AN EMPIRICAL ASSESSMENT OF THE ENTROPY MAXIMISING FAMILY OF INTERACTION MODELS

J.A. WALSH
Department of Geography
Carysfort College, Blackrock
Co. Dublin

R. TOBIN
City Planning Office,
Limerick

ABSTRACT

The parameters, and consequently the performance level, of statistical models of spatial interaction are particularly dependent on the spatial organisation framework which is used to organise interaction data and the way in which the costs of interaction are measured. This paper provides an empirical assessment of the influence of these factors on the performance levels of the entropy maximising family of interaction models, the members of which differ according to the extent of their data requirements. The principal results which emerge are as follows. The level of model performance generally improves as the number of constraints increases. However, the need for a constraint on the average cost of an interaction appears to be relevant at only one level of spatial organisation. The effect of increasing the number of zones in the spatial organisation framework from eight to twenty-two is only a marginal deterioration in the level of model performance. The use of straight line distances in the calibration of these models does not seriously affect their performance levels. However, a logarithmic transformation of the actual distances produces a significant decline in these levels. The paper concludes with a discussion of the general implications of these findings for intra-regional level analyses of activity patterns in Ireland.

In a recent paper one of the present authors introduced the entropy maximising procedure into geographical analysis in Ireland and demonstrated its potential in relation to the analysis of work travel patterns in Co. Limerick (Walsh, 1980). The work reported here extends the analysis discussed in that paper.

An entropy maximising model is one which provides the least biased assignment of probabilities to a set of choices that are subject to some constraints. In interaction models it is usual to make some assumptions about a zoning system, the number of trips that originate from a zone, the number that are destined for each zone and the cost of movement between pairs of zones. Each of these assumptions influences the nature and reliability of the model which is produced in a particular context. This paper explores the effects of each of these assumptions on model performance through an empirical analysis based on data on work trips in Co. Limerick. The specific objectives of this paper are, therefore, to evaluate empirically the role of the constraints in the entropy maximising framework and to examine the effects of scale through the zoning system, and the way in which the costs of interaction are measured, on the performance of different types of interaction models.

The entropy maximising (EM) family of interaction models

This family of interaction models were initially developed by Wilson (1970). The general EM model has the following form:

$$t_{ij}^{*} = (A_i O_i B_j D_j \exp[-\beta C_{ij}]) \quad (1)$$

where $A_i = (\sum_j B_j D_j \exp[-\beta C_{ij}])^{-1}$ (2)

$$B_j = (\sum_i A_i O_i \exp[-\beta C_{ij}])^{-1} \quad (3)$$

and t_{ij}^{*} is the predicted number of trips between zones i and j,

O_i is the actual number of trips that originate from zone i,

D_j is the actual number of trips that are destined for zone j,

C_{ij} is a generalised term for the cost of interaction between zones i and j,

β is a parameter to measure the deterrence effect of C_{ij} on interaction,

A_i and B_j are balancing factors to ensure that the number of trips which are predicted to originate from each zone is equal to O_i and that the number destined for each zone is equal to D_j respectively.

The details involved in the derivation of this model are contained in Walsh (1980).

A number of models can be deduced from the general one, according to differences in the calibration procedure. If both balancing factors, A_i and B_j are held constant then

$$t_{ij}^{*} = K O_i D_j \exp[-\beta C_{ij}] \quad (4)$$

which is known as the unconstrained interaction model. The column and row totals for the matrix predicted by equation (4) need not correspond with those for the actual trip matrix. If only the B_j's are held constant in the calibration then

$$t_{ij}^{*} = K A_i O_i D_j \exp[-\beta C_{ij}] \quad (5)$$

which is known as the origin-constrained model. Similarly, when the A_i's are held constant one obtains the destination-constrained model

$$t_{ij}^{*} = K O_i B_j D_j \exp[-\beta C_{ij}] \quad (6)$$

When neither the A_i nor the B_j distributions are held constant then the model is of the form given by equation (1). This model is said to be origin-destination constrained. If, in addition, a β value can be obtained for the last model which ensures that the average cost of an interaction according to the model is equivalent to an empirically determined average cost then the model is said to be origin, destination and cost constrained.

In the empirical analysis each of these models is calibrated against a data set relating to 4190 work trips in Co. Limerick in 1977. These data are described in Walsh (1980). The calibration procedure used in all instances was the iterative one due to Hyman (1969). Following the identification of the parameters of each model its peformance can be assessed by examining various indicators of how well the predicted interaction pattern corresponds with the actual one. Since there is no general agreement among researchers about which indicator is best to use the following ones are employed here.

(1) The mean and standard deviation of the residuals.

(2) The coefficient of determination, R^2, and the parameters of the regression of the predicted trips on the observed ones. In the case of a good fit the intercept parameter would be close to zero and the slope of the regression line would be close to unity.

(3) The chi-square statistic defined as
$\chi^2 = \sum_i \sum_j (t_{ij} - t_{ij}^*) / t_{ij}^*$, for all $t_{ij}^* \geq 6$

(4) The dissimilarity index, G, defined as
$G = (\sum_i \sum_j | t_{ij} - t_{ij}^* |) . \frac{1}{2}T$
T denotes the total number of trips. The index ranges between zero, representing a perfect correspondence, and 100, representing a situation of maximum possible difference between the two distributions. The magnitude of G represents the percentage of the trips in the predicted matrix that would have to be reallocated in order to replicate the observed trip matrix.

(5) The information gain index I. The index measures the amount of information that is required, relative to what is already known, to alter some prior probability distribution into a posterior distribution. Following Walsh and O'Kelly (1979) the index may be defined as
$I = T^{-1} \sum_i \sum_j t_{ij} \log (t_{ij} / t_{ij}^*)$
It is dimensionless and ranges between zero and $\log (N^2) - 1$ where N is the number of zones.

Results of analysis with eight zone framework

For this part of the analysis Limerick county was subdivided into eight non-overlapping zones (Fig. 1a). The criteria underlying the definition of these zones, which represent local labour markets, are set out in Walsh (1980, p. 42). With this organisational framework 32 per cent of the work trips are interzonal. The O_i and D_j distributions are contained in Walsh (1980), while the actual road distance between zone centroids is the measure used for the C_{ij} distribution.

The goodness of fit statistics for the five members of the entropy maximising family of basic interaction models are contained in Table 1. There is a steady improvement in model performance according to all of the goodness of fit criteria as the number of constraints increases. For example, the G index decreases from 18·26% for the unconstrained model to approximately 14·0% for each of the constrained ones and to 7% for the origin-destination and cost-constrained model. There is very little difference between the results for

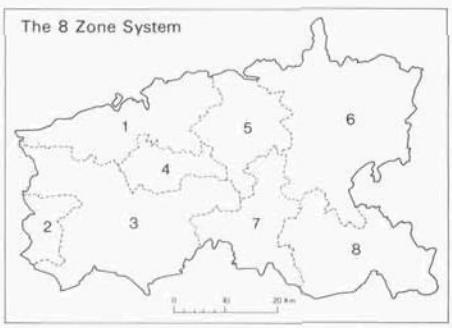

Figure 1a.

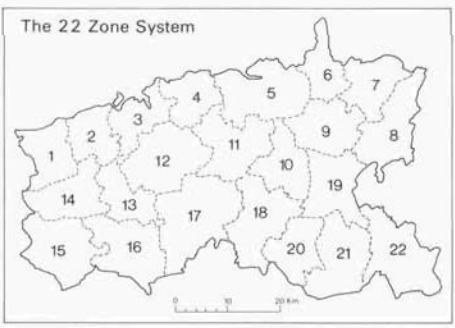

Figure 1b.

TABLE 1

THE GOODNESS OF FIT STATISTICS FOR THE EIGHT ZONE MODELS

	Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 5*
Residual X	0·000	−0·016	0·031	0·000	0·031
Residual S.D.	54·323	31·249	29·816	27·421	13·747
R^2	0·935	0·970	0·998	0·984	0·994
Intercept coefficient	8·864	−1·041	−5·754	−5·903	−0·205
Slope coefficient	0·863	1·016	1·087	1·090	1·003
χ^2	1305·132	1358·688	1017·174	946·868	330·965
G	18·263	13·974	13·962	12·840	6·993
I/I_{Max}	0·045	0·036	0·031	0·027	0·012

Model 1 is unconstrained; Model 2 is origin constrained; Model 3 is destination-constrained; Model 4 is origin and destination-constrained; Model 5 is origin, destination and cost constrained.

Models 2, 3, and 4 according to most of the criteria. Therefore, it seems that in a modelling situation where local labour markets, as defined by Walsh (1980), are the basis of the spatial organisation framework, there is little difference in the overall performance of a model which incorporates information on the distribution of employees in contrast to one which incorporates data on the distribution of employment, or one which utilizes both types of information.

When the β parameter is allowed to vary in the calibration of Model 5 there is a significant improvement in the values of each of the goodness of fit criteria. Clearly then, at this level of spatial organisation, knowledge of the average length of a work trip is an important input into the model. Finally the values of R^2 in Table 1 deserve comment. This is the only statistic which remains relatively constant across the five models. Furthermore, its magnitude is very large in all cases, in fact it is very close to unity for each of the constrained models. The weak discriminatory power of this statistic coupled with the tendency of researchers to equate good model fits with high R^2 values highlights the need for caution in the use of this measure in relation to interaction models.

Results of analysis with twenty-two zone framework

The spatial framework used in this section is based on the Superintendent Registrars' District which represent an administrative system intermediate between District Electoral Division and Rural District (Fig. 1b). The zones vary in size from 29·3 square miles in the case of zone 6 to 68·8 square miles (zone 17). The mean area of the zones is 27·0 square miles while the standard deviation is 11·25. When this framework is used, 55 per cent of the work trips are interzonal. The O_i and D_j distributions are contained in Table 2.

To isolate the effects of the zoning system on the model parameters the first set of calibrations were performed using actual road distances between zone centroids as estimates of the C_{ij} terms. This permits direct comparison with the results summarised in Table 1. The relevant goodness of fit statistics are contained in Table 3.

TABLE 2

THE O_i AND D_j DISTRIBUTIONS FOR THE 22 ZONE FRAMEWORK

Zone	O_i	D_j	*Zone*	O_i	D_j
1	42	0	12	239	243
2	139	80	13	343	703
3	185	428	14	116	0
4	79	0	15	156	70
5	1616	1040	16	212	102
6	106	1154	17	200	0
7	81	0	18	73	219
8	131	0	19	42	0
9	62	0	20	157	151
10	53	0	21	57	0
11	100	0	22	6	0

The general pattern here is very similar to that already observed in Table 1. The level of model performance improves as the number of constraints increases. There is little difference in the G values between models 2 and 3, but the destination-constrained model peforms better according to the criteria of residual standard deviation and the parameters of the regression line. Therefore, when the spatial organisation framework is as detailed as in this instance, it seems that knowledge of the spatial distribution of employment is more useful than knowledge of the distribution of employees in model construction. The results for the origin-destination constrained model are significantly better than those for either of the singly constrained ones. This is in sharp contrast to the situation with the eight zone framework (Table 1). In

TABLE 3

GOODNESS OF FIT STATISTICS FOR THE 22 ZONE MODELS USING ACTUAL ROAD DISTANCES

	Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 5*
Residual $\bar{X}$	0·014	0·036	0·000	0·027	0·009
Residual S.D.	105·680	47·467	26·892	7·727	8·173
R^2	0·707	0·764	0·935	0·992	0·990
Intercept coefficient	5·090	2·421	1·594	0·155	−0·043
Slope coefficient	0·410	0·718	0·815	0·981	1·004
χ^2	3981·026	1171·217	1521·823	320·323	318·877
G	41·632	23·891	22·673	9·474	10·263
I/I_{Max}	0·088	0·020	0·041	0·012	0·011

fact the results for model 4 here are marginally better than those for model 5. Therefore, in this context it seems that knowledge of the average length of a trip does not contribute much to model performance.

The effect of increasing the number of zones on model performance can be seen from a direct comparison of the statistics for each model in Table 3 with those in Table 1. In all instances the statistics indicate that the model performance is better when only eight zones are used. Of course, this is just what one would expect since the fewer zones there are in the first place the smaller the number of degrees of freedom for each trip. In fact when one considers the extra amount of detail which is implicit in the twenty-two zone system the actual model performance is very good. Using either models 4 or 5 approximately 90 per cent of the work trips can be correctly predicted.

TABLE 4

GOODNESS OF FIT STATISTICS FOR THE 22 ZONE MODEL USING STRAIGHT LINE DISTANCES

	Model 2	*Model 3*	*Model 4*	*Model 5*
Residual X	−0·005	0·032	−0·023	0·004
Residual S.D.	35·734	18·336	10·419	9·749
R^2	0·837	0·964	0·982	0·986
Intercept coefficient	1·510	0·984	−0·066	0·161
Slope coefficient	0·825	0·884	1·009	0·981
χ^2	988·117	1040·269	489·574	448·302
G	21·516	20·513	12·781	11·874
I/I_{Max}	0·023	0·028	0·011	0·011

TABLE 5

GOODNESS OF FIT STATISTICS FOR THE 22 ZONE MODEL USING LOGARITHMS OF ACTUAL ROAD DISTANCES

	Model 2	*Model 3*	*Model 4*	*Model 5*
Residual X	−0·009	0·000	0·009	−0·014
Residual S.D.	69·399	75·147	19·906	18·498
R^2	0·623	0·526	0·945	0·945
Intercept coefficient	3·793	4·056	0·642	0·310
Slope coefficient	0·561	0·530	0·925	0·965
χ^2	2340·513	n.a.*	909·331	984·051
G	36·098	46·134	19·260	18·604
I/I_{Max}	0·074	0·142	0·026	0·022

* not available

The effects of using different measures of the distances between zones on model parameters and performance are shown in Tables 4 and 5. The statistics in Table 4 were obtained when the straight line distances between zone centroids were used in the calibrations. Model 1 was not calibrated since its performance has been consistently much worse than that of any of the other models.

The pattern in the table is similar to what has already been observed in Tables 1 and 3. There is hardly any difference in performance between models 2 and 3. However, when the origin and destination constraints are used together there is a considerable improvement. For example, the G index is reduced from 21·5% in the case of model 2 to 12·8% for model 4. Similarly the residual standard deviation is reduced from 35·7 to 10·4. When the β parameter is allowed to vary in the calibration of model 5 there is a marginal improvement in model performance. Therefore, the principal effect of using straight line distances rather than actual ones in the calibrations is to weaken very slightly the performance of each model. For example, in the case of model 4 the G index is increased from 9·47 to 12·78.

The statistics in Table 5 show the results of calibrations when the logarithms of the actual road distances between zone centroids have been used. The general model being calibrated in this instance is of the following form

$$t_{ij}^{*} = A_i\ O_i\ B_j\ D_j\ C_{ij}^{-\beta}$$

and is usually referred to as the power-function model (Openshaw, 1976). The rationale for this model arises out of empirical studies which have shown that commuters sometimes underestimate the length of their journeys, and this is thought to be especially the case for long journeys (O'Farrell and Markham, 1974). In all instances the statistics for these models are much poorer than those contained in either Tables 3 or 4. This is particularly evident when one examines the G statistics for the different models after each calibration. The general pattern of the doubly constrained models being better than the singly constrained ones is repeated in Table 5.

Results of best twenty-two zone model

The best entropy maximising solution for equations (1) to (3) occurs with a β value of 0·25 and parameter sets (A_i) and (B_j) which are contained in Table 6. The A values have been interpreted as measures of the inaccessibility of residents of a zone to employment opportunities. It decreases in value as the locational potential of a zone as an attractor of work trips increases (Walsh, 1980). The A values range from 0·157 in zone 6, which is just east of Limerick city, to 128·09 in zone 22 which is located in the extreme south-east of the county. Generally the values tend to increase with distance from the main centres of employment around Limerick city in zones 5 and 6. Variations in the general pattern are associated with the distribution of employment opportunities. The effects of inaccessibility can be seen from a comparison of data for zones 2 and 15. The centroid of zone 2 is located approximately 23·5 miles from the same point in zone 5 and there are 80 jobs available there. On the other hand, there are 70 jobs available in zone 15, but its centroid is 30·5 miles from zone 5. The A values for the zones are 2·149 and 6·508 respectively.

The B values can be broadly interpreted as measures of the inaccessibility

TABLE 6

PARAMETER VALUES FOR ORIGIN-DESTINATION CONSTRAINED MODELS USING ACTUAL ROAD DISTANCES

Zone	*A*	*B*	*Zone*	*A*	*B*
1	8·525	5·177	12	1·129	8·097
2	2·149	9·034	13	0·902	7·227
3	0·537	21·793	14	10·127	4·091
4	1·626	16·100	15	6·508	4·522
5	0·478	5·599	16	2·954	6·202
6	0·157	29·065	17	2·308	7·083
7	0·797	28·055	18	2·340	7·881
8	4·705	7·593	19	12·704	6·309
9	4·318	7·976	20	4·668	4·861
10	4·754	9·134	21	13·986	4·933
11	2·037	9·046	22	128·091	6·021

of a zone to the workforce distribution. It increases in value as the locational potential of a zone as a source of employees decreases. The lowest values for this parameter, given the distribution of employment in the county and the locations of the zone centroids, occur in zones 1, 14 and 15, which are located on the western boundary of the county, and in zones 20 and 21, which are on the southern boundary with Co. Cork. The highest values occur surprisingly in zones 3, 4, 6 and 7 which are located immediately to the east and west of Limerick city. These large values highlight the enormous difference between the number of jobs which have been provided in these zones and the actual number of industrial workers that reside at these locations. The most serious anomaly was in zone 6 where the ill-fated Ferenka plant was located (Table 2).

The G statistic in Table 3 indicates that this model correctly predicts approximately 90% of the work trips. However, it is worth noting that there are zero entries in 74% of the cells of the actual trip matrix and that the values in a further 14% are less than the residual standard deviation.

Summary and conclusions

This paper has attempted to assess empirically the role of constraints in entropy maximising models and to evaluate the effects of zone system design and interaction cost measurement on the parameters and performance of different models.

Following the introduction of a family of elementary interaction models analyses were undertaken within spatial frameworks consisting of eight and twenty-two zones. While the results of the analyses serve mainly to clarify a number of technical problems that can arise in the application of these types of models, they also provide some insights into the geographical organisation of work trips by manufacturing employees in Co. Limerick.

The results show that generally model performance improves as the number of constraints increases. When the eight zone framework was used the level of performance of the singly constrained and origin-destination constrained models were approximately equal, but a significant improvement

occurred when the constraint on average distance was included. By contrast when the twenty-two zone framework was used there was a significant improvement in the performance of the origin-destination constrained model over the singly constrained ones, while the inclusion of the average distance constraint did not result in any improvement. Therefore, it seems that when a small number of zones, representing local labour markets are used as a framework for organising the data, the average distance constraint is an important input to the model while it seems to be of little importance when a more detailed zoning system is in use. The effect of the number of zones on model performance is that the goodness of fit statistics deteriorate as the number of zones increases. However, the decline in the performance level is only marginal. For example the G statistic increased from approximately seven to ten per cent.

The main implication of these results for intra-regional level interaction studies is that the principal data items, relating to the number of trips originating from and destined for each zone, should be organised within a relatively fine geographical framework. Walsh (1980, p. 47) has already emphasised the existence of a trade-off between the level of spatial aggregation used for data organisation and model performance. The analysis reported in this paper suggests that there may be little lost in terms of model performance if a zoning system considerably finer than the system of local labour markets advocated in the previous paper, is adopted. In fact there may be a considerable gain if one wishes to use this type of model for forecasting exercises, since with a fine zoning system it may no longer be necessary to estimate the average length (cost) of an interaction. This is probably the most difficult data item to estimate in such situations. Therefore, for intra-regional analyses of activity patterns a zoning system which is based on units intermediate in size between district electoral divisions and rural districts is probably most appropriate. However, before reaching a final conclusion on this matter it is necessary that further empirical testing be done in other regions.

The analysis has also shown that there is relatively little difference in the performance level of models which use either actual or straight line distances. This is probably due to the high correlation between these distance measures when a fine zoning system is utilised. The greatest discrepancies between the distance measures is likely to occur over long journeys, but their impact is reduced by the fact that most work trips occur over relatively short distances. When the logarithms of actual road distances were used there was a significant decline in the performance level of each model. This suggests that the actual pattern of work trips in Co. Limerick arises out of a realistic assessment of the distances between residences and work-places instead of assessments based on perceptions which tend to underestimate the length of long trips. These results in relation to distance measurement highlight the usefulness of crude straight line measurements between zone centroids as proxies for more sophisticated forms of 'cost' measurement. When good proxy measures of this type for interaction costs are available they simplify considerably the task of the model builder.

On a technical level the analyses have highlighted some inconsistencies between the various goodness of fit measures. For example, in Table 4 the G statistics for models 2 (origin-constrained) and 3 (destination-constrained) are approximately equal. Nevertheless, the residual standard deviation for model

3 is only about half of that for model 2, but the chi-square statistic for model 3 is larger than for model 2. The most disturbing result in relation to the goodness of fit measures was the weak discriminatory power of the R^2 statistic for differentiating between models. From these analyses it appears that the G statistic is the most appropriate one for assessing model performance and differentiating between competing models.

Finally, the last section of the paper demonstrated the effects of inaccessibility on manufacturing employment provision throughout the peripheral parts of the county. Analysis of the model parameters demonstrated significant discrepancies between the level of employment provision and the availability of employees in some areas. As Walsh (1980) indicates, the parameters which are output from models of this type can be particularly useful for other exercises concerned with subregional planning and management. In the Limerick region the parameters from the twenty-two zone doubly constrained model have been used in studies concerned with estimating the probable distribution of employment arising out of the location of large scale industrial projects, such as Alcan at Aughinish Island, and in a study of the most probable changes in the distribution of commuter traffic if there were changes in the propensity of commuters to travel. The model has also been used in exploratory analyses of future settlement patterns to identify locations where land ought to be acquired by the local authority for residential purposes.

REFERENCES

Hyman, G.M. 1969 'The calibration of trip distribution models', *Environment and Planning*, 1, 105-12.

O'Farrell, P.N. and Markham, J. 1974 'Commuter perceptions of public transport work journeys', *Environment and Planning*, 6, 79-100.

Openshaw, S. 1976 'An empirical study of some spatial interaction models', *Environment and Planning*, 8, 23-41.

Walsh, J.A. and O'Kelly, M. 1979 'An information theoretic approach to measurement of spatial inequality', *Econ. Soc. Rev.*, 10 (4), 267-86.

Walsh, J.A. 1980 'An entropy maximising analysis of journey-to-work patterns in county Limerick', *Ir. Geogr.*, 13, 33-53.

Wilson, A.G. 1970 *Entropy in urban and regional modelling*, London.