Traffic Volume Forecasting Methods for Rural State Highways

Sunil K. Saha and Jon D. Fricker

This study builds on previous efforts found in the field of rural traffic forecasting. The study combines careful statistical analysis with subjective judgment to develop models that are statistically reliable and easy to use. This study developed two different kinds of models—aggregate and disaggregate—to forecast traffic volumes at rural locations in Indiana's state highway network. These models are developed using traffic data from continuous count stations in rural locations as well as data for various county, state, and national level demographic and economic predictor variables. Aggregate models are based on the functional classification of a highway, whereas the disaggregate models are location-specific. These models forecast annual average daily traffic (AADT) for future years as a function of present year AADT, modified by the various predictor variables. The use of both aggregate and disaggregate models will provide more reliable traffic forecasts. The number of predictor variables employed in the models was kept to a minimum. The statistical analysis also found that the predictor variables are statistically significant; no other variables will provide significant predictive power to the models. The models developed in this study provide good R^2 values. More refined statistical techniques reinforce the choice of variables used in the models.

The pattern of traffic growth and projected traffic volumes are prime factors in most analyses of highway projects. The traffic growth factor has a significant effect on highway investment decisions, such as whether to increase the capacity of existing highways and the construction of new facilities when funds are limited. Developing future traffic estimates is not an exact science, dependent as it is on so many hard-to-predict variables. Traffic forecasting procedures should be reasonably easy and economical to perform; they should be sensitive to a wide range of policy issues and alternatives and should provide decision makers with useful information in a form that does not require extensive training to understand.

Estimates of future traffic can be obtained by two very different methods: trend projections and forecasts. Analysts can modify extrapolated trends based on their experience and knowledge of the route, state, or region. With trend projections only traffic data are being dealt with; in forecasting techniques, however, a relationship between traffic and explanatory factors must be established. Traffic forecasting techniques therefore are also concerned with predicting the future values of economic and other measures or indicators of person and vehicle travel.

Traffic forecasting in urban areas has been extensively explored. Urban forecasting methodologies, based mainly on sophisticated computer modeling programs, are relatively advanced. On the other hand, forecasting traffic for individual rural roads, although widely practiced, is still in its early developmental stages. Standardized methodologies for nationwide use have not been established, and state authorities develop their own procedures to accommodate their needs. Various state departments of highways have developed methods of forecasting rural traffic, but very few of them are well documented.

PREVIOUS STUDIES

Four studies may be considered representative of the approaches used in the field of rural traffic forecasting.

Morf and Houska (*1*), in their 1958 study of the Illinois rural highway network, came to the conclusion that the four factors responsible for traffic growth patterns were (1) geographic location, (2) type and width of pavement, (3) proximity to an urban area, and (4) type of service the roadway provides. Their study also indicated that population is the principal component that affects the trend, followed by persons per vehicle, gasoline use, and vehicle miles per vehicle.

In 1982, Neveu (*2*) developed a set of elasticity-based models to forecast rural traffic. Neveu claimed that the type of service the roadway provides (interurban, interregional, rural to urban, urban to rural) is the only factor that had an appreciable effect on traffic growth rates. Multiple linear regression was used to identify factors that best estimated AADT and their respective elasticities. The factors used were population, number of households, automobile ownership, and employment. The roads were classified according to the type of service they provide: (a) interstates, (b) principal arterials, and (c) minor arterials and major collectors. The R^2 values 0.65, 0.77, and 0.20 for road types (a), (b) and (c), respectively, offer an indication of the explanatory power of the data. Two major problems were associated with the model: (1) the difficulty in obtaining projections of the factors/variables and their questionable accuracy at the level needed and (2) the uncertain degree of applicability of the model in certain

School of Civil Engineering, Purdue University, West Lafayette, Ind. 47907.

areas (i.e., whether a specific area is "rural enough" for the model). Neveu used multiplicative constant elasticity in his model, which implies that the effect of the growth in demand on traffic growth will always be the same. The result is that if the model is estimated during a period of high growth rate, future traffic will be overestimated and vice versa. Thus, such models must be recalibrated as often as practicable. Models with variable elasticities are not very common in traffic forecasting. Such model structures involve more sophisticated and expensive analysis.

The Minnesota Department of Transportation (Mn/DOT) (*3*) computes a route-specific growth factor from a trend analysis of the specific route. After determining base year AADT, ten to twenty years (preceding the base year) of AADT counts are taken from traffic flow maps. By linear regression, a line is fitted to the data and that line is extended to the design year. The overall growth is then the difference between design year AADT and base year AADT. Similar graphical plots of AADT against time for all (or several) major highway segments are done along the proposed project. If the growth rates are uniform, a single rate can be applied to the entire project. If not, forecasters must then use judgment in selecting the appropriate rate for each segment based on their knowledge of the project area.

The New Mexico State Highway Department (*4*) has designed a procedure for forecasting heavy commercial (HC) and average daily traffic (ADT) traffic on the New Mexico Interstate system and then calculating the percentage of HC traffic. This process, called "Trend-line," starts with fourteen distinct geographical sectors on the New Mexico Interstate system. Separate forecasting models were developed for each sector. The disaggregate analysis (a separate analysis for each sector) provides better traffic projections than does aggregate analysis (all sectors together). Eight key demographic and economic indicators are used. In the statistical analysis, linear regressions were conducted using heavy commercial ADT (HCADT) and ADT as dependent variables. Historical data covering a period of six years were used. Regression analyses were conducted to find the best-fit equation, leading to equations that had R^2 values over 80 percent.

COMMENTS ON FORECASTING TECHNIQUES

Armstrong (*5, 6*), in his studies of forecasting, concluded that sophisticated extrapolation techniques have had a negligible payoff in terms of accuracy in forecasting. More sophisticated methods are generally more difficult to understand, and they cost more to develop, maintain, and implement. On the positive side, more sophisticated methods may be expected to produce more accurate forecasts and a better assessment of uncertainty. However, highly complex models may in fact reduce accuracy. Armstrong recommended simple methods and the combination of forecast techniques. He suggested starting with the least expensive method(s) and/or the most understandable method(s), and then investing in successively more expensive methods. He proposed that complexities should be avoided unless they are absolutely necessary.

In this study, considerable effort has gone into the statistical analysis of the relationships between traffic volumes and the predictor variables. Each statistical test, however, is available on most standard statistical packages for computers. More important, the resulting model is intended to be easy to understand and to implement.

PREPARATORY STEPS

Functional Classification of Highways

The standard functional classifications of rural highways (*7, 8*) used in this study are (1) rural interstate, (2) rural principal arterial, (3) rural minor arterial, and (4) rural major collector.

The Data

The traffic volume data comprise all usable observations associated with Automatic Traffic Record (ATR) count stations in rural Indiana between 1970 and 1982. These data were analyzed in two ways—by using disaggregate and aggregate techniques. In disaggregate analysis, each station is analyzed separately. Station- or location-specific models for highways with similar characteristics can be developed. In aggregate analysis, stations within a given category of highway will be analyzed as a group, and a model applicable to any highway classifiable within a certain group will be proposed.

Those stations in a functional category the data for which were clearly well out of the range of values for most of the stations in that category were not used in developing an aggregate model. Instead, these stations were "saved" to test the ability of an aggregate model to "predict" their AADT values. Also, because of an inadequate number of data points, a disaggregate model was not developed for some of the stations.

Resulting cases numbered 26 for rural interstates, 39 for principal arterials, 52 for minor arterials, and 37 for major collectors, respectively. The fact that the cases in a highway category were actually drawn from two to four locations, each with up to thirteen years of AADT counts, made careful statistical analysis of utmost importance. A further complication was that traffic forecasting models tend to involve relationships with predictor variables that are themselves forecasts.

The predictor variables (or background factors) used in this study are listed in Table 1. They were chosen because (1) they reflect conditions at the county, state, and national levels that may affect the amount of travel that takes place and (2) their values are fairly easy to obtain through sources accessible to the state highway department. The variables X_7 to X_{13} were used as candidate background factors only in the case of rural interstates and rural

TABLE 1 VARIABLES FOR TRAFFIC FORECASTS

Symbol	Description of the Variable	Type of Variable
Y	Annual Average Daily Traffic (AADT)	------------
X_1	County Vehicle Registrations	Demographic
X_2	US Gasoline Price in cents per gallon, in 1972 dollars	Economic
X_3	Year	------------
X_4	County Population	Demographic
X_5	County Households	Demographic
X_6	County Employment	Economic
X_7	State Vehicle Registrations	Demographic
X_8	State Population	Demographic
X_9	State Households	Demographic
X_{10}	State Employment	Economic
X_{11}	Consumer Price Index (CPI) -- US	Economic
X_{12}	Gross National Product (GNP), in billions of 1972 dollars	Economic
X_{13}	Per Capita Disposable Personal Income (nationwide), in 1972 dollars	Economic

principal arterials. The variables X_1 to X_6 were candidates in all highway categories.

Model Form

Various forecasting models were examined, and an elasticity-based model (*2*) was adopted to relate future year AADT to present year AADT by means of a number of background factors. The general form of the model is as follows:

$$\mathrm{AADT}_f = \mathrm{AADT}_p\left[1.0 + \sum_{j=1}^{n} \frac{e_j(x_{j,f} - x_{j,p})}{x_{j,p}}\right] \tag{1}$$

or, upon rearrangement,

$$\frac{\mathrm{AADT}_f - \mathrm{AADT}_p}{\mathrm{AADT}_p} = \sum_{j=1}^{n} e_j\left[\frac{x_{j,f} - x_{j,p}}{x_{j,p}}\right] \tag{2}$$

where,

AADT_f = AADT in future year,
AADT_p = AADT in present year,
$x_{j,f}$ = value of variable x_j in the future year,
$x_{j,p}$ = value of variable x_j in the present year,
e_j = elasticity of AADT with respect to x_j,
n = number of associated variables.

The elasticity-based model was selected for several reasons. Most important, it was believed that the range of volumes over which the model would be applied would be much greater than that used in developing the model, making a simple linear regression model relating AADT to the background factors directly inappropriate. Second, the use of present year AADT to estimate future year AADT (as a sort of pivot point) would reduce the problem of nonresident travel. Also, the elasticity portion of the model calculates a growth factor directly. (See the right-hand side of Equation 2.)

The AADT values were obtained from the highway department's continuous count program. Only those stations classified as rural were selected for use in the study. This yielded a total of twenty-three stations throughout the state for the four highway categories.

The elasticities and the appropriate background factors are derived from a linear equation that relates AADT to a variety of the factors in Table 1. It can be shown mathematically that, given an equation of the form:

$$Y_i = a + \sum_{j=1}^{n} a_j X_{ij} \tag{3}$$

where

Y_i = value of dependent variable at ith observation;
$i = 1, \ldots, n_1$,
X_{ij} = value of jth independent variable at ith observation;
$j = 1, \ldots, n$,
a = constant term,
a_j = regression coefficient for jth independent variable,
n_1 = observation number,
n = number of independent variable.

Elasticity measures can be estimated by:

$$e_j = a_j\left[\frac{\bar{X}_{ij}}{\bar{Y}_i}\right] \tag{4}$$

where,

e_j = elasticity of AADT with respect to independent variable x_j,
$\bar{X}_{ij}$ = overall mean of the jth independent variable,
$\bar{Y}_i$ = overall mean value of dependent variable.

Thus, using multiple linear regression, the background factors that best estimate AADT and their respective elasticities can be derived.

PRELIMINARY STATISTICAL ANALYSIS

Although multiple linear regression forms the basis for the elasticity model, several other statistical procedures were applied to develop and verify the intermediate model forms. For the sake of brevity and to illustrate the process used, these techniques are described in the context of primarily one highway category: rural major collectors. The figures and tables are, for the most part, the rural major collector portions of exhibits contained elsewhere (*9*). But the figures and tables for rural principal arterials are also given when doing so helps to clarify the underlying techniques.

Preliminary statistical analyses were conducted to identify any possible relationship between dependent (Y) and independent variables (X's), to check the validity of standard regression assumptions, and to justify the combination of stations under the various highway categories.

Homogeneity of Variance

The homogeneity of variance is a property necessary to permit legitimate linear regression analysis. For the variable AADT, this property was checked using the Cochran and Bartlett-Box tests (*10*) by treating the Y's for each station as a group, since there were an equal number of observations in each station or group. For a highway category with an unequal number of observations at the constituent stations, the Burr-Foster Q-test (*10*) is used to check the homogeneity of variance of Y's.

The test results (Table 2) show that the data for rural principal arterials satisfy both the Cochran and Bartlett-Box tests for the homogeneity of variance condition, having equal observations in each station. The Burr-Foster critical q-value (*10*) shows β-levels for rural major collectors between 0.01 and 0.001. Based on the regression analysis, a linear relationship between Y and the X's is feasible. There is no apparent practical or theoretical reason to transform Y. The distribution of Y's at some stations is sparse. Considering all these factors, homogeneity of variance for rural major collectors was accepted at a β-level of 0.001. In fact, homogeneity of variance was satisfied at reasonable β-levels for all highway categories.

Normality

The Shapiro-Wilk test for normality (*11*) was performed on each station separately and after combining stations in a highway category. The results for rural major collectors are shown in Table 3. In this table, small values of W associated with smaller β-levels are significant; that is, they lead to rejection of the normality assumption. This rejection occurred for the aggregate "all stations" case in every category but rural interstates.

Thus the normality assumption on Y does not support the combination of stations for the other highway categories. As will be seen later, however, other statistical tests justify further examination of each proposed model. When the Y's for individual stations are tested for normality, the assumption is satisfied in every case.

Scattergram

The scatterplots of the dependent variable (AADT) against the independent variables show a general linear trend. Two representative plots are shown in Figures 1 and 2. The clusters found in Figure 1 reflect the fact that the data come from thirteen years at each of three count stations, but linearity is still apparent.

CONCLUSIONS FROM PRELIMINARY ANALYSIS

The homogeneity of variance tests, considering each station as a group, shows equal variances between the stations for each category of highway. The normality hypothesis is accepted for each station separately and for rural interstates as a group. The normality test indicates that analysis for each station separately will yield a better model than that for the combination of stations within a highway category. In the scatterplots for the stations—both sepa-

TABLE 2 RESULTS OF THE TESTS FOR HOMOGENEITY OF VARIANCE

A. Bartlett-Box and Cochran Test (Equal Sample Size)					Homogenity of Variance
Highway Category (No. of station or group)	Cochran c [β-level]	Bartlett-Box f [β-level]	Remarks on β-level		
			Cochran	Bartlett-Box	
2. Rural Principal Arterial (3)	0.5686 [0.063]	1.949 [0.143]	$\beta > 0.05$	$\beta >> 0.01$	Checked
B. Burr-Foster Q-Test (Unequal sample size)					Homogenity of Variance
Highway Category (No. of station or group)	Calculated q	Critical q		Remarks on β-level	
		$\beta = 0.01$	$\beta = 0.001$		
4. Rural Major Collector (3)	0.4938	0.4827	0.5543	$\beta = .01 - .001$	Checked

TABLE 3 RESULTS OF THE TEST FOR NORMALITY

Highway Category	Stations(s)	No. of Cases	Shapiro-Wilk W	B-level	Normality
4. Rural Major Collector	(i) All stations	37	0.8051	<0.01	Unchecked
	(ii) 47A	11	0.9167	0.10 - 0.50	Checked
	(iii) 59A	13	0.9143	0.10 - 0.50	Checked
	(iv) 5420A	13	0.8899	0.10 - 0.50	Checked

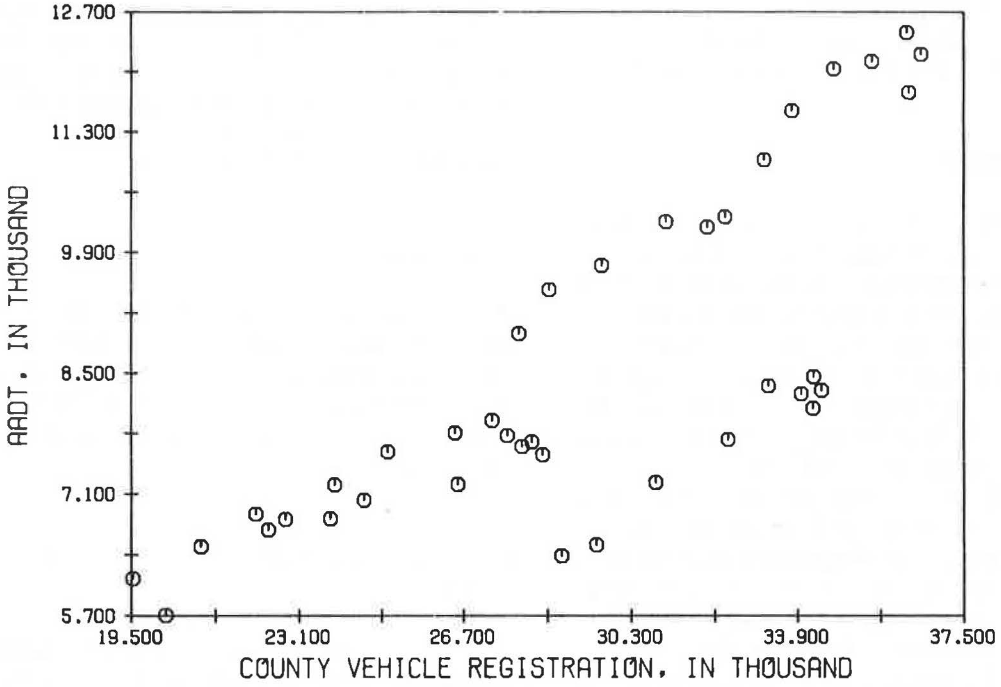

FIGURE 1 AADT versus county vehicle registrations (rural principal arterials).

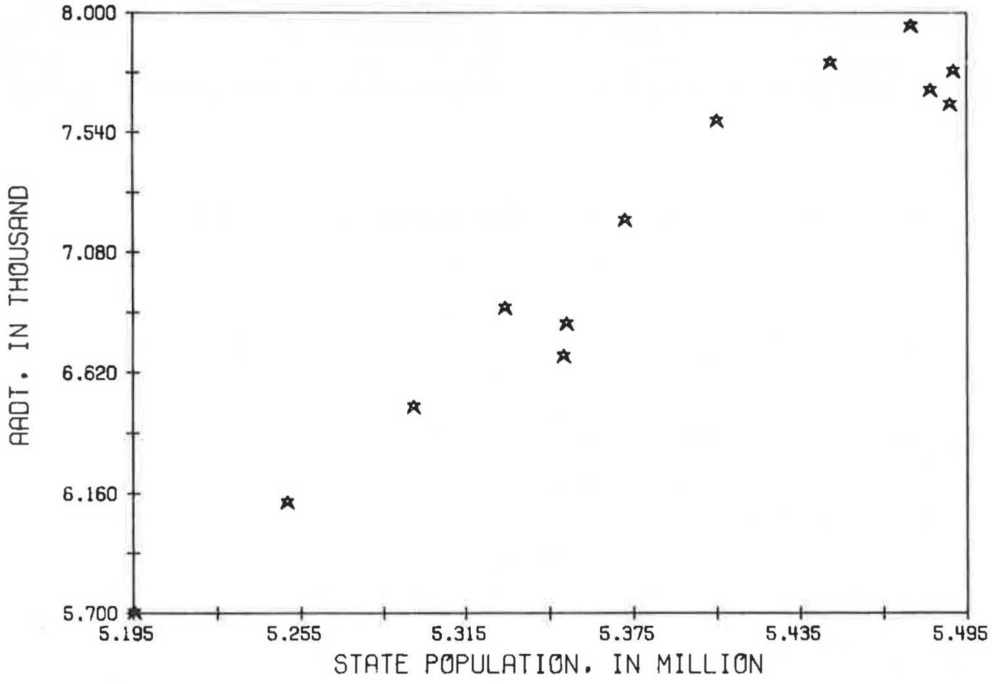

FIGURE 2 AADT versus state population (Station 68A).

rately and together—no recognizable pattern other than linear is noticeable.

Two types of analyses are found appropriate and promising. Aggregate analysis, combining all the stations within a category of highway, is employed to develop an aggregate model for each highway category. Disaggregate analysis of each station separately is also performed, and the resulting models will be location-specific.

AGGREGATE ANALYSIS

In aggregate analysis, a model is sought for each of the four categories of highway, with the count stations pooled by category. The analysis for rural major collectors (and, in some cases, rural principal arterials) is presented here to illustrate the process, but the analysis for other highway categories is similar.

Correlation Matrix

The statistical analysis begins with the study of the correlation matrix for the various factors considered. The correlation matrix for rural principal arterials (Table 4) shows that variables X_4 and X_8 are not highly correlated ($r = 0.251$). The correlation matrix for rural major collectors (Table 4) shows that X_1, X_4, and X_5 have similar degrees of correlation with Y ($0.731 \leq r \leq 0.915$). But the variables X_1, X_4, and X_5 are highly intercorrelated ($r \geq 0.901$). The existence of this multicollinearity does not invalidate a regression analysis; and neither is the absence of multicollinearity a validation of a particular regression model (*12*). Nevertheless, the use of two or more of these highly correlated variables in the same model should be avoided whenever possible. In case of rural major collectors, county employment (X_6) and AADT (Y) are negatively correlated, which is not the expected relationship; so the selection of variable X_6 will not be considered unless supported by other analyses.

Stepwise Regression

Stepwise regression is the most widely used automatic search procedure. It selects one variable at a time for entry into (or removal from) the model, until a desired subset of variables is selected (Table 5). In designing this regression procedure, four parameters—number of steps, F-value to enter (FIN), tolerance, and F-value to remove

TABLE 4 CORRELATION MATRIX (AGGREGATE ANALYSIS)

(2) Rural Principal Arterial

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
X1	.786												
X2	.275	.519											
X3	.398	.639	.827										
X4	.804	.878	.224	.259									
X5	.881	.938	.364	.429	.974								
X6	.633	.901	.543	.692	.667	.764							
X7	.405	.647	.779	.975	.247	.411	.717						
X8	.402	.642	.797	.974	.251	.415	.721	.993					
X9	.401	.640	.830	.9999	.260	.429	.702	.978	.980				
X10	.354	.547	.626	.763	.191	.318	.697	.857	.847	.781			
X11	.373	.602	.865	.974	.260	.427	.654	.905	.912	.972	.676		
X12	.413	.641	.773	.972	.252	.416	.735	.987	.987	.979	.872	.923	
X13	.412	.643	.754	.974	.251	.414	.727	.993	.990	.979	.857	.915	.994

(4) Rural Major Collector

	Y	X1	X2	X3	X4
X1	.766				
X2	.178	.593			
X3	.164	.687	.801		
X4	.915	.901	.354	.341	
X5	.731	.654	.587	.618	.921
X6	-.453	.212	.452	.568	-.163

XI represents X_I, where i = 1 to 13

TABLE 5 STEPWISE REGRESSION SUMMARY (AGGREGATE ANALYSIS)

Highway Category	Case (*) and Parameter	Step	Variable subscript		F value	Significance level	Last step b-coeff.	R^2	Overall F (**)
			Entered	Removed					
RURAL MAJOR COLLECTOR	Case A: Default Parameters	1	4		180.062	0.0	.5988	.837	18.062
		2	6		47.491	0.0	-.0492	.932	233.367
		3	1		2.818	.103	-.0589	.937	164.834
		4	5		1.820	.187	-1.0940	.941	127.151
		5	3		3.125	.087	152.0760	.946	109.102
		6	2		.302	.587	-8.8768	.947	88.921
		Constant term					-303613.9		
	Case B: Alpha = .10 FIN = 2.95 FOUT = 2.80	1	4		180.062	0.0	.2564	.837	180.062
		2	6		47.491	0.0	-.1979	.932	233.367
		Constant term					-4908.47		
	Case C: Alpha = .05 FIN = 4.20 FOUT = 4.10	(SAME AS CASE B)							

(*) A. Default Parameters:
(i) Max. No of Steps = 2 * No. of Independent Variables [For All Cases]
(ii) FIN = .01, FOUT = .005 [For Case A]
(iii) Tolerance level = .001 [For Case A]
B. Tolerance level = 0.1 [For Case B & Case C]
(**) Overall Significance = 0.0

(FOUT)—play important roles in the selection of variables for the models. Case A in Table 5 used the default values provided by the SPSS package (*11*), while cases B and C were defined in terms of the parameter values indicated in the second column of Table 5. For rural major collectors, the case A stepwise regression with default parameters includes all variables, with an R^2 of 0.947. The case B and case C stepwise regressions enter the variables X_4 and X_6 with an R^2 of 0.932. The variable X_4 enters at step 1 with an R^2 of 0.837 in all cases. The inclusion of other variables increased R^2 by only a small amount. It should be mentioned that the order in which a variable enters the regression equation does not indicate its importance. The *b*-coefficients of X_1, X_2, X_5, and X_6 are negative. The negative coefficient of X_2 (gasoline price) is expected. The reason for the negative coefficients of X_1, X_5, and X$_6$, however, is its high correlation with other variables in the model (for example, $r_{1,4} = 0.901$, $r_{1,5} = 0.954$, $r_{4,5} = 0.921$ in Table 4). Since stepwise regression can be misleading, the results of the stepwise regression were compared with other selection criteria (see Table 8) before establishing the final regression equation.

C_P-Criterion in All Possible Regression

The C_P-criterion considers all possible regression models that can be developed from the pool of potential independent variables and identifies subsets of the independent variables that are "good." The C_P-statistic, R^2, and other values for a reasonable number of subsets of variables were calculated with the help of a program called DRRSQU (*13*). Some of those C_P- and R^2-values for rurai major collectors are shown in Table 6. The C_P-criterion is concerned with the total mean squared error (MSE) of the *n* fitted values for each of the various subset regression models. When the C_P-values for all possible combinations of variables in regression models are plotted against *P* (the number of terms in the regression equation), those models with little bias will tend to fall near the line $C_P = P$ (*14*). Models with substantial bias will tend to fall considerably above this line. In using the C_P-criterion, the subsets of *X* variables for which (1) the C_P-value is small and (2) the C_P-value is near *P* are considered for the model. Sets of *X* variables with small C_P-values have a small total mean squared error, and when the C_P value is also near *P*, the bias of the regression model is small. Sometimes the regression model based on the subset of *X* variables with the smallest C_P-value may involve substantial bias. In that case, one may at times prefer a regression model on a somewhat larger subset of *X* variables for which the C_P-value is slightly larger, but that does not involve a substantial bias component. Thus, one should look for a regression model with a low C_P-value about equal to *P*. When the choice is not clear-cut, then it is a matter of personal judgment whether one prefers a biased equation or an equation with more parameters.

Draper and Smith (*14*) recommend the use of the C_P-statistic in conjunction with the stepwise method to choose the best equation. Some statisticians suggest that all possible regression models with a number of *X* variables that is similar to the number in the stepwise regression solution be fitted subsequently to investigate which subset of *X* variables might be best (*15*). The C_P-values for rural major collectors in Table 6 show that the variable set X_3, X_4, and X_5 at $P = 4$ is the best selection, with C_P of 2.40 and R^2 of 0.944. But the selection of X_4 and X_6 at $P = 3$, with $C_P = 7.26$ and $R^2 = 0.932$, is the result of stepwise regression in cases B and C (see Table 5). The variable sets {X_1, X_4, and X_6} and {X_3, X_4, and X_6} at $P = 4$, with C_P of 6.25 and 6.76, respectively, are good for further analysis. Note that X_4 has high correlation with X_5 ($r = 0.921$). Also, X_4 has negative correlation with X_6 ($r = -0.163$), which is not an expected result.

MSE_P or *R_a^2* Criterion in All Possible Regression

The adjusted coefficient of multiple determination $R_a^2$and mean squared error (MSE_P) are equivalent criteria. The users of the MSE_P criterion seek either the subset of X variables that minimizes MSE_P, or the subset(s) for which MSE_P is so close to the minimum that adding more variables is not worthwhile (*15*). The results of using the MSE_P criterion for rural major collectors are shown in Figure 3. The plot of MSE against P indicates that, according to this criterion, the variable set X_3, X_4, and X_5 at

TABLE 6 SELECTED C_P AND R-SQUARED IN ALL POSSIBLE REGRESSION (AGGREGATE ANALYSIS)

Highway Category	Subscripts of Variables in Equation	C_p Values, in same order	R^2 Values, in same order	P
Rural	4 1 5 6 3	58.7, 199.9, 299.8, 414.9, 515.5	.837, 587, .534, .205, .027	2
Major	4 6, 4 5, 5 6, 1 6, 2 4, 2 5	7.26, 13.46, 18.18, 31.48, 46.87, 178.30	.932, .921, .913, .889, .862, .629	3
Collector	3 4 5, 1 4 6, 3 4 6, 1 4 5, 2 4 6, 4 5 6	2.40, 6.25, 6.76, 8.43, 9.12, 9.19	.944, .937, .937, .934, .932, .932	4
	----------	--------	--------	---
	All Variables	7	.947	7

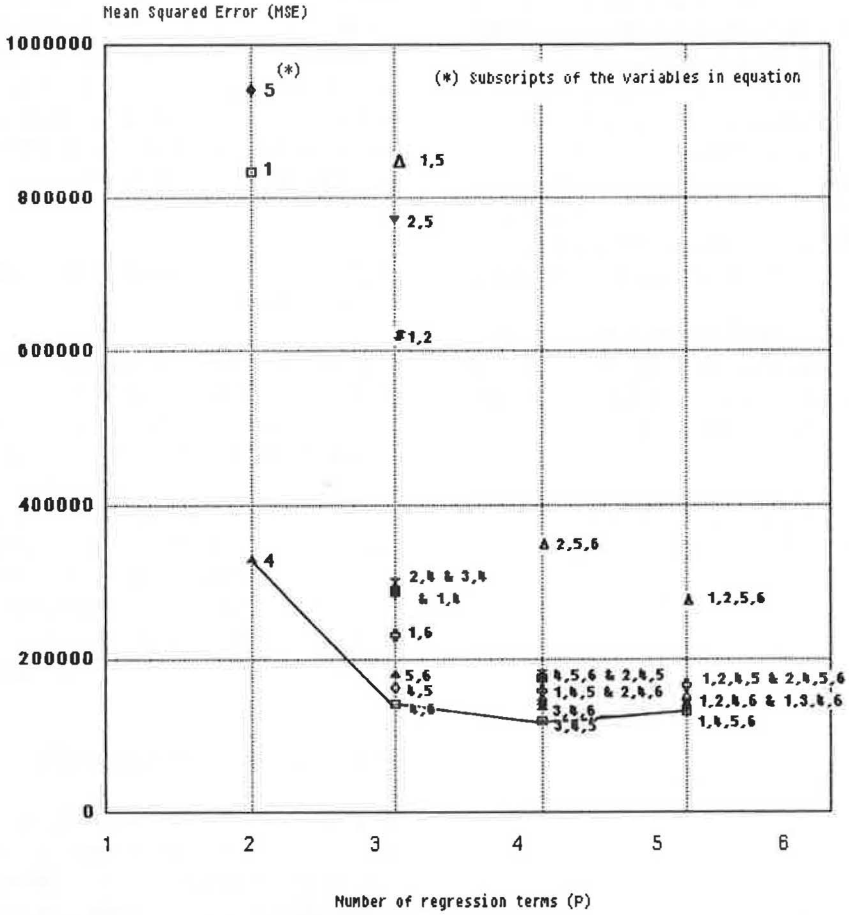

FIGURE 3 *MSE* versus *P* plot for rural major collectors.

$P = 4$ is the best choice. This choice will not be considered further, however, because it involves multicollinearity.

Preliminary Screening of Candidate Variables

The screening of variables to develop the forecasting models was not confined to statistical analysis. The questions listed in Table 7 and the goals in Table 8 provided a basis on which judgment could be applied to the selection process.

Considering all the points discussed above, the following subsets of variables were kept for the final selection process:

1. X_4
2. X_1
3. X_5
4. X_4, X_6
5. X_5, X_6
6. X_4, X_2
7. X_5, X_2

All these choices will provide an R^2 of at least 0.534.

Final Selection of Variables

In the selection of variables for the final regression model, the goals in Table 8 were used as a guide to find the best subset of variable(s) from the preliminary choices for each highway category. Goals 1 to 5 in Table 8 were taken into consideration in the preliminary screening. Final selection of candidate variables from preliminary choices is then made through careful examination of all criteria, subject to subsequent residual analysis and hypothesis testing concerning *b*-coefficients. If the final regression model passes these last two tests, it can be used to build the forecasting model.

In the final selection of variable(s) for a model's equation, the criterion of establishing a high R^2 should not be the only consideration. In addition to high R^2 value, residual plots should also be examined to determine if a "good" fit has truly been obtained. For example, in some situations, data may be quite variable and a large R^2 may not indicate a very good fit. In more controlled situations, however, a relatively small R^2 may indicate a rather good fit (*12*). The value of R^2 will increase if the number of predictor variables increases. Consequently, R^2 is always the maximum for the full set with all predictor variables. So maximizing R^2 should not be the sole selection criterion. However, one can subjectively choose a subset of predictor variables that gives a good value of R^2, so that using any additional predictor variables results in only a marginal improvement in R^2.

TABLE 7 SOME FUNDAMENTAL CRITERIA FOR VARIABLE SELECTION (*12,14*)

1. Are the proposed variables fundamental to the problem?
2. Availability of data (variables).
   (a) Are annual data available?
   (b) Are historical data available?
   (c) What is the most recent year of data?
   (d) Will data be available in future?
3. Cost to obtain the data.
4. How reliable are the data?

TABLE 8 GOALS OF THE ANALYSIS

1. The final equations should explain more than 50 percent of the variation ($R^2 > 0.50$).
2. The C_p value will be lowest and near to P.
3. Mean squared error (MSE) will be minimum.
4. The number of predictor variables should be adequate for each model (*).
5. The selection will respond well to the questions of Table 7.
6. All estimated coefficients in the final model should be statistically significant at an alpha level of 0.05 or 0.10.
7. There should be no discernible patterns in the residuals.

(*) As a general rule, there should be about ten complete sets of observations for each potential variable to be included in the model; e.g., if it is believed that the final practical predictive model should have four X-variables plus a constant, then there should be at least forty sets of observations (n = 40) [14].

Regression on Preliminary Choices and Final Selection

Regression on the preliminary choices was done with the help of the SPSS package (*11*), a summary of which is shown in Table 9. This table shows inconsistencies ("-*b*6") in the *b*-coefficient of X_6 for the preliminary choices $\{X_4, X_6\}$ and $\{X_5, X_6\}$. The choice with X_4 has the largest R^2 and lowest C_P among all the choices with one variable. The choices with two variables without inconsistency in regression coefficients do not provide a significant increase in R^2 with respect to the one-variable choices (see Table 9). Considering these aspects, the variable X_4 is the final selection for further analysis for rural major collectors.

Graphic Residual Analysis on Final Selection

Residual plots were generated by the BMDP package (*16*). These plots were done to verify the validity of each regression model. The normal probability plot in Figure 4 falls reasonably close to a straight line, suggesting that the error terms are approximately normally distributed. The plots

TABLE 9 MULTIPLE LINEAR REGRESSION SUMMARY ON PRELIMINARY CHOICES (AGGREGATE ANALYSIS)

Highway Category	Variable Subscripts in Eqn.(*)	b-coefficient in same order	Inconsis-tencies in b's(**)	R^2	Overall F(***)
	4	.2715	---	.837	180.062
	1	.1991	---	.587	49.693
RURAL	5	.7122	---	.534	40.056
MAJOR	4, 6	.2564, -.1979	-b6	.932	233.367
COLLECTOR	5, 6	.8379, -.3988	-b6	.913	177.805
	4, 2	.2892, -34.5641	---	.862	106.030
	5, 2	.9292, -78.3612	---	.629	28.772

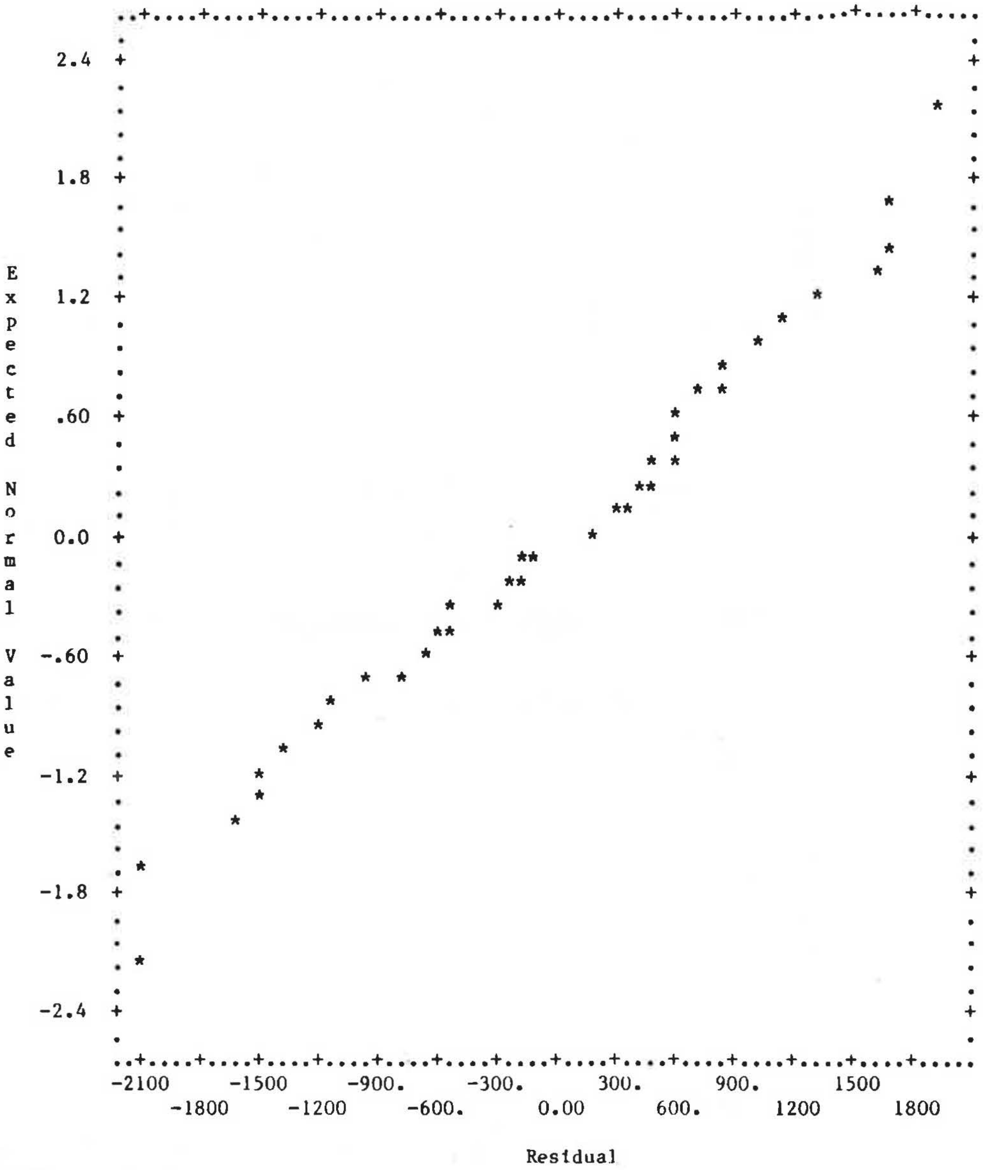

FIGURE 4 Normal probability plot of residuals (rural major collector).

of residuals against the fitted response variable and predictor variable, represented by Figures 5 and 6, indicate no grounds for suspecting the appropriateness of a linear regression function or the constancy of the error variance. The clustering of residuals in some cases (as in Figure 6) is the effect of combining count stations in the analysis. The residual plots against $\hat{Y}$ and the X's do not indicate the presence of any outliers. Residual plots were also generated against variables not included in the model, to check whether some key independent or predictor variables could provide important additional descriptive and predictive power to the model. One such variable is the year (X_3), which has not been included in any model. The plot of residuals against X_3 in Figure 7 does not indicate any correlation between the error terms over time, since the residuals are random around the zero line. Thus, it is confirmed that the most appropriate variable is included in the model and no additional variable will provide significant added explanatory power to the model.

Testing Hypotheses Concerning Regression Coefficients

The overall F-test for the regression relation explains whether the variables in the model have any statistical

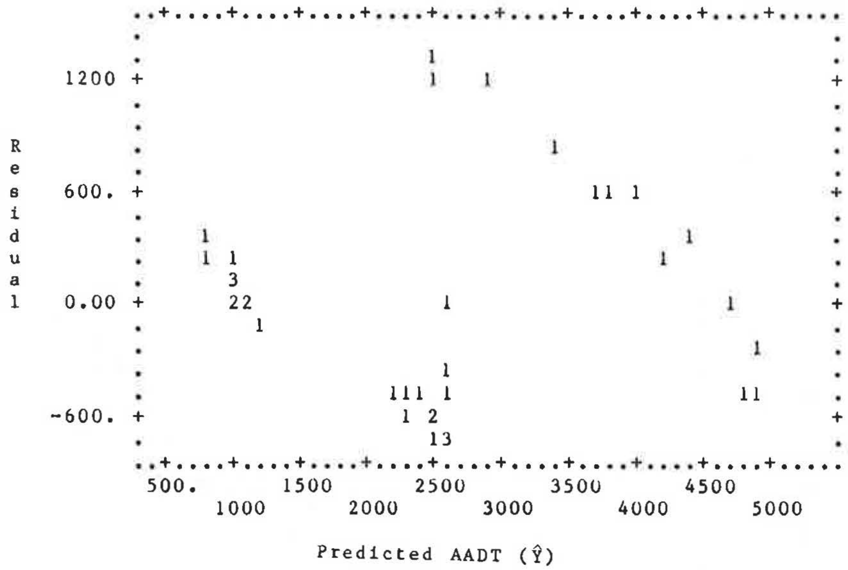

FIGURE 5 Residual plot against $\hat{Y}$ (rural major collector).

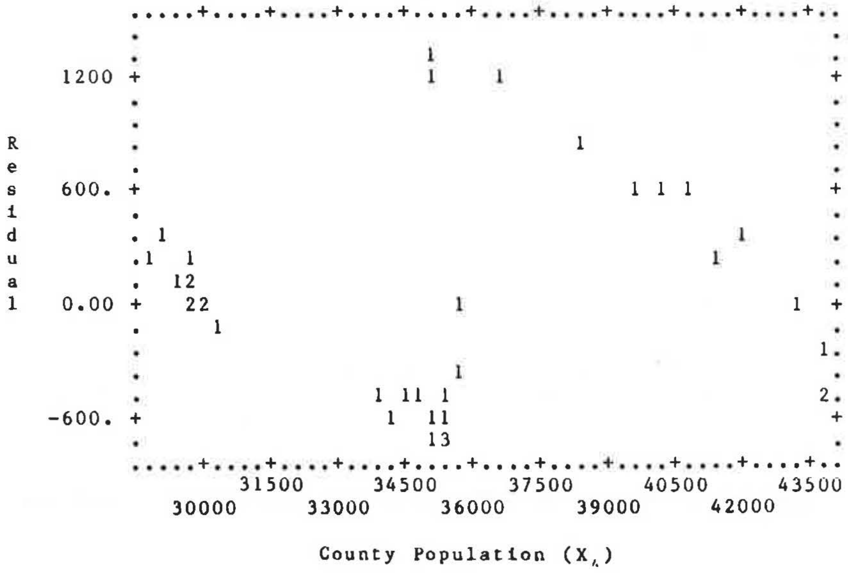

FIGURE 6 Residual plot against X_4 (rural major collector).

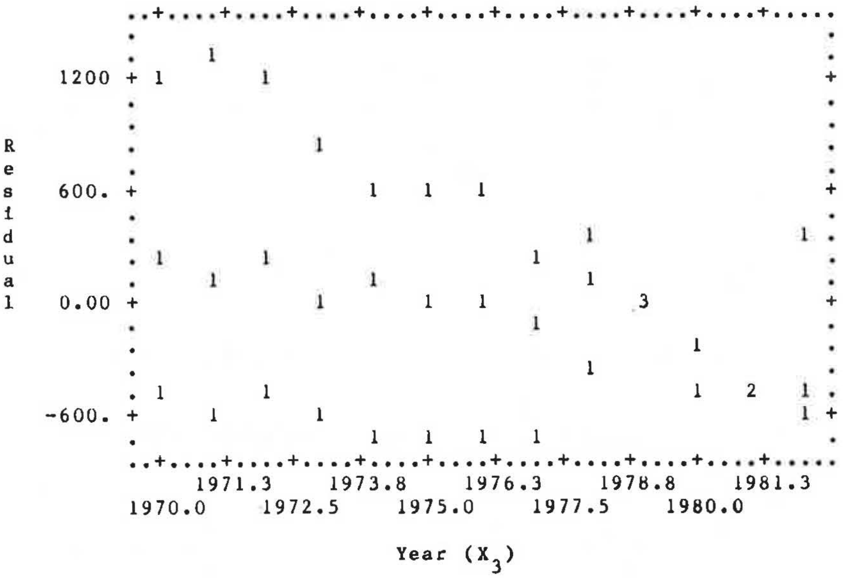

FIGURE 7 Residual plot against X_3 (rural major collector).

relation to the dependent variable. The hypotheses are

H_0: $\beta_1 = \beta_2 = \ldots = \beta_{P-1} = 0$
H_a: all β_k $(k = 1, \ldots, P - 1) \neq 0$.

The test statistic is given by

$$F^* = \frac{MSR}{MSE}$$

If $F^* \leq F(1 - \alpha, P - 1, n - P)$, then H_0 holds, indicating that the variables in the model do not have any statistical relation to the dependent variable. Table 10 shows the result for rural principal arterial and rural major collector at α-levels of 0.05 and 0.10. The test results conclude that hypothesis H_a (i.e., that the relationships between the variables in the models exist) cannot be rejected at an α-level as low as 0.05.

To test the significance of each variable

$(H_0: \beta_k = 0; H_a: \beta_k \neq 0 \text{ for } 1 \leq k \leq P - 1)$

and each subset with more than one variable

$(H_0: \beta_1 = \ldots = \beta_j = 0; H_a: \text{all } \beta_j \neq 0 \text{ for } 1 < j < P - 1)$

a general linear test (*15*) was employed. The applicable F-statistic is shown in Equation 5.

$$F^* = \frac{[\text{SSE}(R) - \text{SSE}(F)]/(\text{df}_R - \text{df}_F)}{\text{SSE}(F)/\text{df}_F} \quad (5)$$

where,

F^* = the F statistic,
SSE (R) = error sum of squares for the reduced model,
SSE (F) = error sum of squares for the full model,
df_R = degrees of freedom of the reduced model, and
df_F = degrees of freedom of the full model.

The reduced model was obtained by dropping the element(s) to be tested under H_0 from the full model. Table 11 shows the summary of the partial F-test results for rural principal arterial and rural major collector obtained at α-levels of 0.05 and 0.10. The results indicate that each variable is significant in the model.

Model Development

The final regression equation for rural major collectors is presented in Table 12, along with the R^2 values, overall F values, t-statistics, and elasticities. Using the elasticities obtained from the regression analysis, the forecasting model was developed for each highway category. The models for all four highway categories are presented in Table 13. Each of the models is relatively simple, containing not more than three variables. The use of these models is also straightforward. The input values are the present year AADT and the present and future year value (the year for which the traffic forecast is needed) of the predictor variables. The data needed to predict rural traffic volumes with these models are readily available at the county and state levels. The performance of the models in Table 13 was tested using data for those Automatic Traffic Record (ATR) stations not used in building the models. The results of the trial forecasts for the rural major collector, shown in Table 14, indicate that the models perform satisfactorily. This table shows the present year used in these trial forecasts. Any closest year for which predictor variables are available could be the present year. In that respect, the

TABLE 10 OVERALL *F*-TESTS FOR AGGREGATE ANALYSIS

Highway Category	Variable Subscripts for Full Model	F*	df_R, df_E (*)	α	Is H_a true for	
					$\alpha = 0.05$?	$\alpha = 0.10$?
2. Rural Principal Arterial	4, 8	40.017	2, 36	<.001	Yes	Yes
4. Rural Major Collector	4	180.062	1, 35	<.001	Yes	Yes

(*) df_R = degrees of freedom for regression; df_E = degrees of freedom for error

TABLE 11 PARTIAL *F*-TESTS FOR AGGREGATE ANALYSIS

Highway Category	Variable Subscripts for		df_R, df_F (*)	F*	α	Is H_a true for	
	Full Model	Reduced Model				$\alpha = .05$?	$\alpha = .10$?
2. Rural Principal Arterial	4, 8	4 8	37, 36 37, 36	4.963 61.223	.025-.050 2.001	Yes Yes	Yes Yes
4. Rural Major Collector	It has only one variable in Full Model.						

(*) df_R = degrees of freedom for SSE or Reduced Model and df_F = degrees of freedom for SSE for Full Model.

TABLE 12 FINAL REGRESSION EQUATION FROM AGGREGATE ANALYSIS

4. Rural Major Collector:	
	AADT = -7048.27 + 0.27151 County Population
$R^2 = 0.887$	t = 13.41872
F = 180.062	e = 3.77379

closest census year is the suggested present year. The forecasted errors are reasonably small in most of the cases and speak well for reliability of the models. The larger forecast errors in some cases are due to low values of AADT, fewer cases, and large variations in response and predictor variables employed in data tables among the stations and counties.

DISAGGREGATE ANALYSIS

In disaggregate analysis, each station has been analyzed separately and a separate forecasting model has been developed for each. The steps in model development and the criteria for variable selection are the same as those in the aggregate analysis.

Model Development

The final regression equations for rural major collector stations are presented in Table 15, along with R^2 values, overall *F* values, *t*-statistics, and elasticities. In case of station 7047A, the negative sign with county population is not surprising. In fact Rush County is losing population over the years, and the AADT values are also decreasing for the period of analysis. This fact establishes the negative sign with county population for station 7047A. Not all the goals of Table 8 have been set in all of the equations of Table 15. However, these equations are the best possible attempts at meeting the goals of Table 8. The station-specific disaggregate models are presented in Table 16. Each of the models is simple, with not more than two variables in any case. The use of these models is also straightforward. The data needed to predict rural traffic volumes with these models are readily available at the county, state, and national levels. The disaggregate model, however, requires a "similarity test" to determine if a segment whose future AADT is desired has characteristics that permit it to use one of the location-specific models (*9*). As a crude test of the model's performance, forecasts of 1983 and 1984 traffic levels were made for comparison against actual volumes for those years. The disaggregate forecasts (Table 17) were much more competitive than were the aggregate forecasts, but the time frame was too short to be conclusive.

SUMMARY AND CONCLUSIONS

The objective of the study described in this paper was to develop a method of forecasting traffic volumes on rural

TABLE 13 AGGREGATE TRAFFIC FORECASTING MODELS

1. Rural Interstate:

$$AADT_f = AADT_p \left[1 + 4.83314 \left(\Delta \text{ State Population}\right)\right]$$

$R^2 = 0.658$

2. Rural Principal Arterial:

$$AADT_f = AADT_p \left[1 + 1.47809 \left(\Delta \text{ County Population} + 2.79623 \left(\Delta \text{ State Population}\right)\right]$$

$R^2 = 0.690$

3. Rural Minor Arterial:

$$AADT_f = AADT_p \left[1 + 0.83377 \left(\Delta \text{ County Household}\right)\right]$$

$R^2 = 0.727$

4. Rural Major Collector:

$$AADT_f = AADT_p \left[1 + 3.77379 \left(\Delta \text{ County Population}\right)\right]$$

$R^2 = 0.837$

(i) R^2 value is for final regression equation.

(ii) Δ represents change in predictor variable with respect to its present value in fraction. For example, $\Delta X = \dfrac{X_f - X_p}{X_p}$, where X_p and X_f denote present and future values of X.

TABLE 14 PERFORMANCE OF AGGREGATE TRAFFIC FORECASTING MODELS (RURAL MAJOR COLLECTOR)

Traffic Count Station	Present Year	Year	Actual AADT ($AADT_a$)	Forecasted AADT ($AADT_f$)	$AADT_f$ - $AADT_a$	Forecast Error in percent (*)
7047A	1790	1971	257	266	9	3.50
		1972	227	296	69	30.40
		1973	233	292	59	25.32
		1974	226	261	55	24.34
		1975	225	271	46	20.44
		1976	231	271	40	17.32
		1977	204	271	67	32.84
		1978	224	271	47	20.98
		1979	294	241	-53	-18.03
		1980	299	236	-63	-21.07
		1981	288	226	-62	-21.53
		1982	272	205	-67	-24.63
30063A	1980	1979	877	752	-125	-14.25
		1981	824	800	-24	-2.91
		1982	767	793	26	3.38
54382A	1980	1979	1159	1062	-97	-8.37
		1981	973	984	11	1.13
		1982	878	1029	151	17.20
200X	1980	1973	8805	6547	-2258	-25.64
		1974	8834	6823	-2011	-22.76
		1975	9002	7155	-1847	-20.52
		1976	9033	7431	-1602	-17.73
		1977	9079	8038	-1041	-11.47
		1978	9457	8535	-922	-9.75
		1979	9636	8977	-659	-6.84
		1981	9226	9197	-29	-0.31
		1982	9004	9308	304	3.38

(*) "+" sign indicates overprediction and "-" sign indicates underprediction.

$$\text{Forecast Error in percent} = \frac{AADT_f - AAADT_a}{AADT_a} \times 100$$

TABLE 15 FINAL REGRESSION EQUATIONS FROM DISAGGREGATE ANALYSIS

Rural Major Collector		
Station 59A:	$AADT = 2772.53 + 0.0481$ County Vehicle Registrations	
	$R^2 = 0.7222$	$t = 5.344$
	$F = 28.563$	$e = 0.36063$
Station 200X:	$AADT = 6557.79 + 0.0503$ County Vehicle Registrations	
	$R^2 = 0.635$	$t = 3.734$
	$F = 13.940$	$e = 0.28407$
Station 5420A:	$AADT = 784.34 + 0.1163$ County Employment	
	$R^2 = 0.681$	$t = 4.842$
	$F = 23.447$	$e = .59744$
Station 7047A:	$AADT = 1122.06 - 0.0433$ County Population	
	$R^2 = 0.521$	$t = -3.462$
	$F = 11.988$	$e = -3.48274$

TABLE 16 DISAGGREGATE TRAFFIC FORECASTING MODELS

Rural Interstate $\{0.706 \leq R^2 \leq 0.925\}$		
Station 172A:	$AADT_f =$	$AADT_p \lvert 1 + 5.24231 (\Delta$ State Population$) \rvert$
Station 3070A:	$AADT_f =$	$AADT_P \lvert 1 - 0.44503 (\Delta$ US Gas Price$) + 7.74428 (\Delta$ State Pop.$) \rvert$
Station 5474A:	$AADT_f =$	$AADT_P \lvert 1 + 6.18172 (\Delta$ County Population$) \rvert$
Rural Principal Arterial $\{0.607 \leq R^2 \leq 0.976\}$		
Station 68A:	$AADT_f =$	$AADT_P \lvert 1 + 0.86979 (\Delta$ State Vehicle Registrations$) \rvert$
Station 134A:	$AADT_f =$	$AADT_P \lvert 1 - 0.43949 (\Delta$ US Gas Price$)$ $+ 0.83878 (\Delta$ State Vehicle Registrations$) \rvert$
Station 173A:	$AADT_f =$	$AADT_P \lvert 1 + 1.47643 (\Delta$ County Vehicle Registrations$)$ $- 0.21371 (\Delta$ US Gas Price$) \rvert$
Station 254B:	$AADT_f =$	$AADT_P \lvert 1 + 0.60300 (\Delta$ County Vehicle Registrations$) \rvert$
Rural Minor Arterial $\{0.525 \leq R^2 \leq 0.867\}$		
Station 25A:	$AADT_f =$	$AADT_P \lvert 1 + 0.90147 (\Delta$ County Vehicle Registrations$)$ $- 0.29365 (\Delta$ US Gas Price$) \rvert$
Station 279A:	$AADT_f =$	$AADT_P \lvert 1 - 0.26635 (\Delta$ US Gas Price$)$ $+ 0.24526 (\Delta$ County Employment$) \rvert$
Station 301A:	$AADT_f =$	$AADT_P \lvert 1 + 0.66731 (\Delta$ County Vehicle Registrations$)$ $- 0.26576 (\Delta$ US Gas Price$) \rvert$
Station 319A:	$AADT_f =$	$AADT_P \lvert 1 + 61456 (\Delta$ County Vehicle Registrations$) \rvert$
Rural Minor Arterial $\{0.525 \leq R^2 \leq 0.867\}$		
Station 42A:	$AADT_f =$	$AADT_P \lvert 1 + 0.49887 (\Delta$ County Vehicle Registrations$) \rvert$
Station 100X:	$AADT_f =$	$AADT_P \lvert 1 + 0.64875 (\Delta$ County Vehicle Registrations$) \rvert$
Station 256A:	$AADT_f =$	$AADT_P \lvert 1 + 0.33059 (\Delta$ County Vehicle Registrations$) \rvert$
Station 262A:	$AADT_f =$	$AADT_P \lvert 1 + 0.28236 (\Delta$ County Vehicle Registrations$)$ $- 0.23256 (\Delta$ US Gas Price$) \rvert$
Rural Major Collector $\{0.521 \leq R^2 \leq 0.722\}$		
Station 59A:	$AADT_f =$	$AADT_P \lvert 1 + 0.36063 (\Delta$ County Vehicle Registrations$) \rvert$
Station 200X:	$AADT_f =$	$AADT_P \lvert 1 + 0.28407 (\Delta$ County Vehicle Registrations$) \rvert$
Station 5420A:	$AADT_f =$	$AADT_P \lvert 1 + 0.59744 (\Delta$ County Employment$) \rvert$
Station 7047A:	$AADT_f =$	$AADT_P \lvert 1 - 3.48274 (\Delta$ County Population$) \rvert$

R^2 value is for final regression equation; subscripts P and f represent present and future years. Δ represents change (decimal) in predictor variable with respect to its present year value. For example, $\Delta X = \frac{X_f - X_P}{X_P}$, where X_P and X_f denote present and future year values of X.

TABLE 17 PERFORMANCE OF DISAGGREGATE TRAFFIC FORECASTING MODELS (RURAL MAJOR COLLECTOR)

Traffic Count Station	Present Year	Year	Actual AADT ($AADT_a$)	Forecasted AADT ($AADT_f$)	$AADT_f$ - $AADT_a$	Forecast Error in percent (*)
59A	1980	1983	4551	4701	150	3.30
		1984	4769	4718	-51	-1.07
200X	1980	1983	9297	9471	174	1.87
		1984	9950	9568	-382	-3.84
5420A	1980	1983	1979	2117	138	6.97
7047A	1980	1983	281	262	-19	-6.76
		1984	273	262	-11	-4.03

(*) "+" sign indicates overprediction and "-" sign indicates underprediction.

$$\text{Forecast Error in percent} = \frac{AADT_f - AAADT_a}{AADT_a} \text{ x } 100.$$

state highways that would be reliable, understandable, easy to use, and well documented. While the reliability of a forecasting model is always difficult to prove in advance, the models in Tables 13 and 16 are each the product of a series of careful statistical tests. The models are understandable: the level and nature of the variables that appear in their final forms have a clear functional relationship. They are not in the model only to provide a better curve fit. They are easy to use: the predictor variables are relatively easy to procure, and the models themselves require only a hand calculator to implement. The two different kinds of models—aggregate and disaggregate—that are offered in the full report (*9*) to forecast traffic volumes at rural locations in Indiana's state highway network can be implemented using any spreadsheet software. And the models are well documented in a report (*9*).

There were (and are) some inevitable problems, however, in developing traffic growth factor values for four functional classes of highways (interstate, principal arterial, minor arterial, and major collector). Although a simple forecasting model was sought, the statistical analyses to develop it were extensive, sometimes complex, and often subject to the analysts' judgment. Most important, there is an even greater need than usual for more data from a larger number of count stations to fill the many gaps in the database used in this study. Fortunately, the Indiana state highway department is installing new, permanent count stations, and a historical record of traffic volumes is now being developed for a wider range of locations.

ACKNOWLEDGMENTS

The work described in this paper was supported by the Federal Highway Administration (FHWA) through the Indiana Department of Highways (IDOH). The authors thank John Nagle of IDOH, the FHWA's Gary DalPorto, and Kumares C. Sinha of Purdue University for their guidance as members of the project's advisory committee. Special thanks are also due Professor Virgil Anderson, who provided advice on the use (and potential abuse) of methods of statistical analysis.

REFERENCES

1. F. T. Morf and V. F. Houska. Traffic Growth Pattern on Rural Areas. *Highway Research Board Bulletin 194*, January 1958, pp. 33–41.
2. A. J. Neveu. Quick Response Procedures to Forecast Rural Traffic. In *Transportation Research Record 944*, TRB, National Research Council, Washington, D.C., June 1982.
3. Minnesota Department of Transportation (MnDOT). *Procedures Manual for Forecasting Traffic on the Rural Trunk Highway System.* Traffic Forecasting Unit, Program Management Division, MnDOT, April 1985.
4. D. Albright. *Abstract on—Trend-Line: Forecasting Heavy Commercial Traffic as a Percent of Average Daily Traffic.* New Mexico State Highway Department, 1985.
5. J. S. Armstrong. The Ombudsman: Research on Forecasting: A Quarter-Century Review (1960-1984). *Interfaces*, Vol. 16, No. 1, January–February, 1986, pp. 89–109.
6. J. S. Armstrong. Forecasting by Extrapolation: Conclusion From 25 Years Research. *Interfaces*, Vol. 14, No. 6, November-December 1984, pp. 52–56.
7. AASHTO. *A Policy on Geometric Design of Highways and Streets*, 1984.
8. M. D. Meyer and E. J. Miller. *Urban Transportation Planning: A Decision-Oriented Approach.* McGraw-Hill Book Co., New York, 1984.
9. J. D. Fricker and S. K. Saha. *Traffic Volume Forecasting Methods for Rural State Highways.* Final Report, Joint Highway Research Project, Purdue University, May 1987.
10. V. L. Anderson and R. A. McLean. *Design of Experiments: A Realistic Approach*; Statistics: Text Books and Monographs, Vol. 5. Marcel Dekker, Inc., New York, 1974.

11. N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner, and D. H. Bent. *SPSS—Statistical Package for the Social Sciences*, 2nd ed. McGraw-Hill Book Co., New York.
12. R. J. Freund and P. D. Minton. *Regression Methods—A Tool for Data Analysis.* Marcel Dekker, Inc., New York, 1979.
13. Purdue University Computer Center (PUCC). *DRRSQU (Program for C_p-Statistic—All Possible Regression): Statistical Program.* Statistical Library, PUCC, Purdue University, West Lafayette, Indiana.
14. N. R. Draper and H. Smith. *Applied Regression Analysis.* John Wiley and Sons, Inc., New York, 1981.
15. J. Neter and W. Wasserman. *Applied Linear Statistical Models.* Richard D. Irwin, Inc., Homewood, Ill., 1974.
16. University of California, Los Angeles (UCLA). *Biomedical Computer Programs (BMDP): P-series.* Health Science Computing Facility, UCLA, 1979.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.