

Traffic Safety Analysis: A Data Mining Approach

J. Michael Hardin, Michael Conerly, and Wade Watkins
Department of Information Systems, Statistics and Management Science
The University of Alabama, Tuscaloosa, Alabama

This work was supported by the University Transportation Center for Alabama,
project 02-115.

Abstract

Our goal was to create an index of the economic impact for traffic crashes across all sixty-seven Alabama counties. This goal was accomplished by the creation of our Traffic Safety Economic Index (TSEI) for Alabama Counties. This index seeks to compare the total economic impact of vehicular crashes between counties for the three types of crashes: property damage, injury, and fatality, as well as an overall comparison. The TSEI helps summarize the information from thousands of crashes in order to facilitate better informed decisions. This concise information will help policy and decision makers better allocate resources to decrease the economic loss due to crashes. Traffic safety educators will also have more evidence as to where they need to focus their energies. In addition, the TSEI will highlight counties with traffic safety problems, spurring further investigation into a specific county's characteristics.

1. Overview

1.1 Introduction and Background

Alabamians, like Americans in general, depend on their cars and other motor vehicles for their transportation needs. Even as people grow older, their reliance on cars and driving continues to give them a sense of autonomy and satisfaction. Yet, as people venture onto the highways they place themselves at significant risk of injury and even death.

According to the report, *Alabama Traffic Accidents Facts, 1998*, a typical Alabama driver has a 57.7% chance of being involved in an injury or fatal crash while driving in an automobile during their lifetime. Further, it was reported that in 1998 1071 people were killed in 956 fatal crashes; one traffic accident occurred every 229 seconds; one person was injured in a traffic accident each 11 minutes and 6 seconds; and, one person was killed every 8 hours and 11 minutes in a traffic accident. The good news in the report is that these numbers were down from the previous year, and recent data from the U.S. Department of Transportation suggests a slight decline in traffic fatalities nationally in 1999 (*Traffic Safety Facts 1999*, U.S. Department of Transportation, National Highway Traffic Safety Administration).

Nevertheless, the cost in lives and injuries and loss of economic productivity is still unacceptably high. As new laws are contemplated to reduce driver distractions and to enforce existing safety measures, new analytical methodologies must be examined for providing additional, in depth analyses of the data being collected from the various existing data systems. Methodologies that help highlight patterns of safety problems, or safety problem clusters, would be of tremendous use to policy and decision makers as well as educators to enable them to formulate specific interventions for these areas.

In our study, we examined methodologies associated with the emerging technology of data mining to detect such patterns or clusters of safety problems. Data mining has become a very valuable data analytic process for detecting interesting patterns and clusters in large data sets. In their recently released book, Hand, Mannila, and Smyth define data mining as: "...the analysis of (often large) *observational* data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (Principles of Data Mining: MIT Press, 2001, p. 1)." Since traffic safety databases are, in general, observational, i.e., they have not been generated as a result of a designed experiment, and large, such an approach was clearly the logical one to pursue.

The TSEI was developed to assist policy and decision makers, and educators in focusing their safety related efforts within counties. The concept of this index is similar to that used by epidemiologists in reporting risk indices for various health issues. The basic index will be a weighted profile composed of a select group of variables. The data mining process was used to assist in the identification of these variables. Other statistical techniques such as multiple regression, analysis of covariance and cluster analysis were also employed. The goal was to produce an index on a countywide basis integrating the injury severity variable(s), e.g., property damage only, injury, and fatality, into the index.

In general, the TSEI will assist researchers and policy makers in summarizing large crash databases to make more informed decisions. This concise information will enable decision makers better allocate resources to decrease the economic loss due to crashes. Traffic safety educators also are given more evidence as to where they need to focus their

energies. In addition, the TSEI highlights counties with traffic safety problems, spurring further investigation into a specific county's characteristics.

1.2 Problem Statement

As stated above, the objective for our project was to employ principles and techniques from the emerging technology of data mining to analyze traffic safety data. Using several publicly available data sets, and working closely with the Alabama Traffic Safety Center at the University of Montevallo, we investigated and developed a traffic safety weighted risk profile on a countywide basis. The intention was to develop a profile or integrated risk assessment tool that incorporated usual crash data as well as county specific factors, such as county population, number of U.S. highway miles contained in the county, etc. A final aspect of our objective was the investigation of the construction of clusters of factors related to crashes using insights from the profiles.

An important data source for our investigations was the Critical Analysis Reporting Environment (CARE) system. The Alabama Department of Public Safety has developed and maintains a centralized database that contains traffic accident data collected from crash reports completed by local police officers and state troopers. CARE was developed by Dr. David Brown and the Computer Science Department of the University of Alabama and provides web-based access to this database along with some basic statistical summary capabilities, primarily one- and two-variable frequency distributions and bar charts. The CARE system allows for frequency distributions, bar charts and two-way frequency tables of the data. The CARE system can be accessed at <http://care.cs.ua.edu>. It is available in both online and downloadable versions. Past studies utilizing the CARE system can also be found at this website. Our project sought to further utilize this valuable data source and to promote its usefulness to policy and decision makers.

1.3 Overall Project Approach

In order to create the desired safety profile, the Traffic Safety Economic Index (TSEI), a method was needed to determine a method to compare crashes in Alabama counties fairly. For example, if we examine a large metropolitan county that has crashes resulting in 4,000 injuries in a year, and we have a small rural county with crashes resulting in 100 injuries in a year, how do we fairly compare these counties when their demographics differ so vastly?

From the traffic safety literature, it is known that demographic variables such as population density and income levels affect the expected number of crashes. For example, the larger the population density, the larger the number of crashes that typically occur. Hence, each county's demographic variables are used to adjust their crash statistics so that a fair comparison can be made of each county's economic impact of vehicular crashes.

As previously noted, one source of data used to create our index (TSEI) was the CARE database. This database contained the information needed to determine the number of property damages, injuries, and fatalities that resulted from crashes during the years 1998 through 2001. The Alabama Department of Transportation provided the number of miles of interstate, U.S. Highway, and state highways per county. These mileages provide an indication of transient inter-county traffic not associated with the characteristics of the county itself. This demographic information is available in the U.S. 2000 census.

2. Methodology

To create the TSEI, we followed these steps. First, we determined the important variables from the data. Second, using these variables, we predicted the number of crashes with property damages, injuries, and fatalities for each county. Third, we computed ratio of the observed crashes to the expected crash statistics based on the model in step two for all three types of crashes. Last, using these ratios, we weighted the economic impact of each type of crash to create the TSEI.

2.1 Determination of Variables

To determine the important variables that predict the amount of property damages, the number of injuries, and the number of fatalities from crashes, we used data mining techniques. Data mining techniques are designed to find information in large databases. The CARE database had about 130,000 observations each year. In particular, we used the data mining techniques of decision trees.

Several general computing algorithms have been proposed in the literature for decision trees. One approach is called CHAID, which uses a chi-square value to measure the worth of a split. Another approach is called CART (Breiman, Friedman, Olshen, and Stone 1984). The classification and regression trees (CART) methodology is a relatively new approach to the problem of predicting a response variable on the basis of several predictor variables. The CART algorithm uses a Gini measure to evaluate the value of data splits.

In typical decision tree analysis, the response variable is categorical, so that the problem reduces to one of classification based on auxiliary information. Predictors can be either categorical or quantitative. At the beginning of a decision tree algorithm, all of the observations are contained in the root node of what will eventually become a binary tree structure. The observations in the root node are split into two child nodes based on the values of a predictor variable. The predictor variable and splitting rule are chosen to make the observations within each node as homogeneous as possible with respect to the response variable. The splitting process continues with the child nodes. The process stops when the homogeneity within terminal nodes in the tree cannot be improved with further splits. The tree can be used to (1) classify (or predict) new observations and (2) estimate conditional probabilities of response categories based on the values of the predictor variables. One advantage of decision tree models is their ability to handle either categorical or quantitative variables and to more easily model complicated interactions than logistic regression models. We used SAS[®]'s Enterprise Miner software for the decision tree analysis in this project. The Enterprise Miner allows the user to select opinions related to the various general decision tree algorithms. In so doing, the Enterprise Miner can be used to construct a CART tree or a CHAID tree, or a tree based on some combination of the algorithms.

Using decision trees we tried to find variables that might help predict county level characteristics associated with the occurrence of a crash. Unfortunately, only a few of these variables were helpful at the county level because of the specificity to the individual crash. However, after interpreting the results, the TSEI augments the CARE database because one can then further investigate a county's problems at an individual crash level.

Using the data from the U.S. 2000 Census, we performed cluster analysis to determine which variables might be redundant. For example, the median household income level for a county and its percent of high school graduates are highly related. If two variables

are highly correlated, then sometimes only one of these variables is needed in the model to predict.

2.2 Regression Prediction Models

After determining the important variables, we built regression models to predict the amount of property damages, the number of injuries, and the number of fatalities that resulted from vehicular crashes. These models were developed using the average of the four years of data, 1998 through 2001. We averaged over the four years of data to smooth anomalous occurrences that may appear in given year of data. For example, if a rural county had a year with an unusual multiple injuries crash, this could severely penalize the county in the index. By averaging over time, the variability is reduced, and hence a better estimate of the county's traffic characteristics can be obtained

To satisfy the theoretical assumptions of linear regression, a transformation was applied to the data. Because of the large difference between Alabama's largest county, Jefferson County, and the rest of the counties in the state, Jefferson County would have had too much influence on the model. To correct for this problem, we transformed the number of property damages, injuries, and fatalities along with population by taking the logarithm of the number of crashes. These transformations clearly improved the linear fit to the data. For example, see Figure 2-1 for the plot of Property Damages versus Population after transformation. A straight-line pattern with a constant variance is apparent in the figure. This approach has been used previously by Page (2001).

After making these adjustments three models were developed. Separate models were built to predict the number of property damages, the number of injuries, and the number of fatalities. In each model a stepwise regression with Mallows' Cp statistic was used (Neter, Kutner, Nachstein and Wasserman, 1996) to determine the best set of important variables. The models performed well explaining almost all of the variability in the data. The models achieved R-square values above 95% for each of the three models. Table 2-1 details the specific models and also includes the values of R^2 and standard deviation. Note that the three models have a number of variables in common as well as a few unique variables. The variables used in these models are defined in Appendix I.

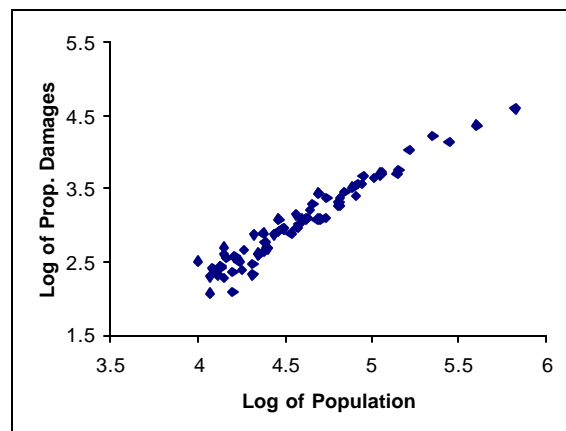


Figure 2-1. Plot of transformed values of property damage and population

2.3 Using Residuals to Adjust Crash Statistics

The regression models in section 2-2 provide estimates for the mean number of crashes for counties with varied demographics and highway usage. The residuals computed with these models, defined as the difference of the observed value and its predicted value, measure how much above or below the expected value a particular county falls. For example, a county with a large positive residual means that the observed crash statistics are greater than what was expected based on the regression models. To adjust for the disparities between counties, we used an analysis of covariance approach (Kleinbaum, Kupper, Muller and Nizam, 1998).

Table 2-1. Final Regression Models

Variables	Property Damage	Injury	Fatality
Intercept	3.03352	2.58719	1.07537
logTPopul	1.34393	1.17074	0.77754
Land Area	-0.00024	N/A	0.00012
Interstate	0.00317	0.00142	0.0037
Median Age	0.05172	0.04418	0.01661
PerEmpl16up	N/A	0.0125	0.00984
MedHHincome	-0.00002	-0.00002	-0.00002
PerAge15t19	0.0138	N/A	-0.05815
PerAge65up	-0.06823	-0.05396	-0.03848
Dry County	-0.0722	-0.049	-0.05256
SV	-0.08472	0.04399	0.08309
SV*Int	0.0074	0.00544	0.00908
SV*PAge15t19	-0.06478	N/A	N/A
US Route	0.00109	N/A	N/A
PerMale	-0.04008	N/A	N/A
Wet City	0.08098	N/A	N/A
PerWhite	N/A	-0.00181	N/A
AvgTravelTime	N/A		0.00788
R ²	96.7%	96.8%	96.0%
Std. Dev.	0.1175	0.0841	0.0674

Thus, we computed the residuals for each county based on the prediction for Alabama County's expected values. Our rationale for this approach is based on the strength of the regression models as evidenced by the high R² values (see table 2.1). Thus, these regressions adequately model the expected crash numbers as a function of these county demographic and highway mileage variables. Hence deviations from these expected values as measured by the residuals, provide indications of the level of traffic safety for each county.

2.4 Constructing the Traffic Safety Economic Index (TSEI)

To create the Traffic Safety Economic Index, we weighted the importance of each type of crash: property damage, injury, and fatality. The weights are based on estimates of the economic cost of each type of crash and the propensity that this type of crash occurs in relation to the other types was required. The *Injury Facts 2001 Edition* provided the following cost estimates:

Property Damages.....	\$6,500
Injuries	\$35,300
Fatality.....	\$1,000,000

For example, the estimated economic cost for each injury that occurs as a result of a vehicular crash is \$35,300.

To estimate the propensity of the three types of traffic crashes we used the average county level for each type. We also required the weights to sum to one so that the average index for the counties would be 100% for ease of interpretation. Therefore, the weights for each type of crash were formed by computing the economic cost of the average county's type of crash divided by the average county's total economic cost. The economic cost is the cost of the type of crash times the average number of crashes. The weights determined are 0.21564, 0.41905, and 0.36531 for property damage, injury, and fatality, respectively. The TSEI was formed using these weights and the information specific to each county's type of crash.

3. Results

The methodology outlined in the previous section was used to analyze the 1998-2001 crash statistics. In this section the specific results will be discussed.

3.1 Crash Data for Alabama

Table 3.1 contains the crash data (averaged over 1998-2001) for all Alabama counties for each type of crash. The average of all counties is also displayed at the bottom of the table. Note that Jefferson County has the highest level of crashes and Lamar County has the lowest levels. This is not surprising since Jefferson is the largest county and Lamar is a small rural county.

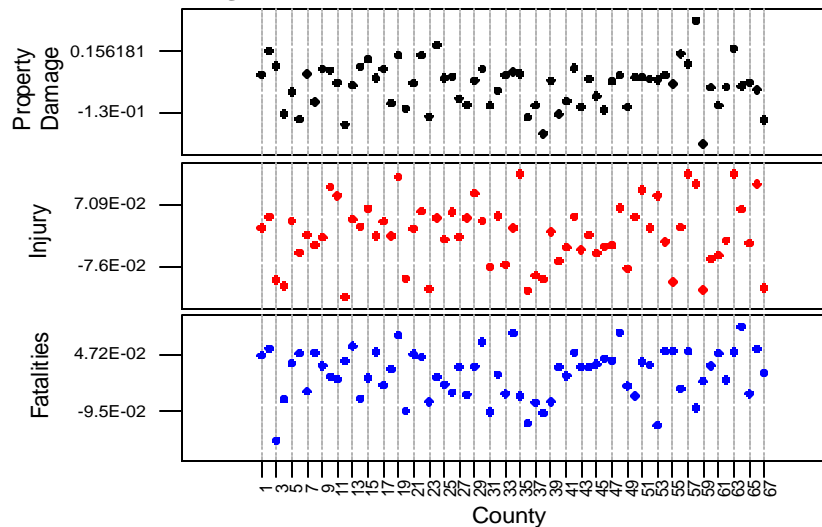
3.2 Regression Prediction Model Results

The final regression models are presented in Table 2.1. In developing these models various diagnostic measures were examined, including normality and residuals plots. No departures from the usual assumption of normality after applying transformations were detected. The residual plots for the transformed data are shown in Figure 3.1. No apparent patterns exist in any of these three plots.

Table 3-1 Average Crash Statistics for Alabama Counties, 1998-2001

County	Property Damage	Injury	Fatality	County	Property Damage	Injury	Fatality
1 Autauga	1103.5	306	11.75	34 Henry	236	101.75	5.25
2 Baldwin	3224	892.25	31.75	35 Houston	3059.5	1046.25	15.75
3 Barbour	594	187	6.25	36 Jackson	867.25	340.5	13.5
4 Bibb	142.25	82	5	37 Jefferson	24968.75	5167	89
5 Blount	830.25	332.5	15.25	38 Lamar	93	65	3.5
6 Bullock	101	70.75	4.75	39 Lauderdale	2227.25	580.25	13.25
7 Butler	548	195.25	9.75	40 Lawrence	495.5	221.75	10.75
8 Calhoun	3249.25	909.5	26.5	41 Lee	3568.5	894	20
9 Chambers	847.75	251.5	9.75	42 Limestone	1577.5	499.5	17
10 Cherokee	365.5	178.75	6.5	43 Lowndes	232.25	118.25	8.25
11 Chilton	847.5	323.5	12.5	44 Macon	568.25	210.75	8.75
12 Choctaw	154.25	77.75	7	45 Madison	8430	2217.25	43
13 Clarke	427.25	190.75	9.75	46 Marengo	218	136.25	7
14 Clay	148.5	75.75	3.5	47 Marion	527	196.25	9.75
15 Cleburne	330.5	129.5	9	48 Marshall	2269.25	671.5	20.75
16 Coffee	899.5	243	8.25	49 Mobile	13766.75	3370.5	68.5
17 Colbert	1546.5	425.25	13	50 Monroe	333.5	153	4.5
18 Conecuh	313.25	135.75	10.75	51 Montgomery	9995.25	2391.75	36.75
19 Coosa	167	103.5	5.5	52 Morgan	3237	810.75	20.25
20 Covington	554.25	214.5	8	53 Perry	117.25	77	2.25
21 Crenshaw	168.25	81	4.75	54 Pickens	220.5	107.25	6.5
22 Cullman	2136.5	594.75	20	55 Pike	742.75	180.75	9.25
23 Dale	786	249	8.75	56 Randolph	333.75	126.5	5.25
24 Dallas	1249	434	10.75	57 Russell	1796.5	602	16.5
25 DeKalb	1405	474	17	58 Shelby	3908.75	854.5	19.5
26 Elmore	1343.75	458.75	13	59 St. Clair	1293.5	446.25	19
27 Escambia	671.5	295	12.5	60 Sumter	264.25	109.5	8.5
28 Etowah	2816	833.75	19	61 Talladega	1779.5	575.5	20.75
29 Fayette	280.25	126.75	5.5	62 Tallapoosa	782.5	283	11
30 Franklin	562.75	214.5	10.75	63 Tuscaloosa	6637.75	1657.75	33
31 Geneva	352	150	6.25	64 Walker	1884	600.75	25
32 Greene	233.5	117	6.25	65 Washington	141.5	92	5.25
33 Hale	218.75	90	4.25	66 Wilcox	158.75	114.5	5.5
				67 Winston	292.25	124.25	8
Overall Mean	1875.24	516.22	14.39				

Figure 3.1 Residuals from Traffic Crash Predictions



3.3 Expected Crash Statistics Results

The Expected Crash Statistics were obtained using the Regression models of section 2.2. These predictions are for the logarithm of the crash numbers. To obtain predicted values on the original scale the predictions were transformed back to the original units. Table 3.2 contains the Expected Crash Statistics for all types of crashes by county.

For Jefferson County the expected number of property damage accidents is 48,3378.2 while for Lamar County the expected number is 205.4. These values can be contrasted with the observed numbers in Table 3-1, which were 24,968.75 for Jefferson and 93 for Lamar County. In both cases, fewer crashes were observed in 1998-2001 than were expected considering the demographic characteristics. On the other hand, Shelby, Tuscaloosa and Walker counties are examples of counties where the observed crashes exceed the expected values from the models. The ratio of these values, observed versus expected, will be used as a measure of traffic safety in the next section.

Table 3-2 Expected Crash Statistics for Alabama Counties

County	Property Damage	Injury	Fatality	County	Property Damage	Injury	Fatality
Autauga	1446.0	387.3	11.5	Henry	273.3	136.7	4.7
Baldwin	3434.4	1089.7	31.5	Houston	4149.8	1037.7	19.9
Barbour	712.4	330.4	10.0	Jackson	1730.3	617.9	18.6
Bibb	259.3	150.5	6.9	Jefferson	48378.2	8183.3	110.5
Blount	1275.9	433.4	15.8	Lamar	205.4	113.3	4.4
Bullock	172.0	117.1	4.4	Lauderdale	3336.0	768.6	16.8
Butler	686.2	290.9	12.5	Lawrence	927.9	355.9	11.3
Calhoun	5920.4	1308.7	24.2	Lee	6397.3	1266.2	21.5
Chambers	1060.8	356.9	10.0	Limestone	1937.2	627.4	17.7
Cherokee	424.7	187.6	7.6	Lowndes	365.2	183.1	8.6
Chilton	1160.7	353.0	13.9	Macon	716.0	299.8	9.4
Choctaw	314.3	158.0	7.4	Madison	14747.2	3202.1	44.0
Clarke	652.4	260.8	9.2	Marengo	397.6	207.1	6.9
Clay	160.1	101.6	4.4	Marion	756.5	284.7	9.7
Cleburne	333.7	158.4	9.5	Marshall	3149.4	794.4	18.5
Coffee	1225.9	341.5	9.1	Mobile	27552.4	5294.2	80.8
Colbert	2045.0	548.1	13.9	Monroe	464.1	196.1	6.6
Conecuh	492.6	197.1	10.7	Montgomery	14638.3	2493.5	36.8
Coosa	154.2	101.1	4.7	Morgan	4602.5	1085.6	22.3
Covington	1070.3	368.4	11.8	Perry	148.3	85.3	3.4
Crenshaw	239.8	111.5	5.2	Pickens	274.5	157.4	5.7
Cullman	2298.5	725.0	19.7	Pike	1092.4	325.1	8.6
Dale	1593.2	452.6	11.7	Randolph	337.6	165.9	6.3
Dallas	1229.8	550.0	11.7	Russell	2177.0	642.8	16.1
DeKalb	1890.8	657.2	19.1	Shelby	3034.4	821.1	26.2
Elmore	1758.8	549.7	15.6	St. Clair	3411.1	816.0	22.3
Escambia	1111.0	414.7	13.3	Sumter	368.7	175.8	9.4
Etowah	5304.0	1030.0	21.7	Talladega	3242.7	897.2	20.0
Fayette	397.1	142.6	5.6	Tallapoosa	1169.6	402.0	11.7
Franklin	690.1	280.1	9.4	Tuscaloosa	6941.1	1617.8	32.4
Geneva	608.2	253.7	8.4	Walker	2863.2	696.0	20.6
Greene	325.5	145.0	6.8	Washington	189.8	130.3	6.2
Hale	274.6	147.4	5.4	Wilcox	218.8	128.0	5.0
				Winston	635.7	246.9	9.2

3.4 Traffic Safety Economic Index Values

The ratios of the observed crash statistics over the expected crash statistics yields a measure of the safety level for all counties for property damage, injury and fatal traffic crashes. These ratios were scaled so that the average values were 100%. This scaling enabled the ratio to be used as a benchmark when comparing counties. All three of these ratios were combined in a weighted average to create the TSEI values. This index serves as a single numeric measure of the overall traffic safety level for each county. The weights used in the index were based on the average economic impact of the three types of crashes. Table 3-3 includes the individual ratios as well as the Traffic Safety Economic Index values for all sixty-seven Alabama counties.

Table 3-3 Ratios and TSEI Values for Alabama Counties

county	Property Damage	Injury	Fatal	TSEI	county	Property Damage	Injury	Fatal	TSEI
Autauga	1.092	1.017	1.101	1.068	Houston	1.101	1.373	0.873	1.077
Baldwin	1.401	1.087	1.147	1.134	Jackson	0.694	0.722	0.744	0.733
Barbour	1.196	0.764	0.675	0.731	Jefferson	0.786	0.784	0.841	0.817
Bibb	0.714	0.742	0.856	0.806	Lamar	0.583	0.769	0.790	0.774
Blount	0.907	1.060	1.050	1.048	Lauderdale	1.027	0.998	0.844	0.911
Bullock	0.680	0.891	1.117	1.012	Lawrence	0.718	0.849	1.032	0.948
Butler	1.098	0.981	0.895	0.937	Lee	0.823	0.917	0.980	0.949
Calhoun	0.818	0.926	1.118	1.031	Limestone	1.167	1.086	1.118	1.107
Chambers	1.156	0.967	1.039	1.015	Lowndes	0.775	0.906	1.031	0.972
Cherokee	1.146	1.280	0.976	1.101	Macon	1.042	0.978	1.031	1.011
Chilton	1.004	1.220	0.961	1.063	Madison	0.870	0.885	1.047	0.977
Choctaw	0.644	0.697	1.070	0.908	Marengo	0.748	0.920	1.083	1.006
Clarke	0.974	1.067	1.162	1.117	Marion	1.016	0.927	1.070	1.012
Clay	1.184	1.025	0.858	0.936	Marshall	1.080	1.140	1.259	1.205
Cleburne	1.283	1.135	0.970	1.046	Mobile	0.775	0.816	0.923	0.876
Coffee	1.051	0.974	1.124	1.063	Monroe	1.061	1.086	0.873	0.963
Colbert	1.159	1.056	0.928	0.987	Montgomery	1.054	1.255	1.063	1.137
Conecuh	0.802	0.976	1.018	0.993	Morgan	1.044	1.017	1.042	1.032
Coosa	1.339	1.349	1.235	1.283	Perry	1.035	1.216	0.736	0.935
Covington	0.762	0.771	0.801	0.788	Pickens	1.082	0.946	1.131	1.057
Crenshaw	0.993	1.015	1.110	1.068	Pike	0.989	0.756	1.134	0.981
Cullman	1.339	1.117	1.094	1.113	Randolph	1.356	1.023	0.911	0.972
Dale	0.703	0.730	0.845	0.795	Russell	1.222	1.370	1.128	1.226
Dallas	1.493	1.078	0.971	1.034	Shelby	1.930	1.297	0.813	1.046
DeKalb	1.051	0.956	0.933	0.947	St. Clair	0.524	0.723	0.947	0.843
Elmore	1.065	1.111	0.889	0.983	Sumter	0.954	0.858	1.037	0.964
Escambia	0.846	0.971	1.031	1.000	Talladega	0.784	0.878	1.115	1.010
Etowah	0.794	1.075	0.880	0.953	Tallapoosa	0.958	0.954	0.958	0.957
Fayette	1.026	1.231	1.033	1.110	Tuscaloosa	1.438	1.372	1.124	1.233
Franklin	1.158	1.058	1.191	1.138	Walker	0.968	1.128	1.301	1.220
Geneva	0.785	0.823	0.798	0.807	Washington	1.000	0.936	0.883	0.908
Greene	0.924	1.088	0.989	1.025	Wilcox	0.928	1.299	1.142	1.194
Hale	1.082	0.832	0.882	0.871	Winston	0.674	0.732	0.997	0.881
Henry	1.129	1.019	1.255	1.158					

After computing the values for each county, the counties were grouped into five tiers based on the value of TSEI. The tiers were defined in Table 3-3 below. We color coded each tier and displayed them visually to allow for patterns to emerge that previously had not been discovered. Unfortunately, no clear patterns were found.

Table 3-4 TSEI Tiers

Tier	TSEI Range
1	TSEI ≤ 80%
2	80% ≤ TSEI < 90%
3	90% ≤ TSEI < 110%
4	110% ≤ TSEI < 120%
5	120% ≤ TSEI

As a final check on the validity of the TSEI, we clustered the counties based on the important demographic variables in the model to determine which counties were the most similar. This method enabled us to group the counties into four peer groups based on demographic variables. We then checked that within each peer group cluster, we had counties from all tiers were represented. This mix of counties indicates that the index values are not biased against counties that were similar in characteristics.

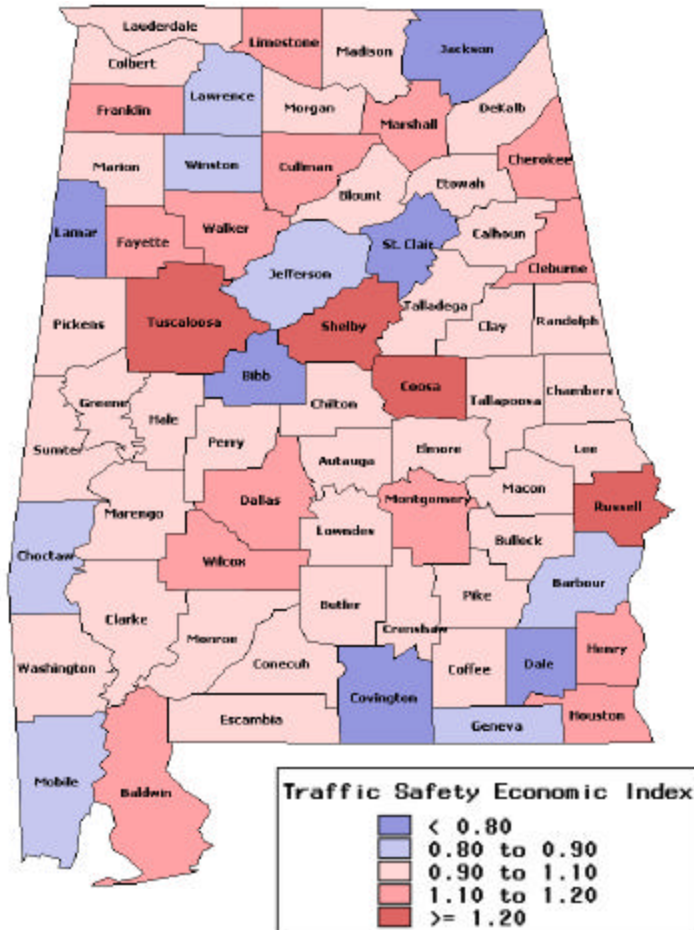


Figure 3-2. Alabama counties grouped by TSEI values

4. Project Conclusions and Recommendations

The economic impact of traffic crashes varies widely across the state of Alabama. This variation in crash occurrence may have some consequence for traffic safety measures in some counties. Those counties with higher than expected economic impacts may need to study traffic patterns within their county to determine whether any corrective actions are needed. It is beyond the scope of this study to determine which traffic crash excesses are due to poor roadway conditions or the necessity of stop signs or traffic lights. The TSEI serves as a benchmark for county-to-county comparisons. The determination of specific precautionary measures to reduce crashes requires study of other factors such as the identification of specific road sections that need work, etc.

A number of extensions and refinements to our methodology could be examined. In particular, techniques for down weighting very unusual values from a county could be considered. At present, an unusually bad crash in a county resulting in a large number of fatalities and damage during the years represented in the data used for analysis can place the county's TSEI value in the upper tier. Several statistical techniques are available that may assist in adjusting this problem. A systematic method for updating the index year after year could be considered. For example, the models may be based on moving averages or weighted moving averages over time to facilitate the addition of another years' crash statistics. Additionally, further explorations of traffic safety data using data mining techniques are warranted.

5. References

- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Belmont, CA: Wadsworth International.
- Critical Analysis Reporting Environment. 1998 through 2001 Accident Database
- Hand, David, Mannila, Heikki, and Smyth, Padhraic (2001) *Principles of Data Mining*, MIT Press.
- Kleinbaum, David G., Kupper, Lawrence L., Muller, Keith E. and Nizam, Azhar (1998). *Applied Regression Analysis and Multivariate Methods*, Third edition, Pacific Grove, CA: Duxbury Press,.
- National Safety Council's Traffic Facts, 2001 Edition.
- Neter, John, Kutner, Michael H., Nachsteim, Christopher J. and Wasserman, William (1996). *Applied Linear Regression Models*, third edition, Boston, MA: Irwin.
- Page, Yves. "A Statistical Model to Compare Road Mortality in OEC Countries". *Accident, Analysis, and Prevention*. May 2001.

Appendix I

Data Dictionary for Variables in the Study

Variables	Description
Land Area	Size of the County in square miles
Interstate	Number of interstate highway miles in the county
SV	Indicator variable that denotes whether the majority of traffic crashes involved a second vehicle
US Route	Number of U.S. highway miles in the county
Wet City	Indicator variable that denotes whether a city permits the sale of alcoholic beverages
Dry County	Indicator variable that denotes whether a county permits the sale of alcoholic
Based on 2000 Census	
logTPopul	Natural logarithm of the total county population
Median Age	Median age for the county based on 2000 Census
PerEmpl16up	Proportion of persons 16 years of age or older in the county that are employed
MedHHincome	Median Household Income
PerAge15t19	Proportion of county residents that are 15-19 years of age
PerAge65up	Proportion of county residents that are 65 years of age or older
PerMale	Proportion of males in the county
PerWhite	Proportion of county residents that are Caucasian
AvgTravelTime	Average commute time for county residents