STRUCTURAL MODEL FOR EVALUATING URBAN TRAVEL RELATIONSHIPS

Edward J. Kannel, University of Illinois at Urbana-Champaign; and
Kenneth W. Heathington, University of Tennessee

Urban travel forecasting equations are typically developed by analyzing only the relationships between several possible explanatory variables and the ultimate variable of interest, trip production. Seldom is the full degree of interaction among explanatory variables such as automobile ownership, household size and income, and accessibility understood. In this paper, a structural model is used to examine the relationships among an entire system of variables rather than just the simple isolated effects. The basic concepts and limitations of the modeling approach are discussed, and models of urban household trip production are evaluated. Several conclusions about the causal structure of urban travel relationships are drawn. The structural model is felt to be an important methodological tool for developing urban transportation theory.

•THE AVAILABILITY of multiple linear regression computer programs has made the linear regression model a popular technique for estimating travel demand in urban transportation planning studies. The user can handle large quantities of data and develop models without fully understanding the procedure or assumptions of the regression model. Evaluation of the model is based more on statistical goodness of fit than on an understanding of the causal structure that exists among the variables examined.

In a traditional trip generation model, evaluation of the relationships among a set of variables would be desirable. A typical trip generation model may use automobile ownership and household size to forecast home-based trip production. However, before auto ownership can be used, a forecast of this variable must be developed. Average family size and income might be used as the explanatory variable to estimate car ownership. These two models do not allow examination of the entire structure of the relationships that exist among family size and income, auto ownership, and trip production.

A complementary analysis technique is discussed that can be used to evaluate the direct effect of income on trip production as well as the indirect effect of income on car ownership rates. The approach discussed has alternatively been referred to as causal models, structural models, path models, and recursive models. These terms will be used interchangeably.

A structural model is a system of equations that allow the analyst to evaluate fully the interrelationships among a system of variables. To be sure, multiple regression analysis allows the analyst to observe the effects of several independent variables, either alone or in combinations, but this is possible only for the single response variable under immediate consideration. Although cross-product terms of selected independent variables may be included to show the effects of interaction among the independent variables, neither the nature of this interaction nor the relative contribution of the component parts can be evaluated. The structural model, on the other hand, uses a set of equations that outline the causal priorities of the variables and permit predictions of how a change in any variable in the system affects other intermediate variables in the system as well as the dependent variable of interest.

DEVELOPMENT AND APPLICATION OF STRUCTURAL MODEL

Pioneer work in the analysis of path coefficients, the parameters of a structural model, was done by Wright in the area of genetics (13). Blalock, drawing on the writings of Simon (10) and Wold (12), provided the major thrust for a study of causal inferences in nonexperimental research of the social sciences (1). Duncan (3), Land (8), and Heise (4) described their studies of causal models and outlined systematic approaches for interpretation of the model.

Analysis of structural models has also received attention in the development of transportation forecast models. Kain used a system of recursive equations to evaluate the interrelationships among variables that affect work trip length (5). He hypothesized a four-stage recursive model in which the decision process is such that the worker first selects an environment in which he wishes to live (space preference). This choice is influenced by factors such as sex, income, and housing prices. The first decision will then affect the choice of car ownership, which in turn affects the travel mode choice. Finally, all three of these affect the length of the work trip. In any of these equations additional variables may contribute to the variation in the dependent variables; however, these variables were taken to be exogenous to the system, and no attempt was made to define the interrelationships among exogenous variables.

More recently, de Neufville and Stafford summarized the difficulties in interpretation that arise when strictly correlational models are used rather than when the causal structure among a set of variables is examined (2). They also evaluated travel demand through structural model studies.

BASIC CONCEPTS OF CAUSAL MODELS

The goal of causal modeling is to develop a set of relationships that correspond to real-world causal processes. Analysis of a causal structure requires specification of a network of causal paths that exist between the variables of interest and identification of the parameters of causation so that the effects of each variable on the other variables in the model can be measured. The mathematical equations that make up the causal structure are a set of recursive equations. [For more detailed discussion of the computational procedures, readers are referred to Heise (4) and Land (8). Also, Blalock (1) offers extensive arguments on the need for caution in making causal inferences and clarifies the conditions under which causal inferences may be possible.] The following four-variable system is an example of a set of recursive equations (Fig. 1):

$$Z_1 = P_{1a}Z_a$$

$$Z_2 = P_{2b}Z_b$$

$$Z_3 = P_{31}Z_1 + P_{3c}Z_c$$

$$Z_4 = P_{41}Z_1 + P_{42}Z_2 + P_{43}Z_3 + P_{4d}Z_d$$

The structural model is composed of three types of variables. Exogenous variables are considered input to the system. It is assumed that these variables are completely determined by other variables outside of the system and that neither the nature of their origin nor the correlation that may exist between these inputs is of concern for the model being considered. Paths between the exogenous variables are represented in the figure as two-headed curvilinear arrows. These paths indicate only that some correlation exists. No direction of causality is assumed. In Figure 1, Z_1 and Z_2 are the exogenous variables of the system.

Measured variables within the structural model are endogenous variables. Unlike exogenous variables, the total variation in the endogenous variables is of interest. The total variation in the endogenous variables is assumed to be completely determined by some linear combination of exogenous variables, other endogenous variables, and some unmeasured residual or error variable. The postulated causal relations among the variables are shown in Figure 1 by unidirectional arrows extending from each determining variable to each variable dependent on it. In the models discussed, it is assumed that there is only one direction of causation; i.e., if X causes Y, Y cannot in turn be a cause of X. Variables Z_3 and Z_4 are endogenous.

Because it is unrealistic to assume that the variation of a system variable can be determined completely by other measured variables of the system, residual variables are introduced. It is assumed that residual variables are uncorrelated with the set of variables immediately determining the variable under consideration and that they have a mean value of zero.

Residual variables are represented on a path diagram by unidirectional arrows. The residual variable paths have alphabetic subscripts to distinguish them from the paths of the measured variables, which have numerical subscripts. For simplification, the residual paths and the paths between the exogenous variables are often eliminated from the causal model diagram. Variables Z_a, Z_b, Z_c, and Z_d are the residual variables in Figure 1.

The exogenous variables, Z_1 and Z_2, are assumed to be completely determined by outside forces Z_a and Z_b, which are either unknown or just not of interest in the analysis. The path coefficients P_{1a} and P_{2b} are equal to one and are not normally included in the model or diagram.

Estimation of the model parameters for each equation revolves around fitting a model to the data so that the amount of variation contributed by the residual variables is minimal. The parameters of the structural model are computed by using the least squares criteria common to linear regression analysis; however, the relative importance of the determining variables is interpreted by using standardized regression coefficients. This standardized parameter or path coefficient P_{ij} is a measure of the fraction of the standard deviation of the dependent variable for which the independent variable is directly responsible. More definitely, it is the fraction that is found if the factor varies to the same extent as in the observed data, all other variables being constant.

The relationship between the regular regression coefficient b_{ij} and the path coefficient P_{ij} is

$$P_{ij} = b_{ij} \frac{\sigma_j}{\sigma_i}$$

where σ_i and σ_j are the standard deviations of the dependent and independent variables. The path coefficients are also referred to as beta coefficients.

Use of the standardized parameter facilitates the computations necessary to evaluate how consistently the model reproduces the interrelationships that exist in empirical data. Using standardized coefficients, we can show that, for the model given in Figure 1, the structural model predicts the following linear correlation between variables Z_4 and Z_1:

$$r'_{41} = P_{41} + P_{42}r_{12} + P_{43}P_{31}$$

where

r'_{41} = predicted correlation for the hypothesized model;
r_{12} = observed correlation between the exogenous variables; and
P_{41}, P_{42}, and P_{31} = path coefficients estimated in regression analysis (8).

The total correlation between Z_4 and Z_1 is composed of three elements. First, there is a direct effect between the variables indicated by path coefficient P_{42}. Secondly, there is an indirect effect, $P_{43}P_{31}$, caused by the influence that variable Z_4 has on Z_3, which in turn influences Z_4. Finally, there is another indirect effect, $P_{42}r_{12}$, that encompasses the correlational effect of the exogenous variables. It must be cautioned that these direct and indirect effects can be interpreted only for the model under study. The direct effect is a true, isolated direct effect only if the other independent variables are orthogonal to the variable being considered and if the effects of all other variables are removed. In practice, these conditions are seldom completely met.

The entire correlation matrix could be reproduced in like manner by using the additional relations derived for the model in Figure 1:

$$r'_{13} = P_{31}$$

Figure 1. Example of four-variable causal model.

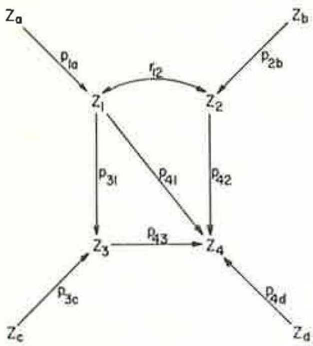


Figure 2. Causal ordering of household travel relationships based on the first model.

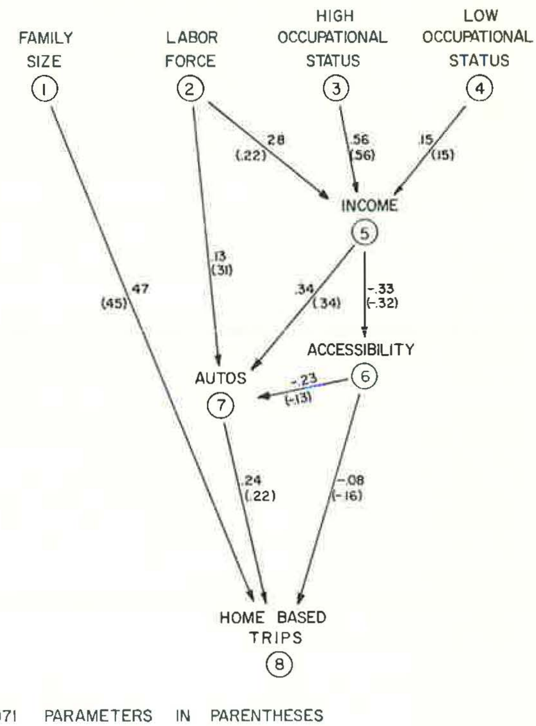


Table 1. Observed correlation matrix of causal model variables.

Variable	Family Size	Labor Force	High Occupation	Low Occupation	Income	Accessibility	Automobiles	Home-Based Trips
1964								
Family size	1.00	0.24	0.01	0.10	0.08	-0.07	0.16	0.52
Labor force		1.00	0.07	0.14	0.34	-0.03	0.26	0.31
High occupation			1.00	-0.83	0.46	-0.23	0.25	0.06
Low occupation				1.00	-0.28	0.15	-0.18	0.05
Income					1.00	-0.33	0.47	0.29
Accessibility						1.00	-0.35	-0.20
Automobiles							1.00	0.35
Home-based trips								1.00
1971								
Family size		0.41	0.12	0.20	0.15	-0.11	0.37	0.55
Labor force			0.16	0.32	0.36	-0.10	0.44	0.33
High occupation				-0.65	0.50	-0.27	0.25	0.24
Low occupation					-0.15	0.10	0.08	0.06
Income						-0.32	0.49	0.31
Accessibility							-0.27	-0.27
Automobiles								0.42
Home-based trips								

$$r'_{23} = P_{31}r_{12}$$

$$r'_{24} = P_{42} + P_{41}r_{12} + P_{43}P_{31}r_{12}$$

$$r'_{34} = P_{43} + P_{41}P_{31} + P_{42}P_{31}r_{12}$$

If a variable explains a significant portion of the variance in a dependent variable, it would exhibit a strong direct effect in the causal model and would logically be considered an important element of the model. However, if a particular variable does not show a strong direct contribution in a given equation, the analyst would not immediately reject the variable from the entire structure. Instead, the importance of the variable in the other equations of the system would be considered. If the variable is significant in other relationships it would be an important variable in the overall system. The advantage of the structural model is that all relationships can be examined simultaneously and the faithfulness of that system in reproducing the empirical relationships in the data set can be evaluated.

EVALUATING THE MODEL

The purpose of developing a causal model is to help the analyst understand the relationships among a set of variables that are important in some behavioral process. An objective in testing these relationships is to obtain a model that adequately reproduces the conditions that occur in empirical data and yet is as parsimonious as possible. For any postulated model, one can compute the correlation matrix and compare this with an observed correlation matrix. In the least parsimonious case in which all possible paths are included, there are no conditions imposed on the model to test the adequacy of that model. Any ordering of the variables results in a reproduced correlation matrix that exactly equals the empirical correlations (8). In this case only the knowledge of the causal priorities determines selection of one model over the other. The problem then is to make an initial determination of the significance of the paths in the model.

If sufficient data are available, analysis of variance models may serve as a starting point for evaluating the significance of variables that might be introduced in the model. The analysis of variance provides a measure of significance not only of the main effects of the variables but also of possible interactions. Because interaction terms are not included in the simplified model, there may be incorrect interpretations from the structural model.

A second method for eliminating paths in the causal models is to retain only those variables that are statistically significant according to the F-test criterion used in regression analysis. However, as the sample size becomes large, path coefficients that make very small contributions to the total variance may be judged statistically significant and retained in the model. For this situation, Land suggests that the analyst choose a minimum value below which a path coefficient is considered substantively insignificant.

Finally, when an over-identified model has been structured, (i.e., a model in which one or more possible paths have been eliminated) additional constraints will be established. A model is judged adequate if it reproduces correlations between the system variables in accordance with the imposed constraints. If these predicted correlations adequately represent the empirical correlations, the analyst might accept this as the best representation of the causal structure or he might check the possibility of eliminating other paths. If the model is not adequate, the analyst either reverts to the previously accepted model or tests some other model in which a different link is eliminated.

APPLICATION TO URBAN TRANSPORTATION PLANNING

The data analyzed in this paper were collected as part of a research project developed to study the temporal stability of household trip production (6, 7). The data were obtained from home interview surveys in Indianapolis. The first survey was conducted in 1964 as part of a traditional urban transportation planning study. The second survey was conducted in 1971. The same families were interviewed in both studies so that variation in household travel behavior, which may be due to family preferences, type of dwelling unit, or location within the urban structure, would be minimized. This sample of households was specifically selected to represent all levels of three principal socioeconomic variables, i.e., family size, automobile availability, and income; however, this simultaneously provided a wide range of other characteristics such as occupational status, educational levels, and location of residence within the urban area.

Although travel relationships at the household level of analysis do not have the apparent statistical strength of those obtained from zonal averages, it was hypothesized that relationships existing at the household level have greater behavioral validity and causal significance and therefore are more temporally stable than zonal model analyses. The stability of these household relationships over a 7-year period was examined in this study.

One of the causal model relationships of interest in this study was an evaluation of the hypothesis that the trip production from households is affected by the accessibility of the household to major activity centers within the urban area. The measures of accessibility used in the study were those developed by Nakkash for the 1964 highway network (9). Relative accessibility variables were determined from the friction factors of the calibrated gravity model. The relative accessibility variable was therefore a function of trip purpose and auto travel time, but was weighted by the amount of a given activity in a zone. Nakkash developed relative accessibility measures for several activity types such as employment, retail floor area, and school floor area, but these are all intercorrelated and only employment accessibility is used in the discussion. This variable tends to decrease with increasing travel time from the central city. Also it must be noted that accessibility measures comparable to the 1964 study could not be reproduced in 1971 because a complete new transportation study was not being conducted. The relative accessibility of each household in 1971 was therefore assumed to be equivalent to the 1964 calibrated values. All statistical tests were performed on the 1964 data.

STRUCTURE OF THE CAUSAL MODEL

The ordering of the causal network was based on a priori knowledge of the variables under consideration and on previous research models (5). Although several simplified models were tested to evaluate items such as the direct and indirect influence of income on trip production, only two models are discussed. The hypothesized formulation is a four-stage recursive model. The model hypothesizes that a family chooses a residential location based on a desire for a certain life-style, quality and style of housing, and preference or need for more or less space. Differences in preferred housing conditions may be shaped by factors such as individual attitudes, age, stage in the family life cycle, and family size. A family that needs more space would tend to locate in lower density areas that are relatively less accessible to major activity centers. The ability to satisfy the desire for housing type and space consumption, however, is controlled by the ability of the family to pay for the desired living style. Thus, income level of the family must be considered. Income might be determined by several factors such as education, occupation, age, and number of working members in the household.

Once the decision about housing requirements and residential location is made, the level of available transportation from that location influences the level of car ownership. Families living in high-density neighborhoods with greater accessibility may have the opportunity to satisfy some of their transportation needs by use of public transportation. Further, because of greater accessibility, more travel needs such as school or social-recreational trips are satisfied by the walking mode. The level of auto ownership may also be affected by family size, labor force, household income, and social status of the family.

Finally, the trip production rate of the household may be affected by any of the variables mentioned. The task is to evaluate and understand the degree of influence a change in one variable has on other variables in the model.

The causal ordering assumed in this study is shown in Figure 2. Only a small number of possible exogenous variables were actually considered in the model. Further, many of the possible linkages that could be included in the model have been eliminated because they were found to be of little importance as explanatory variables.

The correlation matrix for the variables is given in Table 1. The exogenous variables were family size, labor force, and occupation of the head of household. Occupation of the head of household was stratified into three groups and used as a dummy variable in the analysis. The groupings were nongainful, high status, and low status. The high status group was composed of professionals, managers, and salesmen; the low status group contained all other employed individuals. The nongainful dummy class was omitted from the analysis to allow solution of the least squares equations (11).

ANALYSIS OF THE MODEL

Examination of Figure 2 shows that many of the path coefficients are very stable for the 7-year period. The ability of the model to reproduce the correlations that exist among all of the variables can be evaluated by examining Table 2. If the model adequately represents all the existing relationships, these differences should approach zero. Although several of the possible links have been removed, the model does reproduce the correlation matrix quite well. The major discrepancy occurs in its ability to reproduce the relationship between family size and auto ownership in the 1971 data set. The calculated correlation was 0.20, whereas the observed correlation was 0.37. The changing relationship between family size and auto ownership may be due to the maturation process that has occurred. As the families moved through stages of the life cycle and children from the larger families moved out of the household, the relationship between family size and car ownership stabilized. As a result, the linear correlation between these variables increased substantially from 0.16 to 0.37. The same basic change was noted between labor force and auto ownership.

The model was examined for other possible links that might be removed to make the model more parsimonious. Earlier analysis suggested that the effect of accessibility on home-based travel is indirect because of its association with auto ownership (6). A two-way analysis of variance of the 1964 data set indicated that, when ownership and accessibility were tested, ownership was the significant variable and no interaction was found. The path coefficient in the model (-0.08) also indicates that the direct path is small and explains only a small portion of the variance in travel.

A second model, in which the direct link from accessibility to home-based travel has been removed, is shown in Figure 3. The differences in observed and empirical correlations are given in Table 3. The ability of the structural equations of the second model to reproduce the empirical correlation matrix is essentially the same as that of the first model. Only the relationship between trip production and accessibility is altered by removing this causal link. Because new measures of accessibility were not available in 1971, we could not determine whether the difference in the correlation in 1971 was a function of nonmeasurement error or actual changes in the effect of accessibility over time. Because the analysis of variance of the 1964 data found accessibility to be insignificant, the second model was chosen as the final structural model. This model was accepted as the most plausible explanation of the causal relationships among the variables.

SUMMARY

The causal relationships investigated in this study were restricted to models that meet the following assumptions.

1. A change in the dependent variable always occurs as a linear function of changes in the determining variables, and the effects of all other variables are assumed to be held constant.
2. The system contains no reciprocal causations, i.e., the model is strictly a recursive system. Reciprocal causations cause problems in identification. Although methods of treating such models are available, the procedure is more complex and the interpretations are clouded.
3. The causal priorities are sufficiently well known so that the structure of the

Table 2. Differences between empirical and reproduced correlations based on the first model.

Variable	Family Size	Labor Force	High Occupation	Low Occupation	Income	Accessibility	Automobiles	Home-Based Trips
1964								
Family size	—[a]	—[a]	—[a]	—[a]	0.00	-0.04	0.09	0.02
Labor force		—[a]	—[a]	—[a]	0.00	0.08	-0.02	0.12
High occupation			—[a]	—[a]	0.00	-0.08	0.05	-0.01
Low occupation				—[a]	0.00	0.05	-0.08	0.03
Income					—[a]	0.00	0.00	0.10
Accessibility						—[a]	0.01	-0.02
Automobiles							—[a]	0.04
Home-based trips								—[a]
1971								
Family size					-0.04	-0.05	0.17	0.04
Labor force					0.00	0.01	0.00	0.03
High occupation					0.00	-0.10	0.01	0.11
Low occupation					0.00	0.06	0.03	-0.03
Income						0.00	0.00	0.07
Accessibility							0.04	-0.02
Automobiles								0.08
Home-based trips								

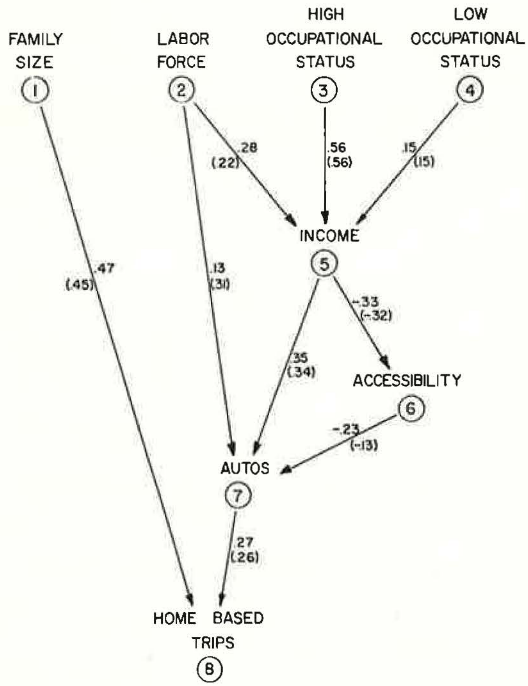
[a]These differences, by definition, must be zero.

Table 3. Differences between observed and reproduced correlations based on the second model.

Variable	Family Size	Labor Force	High Occupation	Low Occupation	Income	Accessibility	Automobiles	Home-Based Trips
1964								
Family size	—[a]	—[a]	—[a]	—[a]	0.00	-0.04	0.09	0.02
Labor force		—[a]	—[a]	—[a]	0.00	0.08	-0.02	0.12
High occupation			—[a]	—[a]	0.00	-0.08	0.05	0.00
Low occupation				—[a]	0.00	0.05	0.08	0.03
Income					—[a]	0.00	0.00	0.12
Accessibility						—[a]	0.01	-0.09
Automobiles							—[a]	0.04
Home-based trips								—[a]
1971								
Family size					-0.04	-0.05	0.17	0.04
Labor force					0.00	0.01	0.00	0.03
High occupation					0.00	-0.11	0.00	0.12
Low occupation					0.00	0.06	0.03	-0.04
Income						0.00	0.00	0.10
Accessibility							0.00	-0.17
Automobiles								0.08
Home-based trips								

[a]For the model specified, these differences must be zero.

Figure 3. Causal ordering of household travel relationships based on the second model.



1971 PARAMETERS IN PARENTHESES

model can be established as the correct ordering of the variables. It is not necessary that all correct paths are known, but the order of causation should be clear.

4. The data are generally measured on an interval or ratio scale, but, as in regular regression analysis, dummy variables may be used if caution is exercised in the interpretation of the results.

5. The usual assumptions of multivariate regression analyses are met.

The analysis model is a simplification of the real-world system and allows the analyst to evaluate the relative direct and indirect effects of the variables within the system. The model examined here is simplistic from the practical standpoint, for not all determinants of travel could be included. The model is simplistic in the statistical sense in that only linear relationships are considered and no interaction terms are specifically introduced for consideration.

Variables that have been found significant in household trip generation analyses were evaluated. The inferences obtained from the analysis indicate that auto ownership and family size have the most direct influence on trip generation rates. Income and level of accessibility to activity centers in the urban area also have an impact on travel, but this influence is indirect because of their influence on auto ownership. As household income increases the families tend to live farther from central employment concentrations and thus have lower relative accessibility to employment. In turn, these households exhibit higher auto ownership rates. The number of household members in the labor force also is a determinant of auto ownership rates, both directly and through its corresponding relationship with household income.

Finally, the effect of occupational status of the head of the household can best be understood by the extent to which it affects household income. Occupational status is an important variable for explaining variations in income, but the direct effects on any other variables in the causal model are negligible.

Although the analysis of causal inferences is subject to practical limitations, when properly used and interpreted it can provide a methodological tool for theory development.

ACKNOWLEDGMENTS

The authors are grateful to the Joint Highway Research Project of Purdue University, the Indiana State Highway Commission, and the Federal Highway Administration for the financial support that made this research possible. The opinions, findings, and conclusions expressed are those of the authors and not necessarily those of any of the supporting agencies.

REFERENCES

1. Blalock, H. M., Jr. Causal Inferences in Nonexperimental Research. University of North Carolina Press, 1964.
2. de Neufville, R., and Stafford, J. H. The Identification of Transportation Demand. In Urban Transportation Innovation (Brand, D., ed.), American Society of Civil Engineers, 1971.
3. Duncan, O. D. Path Analysis: Sociological Examples. American Journal of Sociology, Vol. 72, 1966.
4. Heise, D. R. Problems in Path Analysis and Causal Inferences. In Sociological Methodology (Borgatta, E. F., and Bohrnstedt, G. W., eds.), Jossey-Bass Inc., San Francisco, 1969.
5. Kain, F. J. A Multiple Equation Model of Household Locational and Trip Making Behavior. Rand Corporation, Santa Monica, Memo. 3086, 1962.
6. Kannel, E. J. A Disaggregate Analysis of Urban Travel Behavior. Purdue University, Joint Highway Research Project Rept. 17, July 1972.
7. Kannel, E. J., and Heathington, K. W. Temporal Stability of Trip Generation Relations. Highway Research Record 472, 1973.
8. Land, K. Problems in Path Analysis and Causal Inferences. In Sociological Methodology (Borgatta, E. F., and Bohrnstedt, G. W., eds.), Jossey-Bass Inc., San Francisco, 1969.
9. Nakkash, T. Z., and Grecco, W. L. Activity-Accessibility Models of Trip Generation. Highway Research Record 392, 1972.
10. Simon, H. Models of Man. John Wiley and Sons, 1957.
11. Suits, D. The Use of Dummy-Variables in Regression Equations. American Statistical Association Journal, Vol. 52, Dec. 1957.
12. Wold, H. J. L. Demand Analysis. John Wiley and Sons, 1953.
13. Wright, S. The Method of Path Coefficient. Annals of Mathematical Statistics, Vol. 5, Sept. 1934.