z*-score and *p*-value. A query separates the normalized dataset (based upon the *z*-score) into significant and non-significant clusters (based upon the *p*-value). Based upon the Moran's I autocorrelation results when the manner of collision was the weighted variable, the four lanes dataset was used in the Getis-Ord "hot spot" Analysis. There were not any crashes that were not within the standard deviation of a normal distribution, and there were not any that were statistically significant. As for the two lanes dataset that had the number of vehicles involved used as the weighted variable, there were statistically significant clusters.

The roadway design weighted datasets also had statistically significant clusters in the two lane dataset when using the Getis-Ord "hot spot". The two variables (number of vehicles involved and roadway design) could suggest a common tendency toward crashes overlapping at any location, but it is important to note that any overlap is in the form of the location variables, the overlapping crash designations would depict the same accidents. What is clear is that crashes involving more than one vehicle are occurring more often on the two lane roads at that location because of the roadway design than multiple vehicle crashes on other two lane roadways. Congestion as well as roadway design seems to so far play a role in explaining NCT crashes. If traffic crash patterns are dependent on the roadway, then changing the roadway design should impact the manner of crashes along it. Looking at three major roadway designs used in NCT, cluster tests were performed to identify the worst roadway types for crashes.

Roadway function is planned with capacities in mind. A potentially higher vehicle capacity dictates wider roadways that contain more lanes than roadways designed for fewer cars. Roadway function is also divided into rural and urban subsets. Roadway design is different from roadway function in that roadway design consists of divided and undivided roadways, entrance and exit ramps, and one way features. Some divided roadways contain barriers and some contain median strips, some contain both. Rural areas often have more undivided roadways, NCT included. They are designed to move vehicles from point A to point B. In other words, the roads are designed to move vehicles from peripheral locations to center business districts. Urban roadway design has more divided roadways, with and without barriers that are designed to move and direct many different types of traffic within the central district. These distinctive roadway purposes produce very different crash patterns (Figure 13).

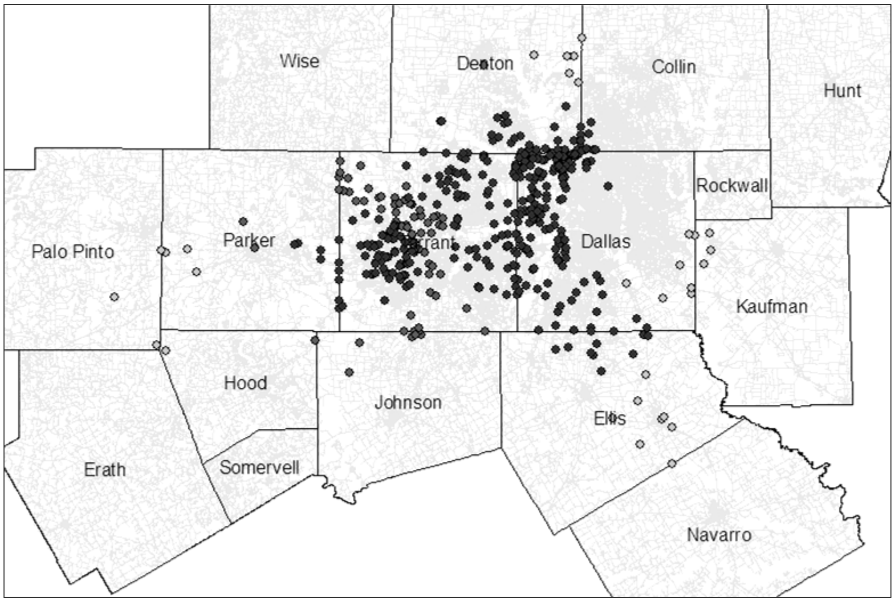


Figure 13. Each colored grouping represents a significant fatal crash cluster distribution after the dataset has been divided into three major roadway design categories (undivided, divided with a median, divided with a median and a barrier).

Summary statistics taken from the datasets that were queried for significant clustering indicate that the collision types most often occurring are collisions that do not involve any other vehicle. Specifically, 58% of the crashes occur with no other vehicles. If research statistics reveal that the collision type is an indication of the lane on which the vehicle was traveling, then most of the crashes in NCT occur off of the road. Almost 10% of the NCT vehicles that were involved in a rear end collision would be on an outer lane. 16% of the NCT vehicles were head on collisions. 3% of the NCT vehicles are involved in a side swiping accident (Figure 14). The remaining class of collision type is other and consists of entrance/exit ramps and one way streets.

NCT roadways are split almost down the middle where roadway design is concerned. Divided roadways account for a slightly higher percentage of the roadways, although urban area roadways contribute more to that percentage because of the density in their numbers. Urban divided roadways are double the urban undivided roadways count; whereas rural undivided roadways are quadruple the rural divided roadway count.

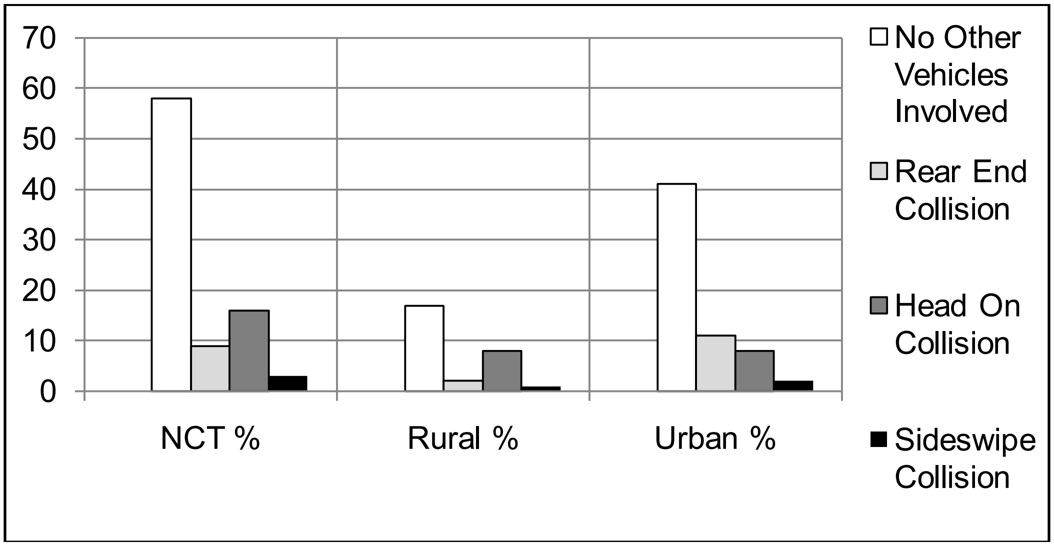


Figure 14. Percentages of fatal crash types on NCT roadways as a whole, and divisions of fatal crash types on urban and rural roadway designs.

A rear end collision (front to rear) is the second most commonly occurring accident for urban roadways accounting for more than 10% of the (fatal) crashes. The urban areas have almost twice as many divided roadways (39%) than undivided ones (20%) while the remainders of the crashes are located on entrance/exit ramps and one ways. Head on collisions account for 8% of the crashes and most often must occur on undivided roadways. Sideswipes are only 2% of urban crashes even though urban areas contain more of the lanes that are built for capacity. In the rural areas, the rear end collision occurs less often (2%) than a head on collision (8%).

A sideswiping accident rates third in all accidents involving more than one vehicle in all three of the categories: urban, rural, and total NCT. Interior lane crashes are predicted to be side-swiping accidents. If NCT roadways contain five or more roadway lanes, then any crash that occurs in one of the middle lanes should most often be a sideswiping crash. This design configuration includes roadways with a middle left hand turn lane. On two lane NCT roadways, the second lane would be the opposite lane and involve head on collisions. In every case, vehicle crashes not involving another vehicle outnumber the other collision types, no matter which lane or road type they are on.

The Moran's I autocorrelation test result of the NCT region indicated that the manner of collision only clustered on four lane roadways, although the Getis-Ord Gi "hot spot" analysis resulted in a no significant cluster finding. The number of vehicles involved in a crash resulted in a clustered pattern only for the two lane roadways, amounting to an inconclusive level of congestion within the metroplex. The overwhelming number of one vehicle crashes in NCT is a variable that does not seem to fit well within other traffic flow research results.

Based on these findings, it seems that NCT crashes cannot be predicted simply by analyzing the congestion level and collision type. This may be a result of NCT opting for a moderately congested traffic flow and density level that will favor a reduction in traffic accidents, but more likely, these influences are in addition to the road design carrying a lot more of the burden than is being recognized in transportation research.

Road Geometry Issues as a Cause of Crashes

To measure randomness, statistical tests begin with an assumption of homogeneity. This means that the tests assume that the environment is the same throughout the testing area. To ensure that the testing process is as unbiased as possible, all testing parameters should be controlled.

The SaTScan scan statistic reveals a different view of NCT than the Global

Moran’s I and “hot spot” analyses. The SaTScan test results indicate significance in traffic fatalities in terms of all the reported traffic crashes in the region. The results are based on quantitative data and are purely spatial. The interpretation makes some assumption, homogeneity, but the interpretation is based on results that are not dependent upon any weighted variables. The SaTScan test reveals the three most significant (*p*<.001) locations where traffic fatality clusters have occurred between the years of 2003-2006. The results are from the SaTScan Bernoulli statistical model (Table 6).

Table 6

SaTScan Results

Cluster	Grid Sq.	Latitude	Longitude	Loc #	*p* value	Obs	Exp
1	21,557	32.9855	-98.2465	3835	0.001	241	97.7
2	176,767	33.3155	-95.8815	1996	0.001	135	42.4
3	138,840	32.2805	-96.4565	3859	0.001	238	126

This probability model offers a “cluster” or “no cluster” (*q* = 1-*p*) result and indicates that the significant Texas Department of Transportation (TxDOT) reported crash location clusters are outside of the metroplex. Although the density of all crashes is highest within the metropolitan area, the proportion of fatal crashes is low compared to the great amount of traffic there. The location of the most clustered fatality group originates from a grid centroid in Palo Pinto County, northwest of the metroplex proper, at north latitude 32.981 and west longitude -98.230. The expected fatality count for the number of crashes in the cluster is 98. The expected fatality count is derived from measurements in the scan taken over the area of the NCT region. The observed fatality count in this cluster was 241; that count makes this area the most clustered in terms of crashes that were fatal. The probability of error that these fatalities are normally distributed is $p < .001$.

Driver behaviors here have been assumed to be equally present in areas where traffic crashes are not as densely distributed as the higher density crash populations. Traffic roadways continue outwardly from busy business districts in diminishing numbers, but the roadways still carry the basic traffic flow variables of volume, density, and speed and are designed with local populations in mind, as are higher density areas. The TxDOT data contain only a few variables such as date, location, and severity of the crash making the FARS dataset a much richer source and one that can be easily queried for additional information. The FARS crashes surrounding the TxDOT cluster centers will be used to detail the characteristics that are generally known to be indicative of traffic crashes (Figure 15). A look at the fatalities within a 24 kilometer radius of the cluster midpoint reveals several interesting points about NCT crashes.

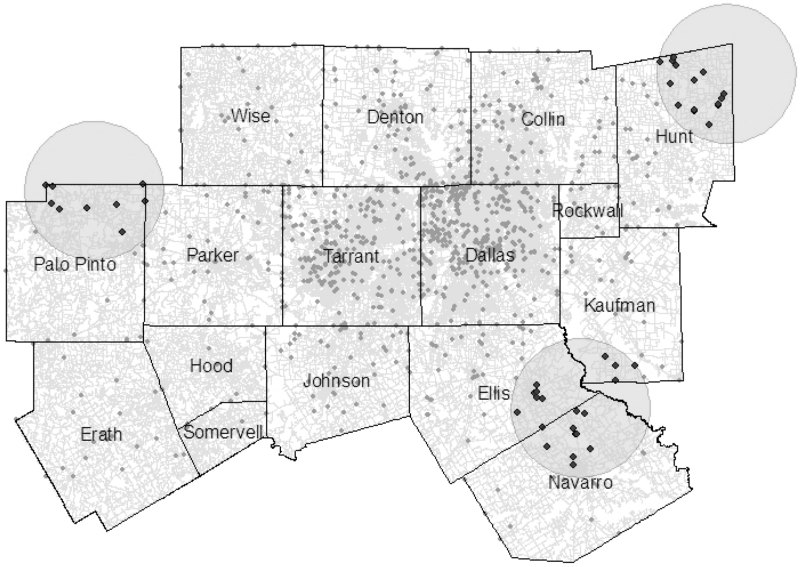


Figure 15. The SaTScan Bernoulli location results of the three most significantly clustered fatality groups in the NCT region.

The SaTScan results indicate clustering in rural areas. The clustered areas reveal the unique geographical differences between the cluster groups and the rest of NCT. The NCT region is comprised of roadways that lead toward the central business locations of each county. Urban roadways in NCT account for 69% of the fatalities in the FARS dataset. The roadway clusters that the SaTScan statistic reveals are reversed. Cluster 1 contains roadways that are 89% rural and 11% urban. Cluster 2 roadway types are 70% rural and 30% urban, and cluster 3 is 80% rural, 20% urban.

NCT roadways may be grouped into two categories. The first category consists of urban roadways. Urban roadways consist of principle and minor arterials, collectors, and local streets. Rural roadways consist of the same variables, only the location and capacity are rural. Each of the roadway types can be divided into categories that describe the flow conditions and density issues of NCT. Flow variables explain the roadway in terms of the number of lanes a roadway has and the geometric road design. The geometric design details the space the roadway consumes. Undivided, divided with a median, divided with a median and a barrier, and one way streets are road design variables. Density may be explained by the number of vehicles in each crash and the collision type. The manner of collision reveals if the driver had to make an unexpected and sudden stop (rear end collision) or if there was no warning to process (head on collision). Except that single vehicle collisions account for more than half of all crashes, the collision type of the clusters is quite different in each of the clusters, as well as in comparison with the whole of NCT.

NCT collision type totals are predominately crashes that do not involve another vehicle. The “no other vehicles” category accounts for 58% of all collisions in the NCT fatality dataset. Once NCT is divided into subsets of rural and urban categories, the rural roadways account for 54% of “no other vehicles” crashes, while the urban roadways account for 60% of “no other vehicles” crashes. The three SaTScan clusters are rural, but mimic the urban and-rural collision types only in that the biggest category of collisions is the “no other vehicles” category. Out of the three cluster groups, Cluster 2 most closely resembles NCT statistics. After that, the clusters differ even amongst themselves. Cluster 1’s “front to front” category is the second largest category. Cluster 2’s second largest most often occurring collision type is front to side, which includes sideswiping crashes, and Cluster 3’s second largest categories are almost equally divided with front to front, front to rear, and front to side (Figure 16).

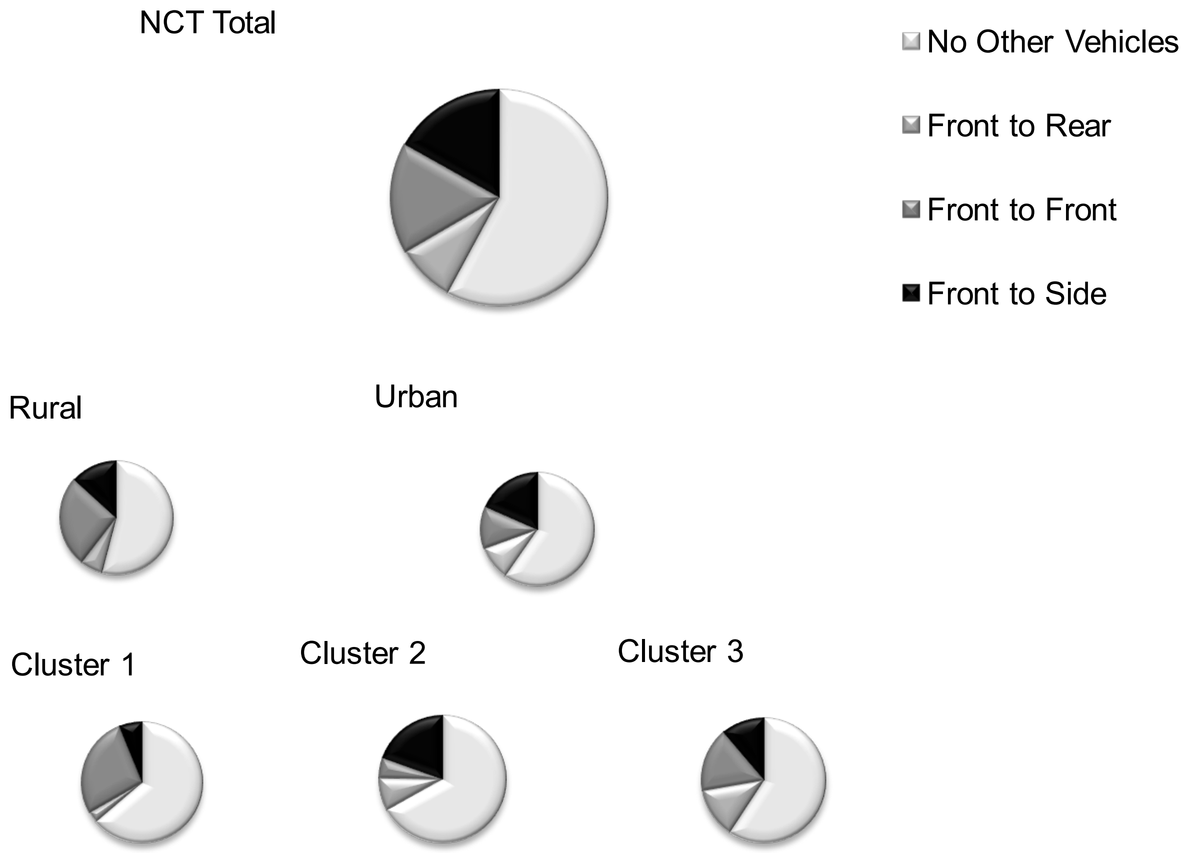


Figure 16. The collision type variable divided by regions: the NCT region as a whole, rural roadways, urban roadways, and Clusters 1-3.

The second category in the density group deals with the number of vehicles involved in each crash, but because the manner of collision category contained 58% of “no other vehicles”, the “number of vehicles involved” is 1.

The flow variables of road design reveal yet another snapshot of NCT (Figure 17).

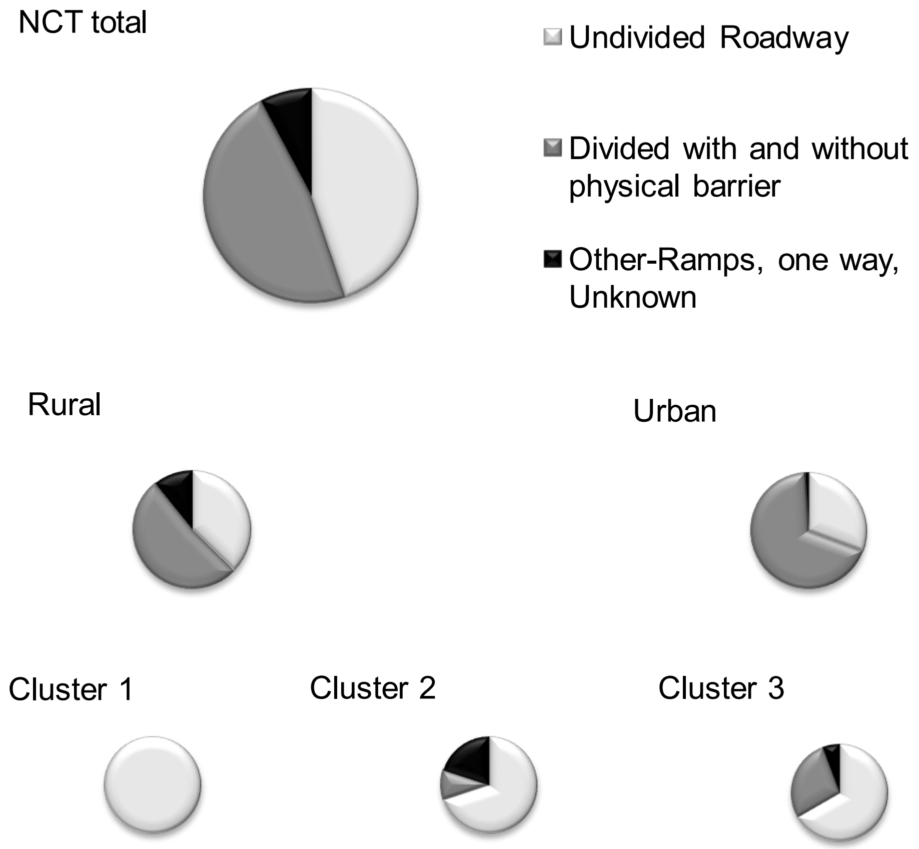


Figure 17. The roadway design type variable divided by regions: the NCT region as a whole, rural roadways, urban roadways, and Clusters 1-3.

For NCT, divided roadways with a median strip or a median strip with a barrier, make up the biggest part of where collisions are happening (48%). Undivided roadways account for 45%. A big part of these statistics are the differences between rural and urban roadways. Each of these categories pulls the overall totals toward the other, leaving the average looking like about half. The urban roadway network is 68% divided roadways, while the rural network's divided roadway count is slightly over one half. The urban undivided roadway accounts for only 31% of urban roadways crashes, while the rural undivided roadways account for 37% of the crashes on rural roadways. There are substantially more crashes on rural ramps also.

The clustered areas of the SaTScan offer a departure from the parent classifications. The road functions varied, but roadways in the three clusters that were not physically divided were the dominant roadway type. The traffic ways not physically divided were in the more rural part of the county areas. Each cluster is overwhelmingly tilted toward undivided roadways. In fact, 100% of Cluster 1 fatalities occurred on undivided roadways. Cluster 2 crashes occurred on undivided roadways 70% of the time, and Cluster 3 fatalities accounted for 67% of the fatal crashes. These figures are in direct contrast with the parent classifications because urban roadways, with 69 % of the roadways, are weighting the NCT averages.

The other flow variable that explains flow crashes is the number of lanes that a roadway supports. With the urban roadway type variable out of the averaging, the cluster percentages more closely resemble the "rural" roadway type statistic. Cluster 1 fatalities are exclusively on two lane roadways. Cluster 2 is made up of crash locations occurring on two lanes 78% of the time; Cluster 3, 89% of the time (Figure 18). Road design in NCT very much point in the direction of bearing responsibility in roadway crashes. Undivided, two lane roadways are almost a hallmark of rural areas. Cluster 1, for example, has 100% of the traffic fatalities on undivided, two lane roadways.

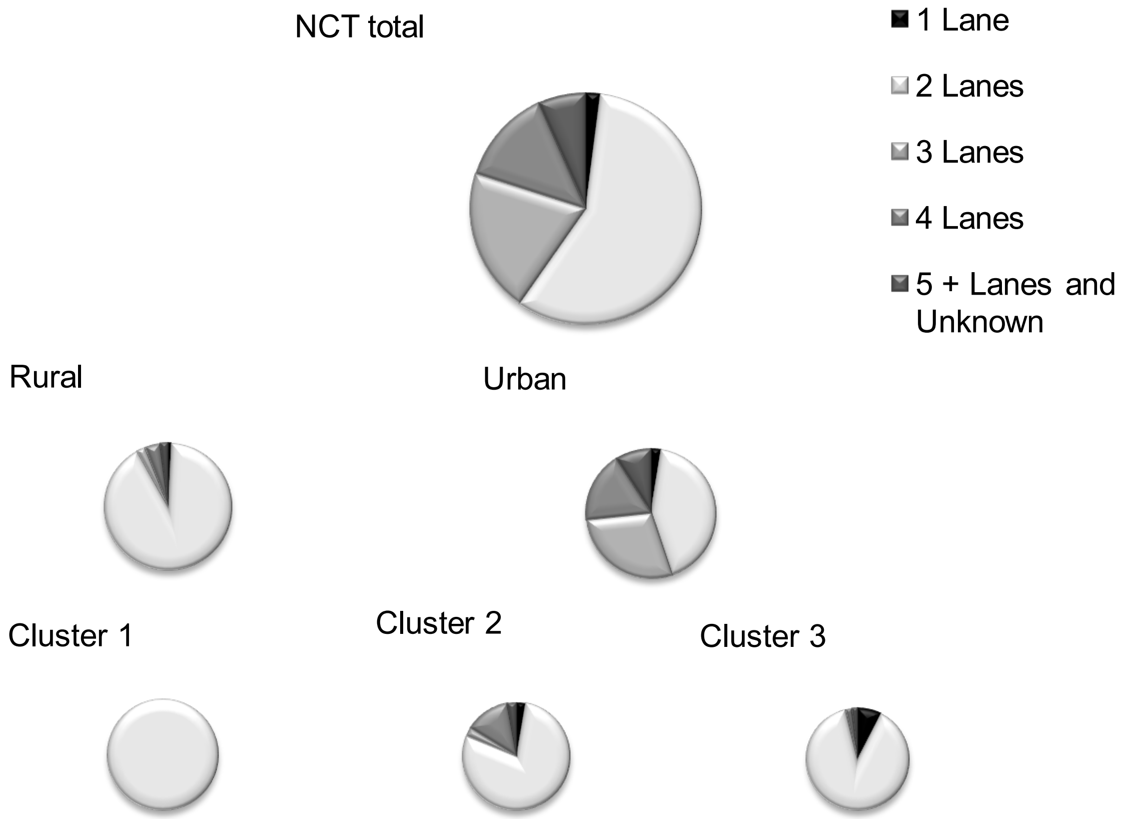


Figure 18. The number of roadway lanes variable divided by regions: the NCT region as a whole, rural roadways, urban roadways, and Clusters 1-3.

This cluster was also measured as the first most clustered spot for traffic fatalities in NCT (Figure 19). The fatalities in rural areas should be a function of the undivided highway (traffic flow) not congested rush hours (density). The severity level indicates low density and high speed. If the crashes occurred in rush hour, there would not be conditions of free flow.

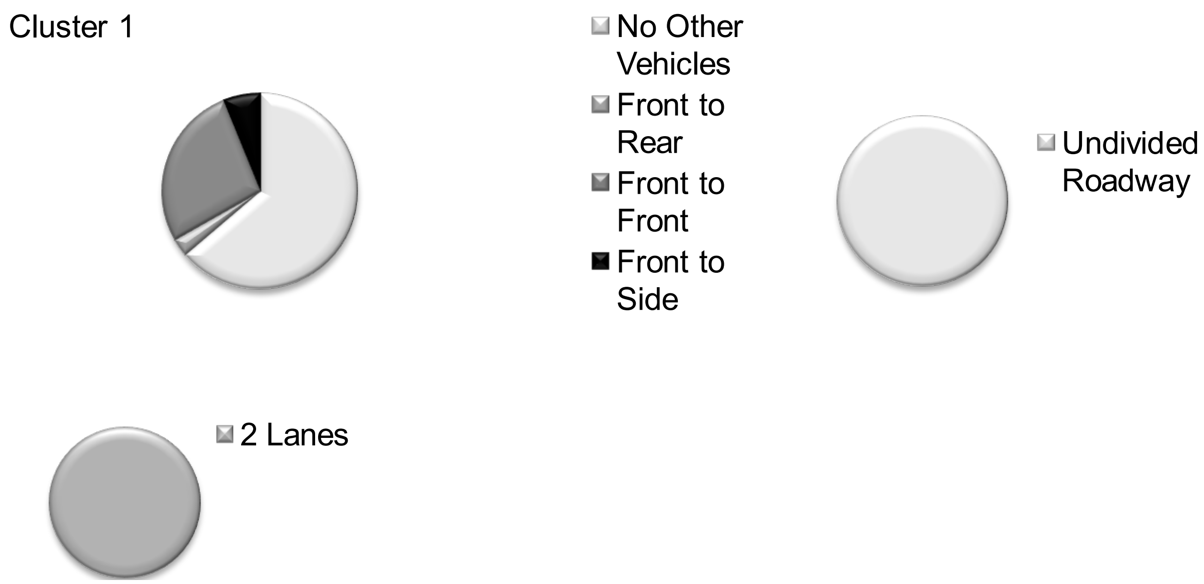


Figure 19. Summary of the results: Cluster 1 data show no resemblance to the parent categories of NCT, urban, and rural roadways, reiterating the reasons for fatality clustering.

Daylight versus Dark Crashes

The preliminary planar Ripleys *k*-function results for fatalities occurring after dark were as expected – significantly clustered – as was the Moran's I autocorrelation test. Density tests were then performed on dark, dark but lighted, and dusk fatality datasets using ArcGIS Spatial Analyst. The density tests spatially analyzed the crash locations to indicate where the specific light condition crashes are the most densely populated. The densest accumulations of crashes invariably occur where the roads are the densest, but the light condition parameters provide a look into areas that may be prone to crashes that are not density driven. The FARS dataset is useful here because of the richness of the data and the severity of the crashes has been controlled. The datasets were eventually pared down to the largest overlapping areas of the dark, dark but lighted, and dusk crashes.

A one dimensional analysis requires a major manipulation of roadway segments. The focus of the SANET test will be to determine if the crashes occurring after dark are clustering and whether or not the clustering is due to flow variables. The Ripley's *k*-function is a test that measures data in multiple distances and provides a normal range for data through the Monte Carlo simulation, and then categorizes the data as normal, dispersed, or clustered. The Ripley's *k*-function bases the designations on how the data behaves throughout the multiple distances. SANET tools rejoin road segments that have been previously segmented and labeled to create the network space required to ensure that sample simulations are placed along the linear surface. The placement of the samples is important so that the resulting network analysis is independent of planar references and is without the standard Euclidean measurements. The points are distributed only along the roadway, by way of the tables that were produced as output during the preparation stage. The tables contain an "either-or" instruction code. It is a binary reference (0, 1) that the simulation will adhere to during the analysis process.

In every *k*-function test performed on the darks and daylight datasets, the planar results were more clustered when compared to the linear tests. That is not to say that they were more accurate than the linear ones. Looking at the tables associated with both tests, both contain a field that stipulates the distance that is used to base the computation. The distance of 141 meters was used for both models. If the starting point of measurement is one data point, the 141 meters spreads outwardly in all directions on the planar application. The possible locations that can be used to place data within this area are higher than placing data within a linear boundary. The reason is that the planar *k*-function has a lot more room to spread a sample, and in turn the clustering seems to be more apparent (Figure 20b). If a sample is restricted within a boundary, the simulated sample is tighter and a cluster is harder to recognize.

The *k*-function cluster tests were performed on the dark, dark but lit, and dusk fatality dataset as well as the daylight fatality dataset. Each dataset contained at least thirty crashes to satisfy statistically normal parameters of a sample. The results reveal clustering, but in different ways. The overall objective of these tests will be to determine if flow variables are responsible for the differences in the daylight and dark data.

The results from the linear *k*-function consist of two output tables. The expected values table lists the distances in increments of 141 meters and contains all of the expected results: the upper and lower confidence envelopes, the mean of all data points, the number of segment lengths analyzed, and the cumulative totals in each 141 meter distance band. The observed data points are in a separate table and also contain the cumulative totals. The linear *k*-function results provide a field that is not found in the planar analyses tables and it identifies the number of segment lengths contained within each distance measurement. That is because a planar space is not prepared in the same way and does not take into account the roadway system. SANET uses methods that are in line with other roadway segmentation methods except that the SANET program uses the roadway as the geographical extent.

In our extracted area of interest, as the distances increase from each analyzed fatality, the average number of roadway segments increase before decreasing (Figures 20a, 20b, and 20c). This could be a result of the roadway types increasing and feeding into the infrastructure, or the roadways may contain many curves and intersections, and so also, the number of roadway segments. The graph indicates that the number of roadway segments peak in the darks dataset with 2,757 segment lengths within the 14th distance iteration, or at 1.25 miles from the fatality, where the number of roadway segments then begin to decline. The graph indicates that most Darks crashes are occurring outside of the densest levels of roadway segments, but within close proximity of high roadway densities or perhaps a high rate of roadway type transitions.

There are 45% less crashes in the daytime at this location, than after dark. The daylight crashes are in areas that do not contain many roadway segments or are in areas that seem to transition smoothly from one roadway type to another, evidinced by the fact that the road segments hover at this segment number level for a considerably lengthy distance. The Daylight crashes peak at 157 segment lengths, but fluctuate around that peak for 20 distance iterations, or 2.82 kilometers (Figure 20b). The number of segment lengths begins to drop afterwards toward zero. The average of fatal crashes occurs very close to the peak of roadway segment levels. The number of segment nodes in the Darks dataset climbs rapidly, reaching over 2,500 compared to the Daylight dataset's 157. From the graphs, it appears that there are more crashes that occur where road densities are at their lowest levels if the crashes occur in the darker hours than when they occur in the daylight.

It is important to separate the meanings of roadway densities and the densities of roadway segments, because we are not looking for factors associated with roadway densities, but for above normal distributions (clustering) of data within the area of interest (Figure 20c).

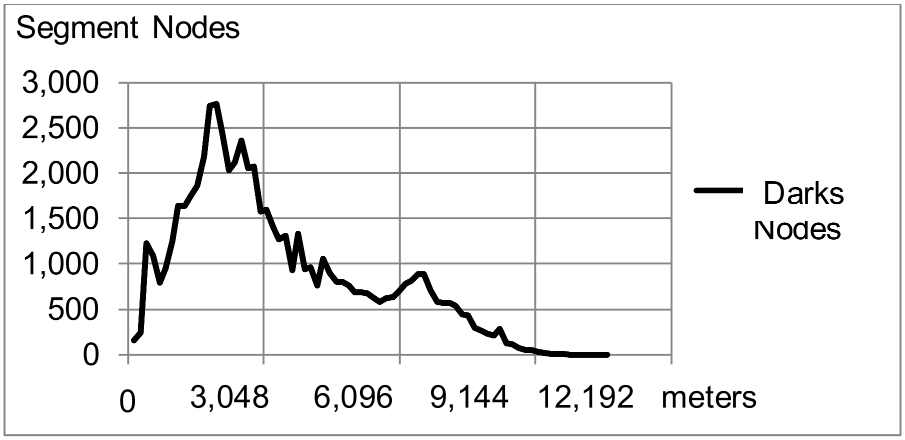


Figure 20a. Darks dataset segments nodes and their behavior.

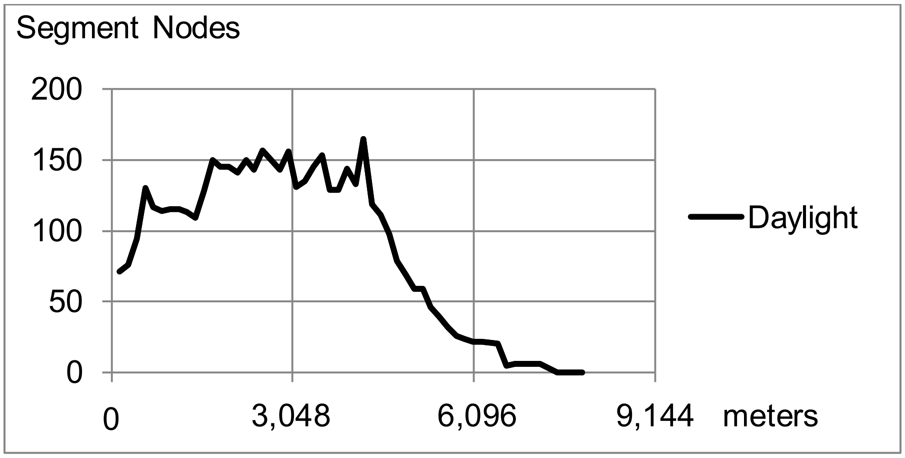


Figure 20b. Daylight dataset segment nodes and their behavior.

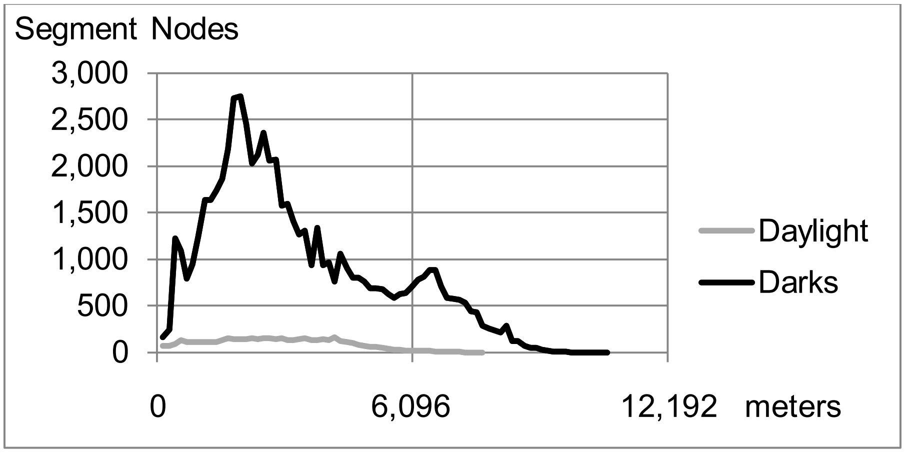


Figure 20c. Darks and Daylight dataset's segment nodes and their behavior.

The darks dataset *k*-function results show an above average, clustered crash distribution. The Monte Carlo simulation distributes a number of random points based upon the input value of the dataset onto the study area of interest. The random points are iterated 999 times per distance calculation. The largest spread of data points between the expected mean and what was observed was at a distance of 2,828 meters (2.82 kilometers). The observed result was an increase from 1,350 data points at the starting point of 141 meters from the fatality. The rise encompassed 19 distance iterations to peak, to then decrease over 63 distance iterations to a low value of 761. The expected cumulative mean at the height of the data spread in the Monte Carlo simulation was 2,090. The highest cumulative observed data points was 4,742.

The positioning of the observed dataset over the upper 5% confidence envelope is what determines if a dataset is significantly clustered or not. The darks dataset indicates significant clustering throughout each of the distances tested. The most clustered distance is at 1,131 meters (1.1 kilometers) from the fatality data point. This is also the distance that contained the most data points, irrespective of cumulative totals. The figure for observed data points at this distance iteration only is 1,246. It is located at the first peak on the graph (the cumulative result being 2,816). The distance decay is evident as the observed totals move closer to the normal 90% range of data (Figure 21).

The daylight dataset results are different from the darks dataset and only indicate significant clustering within the beginning distances tested. The daylight dataset had only thirty-seven fatalities, but enough to assume normality within the sample. The Monte Carlo simulation replicated the sample with 37 random data points 999 times within each distance interval. The daylight dataset contained an above average, clustered crash distribution, but it looks very different from the darks dataset (Figure 22).

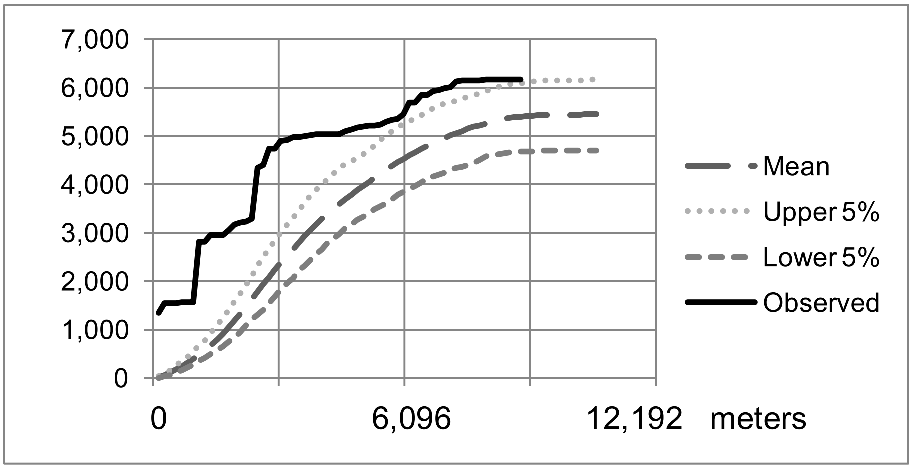


Figure 21. The one dimensional *k*-function cluster results on the Darks dataset. The data are clustered throughout the distance measurements.

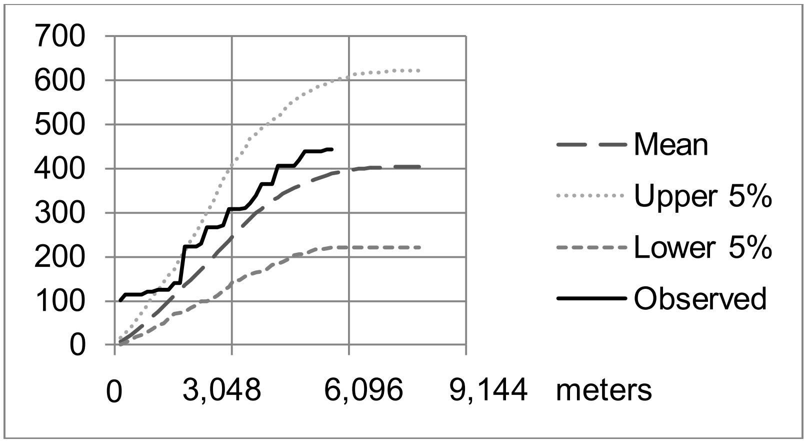


Figure 22. One dimensional *k*-function cluster results on the Daylight dataset. The data are not clustered throughout the distance measurements.

The largest spread of data points between the expected mean and what was observed was at a distance of 928 feet (less than one quarter of a mile average) and was 99.6 data points. The lowest spread of data points between the expected mean and the observed was 16.94 data points at a distance of 5,568 feet(slightly more than a mile) and at only the twelfth iteration out of forty. The differences between the observed dataset and the Monte Carlo upper 5% envelopes indicate a lot of fluctuation in the point densities, but overall, it seems to fall right above the mean and within normal statistical parameters.

The planar *k*-function test embedded in the ArcGIS was used to test the same two datasets for clustering. The resulting outputs contained the upper and lower 5% confidence envelopes and the data point differences between the observed and expected data point counts. The planar *k*-function tests revealed clustering within the extracted NCT area of interest.

The test began at each crash site where 999 random simulations were spread outwardly throughout 10 concentric distance bands to be measured for behavioral clustering. As the distances increased, the averaging of 1,000 data points consistently revealed clustering. Ultimately, the results will be used to determine if flow variables are influencing the differences between daylight and after dark roadway crashes.

The differences between the observed values and expected values were greatest in the first of 141 meter distance bands (Figure 23).

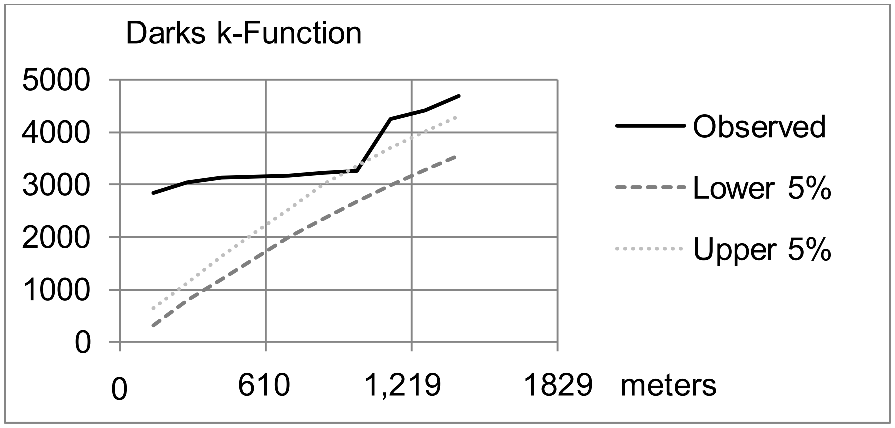


Figure 23. The Darks dataset Two dimensional *k*-function

The observed data point count was 2,842 data points, for a difference of 2,378 data points from the upper 5% confidence envelopes. The observed value decreases with increasing distances until the 7th distance band (1,348 feet), the distance is minimal: 12 feet from the upper envelope in the normal distribution range. The difference values spike slightly at distance band 8 before leveling out.

The Daylight dataset differences between the observed values and expected values were greatest also in the first of 141 meters distance bands. The observed data point count was 3,033 data points, for a difference of 2,569 data points from the upper 5% confidence envelopes. The clustering behavior levels out with increasing distances to distance band 6, about 0.8 kilometers from the crash site, where the dataset becomes a part of the normal distribution (Figure 24).

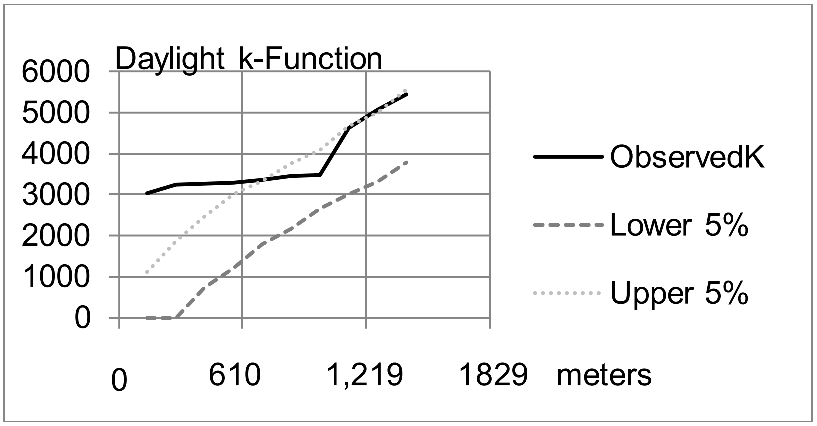


Figure 24. Daylight Two dimensional dataset changing from clustered to normal as distances are increased.

The planar *k*-function results are similar to the linear *k*-function results in that the Darks dataset is more clustered than the Daylight dataset and that the clustering occurs over longer distances. The data within the *k*-function results tables reveal the differences between the two types of cluster tests. Comparing distance iteration by distance iteration suggests that the planar *k*-function does, in fact, over detect clustering patterns. For example, the planar *k*-function distance iteration 5 (707 meters), has 3,371 data points compared to the samelinear *k*-function distance iteration data point value of 114 (Table 7). Both datasets have the same number of fatalities that were used in the analysis. The difference is that the planar *k*-function spread the random simulation of 37 data points 999 times over an area of 707 m^2. The probability of a random data point in a space having at least one big distance measurement (skewing the mean) within the area are a lot higher than the probability of a data point position on the network having an unusually high distance measurement. A bounding rectangle is used in the planar *k*-function test, so theoretically at least one distance measurement of one-half of the distance band distance is possible from any starting point of the iteration.

The answer to the first question as to whether or not the after dark dataset is clustered, the answer is yes, the after dark dataset is clustered. The second question is whether or not the crash clustering is due to flow variables and if the flow variables are responsible for the differences in the daylight and dark data. Flow variables consist of roadway design and the number of lanes each roadway contains. Isolating these variables will give a better understanding of how the data behaves differently. Roadway design is the aesthetic end of how the roadway is laid out on the land; if the roadway will carry vehicles in opposite directions without a divider, how egress methods join the main structure, whether all roadway on a grid are one-ways.

Table 7

Planar versus Linear Results

		Distance (m)	Observed Mean	Expected Mean	Upper 5%	Lower 5%
Daylight	Planar	707	3371	2320	3344	1802
	Linear	707	114	41.47	72	22
Darks	Planar	707	3167	2320	2523	1999
	Linear	707	1570	239	352	164

The data set is completely urban, in contrast to the three rural clusters that carry the most significant fatalities, and there are differences in after dark and daylight crashes. There are more crashes that occur after dark than in the daylight, and the dark dataset fatalities occur more often on roadways with a physical barrier, or without a barrier at all. Egress methods carry a higher percentage of the after dark crashes also. There are exceptions in the distribution; roadways containing a median, or “space” between lanes, carry a lower percentage of the after dark crashes than the daytime crashes. The number of lanes a roadway has makes a difference in how many crashes occur there, but the differences between dark or daylight advantage is small. Three lane roadways take the brunt of the crash statistics with nearly 50%; while four lane roadway crashes are minimal at less than 10% (Figure 25).

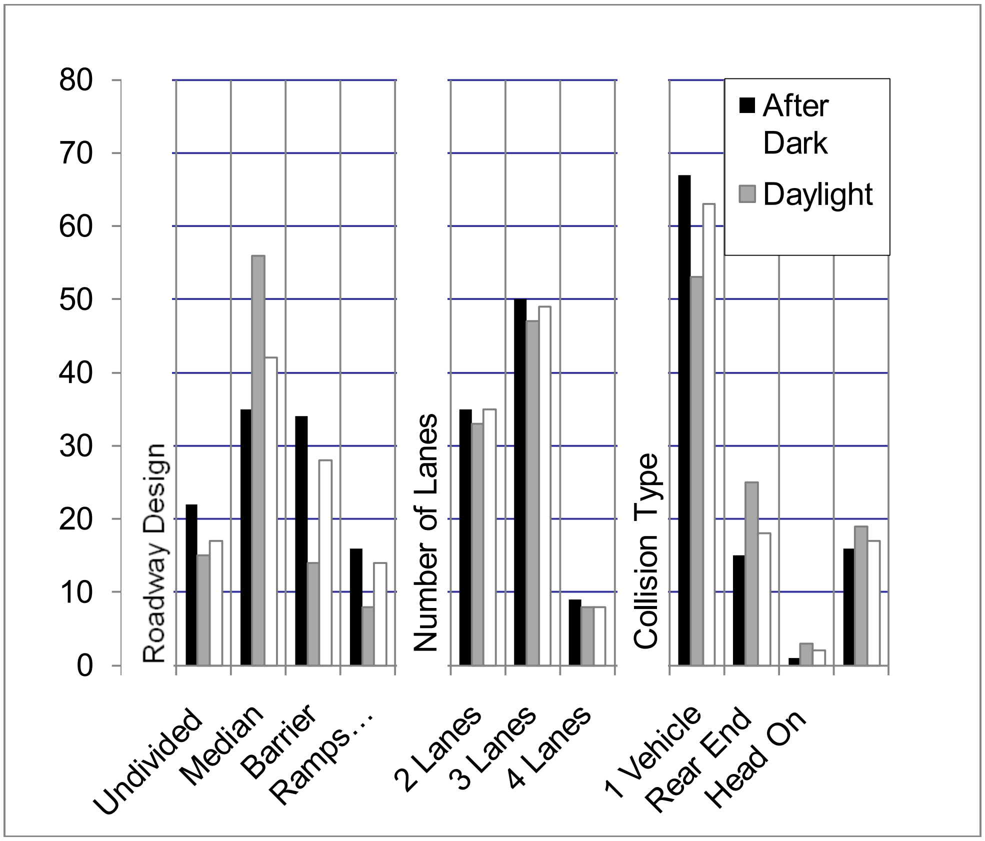


Figure 25. The area of interest roadway designs, numbers of lanes, and collision types compared to the daylight and after dark fatalities.

Collisions that do not involve any other vehicle are the most predominate crash in the extracted area of interest in NCT. This is not surprising because this category has been the most often collision type occurrence throughout the study. After dark crashes have a higher percentage of occurrences than daytime levels, with almost 70% of all nighttime crashes involving only one vehicle and daylight occurrences at 53%. Head on collision is the least occurring collision type for daytime and night time occurrences, at 3% and 1% respectively.

The number of vehicles involved in an accident points more toward density issues than flow ones. Since the majority of crashes involve only one vehicle, it is hard to attribute them to any particular influence of after dark driving. One could surmise, in the case of the after dark crashes, that poor visibility could be an influence. As far as flow variables are concerned, after dark crashes do exhibit flow variable characteristics. The factors present favorable conditions for crashes when combined with involuntary driver behavioral reactions, for instance poor visibility. The fact that NCT experiences most crashes on some sort of divided roadway, the NCT extracted area of interest closely follows the model. The level of discrepancy in crashes occurring on a divided roadway with a median and crashes on roadways divided with a median and a physical barrier is cause to look into driver behavior that may be influenced by outside factors that are combined with the geography. Because the extracted area of interest is urban, it is beneficial to compare the area of interest extraction to urban areas in the rest of NCT. In this way, the differences between dark and daylight datasets can be emphasized without less significant factors coming into play. Undivided roadways account for 20% of roadways in urban NCT and 17% in our area of interest. Roadways divided with a median make up 26% of urban NCT, and 20% in the extracted area of interest. Roadways with a median and a physical barrier constitute 14% in NCT, while the area of interest accounts for 28%. The urban figures take into account both daylight and dark occurrences.

To determine the extent to which flow variables affect night time driving, the day and night crashes should be compared. 54% of NCT crashes occur after dark or at dusk. This figure is less than the extracted area of interest's 69%. Even if the urban crashes are extracted from NCT, the percentage of dark crashes to all crashes is still 53%. When comparing the flow variables for day crashes and flow variables for night crashes, the greatest differences are in the divided roadways. Crashes on roadways with a median only are 63% higher for daylight crashes. Crashes occurring on roadways with a median and a physical barrier are 41% higher for the dark crashes. The area of interest was extracted because the area contained a large population of night time crash densities. Statistical analyses confirmed significant clustering in the dark crashes, and inferential statistics confirm that flow variables are influencing driver behavior, more so in dark situations than daylight situations. The null hypothesis stating crashes occurring at night are susceptible to the same influences that daytime crashes are susceptible to is False. Night time crashes are known to be driven by flow variables, rather than daylight density issues. The reasons to put up barriers, in addition to medians, are safety. If barriers are present, yet the area is prone to a greater amount of casualty, some outside factors are affecting the driver, whether it is visibility or improper regulation.

CHAPTER 5

CONCLUSIONS

Determining the causes of traffic crashes can be accomplished by utilizing analyses that are performed over time to formulate a knowledge base that identifies the precursors of crashes. The tests which were used for this study are used currently in other traffic analyses and the methods used to attain the results are aligned with current research, though specifically regional. The analysis of the predictability of North Central Texas (NCT) crashes was based upon the results of previously published traffic research data, most notably Golob et al. (2004). The results of those tests indicated that traffic flow crashes exhibit certain characteristics that are indicative of their roadway location. Golob et al. (2004) used disaggregated traffic counts, enabling a “real time” traffic analysis. Disaggregated counts were not available for the NCT region, so traffic flow could not be studied in the same way. Based on the findings of the disaggregated traffic flow results and the manner in which traffic was found to move across traffic lanes, an inferential study based on the traffic flow variables was performed. It was determined that the collision type is not an indication of the crash location in relation to the roadway in NCT. A look at the balance of roadway type and design that is unique to NCT precludes a blanket statement categorizing crashes in this way.

The second analysis stemmed from the first in that road geometry seemed to be a better indicator of roadway crashes, and in fact, traffic flow accidents are already known to correlate with traffic flow variables and congestion. The SaTScan program with the Bernoulli test statistic and Monte Carlo simulation verified the roadway design is very much a factor in roadway crashes. The results of this program indicate that serious crashes are most likely to occur on two lanes and undivided roadways; to the extent that the three most clustered areas lie outside the densest areas of roadways that lie within the Metroplex.

The last analysis transformed a standard two dimensional testing environment into a one dimensional linear one. The test results did not offer more information than planar tests – in fact, because a planar test over detects clustering patterns, the one dimensional test seems to provide a more accurate depiction of event clustering. The number of segment nodes that were recorded in the *k*-function analysis back up the test result of clustering. The data points were distributed only between segment nodes contained within the 141 meter distance test environment. There are not clusters on the roadway simply because the roadways are densest at some particular point on the map. The iterated points are measured and averaged amongst themselves and then compared. The process of transformation is a better method of analysis that is also illustrated by the comparison table of dark versus light in linear and planar methods (Table 7). The observed data points and expected data points are consistently higher on the planar method. The spreading of simulated points throughout space increases the probability of a neighboring point within the specified testing distance. The adjustments of planar tests have merit; however, acknowledging that crashes most likely will be within the confines of the roadway system allows a more accurate definition of the environment. Optimally, the variables should be compared in one dimension in order to measure the variability that is based on the local tendencies on a linear space.

Standardized data is crucial for any statistical analyses and working datasets. Standardizing data for inter-departmental communications eases road blocks associated with data sharing and can facilitate data gathering when incorporating a GIS into existing systems (Kulikowski, 2006; Li et al., 2006; Smith, Graettinger, Keith, and Parrish, 2007). Integration of traffic data among existing public and private entities will become increasingly important as the population of the North Central Texas region grows over the next twenty years. As political boundaries become less apparent from the roadway, the need for standardized data will be necessary for interagency sharing and will provide for a smooth workflow. A standardized method for abbreviating road types (State Hwy as opposed to S.H., State Highway, or simply a designation of "377") and capitalization rules and/or spacing rules, would facilitate data entry and the integration of databases. Without standardization rules, the percentage of non workable data entries within the dataset climbs, requiring an item by item reconciliation with any existing dataset already established within the workplace.

Limitations also stem from the large amount of data that was used for this study. The Texas Department of Transportation (TxDOT) data contained around 400,000 accident locations, making analysis extremely time intensive, if not prohibitive, to run unless it was manipulated into manageable sets. An alternative method of reducing the data could have been to decrease the study area, but there were comparatively fewer crashes in the peripheral counties of the NCT metroplex, and in this case the crash count was not noticeably different. SANET computing requirements were also intensive because of the number of segment nodes that were created when dividing the study area. Once a Monte Carlo simulation is processed, the data points reach into the millions for even a small area. The area that was tested was less than 6.5 km^2 and had 4.5 million data points to measure and process.

Traffic crash analysis results are often used when describing traffic movement behavior (for instance: the severity and location of an accident are due to traffic flow variables such as lane position). One goal of this research is to create an awareness of the severity of fatal traffic accidents in NCT and everywhere. 2010 marked the thirty year anniversary of the drinking and driving laws which have resulted in many saved lives. The goal of crash reduction is to thwart increases in fatalities and optimize the methods used to decrease fatality rates. Results from this research could add important data to the existing database of research by providing real numbers, many of which are aggregated into the causes of traffic deaths. A reasonable expectation stemming from research and statistical analysis would be the ability to pinpoint the specific factors that are involved in a large percentage of local traffic crashes, and this study provides evidence toward that cause. Elements within systems move fluidly and should produce a working synthesis of order and progression. This is exactly what we want out of our transportation system. With the progression of GIS applications, fatality research may not simply focus on the analyses of proximity (Kulikowski, 2006; Li et al., 2006) as in the past, but on the inherent processes within the system. The cause of traffic crashes are the result of individual or structural mechanisms, therefore safety issues are a prevalent concern, but focusing on the networks of the many integrated parts of the system, and the mechanisms allocated to each part, a better equipped system may be put into place to save lives. Legal and engineering infrastructures have always played a part in the public roads system, and advances in technology have enabled safety practices and precautions to keep pace with increases in population. Unfortunately, statistical data only measures the amount of dollar damage put upon society as a whole and the toll it takes on the individuals behind the statistics cannot be measured geographically. Efforts in reducing the crash rate help to provide a safety conscientious transportation system on which future growth can build upon. Building roads to keep up with an increasing population should be high on NCT planning board agendas.

REFERENCES

Aguero-Valverde, J. and Jovannis, P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis and Prevention, 38*, 618-625.

Baker, S. and Haddon Jr., W. (1974). Reducing injuries and their results: The scientific approach*. Health and Society*, *52*(4), 377-389.

Cook, P. and Tauchen, G. (1984). The effect of minimum drinking age legislation on youthful auto fatalities, 1970-1977. *Journal of Legal Studies, 13*(1), 169-190.

Crandall, R. and Graham, J. (1984). Automobile safety regulation and offsetting behavior: Some new empirical estimates. *American Economic Review, 74*(2), 328-331.

Eksler, V., Lassarre, S. and Thomas, I. (2008). Regional analysis of road mortality in Europe. *Journal of the Royal Institute of Public Health, 122*, 826-837.

Eksler, V. and Lassarre, S. (2008). Evolution of road risk disparities at small-scale level: Example of Belgium. *Journal of Safety Research, 39*, 417-427.

Fatality Analysis and Reporting System (FARS). http://www.fars.nhtsa.dot.gov , National Highway Traffic Safety Administration, National Center for Statistics and Analysis, Washington DC.

Flahaut, B., Mouchart, M., San Martin, E. and Thomas, I. (2003). Local spatial autocorrelation and the kernal method for identifying black zones: A comparative approach. *Accident Analysis and Prevention, 35,* 991-1004.

Glassman, J. (2004). Administrator's message. National Highway Traffic Safety Administration, National Center for Statistics and Analysis, Washington DC. http://www.NHTSA.dot.gov

Golob, T. and Recker, W. (2004). A method for relating type of crash to traffic flow characteristics on urban freeways. *Transportation Research Part A, 38*, 53-80.

Golob, T., Recker, W. and Alvarez, V. (2004). Freeway safety as a function of traffic flow. *Accident Analysis and Prevention, 36*, 933-946.

Goodchild, M. (2000). GIS and transportation: Status and challenges. *GeoInformatica, 4*(2), 127-139.

Huelke, D. (1968). Automobile accidents: Where we've been, where we are, what needs to be done. *Journal of Risk and Insurance, 35*(1), 61-66.

Kendall, M. (1950). The statistical approach. *Economica, 17*(6), 127-145.

Kulikowski, L. and Bejleri, I. (2006). Building a regional traffic crash data system - bridging the gaps. *Proceedings of the 26th Annual ESRI International User Conference.*

Lewis, H. (1927). Routing through traffic. *Annals of the American Academy of Political and Social Science, 133*, 9-27.

Li, L., Zhu, L. and Sui, D. (2007). A GIS-based Bayesian approach for analyzing spatial-temporal patterns of intra-city motor vehicle crashes. *Journal of Transport Geography, 15*(4), 274-285.

Lin, X, Zhou, X. and Liu, C. (2000). Efficient computation of a proximity matching in spatial databases. *Data & Knowledge Engineering, 33*, 85-102.

Lyons, R., Ward, H., Brunt, H., Macey, S., Thoreau, R., Booder, O. and Woodford, M. (2008). Using multiple datasets to understand trends in serious road traffic casualties. *Accident Analysis and Prevention, 40*, 1406-1410.

Mothers Against Drunk Drivers (MADD), National Office, Irving, Texas. http://www.madd.org

Mountain, L., Fawaz, B. and Jarrett, D. (1996). Accident prevention models for roads with minor junctions. *Accident Analysis and Prevention, 28*(6), 695-707.

North Central Texas Council of Governments. North Central Texas 2030 demographic forecast. http://www.nctcog.org

National Highway Traffic Safety Administration (NHTSA). 2004 national statistics, early edition. http://www.NHTSA.dot.gov

National Highway Traffic Safety Administration (NHTSA). *The economic impact of motor vehicle crashes 2000*. NHTSA Technical Report No. DOT HS 80946. http://www.NHTSA.dot.gov

Okabe, A., Okunuki, K. and Shiode, S. (2006). *A toolbox for spatial analysis on a network*. Center for Spatial Information Science, University of Tokyo, Version 3.0.

Pruitt, C. (1979). People doing what they do best: The professional engineers and NHTSA. *Public Administration Review, 39*(4), 363-371.

Ross, H. (1973). Law, science, and accidents: The British Road Safety Act of 1967. *Journal of Legal Studies, 2*(1), 1-78.

Smith, R., Graettinger, A., Keith, K. and Parrish, A. (2007). Identifying high frequency crash locations: Empowering end-users with GIS capabilities. *Institute of Transportation Engineers* (*ITE) Journal,* 22-27

Texas Department of Transportation. *Crash data analysis and statistics*. http://www.dot.state.tx.us

Taubman-Ben-Ari, O., Mikulincer, M. and Gillath, O. (2004). The Multidimensional Driving Style Inventory-Scale: Construct and validation. *Accident Analysis and Prevention, 36*, 323-332.

Thomas, I. (1996). Spatial data aggregation: Exploratory analysis of road accidents. *Accident Analysis and Prevention, 28*(2), 251-264.

World Health Organization. *Global status report on road safety*. http://www.who.int/violence injury prevention/road safety status

Xie, Z. and Yan, J. (2008). Kernal density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems, 32*(5), 396-405*.*

Yamada, I. and Thill, J. (2004). Comparison of planar and network k function in traffic accident analysis. *Journal of Transport Geography, 12*, 149-158.