TRAFFIC FATALITIES IN MAJOR U. S. CITIES

Olga J. Pendleton
Lindsay I. Griffin
Gary Stevens

Texas Transportation Institute
The Texas A & M University System
College Station, Texas 77843

Abstract

The relationship between the number of traffic fatalities and population in major U.S. cities is examined to determine if this relationship is consistent among all states. A linear regression model was found to adequately represent this relationship based on cities with populations of more than 50,000. Regression diagnostics for identifying states which deviate significantly from the nationwide trend are examined and various states were found to be significantly over or under represented in traffic fatalities as compared to the nationwide trend.

TRAFFIC FATALITIES IN MAJOR U. S. CITIES

Olga J. Pendleton
L. I. Griffin
Gary Stevens

Texas Transportation Institute
The Texas A & M University System
College Station, Texas 77843

INTRODUCTION

It has long been assumed that traffic fatality rates and population density are directly, but inversely related. However, certain geographical areas appear to have higher per capita fatality rates than would be expected from the nationwide trend. In order to identify these areas, a model representing the nationwide trend must be constructed.

The authors' interest in this area began in 1974 when an article published by Griffin showed that there is an inverse relationship between a state's motor vehicle death rate (fatalities per million vehicle miles traveled) and population density. At the time that this article was written it was assumed that the more sparsely populated states were disproportionately characterized by high speed, intercity driving. The more densely populated states were disproportionately characterized by more low speed, intracity driving. States in the South, the West and the Southwest generally displayed sparser population densities and higher motor vehicle death rates. States in the East, Northeast, and Midwest were more densely populated and displayed a lower motor vehicle death rate.

The purpose of this paper is to compare fatality rates in major cities throughout the United States, and to determine if cities in the South, West, and Southwest are over represented in traffic accident fatalities when compared to the rest of the nation. If such a phenomenon can be demonstrated then the simplistic explanation offered for the inverse relationship between a state's motor vehicle death rate and population density may be too simplistic. Perhaps the higher motor vehicle death rate exhibited by many of the states in the South, West, and Southwest result from rapid growths in population - particularly in major urban areas.

METHODS

Traffic fatalities for 1980 were analyzed as a function of population using ordinary least squares regression analysis. All U. S. cities with populations greater than 50,000 for 1980 according to 1980 Census Data were included in the analysis with exeception of Rancho Cucamonga, California and Columbus, Georgia. Traffic fatality data was not available for these cities. These cities will be referred to as major U. S. cities. The number of fatalities for this time period was obtained from the Fatal Accident Reporting System (FARS) of the National Highway Traffic Safety Administration. A total of 4,366 fatalities in 416 cities were included in the analysis. (Table 1 summarizes these data by state.)

A plot of fatality versus population per thousand (Figure 1) suggested a straight line relationship between these variables. Two potential problems with assuming a straight line relationship need to be considered before beginning model fitting.

1. Do the data which are clustered at the lower values of fatalities and population appear to be linearly related, and

2. Are the statistical assumptions required for parametric testing in the simple linear regression model satisfied?

To answer question 1, a plot of the data for major U.S. cities with populations less than 100,000 (Figure 2) was examined. The linearity assumption still appeared to be valid for the less populated major U.S. cities.

Question 2 was examined by verification of the assumption of normal, independent random errors (e) in the linear regression

model

$$y = b_0 + b_1 x + e \qquad (1)$$

where y is the number of fatalities, x is the population (per thousand), b_0 and b_1 are estimates of parameters representing the intercept and slope of the linear relation, and e is normally distributed independent random error.

Typically, count data, such as number of fatalities, do not satisfy these normality assumptions. A potential solution to this problem is to find a transformation which will tend to "normalize" these errors. For example, if the dependent variable represents a poisson distributed random variable, the log of the random variable and hence, the error, will approximate a normal distribution. Other alternatrives such as non-parametric techniques could be considered; however, these methods are generally less powerful than parametric methods, such as least squares, for testing hypotheses about the model parameters. Hence, in situations when the sample size is sufficiently large (as in this study), transformations which will enable parametric testing procedures are generally desirable. In order to determine which transformations, if any, were suitable for this analysis, the method of Box and Cox (1964) was used. Results indicated that no transformation of the variables was required to satisfy the error assumptions required for this model.

In addition to the parameter estimates of the slope and intercept of the fitted line, b_0 and b_1, other statistics of interest in least squares regression are the mean squared error (MSE) and the squared multiple correlation coefficient (R^2). The

mean squared error (residual mean square) is the averaged squared deviations of the actual and predicted values, i.e.

$$MSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(n-2) \qquad (2)$$

where y_i is the number of fatalities for the ith city and $\hat{y}_i$ is the predicted number of fatalities for the ith city according to the fitted model, i=1, ..., n cities. This statistic provides a measure of the goodness of fit of the model. The squared multiple correlation coefficient, R^2, gives the percentage of the total variability in the number of fatalities among cities which is accounted for by the model.

One of the weaknesses of ordinary least squares regression is that the estimates are very sensitive to outliers or extreme values. That is, a single observation can influence the estimated intercept and slope to the degree that the fitted line no longer "fits" the bulk of the data. Figure 3 is a hypothetical example of such a case. These observations often have very small residuals, i.e. deviations from the fitted line, because of their influence. Furthermore, whereas these influential observations are sometimes obvious in two dimensions, in multiple regression and situations with many observations, these influential observations are often hidden. Fortunately, diagnostics have been developed for isolating these observations (Belsley, Kuh, and Welsch 1980). A particular diagnostic used in this study is the studentized residual:

$$t_i = \frac{e_i}{\sqrt{MSE\ (1-h_i)}} \qquad (3)$$

where e_i is the deviation of the actual number of fatalities for the ith city from the predicted value (residual) and

$$h_i = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum_{j=1}^{n} (x_j-\bar{x})^2} \qquad (4)$$

where x_i is the population for the ith city, x is the average population over all cities and n is the total number of cities. The reasoning behind the use of this diagnostic is that it gives a measure of the magnitude of a residual, relative to its variability. This statistic has a t - distribution and hence a statistical test of significance for testing whether or not a particular studentized residual is out of range can be made with (n-3) degrees of freedom. This diagnostic was used in this study to isolate observations or groups of observations which were influencing the least squares estimates.

RESULTS

A least squares regression model of the form

$$y_i = b_0 + b_1 x_i + e_i \qquad (5)$$

where y_i is the number of fatalities for the i th city, x_i is the population per thousand of the i th city, b_0 and b_1 are least squares estimates of the slope and intercept, respectively, and e_i is the i th residual, was estimated for all major U.S. cities collectively and separately for states which had ten or more major cities represented. The parameter estimates of the regression lines, square root of the residual (error) mean square (MSE), squared multiple correlation coefficient (R^2), and number of observations is listed in Table 2. Points of interest in Table 2: are that the slope for Texas cities is higher than any other states, the variability about the fitted line (MSE) is highest in Florida, and the state with the lowest slope is Pennsylvania.

Graphical representations of Table 2 are of interest. In order to visualize the trend for a particular state with respect to the nationwide trend, plots such as Figure 4 may reveal interesting conclusions. For example, in Figure 4 Texas cities only are plotted with respect to the solid line representing the nationwide model. It is of interest to note that all cities of population greater than 110,000 were overrepresented in accident fatalities per capita.

Conceivably, plots for each state could be examined in this manner. This task becomes burdensome with numerous subgroupings,

such as 50 states for example, and is not feasible when many variables are incorporated in the model building process. A more feasible alternative would be to examine diagnostics for identifying extreme or influential observations. One graphical means of doing this is to examine these diagnostics by way of an index plot.

Figure 5 is the index plot for the studentized residuals (3) in this example. the vertical axis represents the studentized residuals centered about zero. The x-axis refers to observation number which is ordered by state alphabetically. The horizontal lines at $\pm$ 1.96 represent the critical t-values (at the 5% level of significance with 413 (416-3) degrees of freedom) defining limits on the ranges of the studentized residuals, i.e. any cities above or below this line are considered extreme. Table 3 lists the number of significant extreme values by state. Note that Texas had the largest number of positive extreme studentized residuals (4/15 = 27%) indicating that Texas major cities as a group have higher fatalities than would be expected based on the nationwide trend (fitted line). These studentized residuals identify certain major cities as "outliers".

The outlier cities, corresponding to Figure 5, are listed in Table 4 in order of magnitude, largest to smallest. The most influential city was New York City which had a lower number of fatalities than would be expected. The only other city with a significantly low studentized residual was Pittsburgh. Chicago, Washington D.C., and Baltimore were influential but have nonsignificant studentized residuals because of the influence of New

York City. This is commonly referred to as the "masking" problem in linear regression, that is, one influential observation obscures others. The only other city which had an extremely large number of fatalities over the expected was Los Angeles. Note that of the six cities with the largest studentized residuals, four were in Texas.

Table 5 lists the sum of the studentized residuals in order of magnitude, by state. Texas has the largest value indicating that Texas' major cities have an extremely large positive deviation from the expected with respect to fatalities. Since the studentized residuals, like ordinary residuals, take on positive and negative values, a state whose cities did not deviate extremely in one direction would be expected to have a small, near zero, sum of studentized residuals.

Another method of representation of these residuals which reveals an interesting geographic distribution of under and over representation of accident fatalities is shown in Figure 6. States whose sum of studentized residuals in Table 4 were positive are shaded darker than those for whom this value is negative. Note the geographical band of states overrepresented in per capita traffic fatalities along the Western, Southwestern, and Southern states. The upper Midwest and East Coast have lower per capita fatalities.

Many theories could be proposed to explain this phenomenon. Capelle (1983) noted that many of these South, West and Southwest states are considered "magnet" areas (Aero Mayflower Transit Company, 1982); i.e. the majority of moves have been into these states. Of the 14 states identified as magnet areas, only three

(Colorado, Arkansas, and North Carolina) were considered underrepresentative of accident fatality by our analysis. Following this train of thought, the percentage change in population from 1970 to 1980 was examined. An arbitrary boundary of 15% was selected and states grouped into two categories, those with greater than or equal to a 15% increase in population between 1970 and 1980 and those with a decrease or less than a 15 percent increase in population in this same period (Figure 7). Note the similarities in states which experienced a large increase in population and those which are overrepresented in traffic fatalities. Disagreement occurred in 12 states as noted in Table 6.

SUMMARY

The purpose of this paper has been two-fold: 1. to develop a model based on traffic fatalities and population for U.S. cities with populations greater than 50,000 and 2. to show how certain statistical measures of deviation from the model can identify subgroups of the data set which deviate in a consistent manner from this model.

The model fit to the data was considered appropriate and adequate as judged by the statistical measures of R^2 and mean squared error. This model revealed that 87 percent of the variability in accident fatality could be explained by the single variable, population, and the mean squared error of 21.2 indicated that this model was accurate in estimating the true average fatality within $\pm$ 2 fatalities with 95% confidence. Assumptions of normality were tested and judged to be reasonable.

Studentized residuals revealed cities which deviated significantly from the nationwide trend represented by this model. Spatial analysis of the geographical distribution of these cities and their effect on the state as a whole with regard to over-or underrepresentation of per capita fatality revealed that states which were overrepresented appeared to be situated in along the Western, Southwestern, and Southern areas. The upper Midwestern and Eastern coastal states were underrepresented with regard to per capita fatalities. A possible explanation for this distribution could be found in examining differences in population growth among the states. Examining the percent change in population between 1970 and 1980 revealed a similar geographical distribution pattern.

In summation this paper has attempted to describe the relationship between traffic fatalities and population. By examining statistical measures of deviations from a fitted model representing the nationwide trend of this relationship, a geographical pattern in the distribution of states whose major cities experienced either higher or lower accidents than would be expected based on the national model was identified. The use of these diagnostics for identifying such patterns, particularly in more complex, multivariable models, should be considered in any model building effort.

References

Box, G.E.P. and D.R. Cox, *The Analysis of Transformations*. Journal of the Royal Statistical Society 26, 211, 1964.

Belsley, D.A., E. Kuh, and R.E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Multicollinearity*. John Wiley and Sons, N.Y., 1980.

Capelle, R.B., *The Missing Dimension In Truck Accident Research: SPATIAL ANALYSIS*. Presented at the meetings of the Association of American Geographers, Denver, Colorado, April 25, 1983.

Griffin, L.I., *The Effects of Population Patterns on the Motor Vehcile Death Rate*. Highway Safety Highlights A(2), 1974.

TABLE 1

FATALITIES AND POPULATION FOR 1980
FOR MAJOR U.S. CITIES

STATE	NUMBER OF MAJOR CITIES	AVERAGE POPULATION*	AVERAGE FATALITIES
ALL	416	178.127	24.5697
ALABAMA	5	176.136	22.8000
ALASKA	1	173.017	25.0000
ARIZONA	6	256.811	57.8333
ARKANSAS	4	87.710	9.2500
CALIFORNIA	81	156.416	23.8642
COLORADO	10	140.878	16.5000
CONNECTICUT	12	86.320	12.9167
DELAWARE	1	70.195	5.0000
D.C.	1	637.651	41.0000
FLORIDA	17	151.231	32.5294
GEORGIA	4	189.352	36.7500
HAWAII	1	365.114	50.0000
IDAHO	1	102.451	6.0000
ILLINOIS	20	220.101	23.2000
INDIANA	10	161.357	26.8000
IOWA	8	91.472	10.1250
KANSAS	5	138.029	19.6000
KENTUCKY	3	185.689	29.3333
LOUISIANA	8	164.417	24.8750
MAINE	1	61.572	7.0000
MARYLAND	1	786.775	66.0000
MASSACHUSETTS	18	111.956	10.7222
MICHIGAN	22	144.259	17.8636
MINNESOTA	5	174.736	18.2000
MISSISSIPPI	1	202.895	27.0000
MISSOURI	7	191.470	31.5714
MONTANA	2	61.761	7.0000
NEBRASKA	2	241.806	27.5000
NEVADA	2	132.715	22.5000
NEW HAMPSHIRE	2	79.400	9.5000
NEW JERSEY	13	108.750	10.2308
NEW MEXICO	1	331.767	49.0000
NEW YORK	13	661.755	56.6154
NORTH CAROLINA	8	128.684	17.7500
NORTH DAKOTA	1	61.308	4.0000
OHIO	19	170.583	23.7895
OKLAHOMA	5	192.514	39.2000
OREGON	3	187.080	30.3333
PENNSYLVANIA	11	253.534	19.3636
RHODE ISLAND	5	87.621	9.0000
SOUTH CAROLINA	4	73.169	7.5000
SOUTH DAKOTA	1	81.343	5.0000
TENNESSEE	5	301.898	62.2000
TEXAS	33	213.717	43.4242
UTAH	4	88.436	14.5000
VIRGINIA	10	150.510	15.4000
WASHINGTON	5	190.393	27.8000
WEST VIRGINIA	2	63.826	10.0000
WISCONSIN	11	125.942	9.0909
WYOMING	1	51.016	3.0000

* PER THOUSAND

Table 2

Least Squares Regression Results

Major Cities	Intercept	Slope $\times 10^3$	$\sqrt{MSE}$	R^2	n
All Major Cities in U.S.	5.19	.109	21.20	.83	416
California	-2.98 *	.172 *	6.40	.99	81
Colorado	-5.76 *	.158 *	3.47	.98	10
Connecticut	1.25	.135 *	3.97	.60	12
Florida	-1.96	.228 *	13.30	.84	17
Illinois	0.69	.102 *	3.31	.99	20
Indiana	2.90	.148 *	5.93	.96	10
Massachusetts	-0.07	.096 *	3.89	.90	18
Michigan	-1.18	.132 *	4.96	.98	22
New Jersey	0.12	.096 *	4.49	.75	15
New York	-1.03	.087 *	4.61	.99	13
Ohio	-5.89 *	.174 *	6.59	.96	19
Pennsylvania	4.09	.060 *	7.92	.94	11
Texas	-6.15 *	.232 *	12.31	.97	33
Virginia	-2.90	.122 *	5.51	.74	10
Wisconsin	-2.13	.089 *	3.65	.95	11

* indicates significantly different from zero at the 5% level

Table 3

Frequency of Influential Observations

by State

State	Number of Studentized Residuals $\geq \pm 1.96$	
	Positive	Negative
Arizona	2	0
California	2	0
Florida	2	0
Missouri	1	0
New York	0	1
Ohio	1	0
Oklahoma	1	0
Pennsylvania	0	1
Tennessee	2	0
Texas	4	0
	15	2

Table 4

Cities With Large Studentized Residuals

RANK	CITY	ti
1	Los Angelas	9.117
2	Houston	9.111
3	Dallas	4.950
4	Phoenix	4.011
5	Fort Worth	3.969
6	San Antonio	3.189
7	Miami	2.885
8	Oklahoma City	2.832
9	Jacksonville	2.693
10	Tucson	2.681
11	Kansas City	2.601
12	Nashville	2.327
13	San Diego	2.301
14	Cleveland	2.098
15	Philadelphia	-4.212
16	New York City	-12.151

TABLE 5

SUM OF STUDENTIZED RESIDUALS
BY STATE ORDERED LARGEST TO SMALLEST

STATE	Σti	STATE	Σti
TEXAS	23.498	DELAWARE	-0.370
FLORIDA	8.744	NORTH DAKOTA	-0.371
ARIZONA	7.010	NEBRASKA	-0.378
CALIFORNIA	6.818	NEW HAMPSHIRE	-0.409
TENNESSEE	5.711	SOUTH DAKOTA	-0.427
OKLAHOMA	3.086	MONTANA	-0.464
GEORGIA	2.071	IDAHO	-0.488
INDIANA	1.917	NORTH CAROLINA	-0.544
MISSOURI	1.836	CONNECTICUT	-0.944
LOUISIANA	0.679	ARKANSAS	-1.036
OREGON	0.679	SOUTH CAROLINA	-1.068
KENTUCKY	0.558	MARYLAND	-1.174
WASHINGTON	0.448	RHODE ISLAND	-1.352
NEW MEXICO	0.364	MINNESOTA	-1.417
NEVADA	0.271	D.C.	-1.587
HAWAII	0.240	IOWA	-1.896
ALASKA	0.047	COLORADO	-1.897
OHIO	0.039	VIRGINIA	-2.912
MISSISSIPPI	-0.012	MICHIGAN	-3.133
UTAH	-0.059	NEW JERSEY	-4.170
KANSAS	-0.143	WISCONSIN	-5.095
WEST VIRGINIA	-0.202	MASSACHUSETTES	-5.653
MAINE	-0.231	ILLINOIS	-5.672
WYOMING	-0.366	PENNSYLVANIA	-7.030
ALABAMA	-0.367	NEW YORK	-17.249

Table 6

States which did not agree in direction of migration and Traffic fatalities

State	percent change population 1970 - 1980 + 15% - 15%	traffic fatality + (overrepresented) - (underrepresented)
Arkansas	+ (18.8)	-
Colorado	+ (30.7)	-
Idaho	+ (32.4)	-
Indiana	- (5.7)	+
Kentucky	- (13.7)	+
Missouri	- (5.1)	+
New Hampshire	+ (24.8)	-
North Carolina	+ (15.5)	-
Ohio	- (1.3)	+
South Carolina	+ (20.4)	-
Utah	+ (37.9)	-
Wyoming	+ (41.6)	-

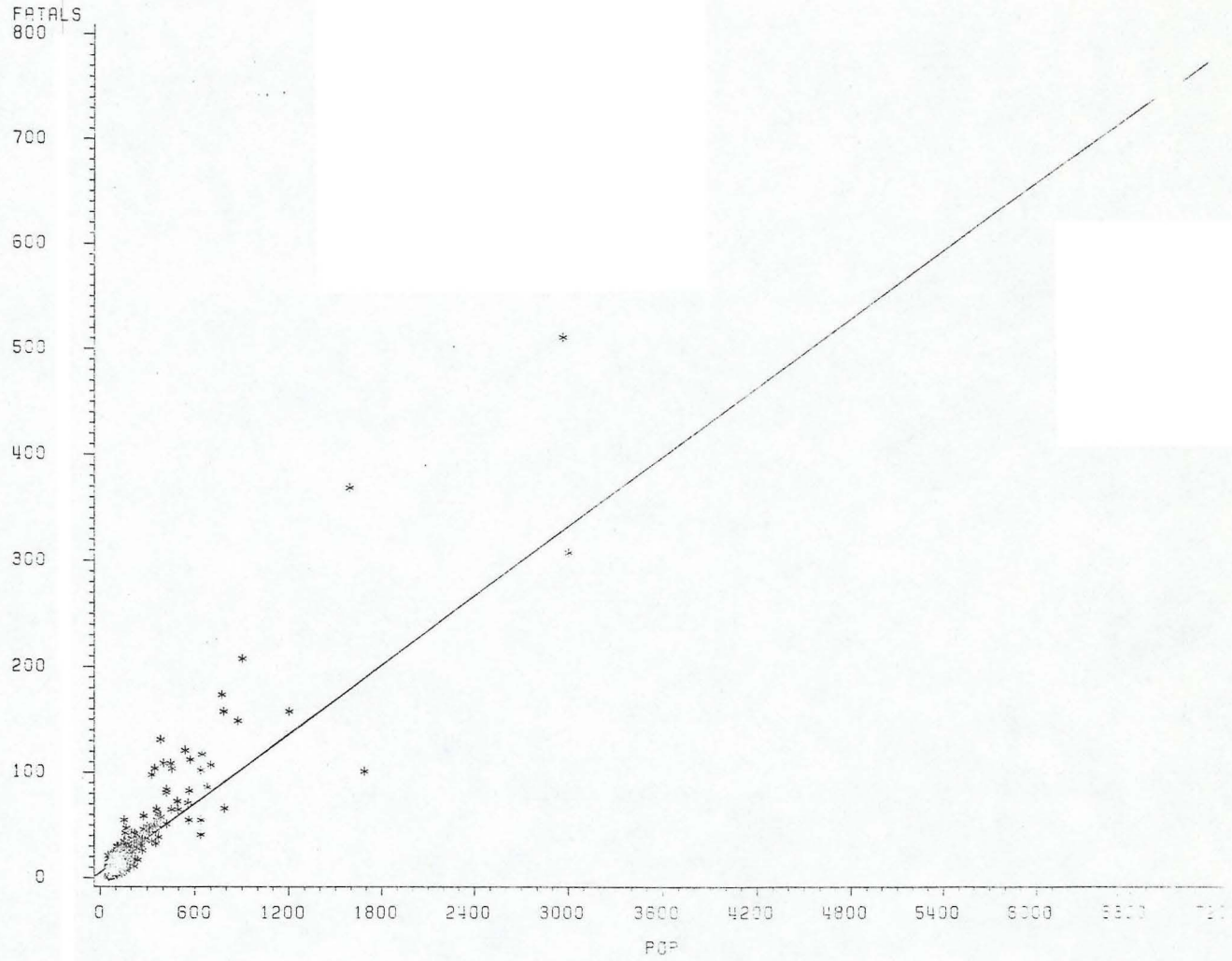

Fig. 1. Fatality vs Population per thousand for U.S. Cities with population greater than 50,000

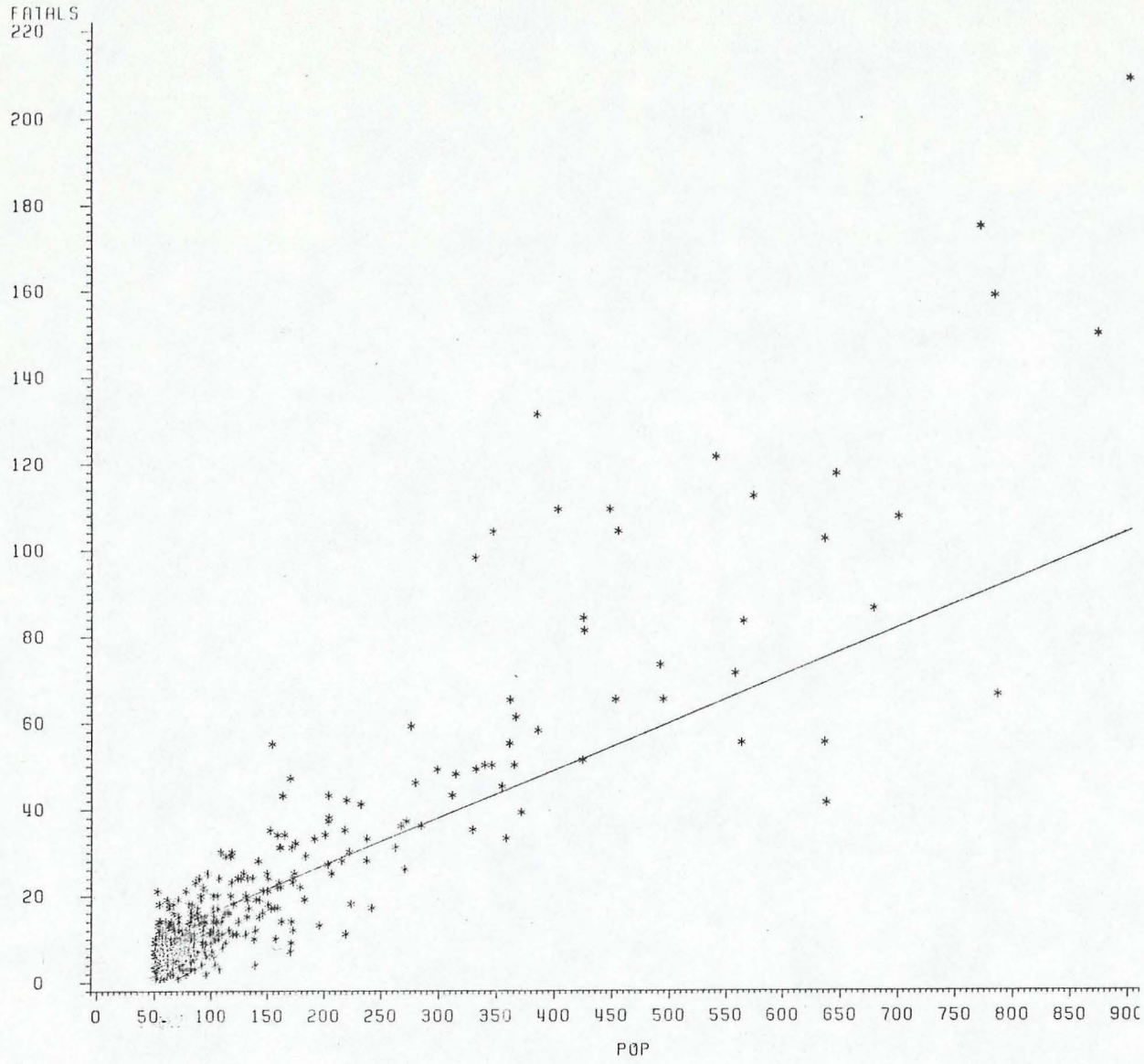

Fig. 2. Fatality vs Population per thousand for U.S. Cities with population between 50,000 and 100,000

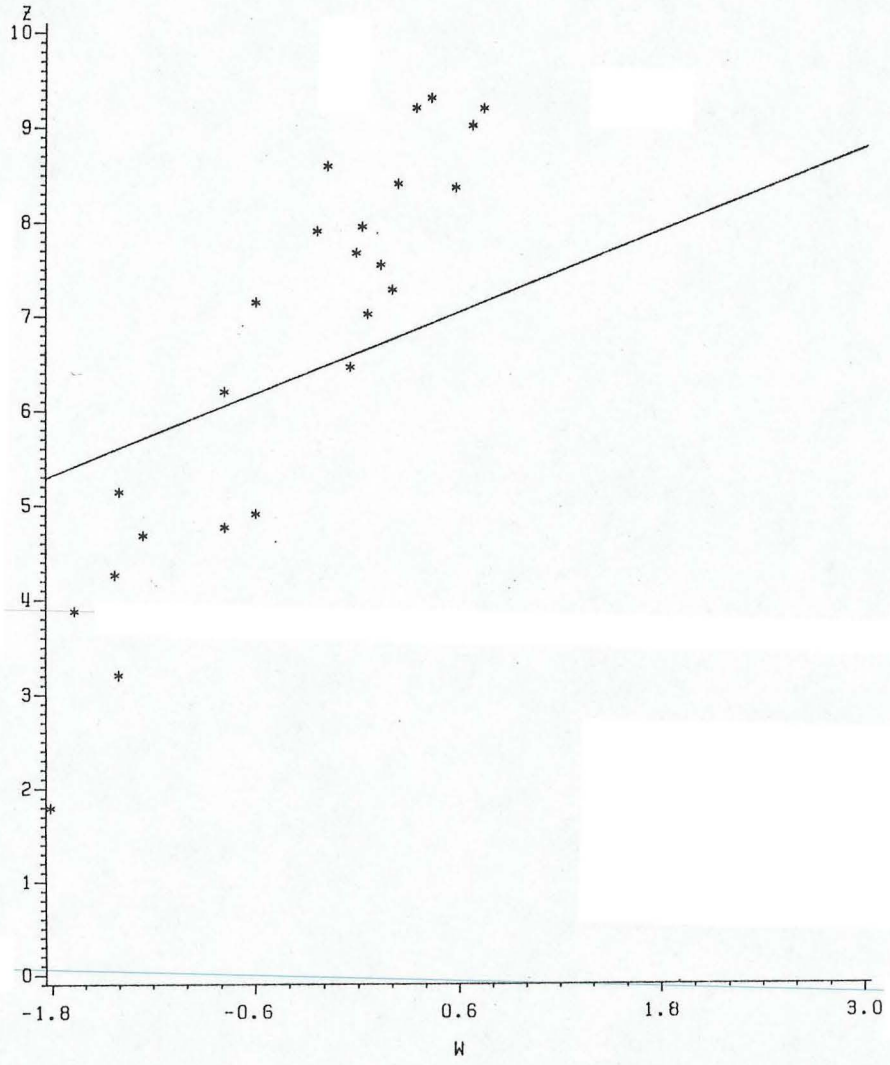

Fig. 3. Hypothetical example of an influential observation

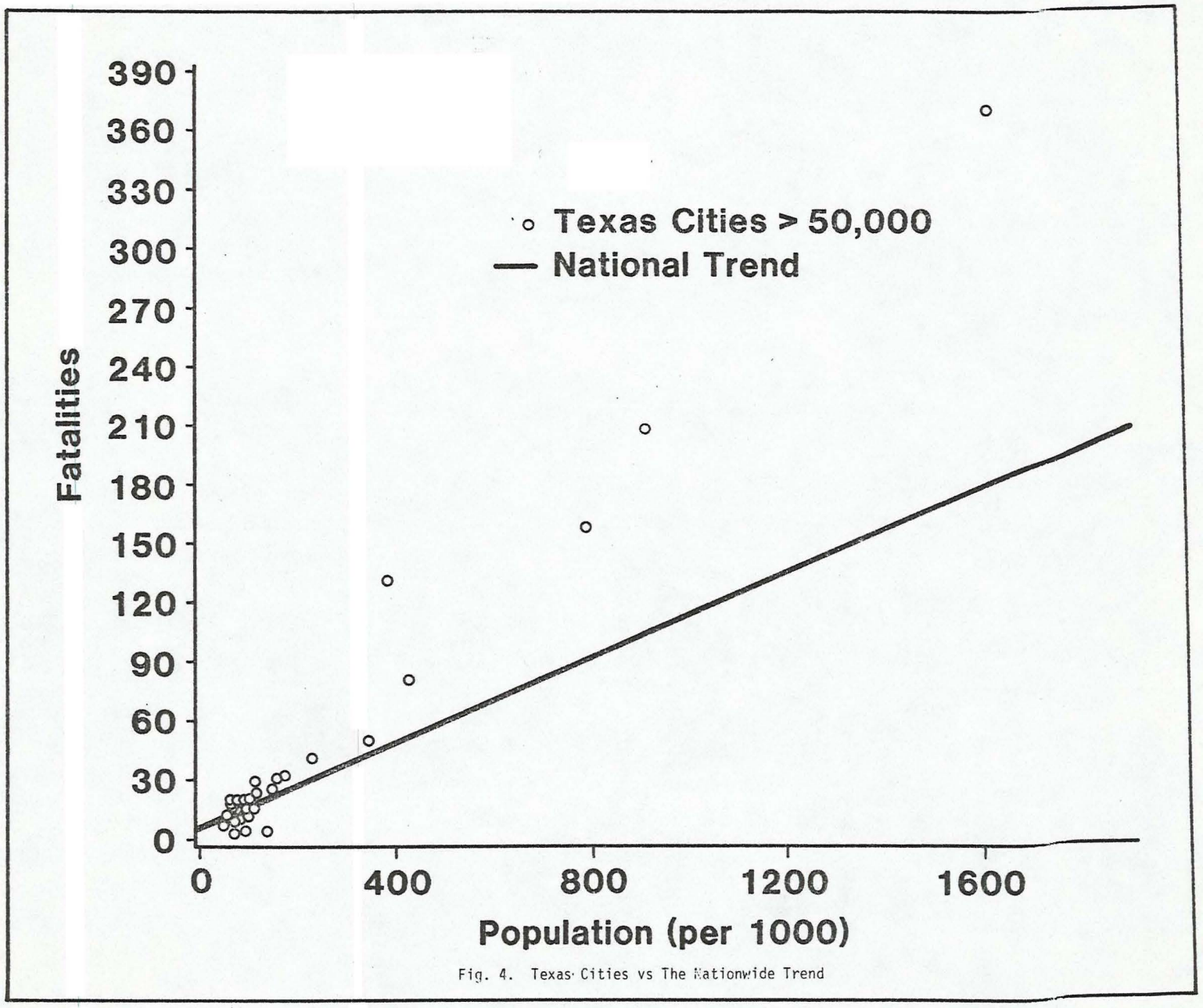

Fig. 4. Texas Cities vs The Nationwide Trend

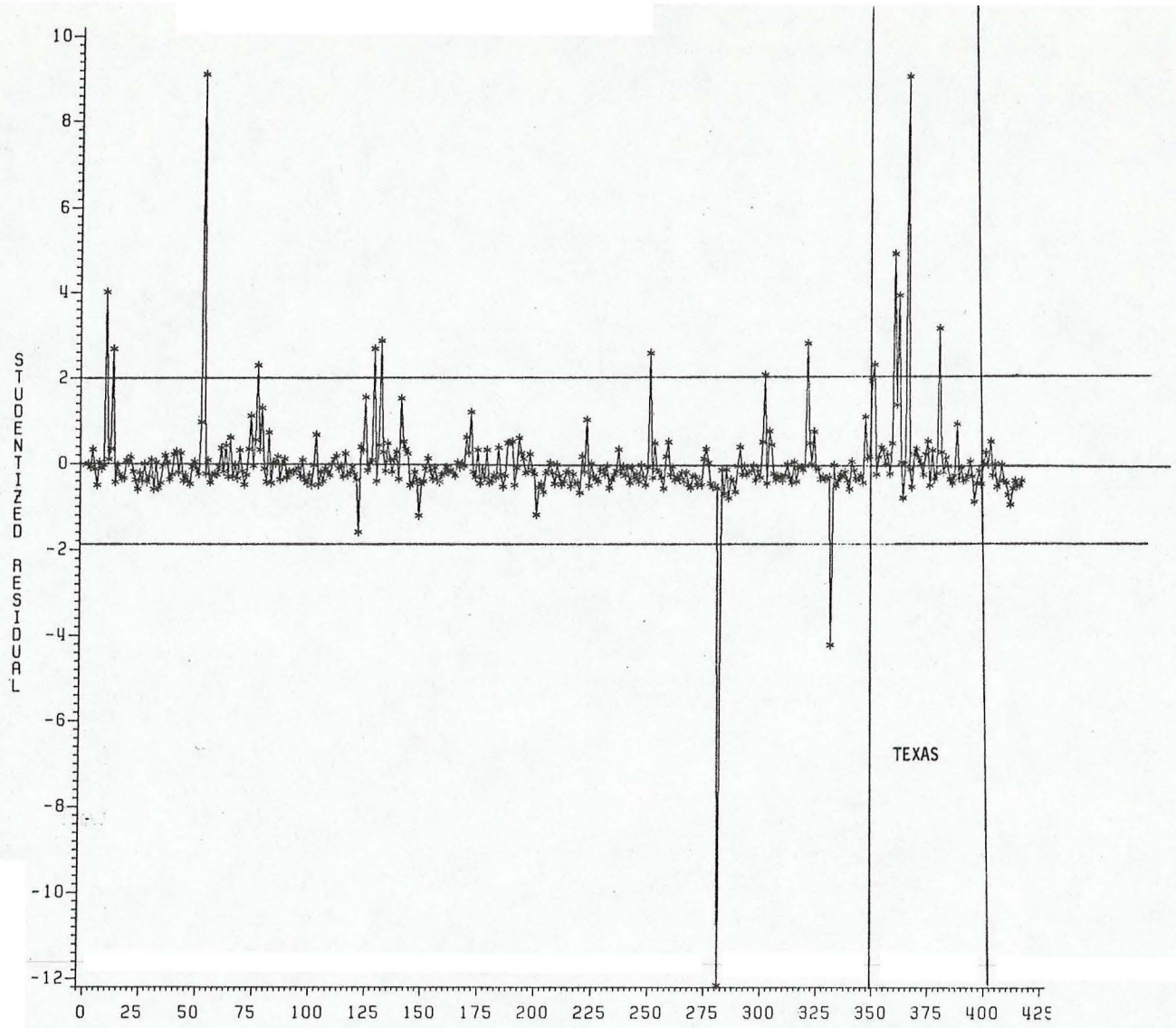

Fig. 5. Index Plot of Studentized Residuals

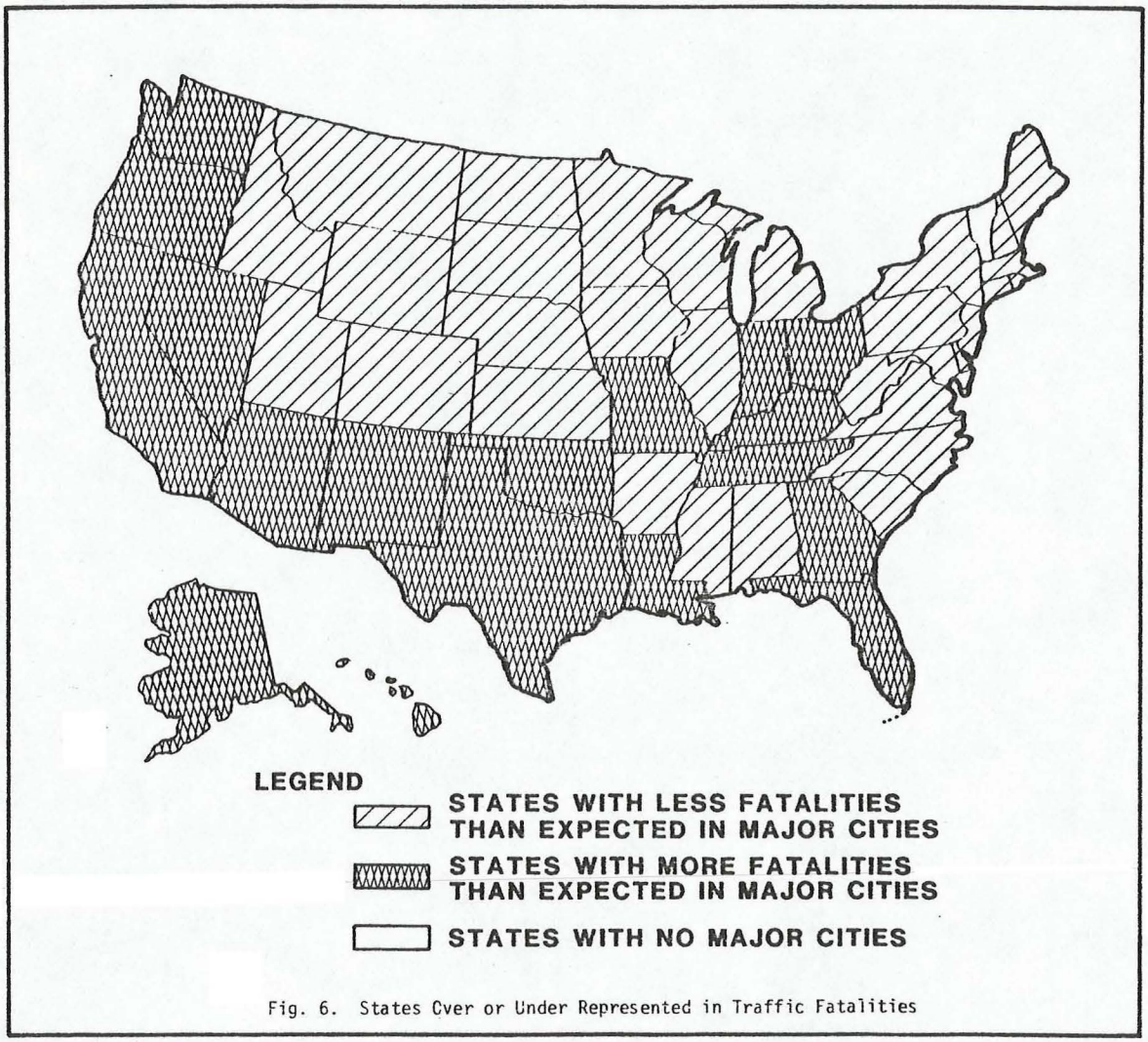

Fig. 6. States Over or Under Represented in Traffic Fatalities

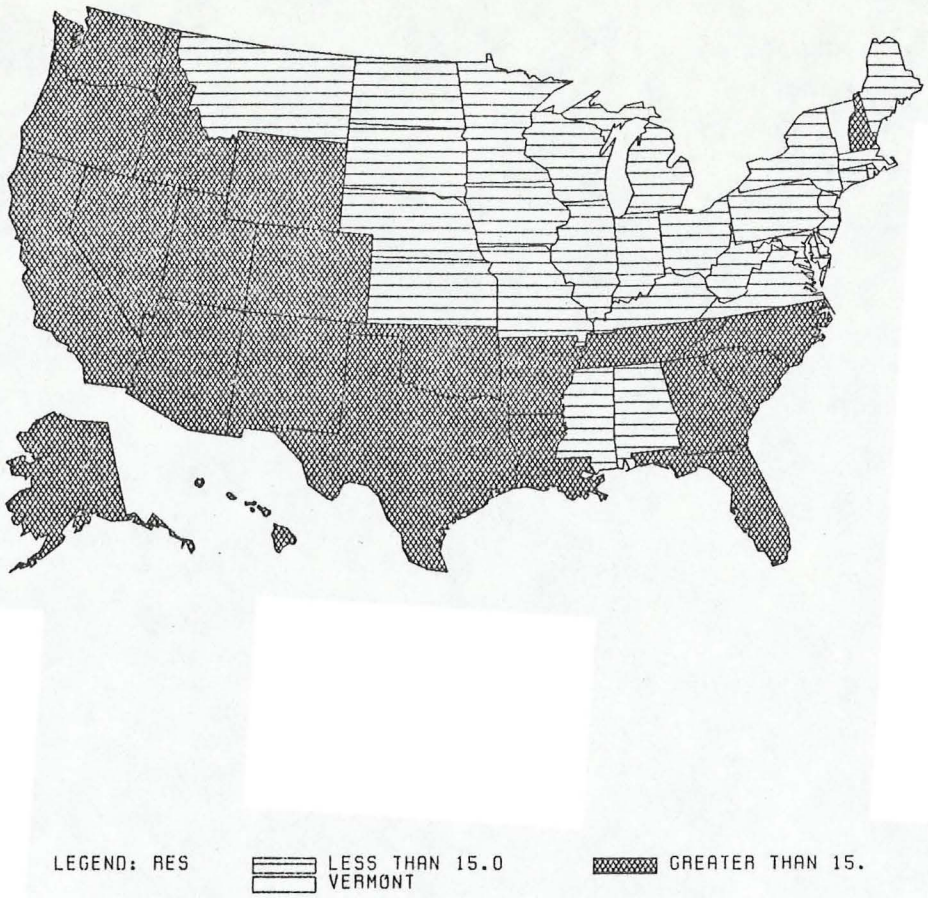

Fig. 7. Percent Change in State Population 1970 to 1980