Factor Analysis of Roadway and Accident Data

JOHN VERSACE, Engineering Psychologist, Chrysler Corp., Detroit

This report is a further analysis of data reported by Schoppert (1). In that study, a number of features characteristic of unit lengths of Oregon two-lane rural highways were related to the occurrence of accidents. Multiple regression equations were obtained to predict where accidents would occur. The data were stratified according to ADT and geography. The present analysis sought to make apparent additional relationships among the roadway features while compressing the results into a small number of more universal notions.

● THERE IS NO ONE CAUSE of automobile accidents. Instead, there are innumerable influences acting at any instant, and for all we know there may even be a residual component of "causelessness," like the uncertainty principle of quantum physics. The fact that there is such a great number of influences should direct us to explore techniques that will seek to find groupings of those influences that have something in common. This common element then would take on a significance of its own and allow us to consider a smaller number of more comprehensive ideas instead of individual influences.

The procedure used in this study is factor analysis. Factor analysis is not a new statistical method, and has been principally applied in the field of mental test construction. A 1958 Highway Research Board report (2) consisted of a very elaborate factor analysis of driver attitudes. It is not so common to find factor analysis applied to investigations like the present one.

FACTOR ANALYSIS

Only the most rudimentary introduction to factor analysis will be discussed here, as there are complete expositions available elsewhere (3, 4, 5), and details appear in the Appendix; for our present purposes, a factor can be thought of as an abstraction or concept which embodies, to varying degrees, a number of aspects of the observable world which have something in common. Thus, one may speak of the factor of human intelligence as an abstraction which embodies a number of measurable entities—ability to remember, ability to do arithmetic, ability to translate verbal problems into algebra, ability to read a map and follow instructions, etc. By invoking the comprehensive notion of "intelligence" to embrace all these abilities we have simplified our understanding. Likewise, traffic accidents can be considered to result from a host of individual influences, but on closer look many of these influences may be found to be so highly interrelated that a few well-chosen abstractions might substitute for the larger number of individual influences.

For example, there may be a tendency to more accidents among those people who have rheumatism or severe eye disorders or poor glare recovery and dark adaptation, or who are unemployed, bald, widowed, or have owned ten cars, etc. But instead of drawing conclusions about the relation of each of these variables to accidents, it might be said that a common factor—that of "old age"—was the important concept to associate with the accidents. There may be additional comprehensive factors, though, that are independent of that of old age and of one another. It is possible that two or three such general factors might account for nearly all of the predictable accidents among people.

If "accidents" is represented by A, and factors by F_1, F_2, etc.,

$$A = B_{1a}F_1 + B_{2a}F_2 + \dots + B_{ua}U_a \qquad (1)$$

is a multiple regression equation in which the fewer generalized factors, rather than all the individual directly observed variables, are the predictors. The coefficients, B_{ia}, are the factor loadings, or coefficients of correlation of variable A with the factors F_i. The factor U is generally called a unique factor, but should be interpreted as representing what is left over after all that is common among the variables is accounted for and represented in the preceding factors. Thus, in the foregoing example, there are some aspects of accidents that are not shared by any of the other variables included in the analysis, and this fact will be represented by a loading on the unique factor U_a.

The coefficients may be regarded as direction cosines of a vector, so the sum of their squares equals unity; that is,

$$1.00 = B_{1a}^2 + B_{2a}^2 + \ldots + B_{ka}^2 + B_{ua}^2 \qquad (2)$$

Statisticians identify these squared loadings as relative variances. Thus, if $B_{1a} = 0.80$, then $B_{1a}^2 = 0.80^2 = 0.64$, and it can be said that 64 percent of the variance in the observations of variable A is accounted for by F_1.

The objective of factor analysis is at least twofold: to determine the numerical value of the coefficients, and to provide a basis to identify meaningfully the factors which emerge.

PROCEDURE

Fourteen variables were chosen for analysis. These covered traffic, road, and accident features, and were observed on nearly 1,400 one-mi stretches of Oregon two-lane rural highways with gravel shoulders. For the purposes of this study, accidents were combined into a single variable. To obtain the accident score, each accident was weighted with a 3 if it was fatal, a 2 if it was personal injury, and a 1 if it was simply property damage. No intimation is intended by this weighting system that these ratios are representative of relative incidence, or of economic or personal loss. It was intended only to provide an accident score which may be somewhat more predictable. Terrain was indicated by a 3-point scale for level, rolling, and mountainous. No attempt to weight this designation was made.

The rest of the procedure was to determine the coefficients of correlation of all variables with one another, to factor analyze this correlation matrix, and to rotate the resultant factor axes to a position which most clearly shows the results.

Some partial correlation coefficients were obtained to clarify the relation of certain roadway features to the accident score. An analysis of variance was also done on the accident score to show the changing influence of access points with different ADT ranges.

RESULTS

The factor results are shown in Table 1. Preliminary results, the correlation matrix, means, and the standard deviations are in the Appendix (Table 3), as is the original unrotated factor matrix (Table 4). Only the first four factors had any substance, and only those are shown. After rotation of the original factor axes to a more meaningful position, the new factor matrix in Table 1 appeared. Table 1 gives the principal results of this investigation.

TABLE 1

QUARTIMAX-ROTATED FACTORS

Symbol	Variable	Factors[a] I	II	III	IV
1 ADT	Average daily traffic	23	66	48	13
2. LA	Lane width	22	27	83	-07
3. SH	Shoulder width	39	30	54	33
4. TC	Terrain code	-70	-15	16	-12
5. SDR	Sight distance restriction	-94	08	02	07
6. LS	Length of structures	-07	13	12	26
7. WS	Width of structures	-06	15	14	30
8. NS	No. of structures	-06	15	08	32
9. NC	No. of curves	-55	-05	01	-02
10. CDW	No. of commercial driveways	-01	66	00	-10
11. RDW	No. of residential driveways	05	46	-11	-16
12. INT	No. of intersections	08	54	04	06
13. CAP	Calculated capacity	93	06	34	01
14. A	Accidents	-04	80	21	12

[a]Entries are decimals (x0.01).

The results in the table should be looked at both vertically and horizontally. Each column represents a factor, and the loadings on each factor are the correlation coefficients between the row variable and the factor. Thus, there is a correlation of 0.93 between the CAP variable and Factor I. Each row can be considered a multiple regression equation. For example,

$$A = -0.04F_{I} + 0.80F_{II} + 0.21F_{III} + 0.12F_{IV} + 0.55U_a$$

The coefficient in the last term was obtained by subtraction because the sum of the squared coefficients is unity.

A set of factor identifications will be proposed, but these are only proposals. There may be better ways of describing them, and the reader is encouraged to make his own interpretations. The largest entries in each column provide the basis for description of the corresponding factors.

Factor I. Capacity

Capacity has a technical definition in terms of the maximum number of vehicles that can use a roadway in unit time. Roads with given calculated capacity, however, may be constructed in areas that have other special characteristics. Thus, it is observed that two-lane rural highways in Oregon that have high calculated capacity are markedly free of sight distance restriction and are not often in mountainous terrain, tend to have fewer curves, and there is some small tendency for the shoulders to be wider. Loadings of ADT and of accidents on the capacity factor are negligible.

Factor II. Traffic Conflict

This factor consists of that complex made up of accidents, traffic volume, commercial and residential driveways, and intersections. Wide lanes and shoulders are also involved, but their contribution is relatively small. Accidents go along with a lot of traffic on the road, especially when there are more interferences from the side. The relative values of the loadings on this factor should not be interpreted as proportional contributions to accidents. The relative contribution to accidents of ADT and commercial driveways depends on the extent to which both are present on the same stretches of highway. Two other procedures, a partial regression analysis and an analysis of variance, reviewed below, help to separate the relative contribution to accidents of each source.

Factor III. Modern Roads

The outstanding feature of this factor is the tendency to wide lanes and shoulders, increased volume, and some tendency for these roads to have a higher calculated capacity. There is a small relation of this factor to accidents. "Traffic volume," an alternate designation for this factor, is more general and may be preferred.

Factor IV. Roadside Structures

This factor is rather weak, accounting for quite a bit less of the total variance than the others. It refers to the presence of roadside structures and wide shoulders, with scarcely visible influence on anything else. It is interesting to note that shoulder width variance is contributed to almost equally by all four factors. In the first factor it accompanies the level terrain and freedom from curves and sight restriction, in the second factor it accompanies the driveways and intersections, and in the third factor it accompanies wide lanes and traffic volume.

It was stated above that the factor loadings for a given variable are direction cosines, and that their squares add up to unity. These squares are also the proportions of variance for which the factors can account in the observations of that variable. The percentage of variance accounted for in each variable is summarized in Figure 1.

Keeping Volume Independent

In the factor descriptions, it was noted that some roadway features accompanied ADT and thus perhaps did not have the significance to the factors that their loadings might suggest. The relation of accidents to each of these features is made less clear by the overlapping presence of ADT. (There are more accidents where the lanes are wider, but how much of this increase is caused by the heavier traffic that accompanies wider lanes?) To find the relation of lane width to accident score, independent of any influence of ADT, the partial correlation coefficient is required. Table 2 shows the gross correlation between accident score and a number of roadway variables, and also the net, or partial, correlation when the influence of ADT was excluded. As expected, there was a very small but inverse relation between accidents and either lane or shoul-

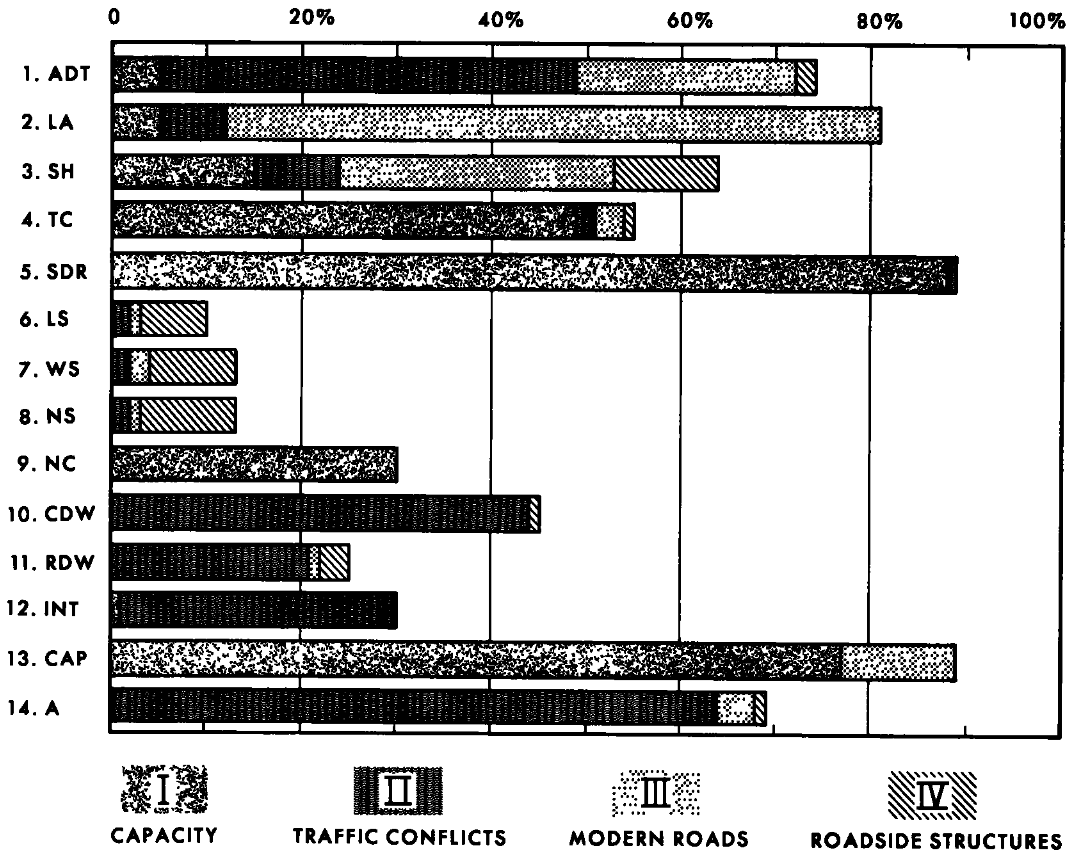

Figure 1. Proportion of total variance accounted for by four factors in each variable.

der width. Since shoulder width is related to lane width, an additional partialling of these effects shows accident score was quite independent of lane width, but not of shoulder width, although the remaining effect was small. Roads with more driveways and intersections also have a greater ADT, but Table 2 shows that the effect of driveways and intersections on accidents still remained appreciable when the influence of ADT was removed.

Traffic Conflict at Low Volume

Although increases in traffic volume and number of accesses lead to more accidents, on the average, it has been repeatedly observed that on those highways where volume is low the accident frequency bears little relation to number of accesses. A statistical interaction is present which is not accounted for in the usual correlational procedures (and hence factor analysis) unless separate analyses are performed on stratified parts of the data. There are many places where statistical interactions might exist in the data collected in this report. In a planned experiment, where one programs the conditions of observation in advance, it is very easy to detect all the interactions. But in a survey analysis, the practical limitations resulting from the

TABLE 2

CORRELATION COEFFICIENTS (r) OF ACCIDENT SCORE WITH SELECTED ROADWAY VARIABLES

Variable	Value of r	
	Gross	Partial[a]
LA	0.370	-0.103
SH	0.308	-0.223
LA[b]	—	-0.029
SH[c]	—	-0.198
CDW	0.525	0.369
RDW	0.343	0.285
INT	0.406	0.237

[a]ADT held constant.

[b]Shoulder width held constant.

[c]Lane width held constant.

disproportionate frequency of occurrence of certain combinations of conditions make clarification of interactions difficult. This particular interaction, however, can be readily shown.

An analysis of variance of accident score as a function of ADT and number of accesses (total of residential and commercial driveways and intersections) per mile was performed, and is presented in the Appendix. Figure 2 shows the chief results: There is a substantial interaction, revealed by the fanning out of the curves with increasing ADT, but the average effects of ADT and number of accesses is still very pronounced over and beyond the interaction.

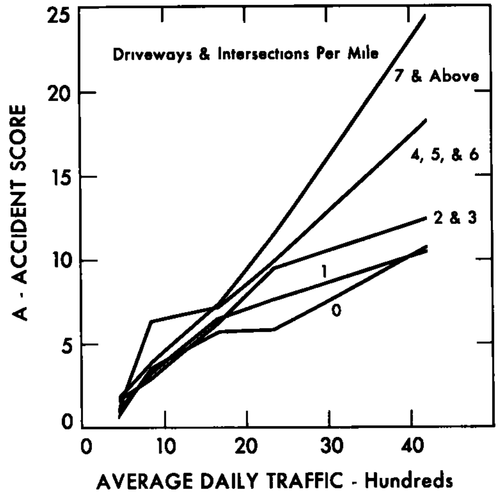

Figure 2. Accident score A associated with ADT and number of driveways and intersections. Points are plotted at the medians of the quintile intervals of Table 5.

DISCUSSION

Study of Figure 1 should promote a number of ideas and hypotheses. Sight distance restriction seems unrelated to anything other than capacity; lane width seems to be associated chiefly with modern roads and only marginally related to accidents. Even that relationship is a positive one, because it is carried along by traffic volume. The same is true of shoulder width. Absence of curves is moderately related to capacity, but otherwise tends to be unique. Accidents seem to be predominantly related to the single factor, traffic conflict.

The relative influence on accidents of the various roadway features is indicated by the partial correlation coefficients in Table 2, although that is not the whole story. It is wrong to suppose that the beta weights of a multiple regression equation are indicators of relative importance if there is any appreciable correlation among the predictors. Thus, of two predictors that are highly correlated and have nearly the same correlation with the dependent variable, the one with the higher coefficient—even if only slightly higher—will take on nearly all the loading in the regression equation. On a subsequent test sample where usual fluctuations might produce a coefficient slightly higher for the other variable, the preponderance of the loading will be on it. This will happen often if the predictors have a factor common to them.

The analysis of variance offers some additional insight into the relative importance of volume and access points. In an analysis where there is homogeneity of variance among the subgroups and where there is an equal number of cases in each subgroup, it is simple to calculate the relative contribution of each effect. In the present instance that would be awkward to do, so the qualitative impression conveyed by Figure 2 must suffice. Some additional insight can be derived from the values of the mean squares in the analysis of variance summary in the Appendix (Table 6) although they should not be considered as directly representative of the relative contribution of each source.

There will be instances when traffic circumstances are so tight that even a good driver in a responsive car on a relatively good road will have an accident. It would seem reasonable to interpret Factor II in terms of the burden placed on the driver, who may be controlling a vehicle of less than optimum responsiveness on a road of less than optimum configuration. This burden would be on his ability to make the correct judgment of the situation and to carry through the appropriate maneuver of his vehicle in order to preserve some measure of equilibrium in the traffic flow. This ability will further depend on the characteristics of his car and the highway. Thus, although the data of this analysis were obtained from roadway features, Factor II, at least, has an element of interpretation in terms of human control and responsiveness. Schoppert (1) concluded, "The number of accidents increases with the number of situations presenting a change in conditions, and therefore requiring a decision on the part of the motor vehicle operator."

The unique factor, U, and the unexplained third of the variance in accidents for which it accounts require some interpretation. The unique factor may not be a single factor at all, but should be considered as the lump sum of all factors which may really be operating (as well as a basic residual of uncertainty) but which were untapped by this study because only the particular 14 variables of Table 1 were chosen for investigation. Thus, if several dozen additional roadway variables had been included, it is probable that some additional factors would have been discovered. Whether the accidents variable would have any significant loadings on these new factors and would thus diminish the present portion of unexplained variance, can only be guessed. Also, the factors already present might change their character somewhat and thus possibly require an expanded identification or interpretation if these new variables have appreciable loadings on them. Again, this is a matter of conjecture. The factor matrix of Table 1, with 4 columns and 14 rows, should be considered as only a corner out of a larger hypothetical matrix consisting of more columns and rows, but of whose complete structure we are ignorant at present. Since the units of observation in this analysis were stretches of roadway, the number of features, in addition to the 14 already used, that could be included in the analysis is limited. It is assumed that no radical alterations of these factors would occur.

It should be understood that this analysis had to do with where accidents occur, and the units of observation were elemental stretches of two-lane rural highway in Oregon. A completely independent study would be designed to find out who has accidents, and in that case the units of observation would be individual people and the variables might be such things as personality, attitudes, medical condition, etc. The percent of explainable variance in accidents among people would be a completely independent determination from that of percent of explainable variance in accidents among places. A great deal of research effort has been expended in the search for personal factors in accident causation, but the results have not been as fruitful. That is, whereas the present investigation shows the few factors that can account for a substantial portion of the variance among places where accidents occur, more than a third of the variance in accidents among people remains unexplained.

The results should be compared with those found by Woo (6) on more than 3,000 mi of varied highways in Indiana. Again, ADT was found to be the variable most highly related to accident occurrence. His study was limited to non-intersectional accidents, and he found a strong relation of accidents to number of access points, but only for ADT less than 999 and between 7,000 and 8,000. Whereas the proportion of unexplained variance found in the Oregon data would imply a multiple correlation coefficient of about 0.82, the indices drawn from the much more heterogeneous highways in the Indiana study gave a multiple correlation coefficient of 0.47.

CONCLUSIONS

Only a single factor emerged from the vast amount of data in this analysis which explained where accidents occurred. Although only highway variables were included in the analysis, this one factor conveys a psychological meaning: There are more accidents at those places where the situation places great demand on the momentary perceptual-decision-motor capacities of the driver. The driver's basic psychophysical capacities are heavily exercised when he must deal with a situation around him that is changing rapidly. This occurs where the traffic friction or conflict is greater, that is, where one encounters more cars and where the flow of traffic is further interfered with by intersections and driveways. Accidents are most frequent in those circumstances. Accident frequency is proportional to the load or rate of demand placed on the driver's basic ability to perceive and cope with the situation.

The results of this analysis are, in general, consistent with other researches, such as those reported by Raff (7). This study has showed how, by application of factor analysis, a great deal of information can be compressed into very compact form (Table 1). An estimate of the explanatory power of the results also was obtained (Fig. 1), and the nature of the situations leading to accidents was made more clear (Factor II). If our only aim is to predict where accidents will occur, the stratified multiple regression techniques of the original report (1) on these data is sufficient and probably preferable. The factor analysis has allowed us to obtain an additional understanding of all the relationships among a great number of roadway variables. The analysis and interpretations in this report are limited to Oregon two-lane rural highways with gravel shoulders, but the obtained factors seem to have a validity that may lead us to expect some degree of universality.

ACKNOWLEDGMENT

Special thanks are due to G. S. Paxson and F. B. Crandall of the Oregon State Highway Department for their cooperation in making the raw data available and for their great interest and informed criticism. T. W. Forbes and J. H. Mathewson, and their colleagues, also contributed some needed interpretation.

REFERENCES

1. Schoppert, D. W.,"Predicting Traffic Accidents from Roadway Elements—Rural Two-Lane Highways with Gravel Shoulders." Oregon State Highway Dept., 1956; HRB Bull. 158 (1957).
2. Goldstein, L. B., and Mosel, N. J., "A Factor Study of Drivers' Attitudes, with Further Study on Driver Aggression." HRB Bull.172 (1958).
3. Holzinger, K. O., and Harman, H. H., Factor Analysis. Chicago Univ. Press (1941).
4. Thomson, G. H., The Factorial Analysis of Human Ability. Houghton Mifflin, N.Y. (1951).
5. Thurstone, L. L., Multiple Factor Analysis. Chicago Univ. Press (1947).
6. Woo, J. C. H., "Correlation of Accident Rates and Roadway Factors." Joint Highway Research Project, Purdue Univ., Final Report (1957).
7. Raff, M. S., "The Interstate Highway Accident Study." Public Roads, 27, 170-186 (1953).
8. Wrigley, C., "The Case Against Communalities." Paper presented before the Psychometric Society, N. Y. (Sept. 1957).
9. Neuhaus, J. O., and Wrigley, C., The Quartimax Method. Brit. J. Stat. Psychol., 7, 81-91 (1954).
10. Walker, H. M., and Lev , J., Statistical Inference. Henry Holt, N. Y. (1953).

Appendix

CORRELATIONS

Three of the variables required some additional adjustment prior to inclusion in the analysis. Since only those roadside structures which occurred could have a length or a a width, and since in most of the one-mi segments there were no structures, it was possible to get a spuriously high intercorrelation among these variables because of all the simultaneously occurring zeros. Therefore, these three intercorrelations were based on only the 252 stretches of road that had structures. However, the data as reported gave the total length of structures in the entire stretch, and the average width. The total length obviously was dependent on the number of structures, so an additional adjustment to remove this spurious element was needed. Pearson's old formula for spurious correlation between indices was used to get the uninflated correlation between number of structures and the average length in each stretch.

$$r_{\frac{a}{b}b} = \frac{r_{ab} v_a - v_b}{\sqrt{v_a^2 + v_b^2 - 2 r_{ab} v_a v_b}}$$

where v = coefficient of variation

Most of the variables had distributions that were highly skewed; the standard deviations were quite large relative to the means and most of the variables were bounded by zero at the lower end. Although normality is not a requisite for the calculation of a correlation coefficient, lack of it might cause difficulty in interpretation. A parallel analysis was performed in which the variables were transformed to approximate normality by taking logarithms and changing frequencies to $(X+0.25)^{2/3}$. Results were substantially the same, so the alternate analysis will not be discussed. The correlations, means, and standard deviations are shown in Table 3.

Some question may be raised about the use of classification numerals as numbers for the terrain designation. Since the numerals were in increasing rank order to correspond to increasing hilliness, it seemed no more objectionable to use them than the conventional procedure of attaching 0 and 1 to qualitative classes in many other contingency correlation tables. Because three classes were used, it must be expected that each will include a wide variety of terrain characteristics, but the average of one class should certainly be clearly different from those of the others. No attempt to weight the classifications was made, however.

TABLE 3

CORRELATIONS, MEANS, AND STANDARD DEVIATIONS[a]

Variable	2	3	4	5	6	7	8	9	10	11	12	13	14	Mean	S.D.
1. ADT	611	618	-238	-142	121	164	139	-159	406	206	354	395	701	20.9	16.7[b]
2. LA		594	-042	-152	104	109	074	-142	166	064	207	519	370	10.3	1.06
3. SH			-261	-310	153	227	181	-241	198	041	265	577	308	4.67	2.64
4. TC				600	039	022	-001	423	-060	-097	-130	-618	-062	0.71	0.71
5. SDR					091	095	096	506	069	-016	-030	-885	077	57.8	35.5[c]
6. LS						092[d]	129[d]	012	023	-025	102	-012	195	4.09	4.87[e]
7. WS							137[d]	017	060	-002	112	003	209	24.9	4.4
8. NS								020	037	012	126	-014	209	1.32	0.59
9. NC									-053	-061	-080	-507	033	1.10	2.15
10. CDW										339	426	023	525	1.33	2.65
11. RDW											272	044	343	1.83	3.56
12. INT												127	406	1.01	1.23
13. CAP													094	4.93	1.49
14. A														7.92	9.43

[a] Correlation coefficients (Cols. 2-14) are decimals; N = 1,391.
[b] Hundreds. [c] Percent. [d] N = 252. [e] Hundredths of a mile.

FACTOR ANALYSIS

The mathematical procedure of factor analysis used in this investigation was that of principal components analysis (3, 4, 5), which consists of finding the characteristic vectors of the correlation matrix. The rank of the correlation matrix was reduced by inserting communality estimates in the diagonal cells instead of 1.00. The communality is the proportion of variance in each variable that is accounted for by all the non-unique factors, F_i. This entry was estimated, in each case, by the squared multiple correlation coefficient, R^2, when each variable was predicted in turn by the remaining 13 variables (8). The analysis yields a factor matrix, where the columns represent the coordinates of the factor vectors in the space defined by the variables, and the rows represent normalized variables vectors in factor space.

The factor matrix resulting from this procedure is one where the first principal component or first factor accounts for the maximum variance, and each successive factor accounts for the maximum of what is left over at each stage. However, the factor space defined by axes that account for the greatest share of the variance is not necessarily the one that produces the most significant results. Therefore, the factor axes are rotated until the projections of the variables vectors upon them meet some criterion of meaningfulness. In this investigation, the quartimax method (9) was used to obtain approximate simple structure while keeping the axes orthogonal. The quartimax method rotates the factor axes until the maximum variance of the squared factor loadings (proportions of variance) is achieved.

TABLE 4

PRINCIPAL FACTORS WITH R^2 COMMUNALITY ESTIMATES

Symbol		Variable	Factors[a]				h^2 [b]	R^2
			I	II	III	IV		
1.	ADT	Average daily traffic	75	41	-05	-02	73	72
2.	LW	Lane width	66	22	-49	-29	81	81
3.	SH	Shoulder width	72	09	-30	17	64	63
4.	TC	Terrain code	-51	42	-30	-18	56	58
5.	SDR	Sight distance restriction	-60	72	-15	03	90	93
6.	LS	Length of structures	10	18	-11	21	10	09
7.	WS	Width of structures	14	21	-11	25	14	12
8.	NS	No. of structures	12	20	-07	28	14	11
9.	NC	No. of curves	-41	35	-12	-03	31	30
10.	CDW	No. of commercial driveways	34	44	36	-12	45	38
11.	RDW	No. of residential driveways	22	23	36	-14	25	19
12.	INT	No. of intersections	38	32	24	03	31	27
13.	CAP	Calculated capacity	83	-52	-09	-06	97	96
14.	A	Accidents	51	62	20	03	68	66
		Root	3.62	2.17	0.85	0.37		

[a] Loadings are decimals.
[b] Sum of the squared loadings in each row.

Table 4 shows the first results of the factor analysis. Quartimax rotations produced the final matrix presented in Table 1 in the text. Other rotational schemes would probably produce somewhat different results.

ANALYSIS OF VARIANCE

A two-way analysis of variance of the accident variable was done, with five categories each of ADT and total entrances. Total entrances was taken as the sum of the number of residential and commercial driveways and intersections in each 1-mi stretch of highway.

Both distributions were highly skewed, so it was not practical to use equal increments for the five successive categories of the independent variables. Both variables were, therefore, fractionated into quintiles, approximately, so that marginal totals of numbers of cases were similar. The points plotted in Figure 2 are at the medians of these quintiles. The ADT and traffic entrances were moderately correlated, so it was impossible to set up a 5 x 5 design with an equal number of cases in each of the 25 cells of the table. Because the cell frequencies varied widely, ranging from 11 to 138, an approximate analysis had to be done (10, p. 381). The procedure amounted to substituting for all the entries in each cell the average for that cell. Table 5 gives the cell frequencies and the category limits for the two independent variables.

The distribution of the accident score was skewed and the cell variances were very heterogeneous, rising with cell averages. Transformation of the scores to log (1+A) stabilized the variance considerably and allowed a within-group mean square to be calculated. Table 6 shows the results of the analysis of the original data and the transformed data.

TABLE 5

NUMBER OF 1-MI SEGMENTS IN EACH TRAFFIC GROUP AND INTERFERENCE CATEGORY

ADT (100's)	No. of Segments with Indicated No. of Driveways and Intersections per Mi					
	0	1	2-3	4-6	7+	Total
0 - 6	102	71	35	19	11	238
7 - 13	65	62	77	54	26	284
14 - 21	30	56	91	77	58	312
22 - 29	63	57	63	47	46	276
30 +	15	23	55	50	138	281
Total	275	269	321	247	279	1,391

TABLE 6

SUMMARY OF ANALYSIS OF VARIANCE OF ACCIDENT SCORE

Source	df	M.S.	M.S. (logs)
ADT	4	140.6	2.753
Entrances	4	19.2	0.287
Interaction	16	6.0	0.036
Remainder	1,366	—	0.012[a]

[a] All effects were highly significant.