Proc. Instn Civ. Engrs, Part 2, 1975, 59, Sept., 545–548

7737 DISCUSSION

A statistical evaluation of the accuracy of a family of gravity models

J. A. BLACK & R. J. SALTER

Mr D. M. Allan, University of Glasgow

This comparative study is a valuable one for the practitioner in this field. The likely instability of the calibration parameter(s) over time, however, should discourage over-refinement of what is only one step in the analysis. In a predictive context the instability problem will always be present, even if a realistic distribution–modal split model is used. The most useful way to extend the comparison of model and survey matrices would seem to be by concentrating attention on the assignment implications. One would be on surer ground, and the results would have more practical value, than when dealing with the statistics of the matrices themselves. However, this would require a fully validated assignment model. Without it, wide discrepancies are bound to appear, as Table 10 shows.

77. The Authors refer twice to the limitations imposed by computer storage (16K). Assignment requirements can undoubtedly present difficulties but distribution algorithms should present fewer problems even in the fully constrained case. With the help of disc or magnetic tape there is nothing to prevent the two matrices—trip and separation—from being partitioned so that a realistic number of zones can be handled without significant increase of calculation time. If necessary, only a few rows of the separation matrix need be input at one time so that it would have been sufficient here to divide the trip matrix into upper and lower halves, accumulating column totals for iteration purposes after each half had been calculated. It is not necessary to work with row and column totals alternately although the Furness method strictly involves this. The unique solution, ably demonstrated by Evans,[10] can also be achieved by continually modifying attraction values only, and it is simpler if the matrices have to be partitioned. The details are given by Edens[55] among others.

78. It seems a pity to allow the restrictions of an elegant calibration routine to prevent a complete study of the distribution models themselves. A simple trial and error procedure, or regression with (finely) grouped data rather than with individual cell values, would have sufficed here although the full technique could have been retained by adapting the partitioning approach described. It is difficult to see why the fully constrained case should present greater difficulty than the others.

Dr I. G. Heggie, Transport Studies Unit, Oxford

The problem the Authors present is quite simple: faced with the choice of four types of deterrence function and with three ways of constraining the solution, what is the marginal benefit of altering either or both of these features?

Paper published: *Proc. Instn Civ. Engrs*, Part 2, 1975, 59, Mar., 1–20.

80. The results of their investigation are intriguing. In Table 4 they compare the fitted and observed origin-destination matrices for the journey to work by car using a simple correlation coefficient as a test of accuracy (R). If this is squared to give R^2, to show the proportion of the variation in the observed data explained by the models, then the simplest model explains 77% of the variation and the best model (exponential fully constrained) explains 86%. The exponential production constrained model likewise explains 85% of the variation and the wholly unconstrained exponential model explains 80%.

81. However, in the absence of firm criteria for evaluating different levels of goodness of fit it is extremely difficult to choose between the various models on the basis of their correlation coefficients. For example, how does one decide whether or not a 3–4% increase in R^2 justifies the use of an exponential deterrence function rather than a simple power one? More important, when the exponential model is used, how does one decide whether or not to constrain the model partially (which increases R^2 by 5%) or to constrain it fully (which increases R^2 by a further 1%)? Such decisions should be based on a comparison between the planning benefits associated with increased accuracy and the added cost of complexity, but this is not easy.

82. This difficulty raises a fundamental issue for modelling practice: what is the most efficient way of assessing the accuracy and performance of alternative planning models?

83. The Authors use three conventional measures for assessing the accuracy of their models

(*a*) they compare the actual and fitted trip length frequency distributions using the root mean square statistic

(*b*) they compare the observed and fitted trip matrices (origin–destination matrices) by means of the correlation coefficient and the χ^2 test

(*c*) once traffic has been assigned to the road network they also compare the observed and assigned network link flows on six major radial routes.

84. Test (*a*) is generally recognized as unsatisfactory. Large estimation errors can be swamped by the aggregation inherent in the preparation of trip length frequency distributions, so that the relationship between the observed and fitted distributions—no matter how they are compared—is rather meaningless.[56]

85. Test (*b*) is more important because correlation coefficients are used in a variety of statistical applications and, as the Authors rightly point out, 'the interpretation of R as a measure of the strength of the linear relationship between two variables is mathematical'. The χ^2 test could be used only subject to a number of limiting assumptions and it showed uniformly poor fits between the survey and all model matrices. However, the correlation coefficient gave absolute values that most engineers would regard as significant, and the variation between the models was large enough to suggest that some deterrence functions were better than others and that constraining the models improved their accuracy.

86. The Authors note that the correlation coefficient lacks sensitivity and this can be shown by calculating R for a typical pair of observed and fitted trip matrices as in Fig. 2. It is based on data of trip generation rates[57] which have a similar pattern to an origin–destination matrix.

87. The correlation coefficient for these two matrices is 0·95 in spite of the fact that errors of ±50% are not uncommon. Clearly when the correlation coefficient drops to 0·85 or less the errors become so substantial that extrapolation or guesswork might be more accurate. Even with $R=0{\cdot}95$ some people may prefer extrapolation.

88. Test (*c*) can also be shown by Fig. 2. The Authors were comparing assigned flows but confined themselves to the major radial routes which, by implication, usually carry the largest flows. The relative errors on the six major flows shown in Fig. 2 are small (only 7% on average). The figures in Table 10 show a similar range of errors. The method of estimating most models (usually by minimizing the sum of the squared deviations) tends to ensure a uniform distribution of absolute errors which

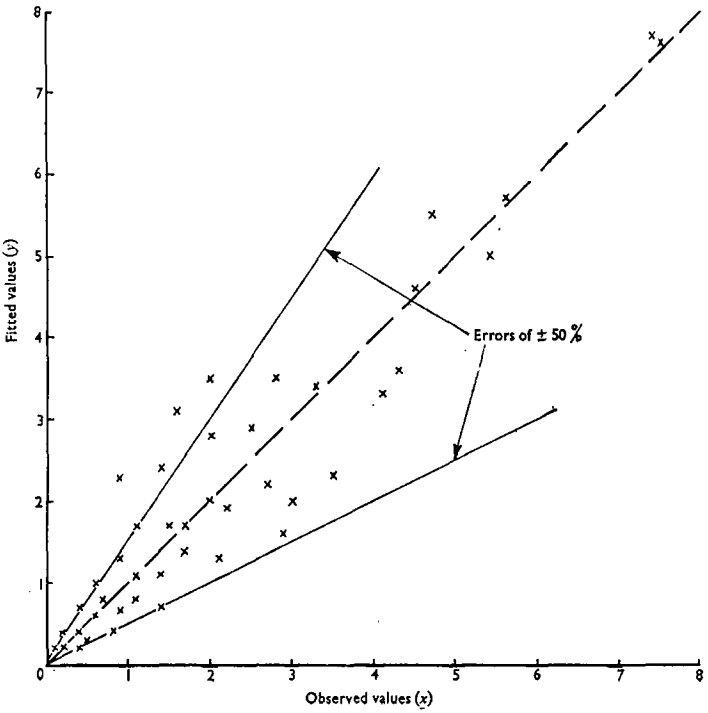

Fig. 2. Correlation coefficient

automatically means that relative errors tend to decline with size. Any comparison between observed and fitted flows which is applied only to large flows will therefore tend to be inefficient.

89. The Paper makes an extremely important contribution to the understanding or the operational performance of different types of gravity model subject to differing levels of constraint. Further work is now needed to develop more efficient ways of measuring the accuracy of traffic models.

Dr Susan Wilson, Department of Statistics, Australian National University

For the type of models used by the Authors the statistically correct test of goodness of fit to use for the trip length distributions and the origin–destination matrices is a χ^2 test. Also χ^2 can be used as a relative measure to compare two models, by examining the probabilities determined from the appropriate χ^2 tables.

91. As χ^3 is an appropriate statistic, it follows from § 71(*d*) that the high χ^2 values calculated indicate that the present trip distribution modelling approach needs refinement. One possibility is the application of Monte Carlo simulation techniques.

92. The Kolmogorov–Smirnov test is a statistically correct procedure which can be used to test between the observed and model trip length frequencies. However, the contingency coefficient[44] is not applicable and the coefficient is meaningless for these models. Its use seems to have developed from the incorrect treatment of the trip matrices as contingency tables.

93. The correlation coefficient, determined between the elements of the survey matrix and the elements of the model matrix, is also inappropriate because the elements of the model matrix are determined from the elements of the survey matrix. This explains why researchers have obtained high values of R, irrespective of whether the model is a good fit or not. Furthermore, it can be shown by examples that R is not a reliable relative measure, although it is difficult to specify the proportion of times in which R will, or will not, indicate the correct relative assessment of models.

94. The root mean square can be used as a relative measure, but should be interpreted with caution as it is sensitive to large deviations of the frequencies from the mean. An alternative measure (related to the χ^2 statistic) that would compensate for this effect is

$$\left[\sum_{j=1}^{n}(O_j-e_j)^2/ne_j\right]^{1/2}$$

where O_j, e_j and n are respectively observed frequencies, expected frequencies and the number of frequency cells.

95. A further technique that could be applicable to the evaluation of gravity models is the log-likelihood method.[58]

Dr Black and Dr Salter

Dr Wilson's explanation of why the correlation coefficient between elements of the model and survey matrices is generally found to be high should be a warning to researchers not to rely on R as a measure of the models' accuracy. **Dr Heggie** also makes this point (§ 87). Although Dr Wilson states that R may not be a reliable measure (§ 93), we found that R ranked the models in the same order as did other measures. Nevertheless, our remarks on the lack of sensitivity of R (§ 71) indicate that the correlation coefficient is a misleading statistic.

97. We would tend to disagree with Dr Heggie that a comparison of the model and observed trip length frequencies is rather meaningless (§ 84). We appreciate that the preparation of such frequency distributions involves an assumption about grouping travel data into class intervals. However, this is a common assumption in all the models we evaluated so we regard the root mean square as a useful measure (although note Dr Wilson's comments in § 94).

98. Dr Heggie questions our concentration on the major radial routes in Bradford. Paragraphs 64–69 were included as being illustrative of an alternative way of considering the accuracy of gravity model results. His point in § 88 is valid and any comprehensive analysis along these lines should cover more roads than are included in Table 10. We did not report our findings in more detail because we did not feel that the all or nothing assignment program was a sufficiently realistic enough model of route choice. Indeed as **Mr Allan** comments (§ 76) without a fully validated assignment model it would be misleading to place any emphasis on our results.

99. We feel the discussion has shed some light on the kinds of statistical methods researchers can use to test the accuracy of trip–distribution models. We are sure the practitioner would rather hear about the accuracy of different models rather than read about new theoretical expressions of the same basic concept.

References

55. EDENS H. J. Analysis of a modified gravity model. *Transp. Res.*, 1970, **4**, 51–62.
56. HEGGIE I. G. *Transport engineering economics*. McGraw-Hill, 1972, 172.
57. PAS E. I. The unit of analysis in the modelling of residential trip generation. Paper presented at the UTSG Conference, 1975.
58. EDWARDS A. W. F. *Likelihood*. Cambridge University Press, 1972.