Geographical Research Forum
No. 4, December 1981

MEASURING SIMILARITY IN THE SPATIAL DISTRIBUTION OF TRIPS

GADI KFIR
The Israel Institute of Transportation Planning and Research, Israel

Many geographical studies are concerned with the spatial distribution of movements. These movements include, among others, personal trips for various purposes, commodity flows, migration flows, telephone calls, and the like. The term "flow pattern" is usually employed in connection with the spatial distribution of movements.When personal trips are considered an important question in the analysis of travel patterns is the degree of similarity in travel behavior between population groups or between zones (e.g. Wheeler, 1970). Various methods are employed by researchers to compare travel patterns, and can be divided into two basic groups. One group is based on the comparison of trip length distribution and the other is based on the analysis of Origin-Destination (O-D) matrices. Some of the most widely used methods are:

A. Methods based on trip length distribution: 1) visual comparison of trip length distribution curves, 2) comparing mean travel
and 3) calibration of gravity model for each group and comapring time or distance parameters (Notes, 1972).

B. Methods based on the analysis of O-D matrices: 1) visual comparison of desire lines, 2) factor analysis of flow matrices (Goddard, 1970, Wheeler, 1970), and 3) application of the transaction flows analysis method of Savage and Deutsh (1960) to movement studies (Wiseman, 1975).

A variety of other methods are employed by others such as Boyce (1965), Kansky (1967), Orrom and Wright (1976) and Smith (1970).

Simplicity is the main advantage of those methods that are based on trip length distribution, but this simplicity is achieved at a cost. The origin-destination matrix contains information on the actual origin and destination of each trip. This information is lost when the O-D matrix is transformed into a trip length distribution, where the two dimensional geographic space is converged into a one dimensional distance space. For applications that involve spatial variations in destination choice, methods that are based on the analysis of O-D matrices should be preferred.

Factor analysis of flow matrices is the widely used method for exploring dimensions of similarity in flow patterns. However, this method has been subjected to criticism, mainly because it lacks theoretical background for this application (Goddard, 1973). This carries us to another important issue. Most of

the studies in this area measure similarity or dissimilarity in travel patterns, but these terms are rarely defined and therefore it is difficult to evaluate the validity of the various measres of similarity.

This article presents a method for measuring similarity in travel patterns based on the analysis of an O-D matrix. A measure of similarity is derived logically from the definition of "travel patterns" and "similarity in travel patterns". The properties of the similarity index are investigated and finally an example of an application is presented.

TRAVEL PATTERNS

A travel pattern is produced by a group of trip makers (Boyce, 1965). Let this group be the trip making population of a given zone. Also let a metropolitan area be divided into N origin zones and M destination zones, so that:

$$i, j, n \in N$$

$$k, l, m \in M$$

Let T_{ik} be the percentage of trips from zone i to zone j, out of all trips that are generated in zone i, then the vector:

$$T_i = [T_{i1}, ..., T_{iM}]$$

will be refered to as the "Travel pattern of zone i" (actually the zone's trip making populatoin). This vector is a row of an O-D matrix, expressed in percentages and the sum of each raw is 100 percent. Percentages, rather than number of trips, are used to enable comparison between zones of different size.

SIMILARITY IN TRAVEL PATTERNS

We turn now to define "similarity in travel patterns". Before proceeding with a formal definition, two new terms will be introduced and two extreme similarity relations will be examined.

Any destination zone m is considered to be a common destination zone for origin zones i and j, if:

$$T_{im} > 0 \text{ and } T_{jm} > 0$$

Let us select randomly one hundred travellers from zone i and denote the number of travelers from this group to zone m by W_{im}. Then it is expected that:

$$W_{im} = T_{im} \quad \forall \; m \in M$$

The common share of two origin zones, i and j, for some destination zone m, is defined to be the number of travelers (out of one hundred in each zone) for which zone m is a common destination. This means that the common share is:

$$\begin{aligned} &T_{im} \text{ or } T_{jm} \text{ if } T_{im} = T_{jm} \\ &T_{im} \text{ if } T_{im} < T_{jm} \\ &T_{jm} \text{ if } T_{jm} < T_{im} \end{aligned}$$

Let us now examine two extreme situations of similarity relations between two origin zones. Let T_i and T_j be the travel patterns of zone i and zone j, respectively. Consider a situation where,

$$T_{im} = T_{jm} \quad \forall \; m \in M$$

Then, according to the definition of travel patterns, zones i and j have *identical* travel patterns and are completely similar. This situation is depicted in figure 1A. It can be seen that in the case of identical travel patterns, the common share of zones i and j for each destination zone equals to T_{im} or T_{jm}, or,

$$W_{im} = W_{jm} \quad \forall \; m \in M$$

This fact will be used later for the definition of similarity in travel patterns.

The other extreme situation is when trips from two origin zones do not share any common destination, or:

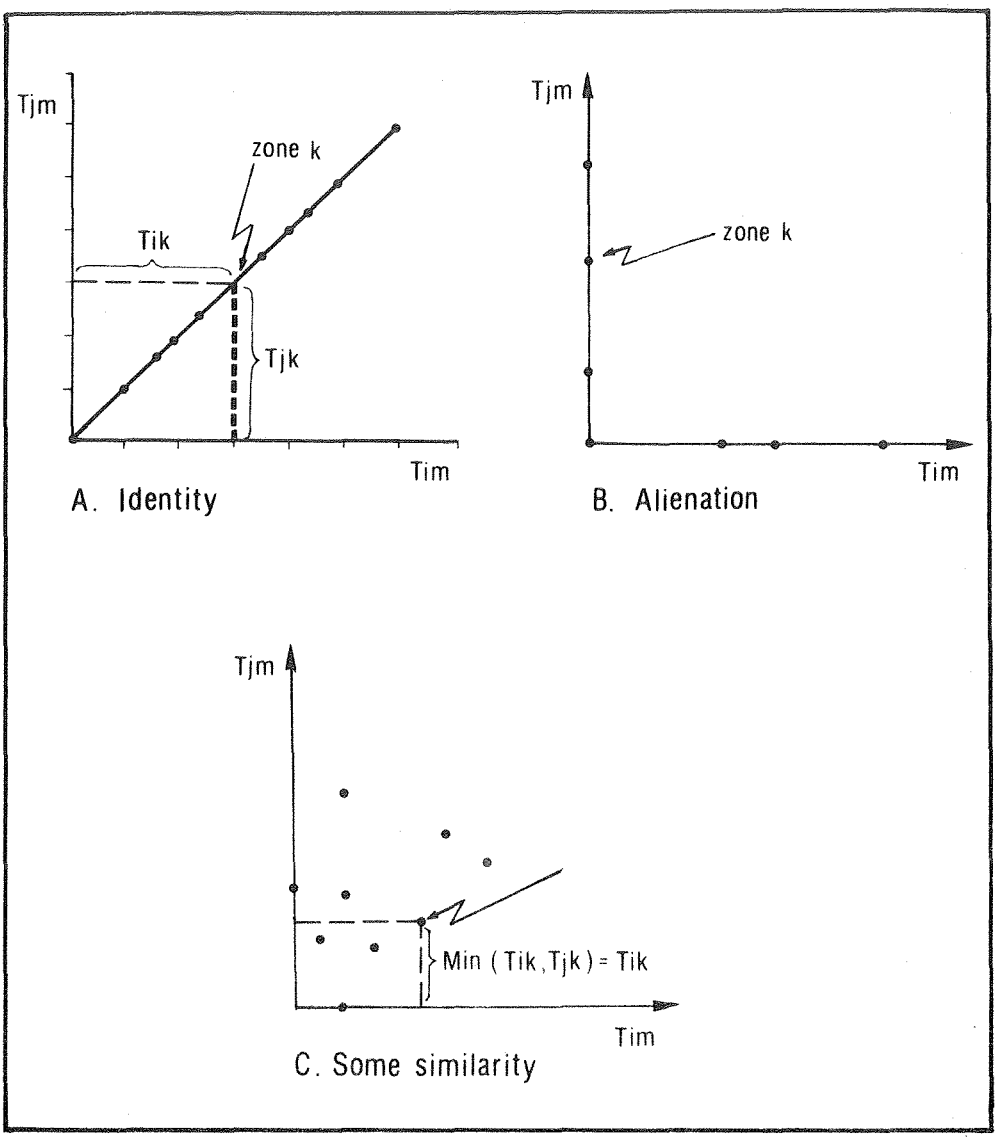

Fig. 1 : Similarity relations between zones.

$$T_{im} = 0 \rightarrow T_{jm} \neq 0 \ \forall \ m \ \epsilon \ M$$

This situation, which is depicted in figure 1B will be called *alienation* in travel patterns. In this case the similarity in travel patterns between the two zones is minimal or zero. Numerous other degrees of similarity may exist between the two extreme situations of identity and alienation. One of them is depicted in figure 1C.

It can be seen that the more travelers from two origin zones share common destinations, the higher the degree of similarity (in travel patterns) between the two zones and vice versa. When less and less travelers share common destinations, similarity decreases and tends towards alienation. Thus similarity in travel patterns can be defined as the degree of sharing common destinations by the trip makers of two zones. The degree of similarity can be measured by the total of the common shares, summed over all possible destinations; that is, the number of travelers (out of one hundred) in one zone that share common destinations with travellers of the other zone, for all destinations.

THE SIMILARITY INDEX

The similarity index, S_{ij} (or S), is directly derived from the definition of similarity:

$$S_{ij} = [\sum_{m} \text{Min}\,(T_{im}, T_{jm})] \,/100$$

The term Min (T_{im}, T_{jm}) equals the common share for some destination zone m. Summing over all destination zones M, we obtain the total common shares. The maximum value of this sum is 100. The division by 100 is done merely to obtain an index that ranges from zero to unity. The similarity index possesses the following properties:

$0.0 \leq S \leq 1.0$
$S = 1.0$ if travel patterns are identical,
$S = 0.0$ if travel patterns are alien,
$S_{ij} = S_{ji}$, and $S_{ii} = 1.0$

It can be shown that:

$$\sum_{m} \text{Min}\,(T_{im}, T_{jm}) = 100 - [\sum_{m} |\, T_{im} - T_{jm} \,|]/2$$

or

$$S_{ij} = 1 - [\sum_{m} |\, T_{im} - T_{jm} \,|]/200$$

In other words, the similarity index is a simple transformation of a well known dissimilarity (= distance) measure, a Minkowski metric where r = 1 (Fischer, 1980). This distance measure is known as Manhattan or City-Block metric in M dimensional space. This metric is widely used in psychology to measure distances between stimuli (Attneave,1950).Economists use a similar measure as an index of dissimilarity of consumption patterns (Kravis, 1958, Musgrove, 1977). The transaction flow analysis, as applied to movement studies (Wiseman, 1975), is also based on this metric. Fischer (1980) analyzes the properties of this metric and compares it to other measures of similarity, including the product moment correlation which is the basis for factor analysis.

THE CONCEPT OF INDIFFERENCE IN TRAVEL PATTERNS

For theoretical reasons there are cases where it is very interesting to compare observed similarity between zones with the similarity that would be expected under certain assumptions. The assumptions depend upon the research objectives. An interesting case is the assumption that there are no forces such as distance that regulate the spatial distribution of trips. Then each zonal trip has an equal probability to terminate in each one of the M destination zones. The probability in this case equals 1/M and the expected value of the similarity index is 1.0.

A more interesting case concerns the assumption that zones (population of the zone) are indifferent to each other (both as origins and destinations) in distributing trips to other zones. In other words, the probability that a trip generated in zone i will end up in a given zone m (P_{im}), is a random variable and $\sum_m P_{im} = 1.0$. This is equivalent to a random division of an interval 0—1 to M segments.

Let $X_1, ..., X_{m-1}$ be M—1 points chosen independently and at random from a uniform distribution in the interval 0,1. Let us denote by X(1), X(2), ..., X(M—1) the above random points rearranged in increasing order (Feller, 1971). Then:

$$a_1 = X(1)$$
$$a_2 = X(2) - X(1)$$
$$\cdot$$
$$\cdot$$
$$a_M = 1 - X(M-1)$$

represent the random probabilities P_{im}. Repeating this process we obtain $P_{jm} = (b_1, b_2, ..., b_M)$.

The joint density function for the a_i's equals (M—1)! (Feller, 1971). It can be shown that the minimum expected value between any pair of a_k, b_k is:

$$E[\text{Min}(a_k, b_k)] = 1/(2M-1)$$

and the expected value of the sum of the minima is:

$$E[\sum_m \text{Min}(a_m, b_m)] = M/(2M-1)$$

i.e., the expected value of the similarity index under these assumptions of indifference is:

$$ES_{ij} = M/(2M-1)$$

and for large enough number of destination zones M, ES_{ij} approaches 0.50. In other words, under the assumptions of indifference 50 percent of the travelers from two zones will share common destinations while the other 50 percent will not. This agrees with a common sense estimation. The standard deviation of ES_{ij} was found to be a function of the number of destination zones. The distribution of ES_{ij} still needs to be investigated; however its standard deviation was computed by simulation. For each value of M (number of destination zones), S_{ij} was computed 500 times, then the mean and standard deviation were computed. The results are shown in table 1.

It can be seen that as M increases, ES approaches 0.5 and it's standard deviation decreases. If the distribution of ES is known, one will be able to test

Table 1: Indifference: Expected Value of the Similarity Index and It's Standard Deviation.

number of destination zones (M)	ES, theoretical M/(2M−1)	ES simulated	standard deviation of ES
2	.666	—	—
4	.571	.582	.167
7	.538	.636	.137
10	.526	.530	.104
15	.517	.520	.089
20	.513	.516	.081
30	.508	.507	.063
50	.505	.507	.050
70	.503	.505	.044
100	.502	.503	.034

whether actual similarity between two zones (S_{ij}) differs significantly from it's expected value under the assumption of indifference. The deviation from indifference can be measured by the affiliation index A:

$$A = (S - ES) / ES$$

The affiliation index possesses the following properties:

$-1 \leq A \leq 1$
$A = 0.0$ indicates indifference
$A = 1.0$ indicates identity, and
$A = -1.0$ indicates alienation

Let us denote the above mentioned method for creating the vectors of probabilities P_{im} and P_{jm}, method A. It is interesting to note that when another method was used, method B, different results were obtained by the simulation. In method B, the random relative frequency of trips was obtained by selecting M independent variables with common uniform distribution. Then each variable was divided by the sum of the M variables to set their sum to 1.0. Simulation showed that in this case the expected value of the similarity index approaches 2/3 as M becomes large. Feller (1971) shows that if the variables in method B are chosen from a common exponential distribution (method C), the resulting expected value is identical to that of method A. These results suggest a possibility that method A represents a process of destination selection for a given number of generated trips, and method B represents a process in which both trip generation and distribution are random. This and other aspects of the indifference still require further investigation.

APPLICATIONS

The similarity index measures resemblance in the spatial distribution of trips. Two or more units that generate movements can be compared with each other. The index can be applied for patterns of attracted trips in which case coloumns, rather than rows of the O-D matrix, are compared. However, in this matrix coloumn totals should sum to 100 percent. Furthermore, the analysis should not be limited to matrices. For example, one may compare population groups according to their frequency of shopping trips to various types of retail outlets in a given period of time.

The similarity index is computed for each pair of analysis zones. As the number of zones increases, it becomes more difficult to study the results (the analysis of N origin zones yields $(N^2 - N)/2$ relevant indices of similarity). One method to overcome this problem of complexity is to compare the travel pattern of each zone to some theoretical pattern, thus reducing the number of similarity indices for analysis to N. However, in this method some information is lost and

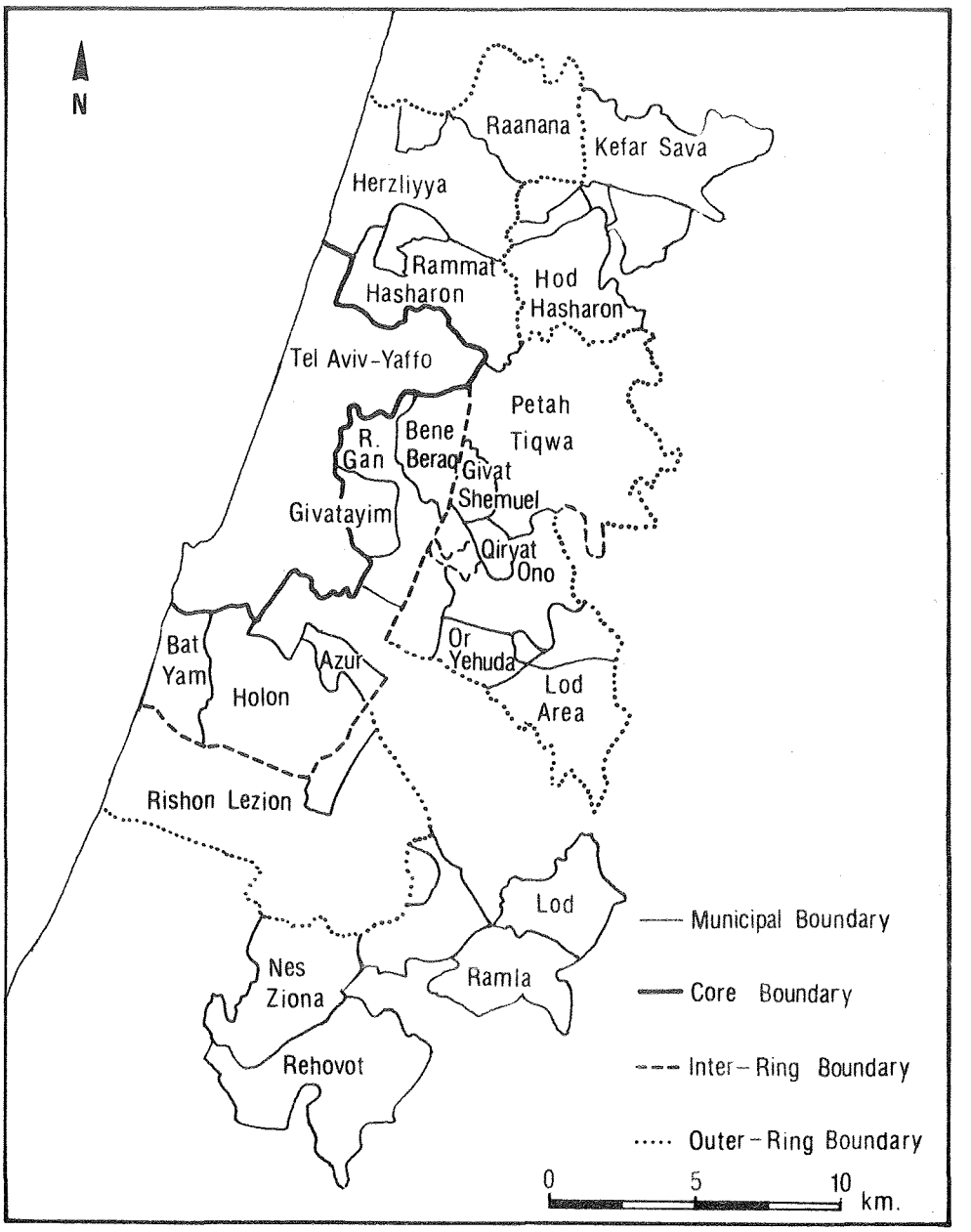

Fig. 2 : Tel Aviv metropolitan area.

still all zones should somehow be compared. A better method with more explanatory power and minimum loss of information involves the mapping of zones in two (or more) dimensional euclidian space according to their mutual similarity, using for this purpose a method of multidimensional scaling. Several methods are available for this analysis, such as Torgersson's (1958) metric method, or non-metric methods such as Smallest Space Analysis (SSA) (Guttman, 1968) or M-D-SCAL (Kruskal, 1964 A, B).

The similarity index was used to investigate the spatial variations of commuting patterns in the Tel-Aviv metropolitan area (Kfir, 1979), using data

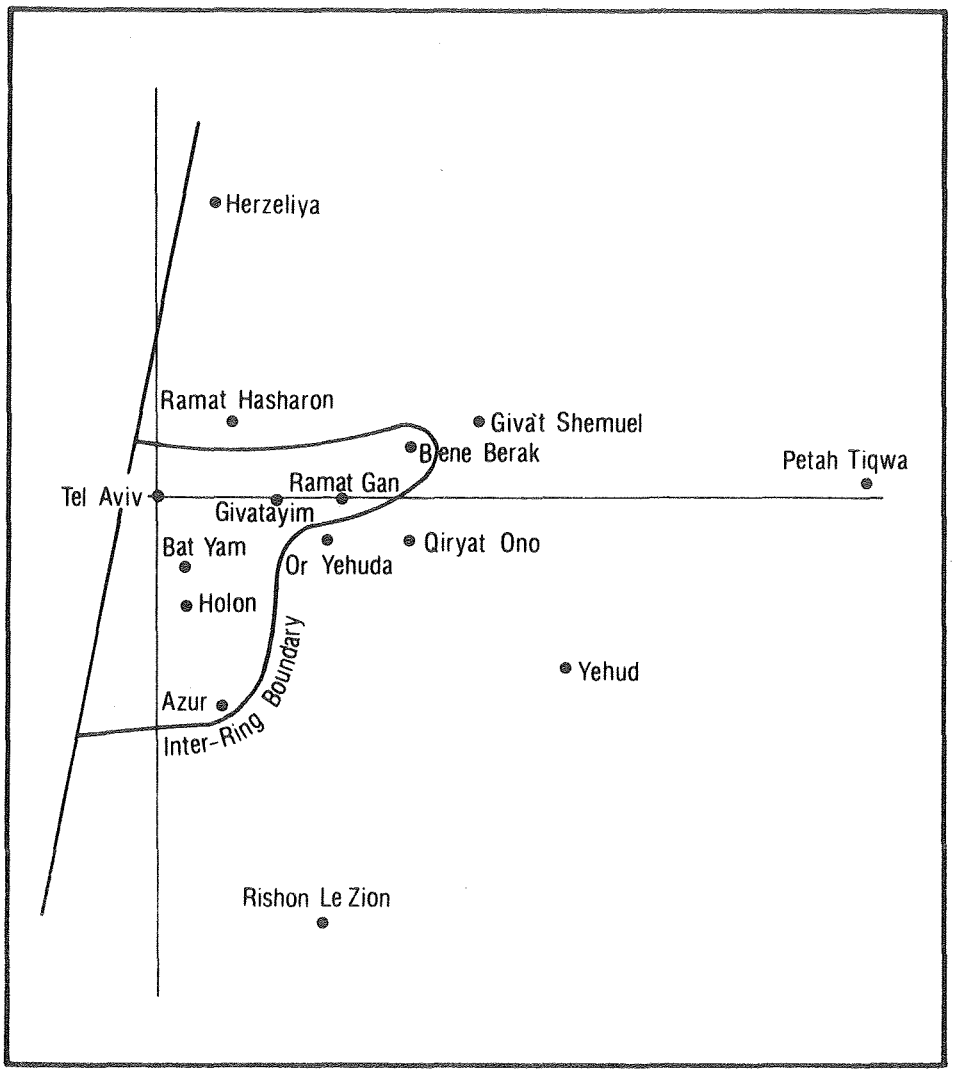

Fig. 3 : SSA Map of Towns in the Tel Aviv Metropolitan Area according to similarity in generated commuting patterns, 1973 (C.O.A = 0.12).

from a 1972/73 travel habits survey. A matrix of similarity indices between the analysis zones was computed from an O-D matrix of trips to work. A regression analysis between S_{ij} and the road distance between the analysis zones (d_{ij}) showed that distance between zones has a great influence on their similarity. The model is:

$$S_{ij} = 0.85 - 0.144(\ln d_{ij}) \text{ with } R^2 = 0.53 (S_{yx} = 0.088 \quad S.E._b = 0.02)$$

as computed with 105 observation units (pairs of zones). Smallest Space Analysis SSA—1 computer program was used to map the zones according to their mutual similarity in commuting patterns. The input to the SSA is a matrix of similarity between zones. The program fits, for every observation (zone), a point in a euclidian space so that:

$$S_{ij} > S_{fn} \rightarrow d_{ij} < d_{fn} \quad (i,j,f,n = 1, \ldots, N)$$

where d is the distance between the observations in the euclidian space. Such perfect solution will not always be possible and the relative goodness of fit is measured by the coefficient of alienation (C.O.A). A C.O.A value of zero means perfect fit. An SSA diagram with a C.O.A. of less that 0.15 is considered to be a good representation of the original similarity data.

Two examples from the Tel-Aviv commuting study (Kfir, 1979) are presented here. The analysis units are the municipalities of the metropolitan area (figure 2). The SSA map of towns according to their mutual similarity in generated commuting patterns is depicted in figure 3. Comparing the SSA map with the geographical set-up of the area reveals a high degree of correlation between the relative location of each town in the similarity space and it's actual location in the geographical space. However, due to the effect of distance friction, towns further away from Tel-Aviv (the central city with 50 percent of the metropolitan employment), are relatively further away on the SSA map. It can be seen that in the Tel-Aviv metropolitan area, location is a prime determinant of generated commuting behavior.

On the other hand, when employment zones are mapped according to their mutual similarity in commuting trip attraction, a different zone arrangement is obtained (figure 4). While location is still an important factor and adjacent zones are grouped, the arrangement of towns in the similarity space does not agree with the real world spatial arrangement. Analysis shows that employment opportunities and location are the major factors that determine attracted commuting patterns. All the zones that are located in the center of the SSA map attract workers from all parts of the metropolitan area. These zones are either employment-intensive or have positive employment balace.

SUMMARY AND CONCLUSIONS

This article presented a method for measuring similarity between movement producing (or attracting) groups, based on the spatial distribution of movements. First we defined "travel patterns" and "similarity in travel patterns". The similarity index was derived from these definitions. This derived index was found to be a simple transformation of a widely used dissimilarity index. The article presents also some initial analysis of the statistical attributes of the similarity index under the assumptions of indifference in travel behavior between groups.

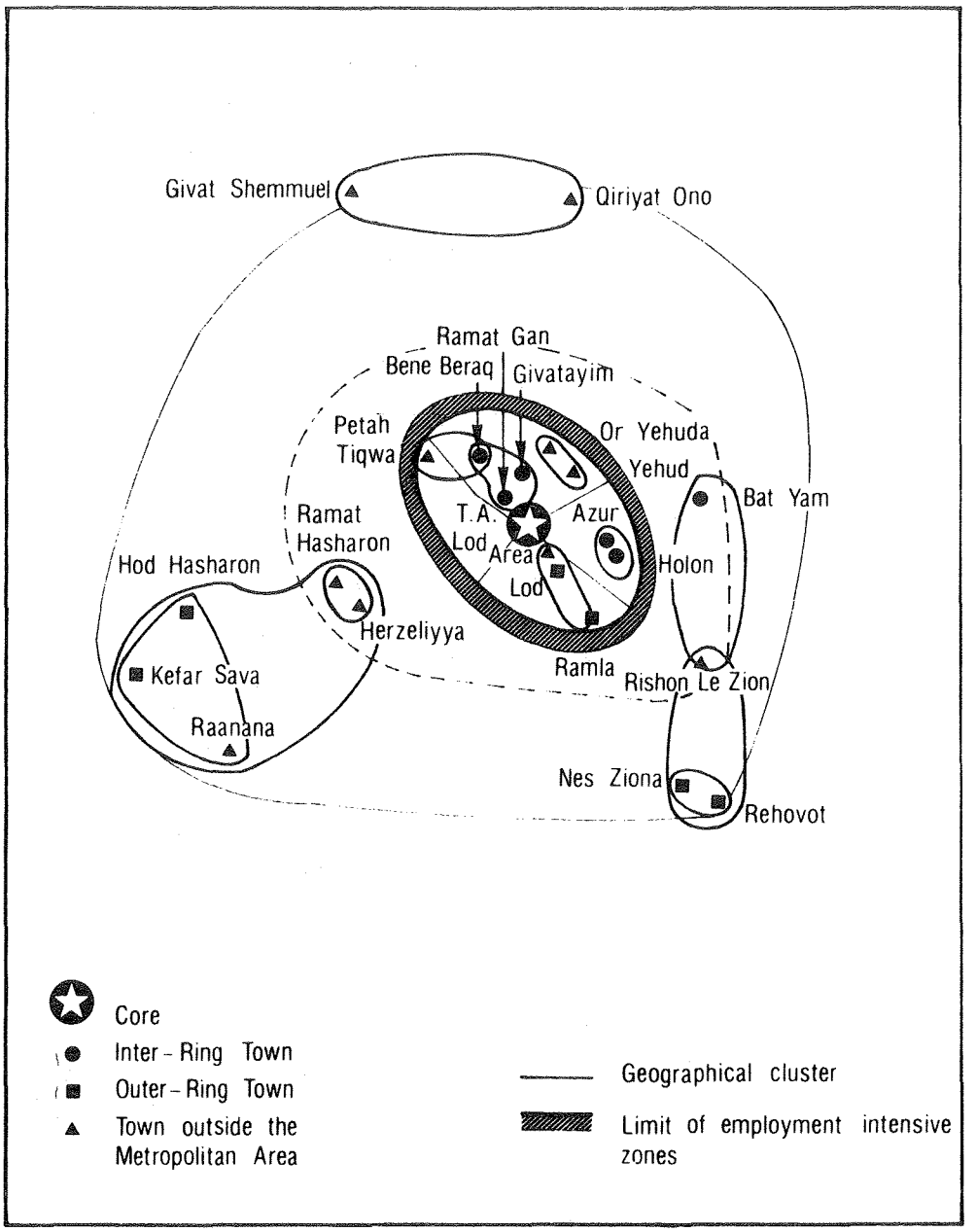

Fig. 4 : SSA Map of Towns in the Tel Aviv Metropolitan Area according to similarity in attracted commuting patterns, 1973 (C.O.A = 0.15).

The similarity index can be applied to a variety of movement studies and, in many cases, may substitute existing methods that suffer from loss of information in abstracting the phenomena of spatial distribution (such as mean trip length), or methods that are applied to movement studies but lack the theoretical foundation for such applications.

REFERENCES

Attneave, F. (1950), "Dimensions of Similarity", *American Journal of Psychology,* 63 pp. 516—556.

Boyce, D. E. (1956), *The Effect of Direction and Length of Person Trips on Distribution Models,* Ph.D. Thesis, University of Philadelphia.

Feller, W. (1971), *An Introduction to Probability Theory and it's Application,* Vol. II, New York: John Wiley.

Fischer, M. M. (1980), "Regional Taxanomy: A Comparison of Some Hierarchic and Non-Hierarchic Strategies", *Regional Science and Urban Economics,* 10, pp. 503—537.

Guttman, L. (1968), "A General Nonmetric Technique for Finding the Smallest Coordinate Space for Configuration of Points", *Psychometrica,* 31, pp. 469—506.

Goddard, J. B. (1970), "Functional Regions Within the City Center: A Study by Factor Analysis of Taxi Flows in Central London", *Transaction, The Institute of British Geographers,* 49, pp. 161—182.

Goddard, J. B. (1973), "Office Linkages and Location", *Progress in Planning,* Vol. 1 part 2.

Kansky, K. J. (1967), "Travel Patterns of Urban Residents", *Transportation Science,* 1, pp. 261—285.

Kfir, G. (1979), *Similarity in Commuting Patterns: A Quantitative Analysis of Spatial Variations in Commuting Patterns in the Tel-Aviv Metropolitan Area,* Ph.D. Thesis, the Hebrew University of Jerusalem (in Hebrew).

Kravis, I. (1968), "International and Intertemporal Comparisons of the Structure of Consumption". In Clark L.H., ed., *Consumer Behavior,* New York: Harper.

Kruskal, J. B. (1964A), "Multidimensional Scaling: A Numerical Method", *Psychometrica,* 29, pp. 1—27.

Kruskal, J. B. (1964B), "Multidimensional Scaling by Optimizing Goodness of Fit", *Psychometrica,* 29, pp. 115—129.

Musgrove, P. (1977), "The Structure of Household Spending in South American Cities, Index of Dissimilarity and Causes of Inter-City Differences", *Review of Income and Wealth,* ser. 73 No. 4, pp. 365—384.

Notes, C. B. (1972), "Access to Jobs and the Willingness to Travel", *Highway Research Record,* 392, pp. 143—146.

Orrom, H. C. and C. C. Wright (1976), "The Spatial Distribution of Journey-to-work Trips in Greater London",*Transportation,* 5, pp. 192—222.

Savage, R. and K. W. Deutsch (1960), "A Statistical Model of the Gross Analysis of Transaction Flows", *Econometrica,* 28, pp. 551—572.

Smith, R.H.T. (1970), "Concepts and Methods in Commodity Flow Analysis", *Economic Geography,* 46, pp. 404—416.

Torgersson, W. S. (1958), *Theory and Methods of Scaling,* New-York: John Wiley.

Wheeler, J. O. (1970), "The Structure of Metropolitan Work Trips", *The Professional Geographer,* 22, pp. 152—158.

Wiseman, R.F. (1975), "Location in the City as a Factor in Trip Making Patterns", *Tijdschrift Voor Economische en Sociale Geographie,* 66, pp. 167—177.