MAKING USE OF BIG DATA TO EVALUATE THE EFFECTIVENESS OF
SELECTIVE LAW ENFORCEMENT IN REDUCING CRASHES

Jenna K. Simandl*, Andrew J. Graettinger, Randy K. Smith, Steven Jones, Timothy E. Barnett
*Corresponding Author

Jenna K. Simandl
Kimley-Horn and Associates, Inc.
1700 Willow Lawn Drive Suite 200
Richmond, VA 23230
Tel: (804) 396-6017; Email: jenna.simandl@kimley-horn.com

Andrew J. Graettinger
The University of Alabama
Civil, Construction, and Environmental Engineering
Box 870205, Tuscaloosa, AL 35487 – 0205
Tel: (205) 348-1707; Fax: (205) 348-0783; Email: andrewg@eng.ua.edu

Randy K. Smith
The University of Alabama
Department of Computer Science
Box 870290, Tuscaloosa, AL 34587 – 0205
Tel: (205) 348-1664; Fax: (205) 348-0219; Email: rsmith@cs.ua.edu

Steven Jones
The University of Alabama
Civil, Construction, and Environmental Engineering
Box 870205, Tuscaloosa, AL 35487 – 0205
Tel: (205) 348-3137; Fax: (205) 348-0783; Email: sjones@eng.ua.edu

Timothy E. Barnett
State Safety Operations Engineer
Alabama Department of Transportation
Montgomery, AL 36110
Tel: (334) 353-6464; Email: barnett@dot.state.al.us

Word Count: 250 (abstract) + 5571 (text) + 1500 (6 tables/figures x 250) = 7321 words
Original Submission Date: 07/30/2015

Paper revised from original submission.
ABSTRACT
State Departments of Transportation (DOTs) across the nation fund selective enforcement campaigns aimed at intensifying law enforcement at certain locations in order to improve traffic safety. At the same time, many states passively collect large datasets such as officer Global Position System (GPS) location tracks. To evaluate the effectiveness of selective enforcement, an approach was developed that employs Structured Query Language (SQL) and Geographic Information System (GIS) technology to mine and integrate police patrol patterns, citations issued, crash occurrences, and selective enforcement periods. This information was analyzed in a relational database within a spatial and temporal analysis framework. The intent was to solely use GIS technology, however the size of the datasets were prohibiting and SQL was used as a Big Data Analytic. The SQL techniques successfully turned over 37 million points of GPS data into 1.3 million points of selective enforcement location information, enabled the geolocation of 72.6% of Electronic Citations (eCitations), and identified 21 selective enforcement areas across the State of Alabama. With Big Data Analytics, GIS technology was reestablished as useful for the evaluation of changes in crash and citation frequencies before and during selective enforcement. Paired difference t-tests confirm the decrease of crash frequency with 85% confidence at urban and rural locations. The analysis of the number of citations at the locations confirmed that citations increased during selective enforcement by an average of 254%. The developed methodology is a successful approach using large datasets for an unintended purpose to make valuable engineering conclusions and data-driven discoveries.

INTRODUCTION
In the age of Big Data, large datasets have become increasingly available to scientists and researchers. The volume of data is growing, the speed of data creation is increasing, and the variety of data, while generally unorganized, is accumulating (McAfee & Brynjolfsson, 2012). Big Data from both experimental and open-source platforms has created a transformative data-intensive science for researchers in science, engineering, and other business disciplines (Jahanian, 2012). Advanced management strategies, computations, and minimization techniques are being developed to turn Big Data into useful information for advanced knowledge, predictions, and decision making (Lohr, 2012; Mayer-Schonberger & Cukier, 2013; Mestyan, Yasseri, & Kertesz, 2013).

Big Data enables a paradigm shift from hypothesis-driven research to data-driven discovery (Lohr, 2012; Jahanian, 2012). Big Data is also proposed as a means to help resolve the Grand Challenges defined by the National Academy of Engineering (Engineering and Physical Sciences Research Council, 2014). A cohesive relationship between data science, Big Data, and domain expertise are required for effective decision making and data-driven discoveries (Provost & Fawcett, 2013). However, the nature of Big Data is that the dataset is too large for traditional data processing and new technologies or techniques are required.

A research project by The University of Alabama for the Alabama Department of Transportation (ALDOT) sought to use Geographic Information System (GIS) technology in order to evaluate the effectiveness of selective law enforcement along state routes in reducing...
crashes in Alabama. Officer patrol route data, Electronic Citation (eCitation) data, Electronic Crash (eCrash) data, and selective enforcement data were to be integrated into one spatial and temporal GIS framework for analysis of citation counts and crash frequencies before and during selective enforcement. However, the large dataset quickly overwhelmed GIS processing tools.

Structured Query Language (SQL) is the standard language for working with relational database tables, and can be used as a Big Data Analytics tool (Cohen, Dolan, Dunlap, Hellerstein, & Welton, 2009). While not as advanced as other data management tools, SQL queries were executed for successful data minimization in order to perform traffic safety analysis on the effectiveness of selective enforcement along state routes in Alabama from August 1, 2010 to July 31, 2011.

A relational database is a collection of data, organized in various tables of attributes (columns) and records (rows) that have a specific, unique relationship with one another. The tables are related through the use of primary keys, a unique attribute in a table, which is included in other tables as a foreign key. SQL query expressions were written and executed in order to translate data in tables to valuable information (Viescas & Hernandez, 2014). The following sections present the logic behind the executed SQL queries used for data minimization of Big Data in order to use GIS tools to evaluate the effectiveness of selective law enforcement. The developed methodology is a successful approach of using large datasets for an unintended purpose in order to make valuable engineering conclusions and discoveries.

BACKGROUND

State Departments of Transportation (DOTs) across the nation fund selective enforcement campaigns aimed at intensifying law enforcement at certain locations in the state in order to improve traffic safety and reduce the number of crashes. Many case studies show that selective police enforcement, whether for a specific negative driver behavior or for a particular high crash location, can have an influential impact on driver behavior, which can successfully reduce the number of crashes in a location (Jonah & Grant, 1985; Vaa, 1997; Erke, Goldenbeld, & Vaa, 2009; Walter, Broughton, & Knowles, 2011). The first edition of the Highway Safety Manual, while focused on engineering analysis and treatment, states that crash frequencies may be reduced through enforcement efforts (AASHTO, 2010).

Large datasets are available for the analysis of the effectiveness of selective enforcement efforts. By integrating police officer patrol patterns, citations issued, crash occurrences, and selective enforcement periods into one spatial and temporal analysis framework, selective enforcement locations can be verified, and citation and crash frequencies in the respective locations can be identified and analyzed. Due to the volume of the data and its constant growth, big data management strategies need to be employed before analysis is feasible.

The Alabama Mobile Selective Law Enforcement Campaign attempts to change negative driver behaviors that contribute to the most severe crashes, including speeding, driving under the influence, and the failure to wear a seatbelt (Peden, et al., 2004; Erke, Goldenbeld, & Vaa, 2009). An agreement between the Alabama Department of Public Safety (DPS) and ALDOT was in
effect between October 2007 and September 2011, employing police officers to work extra-duty
shifts for increased law enforcement efforts on various state and federal roadways.

The over-arching goal of the increased enforcement at selected high crash locations was
to deter negative driver behaviors with an increase in issued citations with the goal of decreasing
the number of crashes. In order to evaluate the effectiveness of selective enforcement, SQL and
GIS technology were used to integrate police officer patrol patterns, citations issued, crash
occurrences, and selective enforcement periods into a relational database and a spatial and
temporal analysis framework. A SQL relational database was used as the Big Data Analytic.
The database and GIS map information provides the means to verify selective enforcement
locations and evaluate the increases and decreases in crash and citation frequencies before and
during selective enforcement.

A relational database of officer patrol routes, citations, and selective enforcement data
was created. In the state of Alabama, State Trooper vehicles are equipped with a Global
Positioning System (GPS) unit that is polled every 30 seconds. The location of the trooper is
recorded along with a timestamp. Officer patrol route location data is a large dataset that is
continually growing. Additionally, Electronic Citations (eCitations) issued by law enforcement
officials in the state of Alabama are stored in a database. These eCitations do not require GPS
coordinates or other structured location data. The eCitations do have an electronic timestamp
indicating when the citation was issued. A temporal join methodology was developed to cross-
reference timestamps between officer patrol route GPS records and issued citations in order to
gеolocate citations. Selective enforcement data for participating officers was obtained from DPS
timesheets and DOT invoices. A methodology to verify selective enforcement locations was
developed by comparing officer patrol routes and dates and hours of selective enforcement shifts.
The steps utilized SQL in order to obtain only relevant data for a one year analysis of selective
enforcement. The relevant data was imported into a GIS, along with crash data from the Critical
Analysis Reporting Environment (CARE) (Center for Advanced Public Safety, 2009), in order to
quantify the number of crashes and the number of citations in selective enforcement locations
before and during the selective enforcement effort. Due to the sensitive nature of crash data,
maps were only generated for internal DOT use. This methodology is extensible to other states
with available officer patrol routes, citation locations, crash data, and selective enforcement time
periods.

METHODOLOGY
A methodology was developed using SQL to geolocate eCitations and verify where selective
enforcement efforts took place in Alabama. This required identifying key locations in the GPS
officer location data and importing that data into GIS technology in order to evaluate the
effectiveness of the selective enforcement campaign in reducing crashes at high crash locations.
Initial work for this project involved digitizing selective enforcement shift information, including
the date and hours worked, for each participating officer from DPS extra-duty timesheets and
DOT invoice files to spreadsheets for use in electronic analysis. Law enforcement officials in
the state of Alabama are given a unique UserID, which is used for identification purposes in both
Simandl, Graettinger, Smith, Jones, Barnett

the GPS officer patrol route and eCitation databases. The unique UserID was identified for each participating officer in the selective enforcement data.

Data processing tasks were executed using SQL queries because of the large datasets and the limited capabilities of GIS tools. The SQL queries described herein are presented in detail in Simandl, 2015. The first task for this project included developing a methodology to cross-reference timestamps between officer patrol routes and issued citations in order to geolocate citations with and without GPS coordinates. Furthermore, a methodology to verify selective enforcement locations was developed by comparing officer patrol routes and dates and hours of selective enforcement shifts. The resulting data from the SQL queries was specific information as opposed to the Big Data in its entirety, and was imported into a GIS along with crash data from the Critical Analysis Reporting Environment (CARE). Only essential data for analysis was imported in order to increase the feasibility of GIS to verify selective enforcement locations and quantify the number of crashes and citations in selective enforcement locations before and during selective enforcement efforts.

**Geolocating eCitations**

Performing a spatial-temporal analysis through a sequence of SQL queries, the time-stamped eCitations and State Trooper location data were cross-referenced in order to accurately locate an eCitation to the nearest GPS point in time. The UserIDs of the officers who participated in selective enforcement were joined to the eCitation data based on UserID, in order to return eCitations written by the relevant officers. The criterion to exclude voided eCitations was also included, as voided eCitations were not officially issued. Over the course of the year being analyzed in this research, including both selective enforcement shifts and regular duty shifts, 475,214 eCitations were issued by the participating officers.

Data for each individual GPS trace point includes the UserID, the latitude and longitude location, and a timestamp recorded in UTC time. The officers who participated in selective enforcement were joined to the GPS data based on UserID, in order to return GPS location points for the relevant officers exclusively. For the year being analyzed in this research, including both selective enforcement shifts and regular duty shifts, 37,628,124 GPS points were recorded for the selective enforcement participating officers.

In order to cross-reference the timestamps of the eCitations and the GPS location points, a SQL query was executed to calculate the time difference between GPS location points and eCitations for a particular officer using the DATEDIFF() command. This query only returned eCitations joined to a GPS location point within 600 seconds in order to maximize the number of eCitations located while also providing reasonable eCitation location accuracies. Out of the returned results, the minimum time difference between GPS location point and eCitation was chosen to signify the most accurate location of the eCitation.

The spatial-temporal analysis, produced through a sequence of SQL queries, located 68.6% of eCitations. Further investigation showed that 49 of the officers (11.6%) participating in selective enforcement efforts did not have any GPS trace data associated with their respective UserIDs, potentially due to lack of GPS equipment and/or faulty equipment. With respect to the
88.4% of officers that had trace data, 72.6% of eCitations were geolocated using the developed methodology. Additionally, 95% of those eCitations were joined to a GPS point with a time difference of only 30 seconds. The resulting dataset of 325,926 eCitations was imported into GIS for visual mapping. Citations that were not successfully located lacked officer location data or there were breaks in the officer trace location data due to lost GPS signals. The loss of eCitation data from the analysis has minimal negative implications to the results. The eCitations that were included represent a conservative understanding of the number of eCitations at various locations.

**Identifying Selective Enforcement Locations**

Selective enforcement DPS timesheets and DOT invoices provided the following information: Date Worked, First Name, Last Name, Middle Initial, Suffix, and Hours. To relate the selective enforcement hours worked with the GPS location data, the total hours worked in a full shift was found. By understanding the relationship between total hours worked and selective enforcement hours worked by an officer on a particular day, specific GPS location points were defined as either “during selective enforcement” or “not during selective enforcement.” The location of the “during selective enforcement,” or overtime work, is vital for the analysis of the effectiveness of selective enforcement. The workflow framework for identifying individual GPS points as “during selective enforcement” is shown in FIGURE 1. FIGURE 1(a) illustrates the steps to calculate full shifts from the GPS officer patrol route data, in order to establish the necessary relationship to join in the selective enforcement data, denoted by the pink arrow. These steps are described in the sub-section titled, Calculating Full Officer Shifts from GPS Location Data. FIGURE 1(b) illustrates the steps to identify individual GPS points as during or not during selective enforcement shifts after the establishment of the relationship between the data, as noted by the pink arrow. These steps are described in the sub-section titled, Defining Individual GPS Points During Selective Enforcement Shifts. The following steps were used to confirm that selective enforcement was completed in the appropriate locations.
FIGURE 1 Workflow framework for (a) establishing a relationship between the GPS data and selective enforcement (SE) data, denoted by the pink arrow; and (b) identifying individual GPS points as during selective enforcement

Calculating Full Officer Shifts from GPS Location Data

A series of SQL queries was used to define the beginning and ending of a full working shift for an officer and calculate the total hours worked for the shift. The LEAD function in SQL accesses the previous row in a column of a table, and the LAG function in SQL accesses the next row in a column of a table. To calculate the time difference in seconds between two successive GPS location points, the LEAD function was used, partitioned by UserID and ordered by GPS Time to access the Next GPS Time in the table. The results of the LEAD calculation are stored in a new column in a table. The DATEDIFF() command created an additional column, recording the difference between the GPS Time column and the newly added Next GPS Time column. An example of this query is shown in FIGURE 2.
FIGURE 2  Example dataset and query results to calculate the time difference between successive GPS points

The beginning and ending of a shift was identified in the table by using the CASE function in SQL, which enforces a “when-then” statement. When the time difference between successive GPS location points was greater than a predetermined time tolerance limit, then the GPS location point was considered to be the beginning of a shift. An example of a large time difference signifying the beginning of a shift is shown in FIGURE 2, highlighted with a red circle. On the other hand, when the time difference between successive GPS location points was less than or equal to the predetermined time tolerance limit, then the GPS location point was considered to be within a shift. To select the time tolerance for shift generation that best represents the GPS data and reasonable shift lengths that officers work, the time difference tolerance of six hours was chosen because six hours minimized shifts of over 24 hours compared to using eight hours, kept an optimum number of seven-, eight-, and nine-hour shifts, and minimized the number of one-hour shifts compared to using five hours.

Using the time tolerance of six hours, the beginning and ending timestamps of a particular shift for each officer were selected and compiled into a new table with new columns of Shift Start and Shift End. A query was used to calculate the difference between these two timestamps, which resulted in the total hours worked for a particular shift and generated a total of 55,402 GPS-identified shifts. The date difference was calculated in seconds and then converted to hours to ensure the precision of the time. Shifts with hours worked between 0.5 hours and 17 hours were considered reasonable, as 0.5 hours is the smallest increment recorded in the selective enforcement invoices and 17 hours provides a conservative inclusion of a maximum possible shift of two back-to-back 8-hour shifts. Out of all the GPS-identified shifts, 94.2% had a reasonable shift length.

Defining Individual GPS Points During Selective Enforcement Shifts

By understanding the GPS point data in terms of full officer shifts, a relationship to the selective enforcement data was established. Using the relationship between hours worked in a GPS-identified shift to the hours worked by an officer for selective enforcement, a series of SQL queries was used to define an individual GPS point as “during selective enforcement” or “not during selective enforcement.” To relate the selective enforcement shifts to the identified GPS
location point shifts, a join was executed in SQL based on UserID and Shift Start date. There were 66\% of selective enforcement shifts accounted for and joined to an appropriate GPS-identified shift. Because of the 49 officers that did not have any GPS trace data associated with their respective UserIDs, 477 selective enforcement shifts were not successfully included. When excluding those particular officers, the GPS generated shifts successfully recognized 72.6\% of the logged selective enforcement shifts. Additionally, this query calculated the difference between GPS shift hours worked and Selective Enforcement shift hours worked. The loss of selective enforcement data from the analysis due to a lack of officer GPS data suggests that additional selective enforcement locations across the state were potentially unidentified or that various locations across the state could potentially have had more police enforcement presence than the analysis shows. However, the verified selective enforcement locations in the study still represent the highest level of selective enforcement presence from the available data across the state between August 2010 and July 2011.

Simultaneously, queries to calculate cumulative time worked in a GPS-identified shift were executed. Using the difference between the GPS-identified shift hours worked and the selective enforcement hours worked in conjunction with the cumulative time worked, individual GPS location points were denoted as “during selective enforcement” or “not during selective enforcement.” The scenarios for the CASE function, a “when-then” command, used in the SQL query and the necessary assumptions for the methodology are shown in TABLE 1.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Assumption</th>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logged selective enforcement hours &lt; GPS shift hours worked</td>
<td>The last ‘x’ hours of the GPS shift are the overtime selective enforcement hours</td>
<td>When Diff &lt;= Cumulative Time, then ‘Yes’</td>
</tr>
<tr>
<td>Logged selective enforcement hours = GPS shift hours worked</td>
<td>Within 0.5 hours is considered “equal”</td>
<td>When Diff is Between -0.5 and 0.5, then ‘Yes’</td>
</tr>
<tr>
<td>Logged selective enforcement hours &gt; GPS shift hours worked</td>
<td>The whole shift is selective enforcement, and the discrepancy can be explained by a loss of GPS signal</td>
<td>When Diff &lt; 0, then ‘Yes’</td>
</tr>
</tbody>
</table>

A final query was executed to create a table of only GPS points during selective enforcement shifts, resulting in a total of 1,647,711 GPS points. The series of SQL queries demonstrates successful data minimization, by turning over 37.6 million GPS data points, into approximately 1.65 million points of useful information.
Verifying Selective Enforcement Locations

Officers participating in the Alabama selective enforcement efforts were allowed up to two hours of paid travel time to selective enforcement locations. Therefore, the GPS points during selective enforcement are not all localized in a particular selective enforcement area. Additionally, police presence in a selective enforcement location does not necessarily mean the presence was statistically significant over other areas of patrol. In order to verify selective enforcement locations, a statistical method employing clustering of spatial data was implemented. To employ the location verification process, the GPS points during selective enforcement and eCitation data were integrated into GIS based on Route-Milepost (Route-MP) along state routes, resulting in 1,283,211 GPS points during selective enforcement and 243,800 eCitations over the course of the year. With Route-MP data associated to each point event, a clustering methodology was executed on the GPS point data using a SQL query technique involving the summation and grouping of point events at each Route-MP with various bucket sizes. This technique provides an efficient Hot Spot Analysis. Hot spot analysis tools in GIS were not used because of the large amount of officer GPS location data, and GIS Hot Spot tools neglect the one-dimensional integrity of the roadway network.

SQL Query Technique to Verify Selective Enforcement Locations

A series of SQL queries using GPS point data identified as “selective enforcement” was implemented to verify selective enforcement locations by finding clusters of points at identifiable Route-Milepost locations. The Route-MP data was incorporated by summarizing the number of selective enforcement points at Route-MP locations using the COUNT aggregate function. Bucket sizes of one-tenth of a mile, one-half of a mile, and one mile were evaluated. To standardize the bucket sizes and confirm appropriate lengths along a particular route in the network, queries to calculate the difference between successive Mileposts and the cumulative difference of Mileposts were employed. The summation of “selective enforcement” points in groups of the three bucket sizes was centered in the middle of each bucket size and recorded. The LEAD and LAG function were used to calculate the number of selective enforcement points prior to and after a particular row of data. A hypothetical example of the logic for the summation query is shown in FIGURE 3.

The Sum column in FIGURE 3 was executed for all three bucket sizes used in the analysis. The sum data was imported back into GIS for thematic mapping using natural jenks for statistical recognition of large sums, or clusters, of selective enforcement locations. Natural jenks organize the data into groups, or classes, that exclusively minimize the variance of the data points included in each class. While the located selective enforcement points may not be the exact area law enforcement was directed to patrol with extra-duty hours, the largest natural jenk was used to identify the areas with statistically high levels of selective enforcement and was therefore considered a selective enforcement location for the analysis.
Three natural jenk classes were used on the one mile bucket summation and grouping analysis values in order to optimize the length of the located and verified selective enforcement locations. The one-tenth of a mile and one-half of a mile analyses generated locations of smaller length, which were in series and relatively close together, whereas the one mile analysis recognized the smaller, close locations as one whole location. The limit value for the one mile analysis was 2885 selective enforcement points at a particular Route-MP. Across the state of Alabama, 26 locations were identified as selective enforcement areas through this procedure.

**Incorporating eCrash Data from CARE**

The Critical Analysis Reporting Environment (CARE) is data analysis software developed by The Center for Advanced Public Safety at The University of Alabama. A data source of crash records from 2010-2014 were imported into CARE, and filters were created to export crash report data for the research time period as a GIS shapefile with GPS coordinates. Two exports were completed: crashes during one year before the selective enforcement year and crashes during the selective enforcement year. The created filters also incorporated highway classifications, and only exported crashes along Interstate, Federal and State routes. A spatial join of the GIS shapefiles of the exported crash data to the Route-Milepost information was performed to ensure consistency of location identification with the selective enforcement points and eCitations.
RESULTS
For a one year study of the Alabama Mobile Selective Law Enforcement Campaign from August 1, 2010 through July 21, 2011, selective enforcement efforts were located at 26 areas along state routes across the state of Alabama. After further investigation of each location using GIS and Google Maps, three locations that were found to have high selective enforcement presence were located at state trooper posts and were omitted from further study. Additionally, two different areas were found to have two separate defined locations of selective enforcement high points, separated by 0.2 miles, and were therefore merged.

Number of Crashes and Citations Before and During Selective Enforcement
The number of crashes and the number of issued citations were evaluated at the 21 locations before and during the selective enforcement research year, including both regular and overtime shifts. The number of crashes and citations for each location over the course of the before and during year are shown in TABLE 2, along with the percent change between the before and during selective enforcement crash and citation counts. A negative percent change for crashes marks a decrease in the number of crashes during the selective enforcement research year, and a positive percent change for citations marks an increase in the number of issued citations. TABLE 2 is ordered by the percent change in the number of crashes at each location, and is sectioned off by negative and positive percent changes. There were 11 locations that experienced a decrease in the number of crashes. These locations had variable changes in the number of issued citations. Notably, Location 9, an enforcement effort between two interchanges, experienced a citation increase of 2015%.
### TABLE 2 Crash and Citation Data for the Selective Enforcement Locations, organized by percent change on crashes before and during selective enforcement

<table>
<thead>
<tr>
<th>Location</th>
<th>Crashes</th>
<th></th>
<th></th>
<th>Citations</th>
<th></th>
<th></th>
<th>Equivalent C Level Crashes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>One Year Before</td>
<td>One Year During</td>
<td>Percent Change</td>
<td>One Year Before</td>
<td>One Year During</td>
<td>Percent Change</td>
<td>One Year Before</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0</td>
<td>-100.00%</td>
<td>93</td>
<td>153</td>
<td>64.52%</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>0</td>
<td>-100.00%</td>
<td>31</td>
<td>23</td>
<td>-25.81%</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>33</td>
<td>19</td>
<td>-42.42%</td>
<td>1512</td>
<td>963</td>
<td>-36.31%</td>
<td>27</td>
</tr>
<tr>
<td>21</td>
<td>22</td>
<td>14</td>
<td>-36.36%</td>
<td>91</td>
<td>257</td>
<td>182.42%</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>7</td>
<td>-30.00%</td>
<td>180</td>
<td>180</td>
<td>0.00%</td>
<td>6</td>
</tr>
<tr>
<td>19</td>
<td>17</td>
<td>14</td>
<td>-17.65%</td>
<td>75</td>
<td>505</td>
<td>573.33%</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>20</td>
<td>-16.67%</td>
<td>268</td>
<td>1133</td>
<td>322.76%</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>13</td>
<td>-13.33%</td>
<td>4</td>
<td>41</td>
<td>925.00%</td>
<td>8</td>
</tr>
<tr>
<td>13</td>
<td>16</td>
<td>14</td>
<td>-12.50%</td>
<td>916</td>
<td>511</td>
<td>-44.21%</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>15</td>
<td>-6.25%</td>
<td>33</td>
<td>698</td>
<td>2015.15%</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>16</td>
<td>15</td>
<td>-6.25%</td>
<td>214</td>
<td>826</td>
<td>285.98%</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>22</td>
<td>0.00%</td>
<td>886</td>
<td>1243</td>
<td>40.29%</td>
<td>11</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>2</td>
<td>0.00%</td>
<td>33</td>
<td>46</td>
<td>39.39%</td>
<td>6</td>
</tr>
<tr>
<td>20</td>
<td>9</td>
<td>9</td>
<td>0.00%</td>
<td>144</td>
<td>343</td>
<td>138.19%</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td>15</td>
<td>15.38%</td>
<td>311</td>
<td>223</td>
<td>-28.30%</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>43</td>
<td>50</td>
<td>16.28%</td>
<td>35</td>
<td>170</td>
<td>385.71%</td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>9</td>
<td>28.57%</td>
<td>371</td>
<td>365</td>
<td>-1.62%</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>13</td>
<td>19</td>
<td>46.15%</td>
<td>229</td>
<td>568</td>
<td>148.03%</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>30</td>
<td>57.89%</td>
<td>140</td>
<td>161</td>
<td>15.00%</td>
<td>7</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>5</td>
<td>66.67%</td>
<td>27</td>
<td>91</td>
<td>237.04%</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>34</td>
<td>161.54%</td>
<td>868</td>
<td>1599</td>
<td>84.22%</td>
<td>8</td>
</tr>
</tbody>
</table>

Severity weighted crash counts are also presented in TABLE 2. The weighted crash counts were included in the analysis of the effectiveness of selective enforcement in order to draw conclusions on any changes in the severity of crashes. For instance, the number of severe crashes may have decreased, while the number of total crashes may have remained constant or even increased. The severity weighted crash counts for the 21 selective enforcement areas before and during the selective enforcement year are presented in terms of equivalent C level crashes, obtained by using the “KABCO” injury scale (Federal Highway Administration, 2011).

**Evaluating the Effectiveness of Selective Enforcement**

The effectiveness of the selective enforcement efforts were evaluated in terms of crash frequencies and the number of issued citations. A paired difference t-test was used to determine whether the mean, or expected value, of the differences between the before and during crash frequencies and citation counts yielded a significant change. A positive value change for crashes...
will conclude that the number of crashes decreased during the selective enforcement efforts. The null and alternative hypotheses for the before and during crash number tests are,

\[ H_0: \mu_d \leq 0 \quad \text{and} \quad H_1: \mu_d > 0 \]

where \( H_0 \) is the null hypothesis, \( H_1 \) is the alternative hypothesis, and \( \mu_d \) is the difference in the means. A negative value change for citations will conclude that the number of issued citations increased during selective enforcement. The null and alternative hypotheses for the before and during citations number tests are,

\[ H_0: \mu_d \geq 0 \quad \text{and} \quad H_1: \mu_d < 0 \]

If the null hypothesis is rejected based on the t-test statistic and a corresponding P-value less than a particular significance level, than the alternative hypothesis is accepted, meaning there was a statistical change in the number of crashes or citations at the various selective enforcement locations. For example, at a significance level of 0.05, the null hypothesis is rejected with 95% confidence and is considered statistically significant. The paired difference t-test determines whether a change in the mean values of two samples is statistically significant, interpreted by the resulting P-value. The actual mean values of the samples do not indicate the results of the test since the variance of the datasets may be differing.

The 21 locations were analyzed by area type. Locations were considered rural if the population of the location municipality was less than 5000, while locations were considered urban if the population of the location municipality was greater than 5000. The paired difference t-test results for the total number of crashes, the number of issued citations, and the number of equivalent C injury level crashes for urban and rural selective enforcement locations are shown in TABLE 3.

For urban selective enforcement locations along state routes in Alabama from August 1, 2010 through July 21, 2011, the p-value for the crash frequency data was 0.148. The null hypothesis (the number of crashes remained constant or increased) can be rejected at a significance level of 0.15, or with 85% confidence. While the actual mean value of the crash frequency went up during selective enforcement, from 14.3 to 17.2, the variance also increased. For rural selective enforcement locations, the p-value for the crash frequency data was 0.122. The null hypothesis can be rejected at a significance level of 0.15, or with 85% confidence. The results confirm that crash frequencies were reduced at both urban and rural selective enforcement locations, with 85% confidence. While not considered statistically significant at a significance level of 0.15, a marked decrease or trend was identified. The number of issued citations increased during selective enforcement for both urban and rural locations, with rejections of the null hypotheses with low-p values of 0.040 and 0.103, respectively. The results confirm that the number of issued citations was increased at selective enforcement locations. The results for the equivalent C severity level crash frequencies at both urban and rural locations did not constitute the rejection of the null hypothesis. Therefore, when analyzing urban and rural locations exclusively, it is plausible that selective enforcement was not effective in reducing the severity of crash occurrences.
In summary, there is evidence of trends between selective enforcement and the decrease of crashes by analyzing urban and rural locations separately, as opposed to collectively. Analyzing the locations by area type explains some of the variance of the data, and shows some improvement to understanding the effectiveness of selective enforcement. However, the decrease in the severity of crashes was inconclusive. In each paired difference test performed, the results confirm that the number of issued citations increased during the selective enforcement efforts. Overall, there was a statistically significant increase in citations at all identified selective enforcement locations, while there was a marked decrease in the number of crashes. An analysis that included crash and citation frequency collectively to directly examine the relationship between the two was not included in this study. The goal of this work was to evaluate the effect of selective enforcement on the number of citations, and more importantly the number of crashes, before and during selective enforcement. The characteristics of the selective enforcement locations, such as number of lanes, average annual daily traffic, and even safe pullout zones for officers, varied greatly and have a large effect on evaluating correlations between citations and crashes across all selective enforcement areas.

**TABLE 3 Paired Difference t-test Results for Urban and Rural selective enforcement locations along state routes in Alabama**

<table>
<thead>
<tr>
<th></th>
<th>Urban Locations</th>
<th></th>
<th></th>
<th>Rural Locations</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Crashes</td>
<td>Citations</td>
<td>Equivalent C Crashes</td>
<td>Crashes</td>
<td>Citations</td>
<td>Equivalent C Crashes</td>
</tr>
<tr>
<td></td>
<td>One Year Before</td>
<td>One Year During</td>
<td>One Year Before</td>
<td>One Year During</td>
<td>One Year Before</td>
<td>One Year During</td>
</tr>
<tr>
<td>Mean</td>
<td>14.3</td>
<td>17.2</td>
<td>181.1</td>
<td>357.8</td>
<td>6.9</td>
<td>6.7</td>
</tr>
<tr>
<td>Variance</td>
<td>151.122</td>
<td>257.956</td>
<td>66137.21</td>
<td>222289.1</td>
<td>34.9889</td>
<td>30.9</td>
</tr>
<tr>
<td>Observations</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Hypothesized $\mu_d$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Deg. of Freedom</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>t-Statistic</td>
<td>-1.111</td>
<td>-1.967</td>
<td>0.198</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td>0.148</td>
<td>0.040</td>
<td>0.424</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Paper revised from original submission.
CONCLUSIONS
For the analysis of the effectiveness of selective law enforcement in Alabama, Structured Query Language (SQL) was used to employ data minimization techniques necessary to turn a Big Dataset into valuable information. The SQL techniques successfully turned over 37 million points of GPS data into 1.3 million points of selective enforcement location information, enabled the geolocation of 72.6% of Electronic Citations (eCitations), and identified 21 selective enforcement areas across the State of Alabama. With Big Data Analytics, Geographic Information System (GIS) technology was shown to be useful for the evaluation of the increases and decreases in crash frequencies and the number of issued citations before and during selective enforcement.

Officer patrol routes, geolocated citations, crash data, and selective enforcement time periods were integrated into one cohesive source of spatial and temporal data for the analysis. The processing tasks in this research were essential to the evaluation of selective law enforcement and the results, exemplifying data-driven discovery. Paired difference t-tests confirm the decrease of crash frequency with 85% confidence, at urban and rural locations separately. The analysis of the number of issued citations at the locations confirmed that citations increased during the selective enforcement year by an average of 254%. While a statistical increase in the number of issued citations does not necessarily cause a statistical decrease in the number of crashes, the increase seems to have some effect on the number of crashes at various locations during the selective enforcement effort year in the study. Future work will include analyzing crash frequencies several years after the selective enforcement effort to potentially draw conclusions on any lasting effects, in attempts to understand possible changes in driver behavior. However, continued selective enforcement campaigns or a high presence of selective enforcement police presence in the same locations may cause potential problems for interpreting these results. Future work will also involve reducing crash frequency at high crash locations through selective enforcement campaign recommendations such as implementing steps to locate crash hot spots for selective enforcement efforts, developing crash modification factors, and calculating a return on investment for selective enforcement campaigns.

ACKNOWLEDGMENTS
This project was funded by the Alabama Department of Transportation. Special thanks for their contributions to this work go to Beau Elliott and Jesse Norris, The Center for Advanced Public Safety, and Brittany Shake, Graduate Research Assistant, The University of Alabama.

REFERENCES


Paper revised from original submission.