A MULTIMODAL LOGIT MODEL OF MODAL SPLIT FOR A SHORT ACCESS TRIP

Paul Inglis, De Leuw Cather and Company of Canada, Ltd.

The paper involves a discussion of research into the use of a logit mathematical formulation to model modal split. The model investigates the extension to the multimodal situation in terms of both user and system variables. Time and cost difference are the major variables introduced for the systems; income, age, and a rush-hour dummy are other variables entered. The trip modeled is a short commuter trip, which is a part of a longer overall trip and involves the access to a commuter rail station only. The line-haul portion is not considered. The mathematical formulation is probabilistic in nature and can be used with any number of choices of mode. Four choices were available for the research data. Two methods of aggregating, summing probabilities or taking absolute choices, were discussed and tested. The value of time can be developed by the use of time and cost coefficients derived for the model. A value was found that was similar to some of the previous values but of lesser magnitude than values found for longer trips. The model coefficients themselves were derived by a computer program that employed a maximum likelihood technique to iterate to significant values.

•STUDIES to improve the flow and direction of traffic volumes have been conducted for many years. As early as 1844, traffic counts were being made in France. Yet it was not until federal legislation in the United States in 1944 that transportation planning, in the form of origin-destination studies, evolved in a recognizable form. Only since 1955 has there been any advance beyond simple extrapolation of past trends. Modern analytic predictive planning, then, has been developing for only a relatively short span of less than 20 years. From the relevant technology, there has evolved a relatively standard format for predicting future flows. This has been called the "urban transportation planning" (UTP) package.

The package generally comprises 4 models: trip generation, trip distribution, modal split, and network assignment. Martin, Memmott, and Bone (15) and Davis (5) discuss the UTP package in full but cannot agree on a precise sequence for the 4 models. The first authors prescribe modal split to remove the transit riders before distribution and assignment, while the latter inserts modal split after distribution to try to achieve a total picture.

This paper concerns only one part of the UTP package: the choice of mode for a relatively short journey.

Since the late fifties, major transportation studies have been carried out in almost all major cities in North America. Each study was required to build its own models for future prediction. A definite advantage can be gained if some degree of standardization can be obtained. That has not previously been possible because of the nature of the explanatory variables employed in the models. Early work by Wynn (25), Carroll (4), and Adams (1) placed a large emphasis on urban land use and zoning. That was followed by Chicago Area Transportation Study reports by Howe (12), Biciunnas (3), and Sharkey (20, 21) and a Milwaukee study report by Hadden (9). They placed emphasis only on socioeconomic measures and activity levels derived from urban zonal theory.

Publication of this paper sponsored by Committee on Social, Economic and Environmental Factors of Transportation.

This meant that changes in the systems would not be reflected in the results of the models. A second shortcoming in the use of variables such as income and car ownership was the continual inflationary trend as the standard of living increased. Both of the above groups used linear techniques in the model formulas and aggregate groupings in zones as the basic unit for prediction.

The use of high levels of aggregation resulted in a very high variance within the zones, especially when short trips were being considered. To illustrate, a trip from zone A to adjoining zone B could vary from several blocks to more than a mile. Also, generalized zonal activity could override important small pockets of different types of land use.

By the late fifties, interest in system variables had begun to rise. Large studies in Washington, Chicago, San Francisco, Toronto, and Philadelphia resulted in the definition of a set of diversion curves for modal-split prediction. The models related either time differences or time ratios to the percentage of total trips diverted from one mode to the other. The techniques used are well documented by Quinby (18), Hamburg and Guinn (10), and Hill and Von Cube (11). Quinby recognized the mathematical inferiority of regression in this case and proposed that Pearl-Reed logistic curves be fitted before he finally settled on a Gompertz exponential curve formulation. The above works led to a large set of diversion curves illustrating the diversion to transit with a change in travel time ratios, for different cost ratios and income levels. They were developed by Traffic Research corporation and documented by Deen, Mertz, and Irwin (7). However, the portions of the curves of highest predictive value were also the areas of greatest uncertainty, for no corroborating observations were available for the predictive areas.

Errors in those earlier models were potentially very high. Much of the error in prediction was attributed to the level of aggregation at which the models were built. Working at the zonal level did not allow the large variance within the zonal populations to be accounted for. Reducing to the basic component, the individual user, overcomes that problem. The model can then be aggregated to any level desired with less chance of variance errors occurring.

Modal split provides a good point to start from in redeveloping the UTP package at the disaggregate level. Data are relatively easy to gather, the result can be easily measured by survey, and the variables of influence can be defined.

Once the modal-split model has been completed, it is anticipated that the use of a similar technique and mix of user and system characteristics will allow the whole package to be integrated. Several authors who have done work with stochastic models of modal split include Warner (25), Quarmby (17), and Lisco (14). Currently 3 separate classes of models exist, depending on the type of statistical technique: discriminant analysis, probit analysis, and logit analysis. The use of linear regression has been largely superseded because of the possibility of predicting negative probabilities and values greater than one.

Discriminant analysis (8) is based on the existence of overlapping normal subpopulations that are distinct in the decision sense. By identifying attributes that can account for the difference in choice, a function that discriminates among the populations can be developed. Models by Quarmby (17) and McGillivray (16) employed user characteristics as well as system characteristics, but the set of variables and their form are still a subject for much debate and research. Probit analysis was first suggested for modal split by Warner (25), who rejected it as computationally too complex. Lisco (14) was the first to use this method successfully for his economic studies on the value of time, a result derived from the cost and time coefficients of his modal-choice model. Lave (13) has since built other modal-split models by using this mathematical form. It is a technique that requires a normal distribution of threshold values, which may not always be a valid assumption.

Logit analysis was developed by Thiel (24) in the multinomial sense. Stopher (23) limited his use of this technique to binary models in his research at Northwestern University. Rassam, Ellis, and Bennett then developed a multimodal aggregate model of modal split for access to a Washington airport.

This paper attemps to document the next step: the development of a multimodal, stochastic, disaggregate model of modal split for a short journey.

THE MODEL

In the past, models of modal split have usually been restricted to 2 dimensions. When more than a binary choice exists, a set of binary models is required to produce the final result. The design of this set of models requires that some step-by-step choice process be presumed among the modes. If this were not the case, a great number of models would result, for each mode would be compared individually to each of the alternate modes. It is then desirable to be able to compare all modes at once, without having to make an a priori decision on which modes are to be directly compared. The substitution of an n-dimensional model to replace the binary system should be a progressive step. A brief outline is given below of the form of the logit model.

The choice of a modal-split model is based on the premise that the probability of using a particular mode is a continual function whose dependent variable p ranges from 0 to 1 according to some function of the sociological traits of the user and the characteristics of the modes involved. Thus, as any of these variables change, so does the value of the function and hence the probability. The use of a simple linear relation is rejected because of the bounds imposed by the 0-to-1 range. The function should be asymptotic to both of those limits. This can be done using a logit formulation. The binary logit relation may be generalized as

$$p = \frac{e^{G(x)}}{1 + e^{G(x)}}$$

$$(q = 1 - p)$$

$$q = \frac{1}{1 + e^{G(x)}}$$

where G(x) represents a function to describe the response relation to a particular mode. It may be formulated as

$$G(x) = \text{constant} + \sum_{k=1}^{N1+N2} a_k x_k$$

where N1 is the number of system-dependent variables (such as time or cost) and N2 is the number of system-independent variables (such as age or income).

A simple mathematical manipulation in terms of variable differences illustrates the symmetry of the formula. The system-dependent variables are used in terms of differences to reflect the advantage of one over the other while only a single function is dealt with. That symmetry may be extended to the multidimensional form by proposing

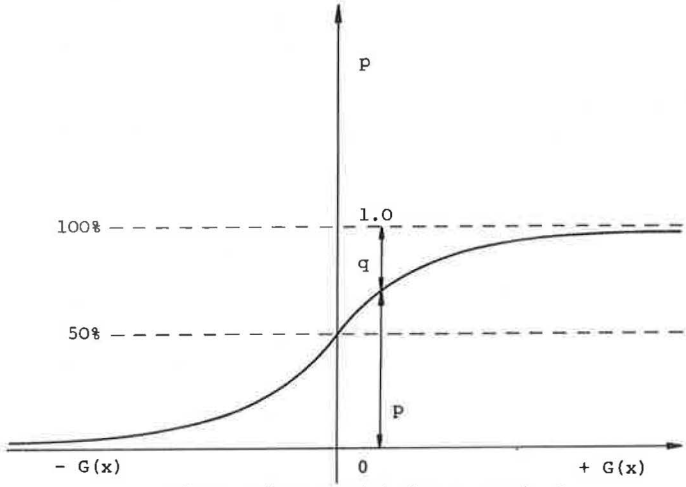
$$p_1 = \frac{e^{G1(x)}}{\sum_{m=1}^{M} e^{Gm(x)}}$$

A pictorial representation is given in Figure 1 of a binary-choice situation.

The major problem, given the relation above, is the estimation of the coefficients in each equation. That was done by the use of a maximum likelihood estimator procedure.

The estimation technique requires that a base mode be established. The generalized formulation above does not lend itself to the maximum likelihood estimator technique. For that reason, the model is developed in terms of (n - 1) different modes, and the last mode is determined by the limitation that the sum of the probabilities is one. To involve the system variables of the last (or base) mode requires that these particular variables be somehow related to the base mode. That can be done by using differences or by using ratios. Both techniques are valid; however, for this model, the former was

Figure 1. Typical 2-dimensional logit curve.



NOTE: Given a G(x) for a particular observation, p and q are the resulting probabilities.

chosen. To have a model that can be realistically analyzed and that conceptually follows the formulation stated above required that the coefficient of the system variable be kept constant for each mode. Also each system variable must be related to each mode. For example, a comfort rating could not be entered into the bus mode if it were not entered in the car mode or in every other mode. On the other hand, the user characteristics need not be entered in each mode equation and should not have like coefficients unless there is a like correlation.

Several advantages are attributable to this technique, particularly in terms of theoretical assumptions. There is no assumption of normality to be met in the G(x) function, resulting in a more generally applicable model. Also the use of a probabilistic sum for aggregation gives a better conceptual idea of the true process that occurs in this mechanism.

To make this model operational requires some aggregation because collection of the information for each trip would be prohibitive. The methodology for this aggregation is still a question for review and testing. There are 2 distinct possibilities. First, if our sample is 10 percent, then each datum is a proxy for 10 other members of the population. It may be assumed that all 10 will make the same absolute choice as the proxy, i.e., the mode of highest probability according to the model. Thus, to aggregate, one multiplies the result by the sample ratio. Second, again if the sample is the same, the probabilities for the individual mode choices and not the ultimate choice are considered. To aggregate, one obtains the sums of the probabilities and multiplies those sums by 10. This better describes the behavioral process because we are dealing with human beings who can and will change their habits in this nonexact manner. The scope of this work included only a minor attempt to determine the superior methodology. That is a matter for further study.

DATA BASE AND VARIABLES

The data used to derive and test the models have been titled the "suburban station access" data. They were collected from people using the Chicago and Northwestern Railroad suburban routes from the northwest corridor into Chicago. The trips that were modeled were not trips to the center of the city but rather shorter access trips between home and the commuter station. The 4 modes involved in the study were walk, drive and park, driven, and bus. A final set of 117 observations was used to build the models with a good balance of each modal user. A second set of 400 observations was used to test the models. The second set had a different modal user mix and slightly different geographical area. Those 2 factors would help in providing a good test of the technique.

Probably the most important task for the model builder is the choice of variables to be used in the models. Below is a brief discussion of the variables that were available from the given data.

Cost

Because cost is the measure of almost all other goods and services, it should be significant in the decision to use a mode of travel. However, true costs are seldom evaluated by the user; rather he sees only direct costs such as parking, gas, and tolls. His decision, then, is based only on this perceived cost. Betak (24) feels that this variable could prove much more important if the costs of the different modes were economically substitutable; however, the investment in a car, for example, is basic to a family and not related directly to a particular trip. That is a system variable and as such is entered as a difference between the mode and a base mode. Because the walk mode has no monetary cost associated with it, it is used as the base mode.

Time

Time is another important system characteristic. Attitudes toward time vary according to the activity during a given period. Therefore, time was entered for waiting, walking, parking, and access time separately; but, except for walking time, no advantage was obtained over total travel time. Overall time, then, is entered as a difference variable in the model.

Time of Day

A rush-hour dummy variable was entered to try to ascertain any difference in attitude between the morning rush period and the rest of the day. The dummy has a value of 1 if the trip is taken in the rush period and 0 otherwise. It is entered linearly as were the previous variables. Thus, a separate factor is entered for the rush-hour period.

Age

Age was entered in 2 forms. Age was divided into 4 groups—less than 25, 26 to 45, 46 to 65, and older than 65—in an attempt to eliminate the linear effect of entering age directly into the model. This was done by using 3 dummy variables each taking either 1 or 0 value with a maximum of 1 variable having the value 1. It was hoped that this stratification would delineate different attitudes toward the separate modes at different age levels. Age was also tried as a linear variable.

Income

Income was treated in the same way as age. A linear correlation should not be expected between a variable such as this and choice, so that a set of 5 dummy variables was put forward. The groups were less than \$5 thousand, \$5 to \$8 thousand, \$8 to \$12 thousand, \$12 to \$17 thousand, \$17 to \$25 thousand, and greater than \$25 thousand per year. Some researchers have used income as a combining variable, especially with items of cost (for example, de Donnea, 25). Time limitations prevented that idea from being tested with respect to the multimode case.

Car Ownership

Car ownership was entered when an automobile was involved in the mode, i.e., for driving or riding. It was added linearly, but an argument could be made for making it a preemptive variable for the drive mode. That is, if the user does not own a car, he cannot possibly drive to the station.

Sex

The coefficient of the sex variable reflects the attitude of the female relative to the male toward the mode involved, for it is in the form of 0 or 1.

Stage in the Family Life Cycle

There was a high expectation for this set of variables. The set comprised 3 dummy variables that categorized the population sociologically according to 4 different classes. The first variable takes the value of 1 if the user is unmarried and living at home. The second has a value of 1 if he is unmarried and independent or if he is married with a spouse who does not compete for the use of the car. The third has a value of 1 if the user is married and has a spouse who goes to work separately. All others are together in the fourth group by reason that they will respond 0 to all of the above variables. The variables were entered to try to model the user characteristics and their relative traveling-mode priorities. It was felt that there might be some interrelation between this variable and the ownership variable.

Trip Purpose

The trip purpose variable could be used to describe the different economic demand generated by the separate trip purposes. That would involve stratifying and thus complicating the model, making it computationally impossible for the capacity of the program used unless separate models are built for each purpose. Thus, a constant demand function was assumed with respect to trip purpose, and the variable was not included.

THE RESULTS

Below is an outline of the most significant model obtained in runs of the maximum likelihood program to develop the logit model coefficients. Variables were eliminated if the t-value (in brackets below) obtained was not significant at the 0.90 level, and the program was rerun. Some of the variables (stage in the family life cycle and age) mentioned in the previous sector failed to be significant at all as a result.

For walk mode,

$$G_1(x) = 0 \text{ (base mode)}$$

For drive mode,

$$G_2(x) = \underset{(5.03,\ 0.9995)}{-0.114\ \Delta C} - \underset{(5.14,\ 0.9995)}{0.00421\ \Delta t} + \underset{(3.29,\ 0.995)}{0.238\ \text{FRICT}} - \underset{(214,\ 0.975)}{0.0123\ \text{WALK PLS}} + \underset{(3.32,\ 0.995)}{1.49\ \text{CAR}}$$

$$+ \underset{(1.57,\ 0.900)}{0.0108\ \text{AGE}} - \underset{(2.22,\ 0.975)}{5.57\ \text{ID1}} - \underset{(3.28,\ 0.99)}{7.42\ \text{ID2}} - \underset{(3.80,\ 0.999)}{7.97\ \text{ID3}} - \underset{(3.53,\ 0.995)}{7.09\ \text{ID4}}$$

For driven mode,

$$G_3(x) = 0.114\ \Delta C - 0.00421\ \Delta t - \underset{(3.45,\ 0.995)}{2.26\ \text{RUSH}} + \underset{(1.97,\ 0.950)}{1.169\ \text{ID4}}$$

For bus mode,

$$G_4(x) = -0.114\ \Delta C - 0.00421\ \Delta t - \underset{(1.85,\ 0.950)}{1.34\ \text{RUSH}} + 2.025 \text{ (a constant)}$$

Likelihood ratio test, 120.06 with 15 deg of freedom; proportionate pseudo R-squared, 0.686

In the statements given above, ΔC is cost difference; Δt is time difference; ID1, ID2, ID3, ID4 are income dummy variables; FRICT is $\frac{1}{2}$ cost of parking + walking from parking at 6 cents/min; WALK PLS is the walking time in sec; CAR is car ownership; AGE is age; and RUSH is time of day.

The likelihood ratio test is highly significant even as high as 0.999, indicating that the model is valid. The proportionate pseudo R-squared statistic is only an approximation and not a true R-squared value. The latter value was used for comparison of models within the research program only and should be taken only as a rough guide and not as absolute.

A multiple F-value of 6.954 was obtained in a secondary evaluation of the model (on the larger data set). (A value of 3.38 is significant at the 0.999 level.)

From this set, the mean true proportions and those predicted by the model when the predicted probabilities were summed and when the exact count was made from the individual decisions were as follows:

Mode	Mean True	Summed Probabilities	Exact Count
Walk	0.135	0.112	0.078
Drive	0.293	0.217	0.208
Driven	0.469	0.398	0.453
Bus	0.103	0.273	0.191

This indicates that 17 percent were misplaced for the probabilities sum and 8 percent were misplaced for the absolute count. Unfortunately, for both cases, all of those misplaced were put on the buses. That indicates that $G_4(x)$ is overestimating and may be attributable in part to the constant being placed in that modal sector. Throughout the research, the placing of the constant had a definite effect on the model. Further research is required to investigate the sensitivity of the constant.

A secondary result of the model as constituted is the derivation of a value of time. In this case, it is a value of time saved. From the above coefficients of time difference and cost difference, we calculate the value $1.33/hour. Based on a 2,000-hour work year and the average yearly wage of the data set ($11,000), this value is 24.1 percent of the wage rate. That is only about half the value derived by Lisco (14) in his probit analysis. Stopher (22) found the value to vary from 0.33 to 0.14 depending on the salary. The above value falls in the middle of that range.

CONCLUSIONS

The most important conclusion that can be made is that it is possible to build a significant modal-split model by using this technique. Some further work should be done with respect to the constant.

The value of time for the short journey seems to be less than similar values derived by Lisco for longer trips in the same geographical area.

Many of the more definite conclusions concern the choice of variables for the models. The most disappointing variable was stage in the family life cycle, which failed to be significant in any of the sectors. Perhaps it would reflect a large importance when related to the longer overall trip.

Once again, time and cost proved to be significant as explanatory variables. That was expected. What is slightly surprising is that cost attained the same level of significance as time. It was expected from previous studies that time would be the more important variable. The difference in significance between the two, however, was only minimal. It can, therefore, be concluded that time and cost are both highly important in consideration of this modal choice.

One convenient advantage of this method is the unavailable mode situation. If, for example, the bus were not available to a user, it could be assigned an arbitrarily large time difference so that its probability were reduced to a very small value approaching zero.

It is not possible to claim that this is an operational model. The variable set should be refined, the problems with a constant should be studied, and a definite statistical advantage should be established over other modeling techniques. If the evening peak is used, there may be a whole different set of variables, for the results of the morning peak put definite limits on the evening return trip. Given that the problems can be resolved, the next step is an extension to include the other 3 steps in the urban transportation planning package.

ACKNOWLEDGMENT

This paper is the result of research conducted in the Department of Civil Engineering and Engineering Mechanics at McMaster University, Hamilton, Ontario, under the supervision of Peter Stopher. The data base used in the research was collected in suburban north Chicago by the Chicago Area Transportation Study.

REFERENCES

1. Adams, W. T. Factors Influencing Transit and Automobile Use in Urban Areas. Public Roads, Vol. 30, 1959.
2. Betak, J., and Betak, C. Urban Modal Choice: A Critical Review of the Role of Behavioural Decision Variables. Chicago, 1969.
3. Biciunnas, A. E. Modal Split of First Work Trip. CATS Res. News, Chicago Area Transportation Study, Vol. 6, No. 1, 1964.
4. Carroll, J. D. Relation Between Land Use and Traffic Generation. BPR Res. Digest, No. 1, Feb. 1954.
5. Davis, H. E. Urban Transportation Planning: Introduction to Methodology. In Defining Transportation Requirements, ASME, 1969.
6. de Donneau, F. Probit and Logit Modal Split Models for Rotterdam. Netherlands Economic Institute, Rotterdam, Dec. 1970.
7. Deen, T., Mertz, W., and Irwin, N. Application of a Modal Split Model to Travel Estimates for the Washington Area. Highway Research Record 38, 1963, pp. 97-123.
8. Fisher, R. A. The Use of Multiple Measurements in Taxonomic Problems. Ann. Eugen. Lond., Vol. 7, 1936.
9. Hadden, J. K. The Use of Public Transportation in Milwaukee. Traffic Quarterly, Vol. 18, 1962.
10. Hamburg, J. R., and Guinn, C. R. A Modal Split Model: Description of Basic Concepts. New York State Dept. of Public Works, Albany, 1966.
11. Hill, D. M., and Von Cube, H. G. Development of a Model for Forecasting Travel Mode Choice in Urban Areas. Highway Research Record 38, 1963, pp. 78-96.
12. Howe, J. J. Modal Split on CBD Trips. CATS Res. News, Chicago Area Transportation Study, Vol. 2, No. 12, 1958.
13. Lave, C. A. Modal Choice in Urban Transportation: A Behavioral Approach. Stanford University, PhD dissertation, 1968.
14. Lisco, T. E. Value of Commuters' Travel Time: A Study in Urban Transportation. Univ. of Chicago, PhD dissertation, 1967.
15. Martin, B. V., Memmott, F. W., and Bone, A. J. Principles and Techniques of Predicting Future Demand for Urban Area Transportation. M.I.T., Cambridge, Res. Rept. 38, June 1961.
16. McGillivray, R. G. Binary Choice of Transport Mode in the San Francisco Bay Area. Univ. of California, Berkeley, PhD thesis, 1969
17. Quarmby, D. A. Choice of Travel Mode for the Journey to Work: Some Findings. Jour. of Transp. Econ. and Policy, Vol. 1, No. 3, 1967, pp. 273-314.
18. Quinby, H. D. Traffic Distribution Forecasts: Highway and Transit. Traffic Eng., Vol. 31, No. 5, 1961.
19. Rassam, P. R., Ellis, R. H., and Bennett, J. C. The N-Dimensional Logit Model: Development and Application. Peat, Marwick, Mitchell and Co., Washington, D.C., unpublished, 1970.
20. Sharkey, R. H. A Comparison of Modal Stratification of Trips by Distance From the CBD. CATS Res. News, Chicago Area Transportation Study, Vol. 2, No. 5, 1958.

21. Sharkey, R. H. The Effect of Land Use and Other Variables on Mass Transit Usage in the CATS Area. CATS Res. News, Chicago Area Transportation Study, Vol. 3, No. 1, 1959.
22. Stopher, P. R. Report on the Journey to Work Survey. Res. memo, London, Jan. 1968.
23. Stopher, P. R. A Probability Model of Travel Mode Choice for the Work Journey. Highway Research Record 283, 1969, pp. 57-65.
24. Thiel, H. A Multinomial Extension of the Linear Logit Model. Internat. Econ. Rev., 1969.
25. Warner, S. L. Stochastic Choice of Mode in Urban Travel: A Study in Binary Choice. Northwestern Univ. Press, Evanston, Ill., 1962.
26. Wynn, F. H. Land Use Relationships With Travel Patterns. BPR Res. Digest, Nov. 1955.