PREDICTION WITH DISAGGREGATE MODELS: THE AGGREGATION ISSUE

Frank S. Koppelman, Massachusetts Institute of Technology

This paper describes the problem of aggregation in forecasts of travel behavior under conditions in which aggregate behavior is the accumulation of travel choice decisions by individuals or households. Failure to deal with this problem, which is explicit in the use of disaggregate models and implicit in the use of aggregate models, leads to predictions that have biases related to the heterogeneity of the group for which the prediction is made. Alternative approaches to the development of unbiased aggregated forecasts based on disaggregate choice models are described. The importance of forecasting the distribution of characteristics that influence individual or household choice is cited. The advantages of an explicit aggregation procedure are identified with respect to sensitivity to changes in the distribution of choice-influencing characteristics and to improvement in the sensitivity to changes in the average values of these characteristics. Directions for future research to overcome the aggregation problem are identified.

•DISAGGREGATE choice models have had rapid development in recent years. The improved understanding they have provided of the decision process influencing individual behavior has contributed to the refinement and modification of theories of travel behavior. More recently, attention has been directed toward the use of disaggregate models for the prediction of aggregate travel behavior. This approach to obtaining aggregate predictions is based on the principle that the travel behavior of large groups is the manifestation of the travel choice decisions of numerous individuals or households. The problem associated with aggregate predictions based on disaggregate models is the development of a procedure for expanding individual choice estimates over the population of interest to obtain a reliable, unbiased description of group behavior.

The construction of an aggregate forecasting model based on a disaggregate model depends on both the form of the disaggregate model and the shape of the multivariate distribution of characteristics that influence travel choice. If the underlying disaggregate model is linear over the range of interest, the aggregate forecasting model will have the same linear specification; averages of the variables will be substituted for the individual values. However, if the disaggregate model is nonlinear, the disaggregate functional specification, in which averages of the independent variables are substituted for individual values, will give a biased forecast of the average of the dependent variable, except in the special case where the population is homogeneous with respect to those characteristics that influence the choice under study. This is shown with an example in the following section.

In principle, the transformation of a disaggregate model into an aggregate forecasting model can be accomplished by integrating the relation over the distribution of the choice-influencing characteristics. In general, the explicitly aggregated forecast model will contain parameters of the relevant distributions as well as parameters of the choice process. Such models will therefore be adaptable to forecasting under conditions where different distributions prevail or where the distribution structure is expected to change over time.

On the other hand, a model that is calibrated with aggregate data and that does not explicitly take account of the distribution of choice-influencing characteristics will have biased coefficients and will be valid for forecasting only if the distribution of characteristics for the forecast situation is reasonably similar to the distributions in the groups on which the model was originally calibrated.

The transformation of a disaggregate choice model into an aggregate model by mathematical integration may be an intractable problem, depending on the form of the disaggregate model and the shape of the relevant distributions. However, to obtain approximate aggregate forecasts by use of numerical integration or grouping techniques is always possible. Any transformation method will require the forecast of the distribution of relevant characteristics in addition to representative values. Even if no forecast distributions are available, judgment may be used in a Bayesian sense to suggest modifications to existing distributions.

AGGREGATION PROBLEM

Consider a disaggregate model describing the probability of a decision-making unit, either individual or household, choosing an alternative from a set of possible alternatives (such as one of several modes to work or one of several destinations for a weekly shopping trip). The general form of this model is

$$P_t(i:A) = f(U_{jt}, \text{ all } j \text{ in } A) \tag{1}$$

where

$P_t(i:A)$ = probability of decision unit t choosing alternative i from the set of alternatives A,
$f(\ )$ = function of the enclosed arguments, and
U_{jt} = utility of alternative j to individual t.

For the purpose of this discussion we will assume that the utility of each alternative for individual t is a linear function of the attributes of that alternative. (We will refer to the decision unit as an individual henceforth. However, the discussion applies equally to any behavioral unit. The linear assumption does not place a significant constraint, for nonlinear relations may be expressed by defining attributes in terms of logarithmic, exponential, or power functions, and interaction of variables may be represented by creating variables that are functions of groups of attributes.) That is,

$$U_{jt} = \sum a_m X_m^{jt} \tag{2}$$

where

X_m = value of attribute m of alternative j for individual t, and
a_m = parameter that describes the influence of the associated variable on the utility value. (The assumption that parameters a_m are identical for each household will be used throughout. Differences in parameters representing differences in behavior may occur for different market segments. Aggregation over different market segments is discussed in a later section.)

In the special case where the choice model applies to a binary (2-choice) situation and the function of utilities is the difference between the utilities, that is,

$$\begin{aligned} P_t(i:A) &= f(U_{it},\ U_{jt}) \\ &= U_{it} - U_{jt} \\ &= \sum_m a_m(X_m^{it} - X_m^{jt}) \end{aligned} \tag{3}$$

it can be shown that the expected proportion of individuals who will choose alternative i is equal to the probability of choosing i for an individual who faces the average of the attributes of each alternative. That is,

$$\begin{aligned}\bar{P}(i{:}A) &= P_{\bar{t}}(i{:}A)\\ &= \bar{U}_i - \bar{U}_j\\ &= \sum a_m(\bar{X}^i_m - \bar{X}^j_m)\end{aligned} \tag{4}$$

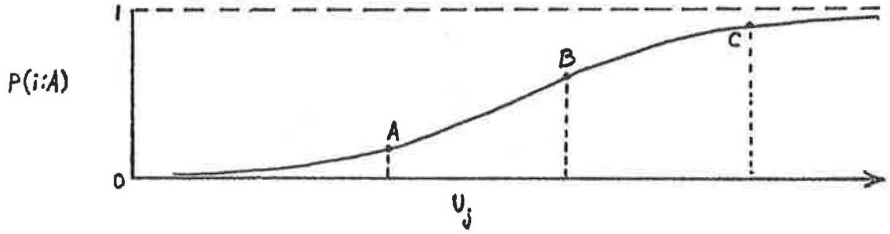
where

$\bar{P}(i{:}A)$ = expected proportion of people choosing alternative i,
$P_{\bar{t}}(i{:}A)$ = probability of choosing i for an individual facing average attributes for all alternatives,
$\bar{U}_j$ = average utility of alternative i, and
$\bar{X}^i_m$ = average value of attribute m of alternative i.

The aggregate model in Eq. 4 is identical to the disaggregate model of Eq. 3; average values of all the attributes are entered in place of the individual values. In this case, the influence of any attribute on the aggregate proportion choosing an alternative can be fully represented by the average value of the attribute to the group under study. The aggregate relation, Eq. 4, would give unbiased predictions for expected choice proportions. Unfortunately, the structural requirements of disaggregate choice models (probability of any choice must be within the 0 to 1 range) requires that the choice model be a nonlinear function of the relevant utilities. In this case, it can be shown that the corresponding nonlinear aggregate function with average values used to replace individual values will give biased results unless the individuals in the group are homogeneous with respect to all of the characteristics that influence the choice (binary or multiple) under study (1). That is, the average of the function (the average probability) is not equal to the function of the averages (the probability for an individual facing average attribute values). For example, consider the logit formulation of the binary choice model,

$$P_t(i{:}A) = \frac{e^{U_{it}}}{e^{U_{it}} + e^{U_{jt}}} \tag{5}$$

which can be represented as a function of $(U_{it} - U_{jt})$ by the following diagram:



(The binary choice logit model is used for ease of discussion. Essentially identical results may be obtained for the multinomial logit model.) The probability associated with any value of $U_{it} - U_{jt}$ for a single individual may be read directly from the graph, and the influence of a small change in U_{it} or U_{jt} is a function of the slope of the curve at the point of interest. If we consider a population with average utilities $\bar{U}_i$ and $\bar{U}_j$ represented by point B and assume that all $U_{it} = \bar{U}_i$ and all $U_{jt} = \bar{U}_j$, the average probability of choosing i and the sensitivity of that probability to changes in the difference between the utility functions will be identical to that for one individual represented by B. However, if the true population consists of subgroups represented by points A and C, both the estimated average probability and the sensitivity to changes in the attributes of an alternative will be biased. This analysis can be extended to multiple subgroups or continuous distributions of group members with similar results.

AGGREGATE PROBABILITY ESTIMATES

The best estimate of the proportion of the population that will choose alternative i from set A is simply

$$\bar{P}(i:A) = \frac{1}{N} \sum_{t \in N} P_t(i:A) \tag{6}$$

which is the average of the expected response probability of each individual in the population. (Equations 6 and 7 apply equally to binary- and multiple-choice situations.) Similarly, the expected change in the proportion choosing i due to a change in the value of one of the attributes of any of the alternatives in set A, say j, will be

$$\Delta\bar{P}(i:A) = \frac{1}{N} \sum_{t \in N} \frac{\partial P_t(i:A)\, \Delta X_n^{jt}}{\partial X_n^{jt}} \tag{7}$$

But the change in the selected attribute, ΔX_n^{jt}, and the responsiveness to change, $[\partial P_t(i:A)]/\partial X_n^{jt}$, may be different for different individuals. Since the responsiveness to change depends on the probability prior to the change, solution of the estimation problem for perfect prediction requires knowledge of the distribution of the choice probabilities, or all the attribute values from which the choice probabilities are determined, and of changes in X_n^{jt}. Obviously it will not be feasible to predict these values for each individual and to explicitly aggregate the results as implied by Eqs. 6 and 7.

The condition for consistent aggregation with nonlinear functions, homogeneity of individuals in the group, suggests that one method of approximating this representation is to group individuals in categories such that the assumption of a representative value of individual utility is an acceptable approximation for all individuals in the group. In this case,

$$\bar{P}(i:A) = \frac{1}{N} \sum_{T=1}^{NG} N_T P_T(i:A) \tag{8}$$

and

$$\Delta\bar{P}(i:A) = \frac{1}{N} \sum_{T=1}^{NG} N_T \frac{\partial P_T(i:A)}{\partial X_n^{jT}} \Delta X_n^{jT} \tag{9}$$

where

N = total number of individuals;
NG = number of groups;
N_T = number of individuals in group T;
$P_T(i:A)$ = probability function for the representative individual in group T;
$[\partial P_T(i:A)]/\partial X_n^{jT}$ = derivative of the probability response function with respect to a change in any attribute, X_n^{jT}, for the representative member of group T; and
ΔX_n^{jT} = representative change in attribute X_n^{jT} for group T.

In cases where the attribute change is not uniform, the selection of groups should provide a reasonable degree of homogeneity of this change within groups as well as for the attributes influencing individual choice probabilities. A variety of methods for grouping households can be suggested. Generally it will be most understandable to group them according to the variables that are relevant to the choice under study. The degree of stratification for each of the variables will depend on the range of the variable and the size of the parameter associated with it. In cases where the distribution can be described by continuous functions, the grouped summation may be replaced by a numerical or mathematical procedure (2, 3).

DISTRIBUTION OF INDIVIDUAL CHARACTERISTICS

The models of aggregate behavior represented by Eqs. 8 and 9—or corresponding methods based on numerical or mathematical integration—make explicit the dependence of the estimate on

1. The individual's response to the characteristics he or she faces (the individual choice function), and
2. The distribution of individuals according to their characteristics and the characteristics they face.

This explicit representation is the basis of 2 major advantages of explicitly aggregated predictive models based on disaggregate analysis over aggregate forecasting models based on correlative analysis of aggregate data.

1. Improved sensitivity to changes in individual behavior due to changes in environmental characteristics including policy controlled variables, and
2. Sensitivity to changes in the distribution of the characteristics that the population has or faces.

The structure of the aggregate models represented by Eqs. 8 and 9 requires a complementary population distribution model to be employed in conjunction with the disaggregate choice mode. Although this adds a potentially complicating dimension to the application of disaggregate models, it should be possible to develop simple models of distribution based on assumptions that are at least as good as the implicit assumptions embodied in models based on aggregate travel data. In addition, the possibility of developing improved representation of population distributions can be explored.

NEED TO FORECAST CHARACTERISTIC DISTRIBUTIONS

Each of the possible methods for obtaining unbiased aggregate forecasts of travel behavior requires explicit representation of the distribution of the characteristics of the population and the travel choices they face. This requirement places increased demands on the forecast of explanatory characteristics. Obviously it will be difficult to develop models capable of forecasting joint distributions of a wide variety of population and travel choice characteristics. The critical issue for modeling strategy is to improve the quality of distributional forecasts to a level that is compatible with the quality of other elements in the overall forecast process. Decisions must be made as to the methods of representation of the required distributions including assignment to groups versus continuous representation. In addition, those distributions that are to be represented with the greatest level of detail and accuracy must be identified. Primary attention should be directed toward improving the quality of forecasts for those distributions to which the required aggregate forecasts will be most sensitive.

The criteria for these decisions must be related to the objectives of the analyses to be performed, but would presumably include evaluation of the expected bias and standard error of the aggregate forecast. The levels of satisfaction of these criteria will be interrelated. For example, increasing the number of dimensions along which the population is stratified will tend to reduce the aggregation bias but may also increase the standard error of the aggregate forecast.

Decisions on the distribution forecast procedure to be used will depend on the particular situation under study, the type and quality of population distribution forecast models available, the range of the nonlinear function included in the disaggregate model, and the robustness of simplifying assumptions concerning the shape and interdependence of these distributions.

ALTERNATIVE DISTRIBUTIONAL FORECAST PROCEDURES

A range of procedures may be considered for use in forecasting the required distributions. Such procedures represent different assumptions concerning the process that underlies the development of the observed distribution (for example, the co-location aspects of residential location choice for different income groups) and the degree to which simplifications may be introduced. The acceptability of alternative forecasting procedures depends on their conformity with the underlying distribution process and the robustness of the aggregate behavioral forecasts to the simplifications used. Three general approaches are described below.

1. The simplest distribution forecast procedure would be to project the existing distribution in a zone unchanged over the period of interest except for already planned or in process changes that can be explicitly identified. This assumption will be best for short-term predictions. However, even for longer periods this assumption—with modifications based on available information and judgment—can provide better aggregate forecasts than those that could be obtained through the use of conventionally developed aggregate modes. For example, the near-term effect of a change in public transit service could be based on the existing distribution of household and highway service characteristics.

2. Another procedure would be to assume that the distribution of the population is systematically related to a small number of indexes (means, for example) that might be readily forecast. For example, one might assume a gamma distribution of income, 1 parameter (defining the shape of the distribution) fixed and the scale parameter determined from the mean (4) or both parameters simply related to the mean or more generally to directly predict both parameters of the distribution.

3. A more sophisticated procedure would be to develop a transition matrix for "growing" households from inception, through various life-cycle stages, to dissolution, including information on relevant characteristics. Such an approach would be most appropriate for relatively large areas where the effects of migration are relatively unimportant.

INTEGRATION OF HOUSEHOLD AND SPATIAL CHARACTERISTICS

The distribution of spatially related characteristics facing the household must be considered as well as the distribution of socioeconomic or household characteristics. Characteristics that are neighborhood or transport-system specific are examples. This suggests the need to spatially assign the household-characteristic distribution and to develop joint distributions over household and spatial characteristics. Obviously, this step will be much simplified if an assumption of independence can be justified between household and spatial distribution or if the dependence can be simply specified.

However, to argue that the distribution of household and spatial characteristics is related through the household location choice process is more reasonable. This relation could be modeled by first forecasting the regionwide distribution of household characteristics and then assigning households with specific characteristics to geographic locations as part of a residential location choice model that explicitly accounts for geographic, neighborhood, and transportation service characteristics.

In general, the entire problem of forecasting interrelated distributions of population and spatial characteristics could be simplified by designing spatial groupings (zones or districts) so as to highlight differences that are relevant to the analysis in question. Considering the spatial sensitivity of out-of-vehicle travel time and access to transit, for example, it would be useful to explicitly identify areas that are, or would be in the future, highly differentiated in terms of accessibility to transit service. Geographic aggregation of the population for areas with common service characteristics would simplify the aggregate prediction problem when compared to present zonal groupings.

MARKET SEGMENTS IN BASE MODELS

To this point, we have explicitly assumed that travel choice behavior can be represented by a single disaggregate model. That is, we have assumed that all groups of the population have the same behavioral response when they are confronted by identical conditions. However, in many cases the population will have to be segmented and different disaggregate models developed for each market segment. In this case, prediction requires the explicit distribution of the population into these market segments, and all further distributions must be conditional on them. The aggregation procedure would be applied to each market segment and then aggregated over all market segments. Suitable market segments might be related to household life-cycle, occupation of primary wage earner, or other characteristics that may be expected to influence taste patterns with respect to travel behavior.

SUMMARY DESCRIPTION OF ANALYSIS AND PREDICTION

Development of a behaviorally sensitive aggregate forecasting model based on individual or other behavioral unit responsiveness to external characteristics requires the development of procedures for forecasting the distribution of these external characteristics and the characteristics of the household and the development of the underlying disaggregate choice model.

The proposed procedure for obtaining aggregate predictions may be divided into 4 stages. The first stage is to analyze existing data to obtain a disaggregate travel choice model and a household characteristics distribution model. The second stage is to forecast future distributions of population characteristics. The third stage is to define alternative distributions of transportation service characteristics based on policies to be tested, and the fourth stage is to predict aggregate travel behavior.

Once the models have been developed (stage 1) and the distribution of population characteristics for the area has been predicted (stage 2), stages 3 and 4 only have to be repeated to test additional transportation service alternatives.

RESEARCH DIRECTIONS

The preceding discussion indicates that the development of behaviorally sensitive aggregate forecasting models depends on the availability of models for the prediction of the distribution of characteristics that influence travel behavior and the prediction of the probability of disaggregate travel choices when disaggregate characteristics are known. This suggests that, in addition to the ongoing research directed toward the improvement and extension of disaggregate choice models, research must be directed to the development of models that may be used to predict the multivariate distribution of population and service characteristics that influence travel-choice behavior. Specific areas of research are

1. Analyze existing distributions of population characteristics to identify their shape and interdependence;
2. Develop procedures to forecast parameters of the identified distributions on a spatially specific basis;
3. Identify the relation between the distributions of population and transportation service characteristics, taking account of the potential development and application of disaggregate models for household location choice;
4. Develop and apply procedures to test the robustness of simplified descriptions of characteristic and service distributions;
5. Identify those forms of choice models and distribution representations that are amenable to mathematical integration; and
6. Develop criteria to be used in the comparison of aggregate forecasting models based on disaggregate and aggregate analyses, perform a full-scale test of alternative aggregate forecasting procedures, and identify the circumstances under which the different procedures should be used.

ACKNOWLEDGMENTS

The contents of this paper have benefited from extensive discussion with Moshe Ben-Akiva and Steven Lerman, M.I.T., and Charles Lave, University of California, Irvine.

Earlier work on this problem was done under the auspices of Cambridge Systematics, Inc., under funding of the U.S. Department of Transportation for research into automobile ownership behavior.

REFERENCES

1. Green, H. A. J. Aggregation in Economic Analysis. Princeton Univ. Press, Princeton, N.J., 1964.
2. Kanafani, A. An Aggregative Model of Trip-Making. Transportation Research, Vol. 6, 1972, pp. 119-124.
3. Westin, R. B. Predictions From Binary Choice Models. Journal of Econometrics, April 1974.
4. Wooten, H. J., and Pick, G. W. A Model for Trips Generated by Households. Journal of Transport Economics and Policy, 1967.