Verification of Land Use Forecasting Models: Procedures and Data Requirements

DAVID E. BOYCE and ROGER W. COTE
Battelle Memorial Institute, Columbus Laboratories, Columbus, Ohio

A review of the available land use forecasting models suggests that verification and evaluation of these models is effectively limited because confidence statements on the forecast variables are not available from the models. This paper proposes possible procedures for providing confidence statements for two types of models. The data requirements for verification studies of the type proposed are described.

•DURING the past 5 years, model development in urban transportation planning has been oriented primarily to the problem of land use forecasting; that is, the allocation or distribution of urban activities to small analysis zones at future points in time. The publication of a special issue of the Journal of the American Institute of Planners, May 1965, on urban development models marks the completion of the first round of intensive model development. Reported in this Journal issue and elsewhere [e.g. Lowry (10), Hill et al. (5), and Donnelly et al. (1)] are the results of more than 10 separate studies whose objective was the formulation, development, and testing of land use forecasting models. These models represent a wide variety of approaches and orientations to model development and incorporate much of what is known and theorized about urban structure and growth. Although a great deal remains to be done before even crude forecasts of land use will be available on a production basis (in the sense that urban trip forecasts are now prepared), it is clear that a foundation for future model development and refinement has been provided.

This being the case, it is timely to ask

1. How well do the current models perform?
2. What procedures, tests, and criteria are available to evaluate their performance?

A review of current models indicates that the first question may be answered only superficially, for two reasons: (a) these models cannot be verified in a strict sense because their formulations do not provide a confidence statement about the relationship between the observed and predicted values; and (b) the data required for testing often exceed the data available, a condition that has perhaps properly reduced the priority of model verification in past studies. However, expected developments in data collection and manipulation (2) suggest that future studies will have at their disposal sufficient data for model testing and evaluation.

For these reasons, this paper does not attempt to evaluate these land use forecasting models, or to review the evaluation procedures in use. Rather, one requirement of verifiability is identified, and two techniques for satisfying the requirement are proposed. Then, data requirements for verification of models are reviewed. The proposed model verification and evaluation procedures should lead to an increased ability to verify forecasts and provide a basis for evaluating the model itself before test forecasts are prepared.

Paper sponsored by Committee on Land Use Evaluation and presented at the 45th Annual Meeting.

REQUIREMENTS AND PROCEDURES FOR MODEL VERIFICATION

Requirements for verification of forecasting models are discussed by Theil (11). Verifiability is defined as that characteristic of a model which makes it possible to conclude, after a certain time and in an unambiguous manner, whether the forecast has turned out to be correct or incorrect. It must be possible also to verify the procedure by which the prediction itself was derived. From this definition, Theil derives a series of requirements of forecasts and forecasting models for verification. These requirements are quite basic, and include such concepts as the explicit statement of the model, specification of the time to which the forecast refers, and internal consistency of multiple forecasts, requirements satisfied by nearly all land use forecasting models.

A requirement not satisfied by most land use forecasting models concerns the specification of the relationship between the forecast and the actual outcome. For the forecast to be verified at all, it is necessary that a probability statement relating the forecast and observed values be given. For point predictions, it is not expected that the prediction will coincide with the observed value, so it is necessary to specify the distribution of error around the prediction if an evaluation is to be made. For interval predictions, the probability that the observed value will fall within the predicted interval is required.

Inasmuch as the land use forecasting models already mentioned do not provide, as part of the forecast output, statements concerning the probability distribution of forecast variable, it is not possible to evaluate the performance of these models. It should be noted that measures of accuracy such as the coefficient of determination (R^2) and the root-mean-square error are often applied to compare observed outcomes with predicted values from these models, or to test how well the calibrating data fit the model. Definitions and analyses of these accuracy measures may be found in Irwin and Brand (6), as well as Theil (11). For comparing the performance of alternative models, these accuracy measures may be of limited value. For evaluating a single model, however, in the absence of any confidence statement about the forecast variable, there is little basis for evaluating the magnitude of these accuracy measures.

In addition to permitting model verification, the specification of confidence statements on the forecast variable permits the model to be evaluated in another sense. Since the land use forecast is an input to trip forecasting models, the distribution of the error about the forecast must be sufficiently small so that these inputs are meaningful. Generally, the acceptable level of error for land use forecasts as an input to the trip-forecasting procedures can be determined. As presently formulated, land use forecasting models do not provide for such confidence statements, so the forecasts cannot be evaluated, even on this basis.

Two procedures for providing confidence statements for land use forecasting models are described. These procedures have shown potential merit in preliminary investigations. The first procedure applies to the class of models for which the parameters are estimated outside the model and independently of each other. Such models are hereafter referred to as "nonstatistical" models. An excellent example is the Pittsburgh model developed by Lowry (10); another model of the same type is the direct trip allocation model developed by Lathrop et al. (9), which is used for explanatory purposes here.

Models in which the parameters are estimated simultaneously so as to maximize the goodness-of-fit to the data are a second class of models. These "statistical" models, such as the simultaneous equation regression model of Hill et al. (5), commonly employ one of the least-squares techniques to estimate the model parameters. The selection of the regression estimation technique to be used is not discussed here. The problem of interest is how to provide confidence statements for the forecasts.

Confidence Statements for Nonstatistical Models

The objective of this procedure is to provide confidence statements for the forecast variables. A complete specification of the probability distribution of the forecast variable constitutes the most desirable and complete type of confidence statement that can be made. If this is not possible, a specification of the variance of this distribution and the type of distribution will permit precise confidence statements to be made. A less

desirable, but still quite useful, statement is the specification of a confidence interval and the probability that the observed value will fall within this interval. The procedure produces confidence statements of any of the above types, depending on the form of the confidence statements on the variables and parameters that are inputs to the model.

The direct trip allocation model (9), an example of a nonstatistical model, is used to illustrate the procedure. This model is defined by the equations

$$A_j = A\left[e^{-\ell O^{(j)}} - e^{-\ell(O^{(j)} + O_j)}\right], \quad j = 1, 2, \ldots, N \tag{1}$$

where

A_j = the amount of activity to be allocated to zone j;

A = the total amount of activity to be allocated within the study area;

ℓ = probability of a unit of activity being sited at a given opportunity;

O_j = opportunities in zone j;

$O^{(j)}$ = the opportunities for siting a unit of activity rank-ordered by access value and preceding zone $j = \sum_{i=1}^{j-1} O_i$; and

N = the number of zones in the study area.

On the basis that the parameters ℓ, O_1, O_2, . . . , O_N, are to be estimated outside the model and independently of each other, the following approach is proposed. First, it is asserted that it is unrealistic to attempt to produce point estimates of the parameters. Specification of point estimates requires more than is known about these parameters and at the same time discards valid information on the range and variance of the parameter value. One can, however, hope to obtain an interval of values for a parameter and a corresponding probability measure of that interval.

A further extension of this approach leads to a probability distribution on the range of possible values for each parameter. Evidently, the most meaningful way to express the available information about the parameters is to state that the quantities ℓ, O_1, . . . , O_N are independent random variables with distributions determined outside the model. These distributions are not necessarily independent, and in fact may be highly interdependent. Conceptually, this interdependence could be satisfied by specifying multidimensional distributions on the input variables. Operationally, however, this approach is not realistic for more than two or three variables. Therefore, as in the original model itself, the assumption of independence is necessary to the solution of the problem. The variables A_j, $j = 1, \ldots, N$, being functions of the random variables ℓ, O_1, . . . , O_N, are also random variables with distributions made up of the same parameters because of the functional relations comprising the statement of the model.

In general, it is quite difficult to determine analytically the probability distribution of a random variable that is a function of several other random variables; however, one can develop an empirical solution using Monte Carlo techniques (7).

Empirical distribution of the random variables, A_j, can be constructed as follows:

1. Perform a sequence of independent sampling experiments that yields an observation $\left(x_o^{(1)}, x_1^{(1)}, \ldots, x_j^{(1)}\right)$ in which $x_o^{(1)}$ is a random number generated by the probability distribution of the random variable ℓ, and $x_i^{(1)}$, $(i = 1, \ldots, j)$ is a random number generated by the distribution O_i.

2. According to Eq. 1 compute

$$A_j^{(1)} = A\left[e^{-x_o^{(1)} \sum_{i=1}^{j-1} x_i^{(1)}} - e^{-x_o^{(1)}\left(\sum_{i=1}^{j-1} x_i + x_j\right)}\right] \tag{2}$$

3. Repeat these first two steps T times and obtain a set of T random numbers $\left(A_j^{(1)}, A_j^{(2)}, \ldots, A_j^{(T)}\right)$ from the distribution of A_j.

4. Organize these T random numbers into a frequency distribution and study its properties, in particular its mean and variance. This empirical distribution is an approximation of the exact, theoretical distribution of A_j.

5. Repeat Steps 1 to 4 for each j = 1, ..., N.

Following the development as outlined, one obtains probability distributions for the A_j's, rather than point predictions. In this manner, the amount of uncertainty associated with the individual parameters after they have been estimated is retained and translated into the uncertainty about the A_j's. One can evaluate the quality of the prediction in terms of the dispersions of the distributions of the A_j's, the dispersion or variance of the distribution being a measure of the precision of the forecasting model. A model that yields a predicted distribution for A_j with small dispersion permits one to establish upper and lower bounds on A_j that are fairly close with high probability. On the other hand, a model that results in a relatively flat distribution for the predicted variable restricts one to establishing upper and lower bounds that are close only with low probability; the higher the probability desired, the lower the precision obtainable.

The Monte Carlo approach also provides a means of tightening the forecast distributions. By carrying out this distribution sampling experiment under a variety of specifications of the distributions of the independently estimated parameters, it is possible to obtain an indication of how much the individual parameter distributions are contributing to the dispersion of the forecast variable. Thus, by working "backwards," one can obtain the required specifications for the input distributions. These specifications can then indicate goals to be met when estimating the parameters and determine where it is most economical to invest estimation effort.

Confidence Statements for Statistical Models

Confidence statements for the dependent variable in regression models (the only statistical models considered here) may be derived for the classicial treatment of univariate multiple regression models as follows. It is assumed that the mean of a random variable, $y_{\underline{x}}$, denoted by $m(y_{\underline{x}})$, is a linear function of the k elements of a nonrandom vector $\underline{x} = (x_1, \ldots, x_k)$. A sample of n independently observed points $(y_i, x_{1i}, \ldots, x_{ki})$, $i = 1, \ldots, n$, is used to obtain the maximum likelihood estimates of the parameters σ, α, $\beta_1, \ldots, \beta_k$, where the linear hypothesis about the relation of y to $(x_1, \ldots, x_k)$ is

$$m(y_{\underline{x}}) = \alpha + \sum_{j=1}^{k} \beta_j x_j \tag{3}$$

and where σ is the standard deviation of $y_{\underline{x}}$ about the mean function at any point $x_1, \ldots, x_k$.

The application of the general method of confidence limits to this model yields the interval (Y_T, Y^T), which is a γ percent tolerance interval of confidence level p/100, if the probability of this interval (Y_T, Y^T) containing γ percent of the possible values of $y_{\underline{x}}$ is equal to p/100. The interval (Y_T, Y^T) is determined by the formula

$$\hat{y}_{\underline{x}} \pm k S_{\hat{y}_{\underline{x}}} \tag{4}$$

where $\hat{y}_{\underline{x}}$ denotes the estimate of $m(y_{\underline{x}})$ and $S_{\hat{y}_{\underline{x}}}$ denotes the standard error of the estimate $\hat{y}_{\underline{x}}$. This interval specifies a range within which γ percent of the possible values of $\hat{y}_{\underline{x}}$ will occur with probability p/100. The number k depends on γ, p, and the sample

size, n. Once the values of $\hat{y}_{\underline{X}}$, $S_{\hat{y}_{\underline{X}}}$, and $k(\gamma, p, n)$ have been determined and used in Eq. 4 to obtain the interval (Y_T, Y^T), one can state with p percent confidence that γ percent of the observed values of the random variable $y_{\underline{X}}$ lie within k standard errors of the prediction, $\hat{y}_{\underline{X}}$.

As an example, consider the problem of forecasting the number of households per zone, $m(y_{\underline{X}})$, and evaluating the forecasting model by comparing this predicted value with the observed outcome. If the model is valid, the observed outcomes will fall within the 95 percent tolerance limits (Y_T, Y^T) with probability 0.99, for example. An acceptable model is one for which observed values are so located, and the tolerance limits are sufficiently narrow, at an acceptable level of probability, to provide meaningful inputs to the trip forecasting methodology.

The confidence statements for this classical regression model are limited in that the problem formulation does not refer explicitly to forecasting future values of the dependent variable. An alternative approach to confidence statements, described by Zellner and Chetty (12), specifically considers this prediction problem. In this Bayesian formulation, the probability distribution of the next q observations is derived on the basis of the past n observations. Procedures for employing this distribution to make inferences about the future values of the dependent variable are developed. Specifically, the variances of the predictions is a function of the original n observations as in the classical treatment, plus the q observations on the independent variables for the forecast period. From these variances, confidence intervals can be constructed for each of the q predictions.

Data Requirements for Model Verification

In general, land use forecasting models use cross-section data, that is, observations on the model variables for each analysis zone for a given point in time.

Data requirements for model verification may be considered in terms of two classes of forecasts—incremental growth estimates and total estimates of activity. The incremental model estimates of the increment of growth by zone over a specified time period, whereas the total activity model estimates the total activity for a given point in time. Most land use forecasting models are of the incremental type in that they incorporate the land use pattern for some base year into the model. However, models that allocate the total activity, such as Lowry's Pittsburgh model, have been developed.

Verification studies of incremental models require a minimum of two sets of cross-section data, in addition to the data used for model calibration. These models may require data on specified activities for the second as well as the first time period. For example, an incremental residential and retail model might require inputs of the total activity patterns for the base year and the basic employment pattern for the forecast year. The ability to test the reliability of these models increases in proportion to the number of cross-sections of data available.

The total allocation models, on the other hand, require only one set of cross-section data for verification purposes. These models specify certain activities as inputs and estimate the total amount of the remaining activities. Although this type of model requires fewer data, it is much more demanding in that the total structure of activity location, which may reflect past location decisions made under outmoded technologies, is to be reproduced. In view of the vast changes in transportation and building technology over the past several decades, this requirement may be indeed difficult. Inasmuch as the principal interest in land use forecasting is the preparation of a land use forecast based on the existing land use pattern, it is expected that the incremental model will continue to dominate the course of model development.

Model verification studies of the types proposed depend on data inputs over and above the cross-section data described earlier. In particular, for the Monte Carlo studies, it is necessary to determine the probability distribution of the model parameters. In some cases, this may only require a specification of the measurement error for the data. If the data are based on a sample, as contrasted with an inventory, the specification of the sampling error is required. Several models also include parameters which cannot be measured directly, such as the probability of siting, ℓ, in Lathrop's direct

trip allocation model. Nevertheless, the specification of the probability distribution of ℓ is required for these verification studies. A logical choice in this case is the β distribution because it has a range from 0 to 1, as does the probability ℓ. The specification of this distribution and its parameters should be regarded as an assumption about ℓ.

In summary, data requirements for model verification studies involve more than the availability of a sufficient number of cross-sections, a demanding requirement in itself. Also needed is the information necessary to characterize the distributions of the observed variables and parameters, and to support assumptions on other parameters. Fulfillment of these requirements will demand more and more concern with sampling methods, measurement of sampling error, analysis of data characteristics, and related data problems by urban transportation studies.

CONCLUSION

The preliminary research reported here suggests several procedures for evaluating land use forecasting models by providing for confidence statements on forecasts. Through the use of these procedures it can be determined whether these models are providing forecasts of sufficiently low variability, let alone accuracy, to be of value. In addition, the Monte Carlo technique provides a means of exploring the properties of the models themselves.

It is also clear from the brief analysis of data requirements that verification of these models depends heavily on many more data than are currently available. However, it is expected that this situation will be alleviated considerably in the next few years; hopefully, the verification techniques necessary to fully exploit these data will be developed in the interim.

REFERENCES

1. Donnelly, T. G., Chapin, F. S., Jr., and Weiss, S. F. A Probabilistic Model for Residential Growth. Institute for Research in Social Science, Univ. of North Carolina, Chapel Hill, 1964.
2. Garrison, W. L. Urban Transportation Planning Models in 1975. Jour. of the Amer. Inst. of Planners, Vol. 31, No. 2, 1965.
3. Hald, A. Statistical Theory With Engineering Applications. John Wiley, New York, 1952.
4. Harris, B., ed. Urban Development Models: New Tools for Planning. Spec. Issue, Jour. of the Amer. Inst. of Planners, Vol. 31, No. 2, 1965.
5. Hill, D. M., Brand, D., and Hansen, W. B. Prototype Development of a Statistical Land Use Prediction Model for the Greater Boston Region. Highway Research Record 114, pp. 51-70, 1966.
6. Irwin, N. A., and Brand, D. Planning and Forecasting Metropolitan Development. Traffic Quarterly, Vol. 19, No. 4, 1965.
7. Johnston, J. Econometric Methods. McGraw-Hill Book Co., New York, 1960.
8. Klein, L. R. A Textbook of Econometrics. Row, Peterson and Co., Evanston, Illinois, 1953.
9. Lathrop, G. T., Hamburg, J. R., and Young, G. F. An Opportunity-Accessibility Model for Allocating Regional Growth. Highway Research Record 102, pp. 54-66, 1965.
10. Lowry, Ira S. A Model of Metropolis. The RAND Corp., Santa Monica, 1964.
11. Theil, H. Economic Forecasts and Policy. North-Holland Pub. Co., Amsterdam, 1961.
12. Zellner, A., and Chetty, V. K. Prediction and Decision Problems in Regression Models from the Bayesian Point of View. Jour. of Amer. Stat. Assoc., Vol. 60, No. 310, 1965.