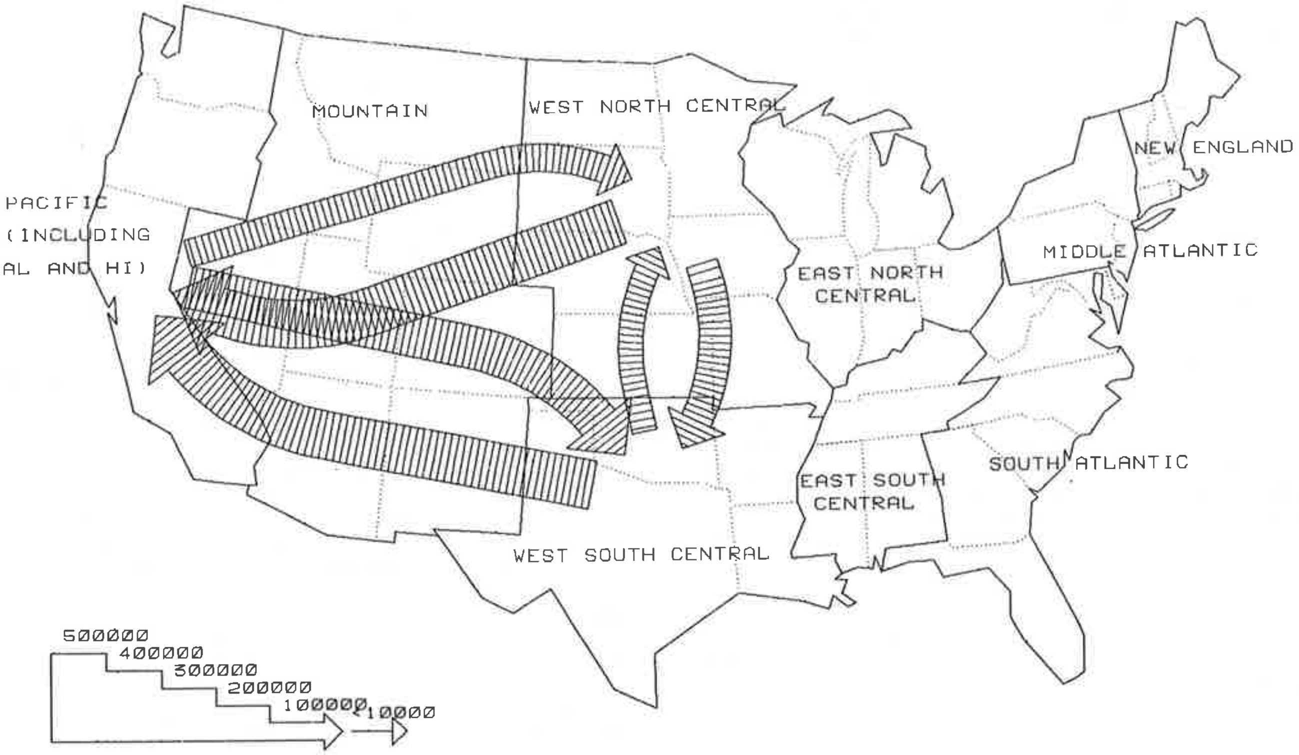
Figure 15. U.S. migration, 1965-1970.



ACKNOWLEDGMENT

The research described in this report was funded in part by a grant from the National Cancer Institute to the Fred Hutchinson Cancer Research Center in Seattle, Washington.

REFERENCES

1. W. Tobler. A Geographic Flow Mapping Program. Department of Geography, Univ. of California, Santa Barbara, 1979.
2. R.I. Wittick. A Computer System for Mapping and Analyzing Transportation Networks. Southeastern Geographer, Vol. 16, No. 1, May 1976, pp. 74-81.
3. N. Kadmon. Komplot: Do-It-Yourself Computer Cartography. The Cartographic Journal, Vol. 8, No. 2, 1971, pp. 139-144.
4. D.A. Plane. Estimation of Place-to-Place Migration Flows from Net Migration Totals: A Minimum Information Approach. International Regional Science Review, Vol. 6, No. 1, 1981, pp. 33-51.

# Graphical and Mathematical Methods in the Analysis of Urban Gasoline Consumption

ROBERT M. RAY

**Some preliminary results are presented of a study that focuses on urban form efficiency. Results are based on a new method for analyzing dimensions of urban form. This method employs linear programming as a technique for measuring the spatial congruence of urban population and activity distributions. By using daytime employment distributions derived from data of the travel-to-work supplement of the annual housing survey (1975-1977) and 1970 census population data, the spatial congruence of population and employment is computed for 23 standard metropolitan statistical areas (SMSAs). A statistically significant positive correlation is found between these measures and per capita gasoline consumption across these SMSAs. The method employed requires no assumptions regarding travel behavior within SMSAs. Thus, like graphical models, it offers a purely descriptive means of analyzing urban form. The method also permits measurement of some of the spatial consequences of social distance relations among urban populations. Thus, it may provide an effective means of combining both functional and social distance factors within a common framework for descriptive analysis of urban spatial structure.**

This paper reports some preliminary results of a study under way at the Research Triangle Institute for the U.S. Department of Transportation (DOT) that looks at alternative methods for comparing the spatial structure of U.S. cities. Both quantitative and graphical methods are being used in a productive manner. A primary objective of this research is to explore the relation between urban form and gasoline consumption. We hope that, by identifying characteristics of spatial structure that correlate strongly with rates of gasoline consumption, we will better understand this important aspect of urban form. This should prepare us to encourage more efficient courses of urban development through plans and policies.

## PROBLEM

Urban spatial structure (urban form, spatial organization, order) is an extremely complex concept that does not lend itself easily to quantitative treatment. Where it has been entered into statistical analyses, it has often been poorly represented by available variables and measures.

Proxy boundaries of urban space, such as standard metropolitan statistical area (SMSA), urbanized area, place, central business district (CBD), and major retail trade area are borrowed directly from the U.S. Bureau of the Census. The criteria used by the Bureau for these definitions may be largely irrelevant to the purposes of the individual researcher. However, even the concept of density, as it is employed in most studies, depends in a fundamental way on these definitions. For U.S. cities, centrifugal forces of accessibility, telecommunications, and municipal fragmentation have greatly reduced the usefulness of these concepts.

Given the decentralization and general defocusing of activity distributions in U.S. cities, various measures of spatial dispersion might seem appropriate for describing urban spatial structure. A good comprehensive presentation of the variety of measures available has been given by Neft (1). Unfortunately, the majority of these measures (like the concept of variance itself) assumes unimodal distributions. Since multimodality, multinucleation, or at least acute lumpiness is a frequent property of urban landscapes, the value of most of these concepts is also limited. They transmit ambiguous messages.

Other difficulties concern the measurement of areal association (spatial congruence, dissimilarity, proximity) between urban population and activity distributions. Here, a serious problem is that most measures depend in a fundamental way on the geographic scale of subareas (tract groups, tracts, blocks) selected for spatial sampling. Duncan and others (2) provide a good review of problems here. Analyses conducted at the tract level yield different results when conducted at the block level. Furthermore, most of these methods do not permit us to measure spatial associations between items reported by different area units. This condition has impacted critically much urban research where often important variables are collected by independent agencies by using different data-reporting systems.

## PROMISE OF GRAPHICAL AND SPATIAL ANALYSIS METHODS

All of the above difficulties suggest that an important role will exist for computer graphics and computer cartography in the study of urban space for some time to come. Since so many interesting aspects of spatial structure such as complexity, dispersion, and lumpiness are more easily discussed with visual displays, and since such spatial properties have shown themselves to be numerically (as well as verbally) difficult, complete maps are often the only means sufficient for characterization of urban space. Computer graphics provide a practical method for producing such displays in the quantities and time frames needed.

Computer cartography permits visual analysis of spatial variation across cities; however, it does nothing per se toward increasing our ability to treat such variation quantitatively. For this we still require numerical methods. Nevertheless, graphics has a role to play in the development of such methods.

Mathematically, the description of urban spatial structure is akin to the problem of feature extraction in visual pattern recognition research. To gain confidence that a specific numerical index consistently measures some characteristic of spatial structure (say, for example, distribution multimodality), we need to compute the index for a large number of sample cases and compare numerical outputs with visual displays of data inputs. We hope that the index yields automatically a rank ordering of all samples that agrees closely with the rank ordering obtained by visually comparing all samples with respect to the spatial feature. Again, computer cartography provides a means for displaying such samples in comparable format and in the quantities needed.

In our present research, we are examining the usefulness of several new measures of spatial pattern properties that seem particularly well equipped for analysis of urban space (3). In particular, we are attempting to demonstrate that, with an appropriate redefinition of system variables, some of the same mathematical models used for transportation network analysis may be used profitably for more general analysis of spatial relations between distributions.

Specifically, the models derived appear to yield unambiguous measures of the extent of spatial congruence between population and activity distributions. Because gasoline use is usually associated with the movement of people between activities, per capita rates of gasoline consumption will be positively correlated with the distances that separate activity distributions.

One of the most important spatial relations that defines the form of U.S. cities and determines greatly their gasoline consumption is the distance between place of residence and place of work. This one relation is probably more important than we would initially think because work places for many are often the shopping, eating, and entertainment places for many others. Thus, we suspect that much of the essential character of U.S. cities is contained in the information conveyed directly by population and employment distributions. Cities that have revitalized inner-city neighborhoods are, in general, cities with stable concentrations of employment in downtown areas. On the other hand, lively daytime downtowns become quite dead at night if at 5:00 p.m. everyone returns home to the distant suburbs.

## METHOD SELECTED FOR USE

Let X (nx1) and Z (nx1) be discrete probability distributions of some two urban variables over some area sampling frame. For example, X might be the distribution of total population by place of residence across census tracts and Z might be the tract-level distribution of total employment by place of work. In this case, $x_i$ is the probability that a randomly selected person lives in tract i, and $z_j$ is the probability that any worker works in tract j.

Let C (nxn) be a matrix whose elements ($c_{i,j}$) are some function of the distance between tracts i and j. Then, a measure of the overall distance that separates the two spatial distributions X and Z may be formulated as follows:

$$D = \min \sum_{i=1}^{n} \sum_{j=1}^{n} q_{i,j} c_{i,j} \qquad (1)$$

subject to

$$\sum_{j=1}^{n} q_{i,j} = x_i \quad i=1,\ldots,n \tag{2}$$

$$\sum_{i=1}^{n} q_{i,j} = z_j \quad j=1,\ldots,n \tag{3}$$

$$q_{i,j} \geqslant 0 \quad i,j=1,\ldots,n \tag{4}$$

Equations 1-4 will be recognized by many as a special case of linear programming known as the transportation problem (4,5). If X and Z are given in terms of absolute supply and demand quantities instead of probabilities, and C is defined in terms of network travel costs (distances, times), then the problem is to determine a set of flows from supply sites to demand sites that minimizes aggregate travel costs. The transportation model of linear programming finds wide application in urban planning for both descriptive and normative models of traffic flow between activities.

Our interest in this model, however, stems not from its ability to describe or prescribe efficient traffic patterns but rather from its usefulness as a method for measuring proximity relationships generally among spatial distributions. If X and Z are probability vectors for any two spatial distributions and C is a matrix of Euclidean distances between area units, then D may be interpreted directly as a measure of the spatial congruence that exists between the two distributions.

Note that such a measure has an interpretation of distance in the abstract that is completely independent of any material flow between the two distributions. Thus, in our example above that measures the distance between total population and total employment, it makes little difference that many people never go to work. The measure has a mathematical meaning that is valid apart from any assumption of traffic between the two distributions. Namely, the measure reports simply the physical congruence of the two distributions. It does this in a manner completely neutral to any real or simulated patterns of home-to-work commuting. Similarly, such a measure may be used to quantify the degree of residential distance between racial, ethnic, income, and employment groups in a city in cases where it is meaningless to think of traffic flows between distributions.

If in Equations 1-4 we define C as a matrix of squared Euclidean distances between area frame units, we arrive at an index of interdistribution distance with still more interesting mathematical properties. We call this measure pattern distance. This least-squares index can be shown to be decomposable into a set of additive terms that express the squared distance between the centroids of the two distributions X and Z, the spatial variance of X, the spatial variance of Z, and a fourth term that represents the composite covariance of spatial coordinates matched optimally between the two distributions by the matrix Q. Thus, pattern distance measures the extent to which two spatial distributions share the same location, the same dispersion, and the same shape. The square root of this measure can be translated directly into units of physical distance (e.g., miles or kilometers).

Pattern distance may be viewed as an extension of an index proposed earlier by Bachi (6). By measuring the distance between spatial distributions of population before and after migration, Bachi developed an index that has identical decomposition properties that he called quadratic averages of squared distances. However, Bachi did not specify tight limits for the range of his measure. Here, linear programming allows us to determine for any two distributions the minimal feasible value of Bachi's distance.

Both measures have the desirable property that they depend only in a minor way on the number and size of subareas of the area sampling frame. In fact, both may be computed between pairs of distributions sampled with respect to different frames, provided that both frames can be described by using a common coordinate system. Of course, neither measure requires any assumption concerning the unimodality of distributions.

Our present research concerns the relation between urban form and gasoline consumption across U.S. cities. We would like to show that urban spatial structure can be meaningfully characterized in terms of pattern distance relations between population and employment distributions and that differences in spatial structures so characterized can be related statistically to variations in gasoline consumption. In posing the problem this way, it is our hypothesis that the spatial patterning of phenomena within cities directly conveys much information about spatial efficiency. This pattern analysis method permits us the convenience of macroscopic study of spatial systems despite incomplete microscopic specification. Thus, it functions much like graphical methods and allows comparative analysis to proceed in the absence of complete data and operational behavioral models.

## DATA

To carry out our study we needed data that describe population and employment distributions in comparable format across some sample of U.S. cities. We also needed data that indicated comparable levels of gasoline consumption for these same cities.

Decennial census data, of course, are our best source of data on population distributions. Since 1980 census data were not available to our project, we are using 1970 data. Our access to these data has been greatly aided by the special area profile tapes created by the Bureau in conjunction with the urban atlas project. These summary files provide a convenient source of tract-level 1970 census data for a large number of cities.

Data on employment distributions across cities are not as easily found. Problems in this area, however, should diminish greatly in the future as a result of the travel-to-work items included on the 1980 census questionnaire and new methods adopted by the Bureau for place-of-work address coding. (In 1970, similar travel-to-work information was sampled but, due to questionnaire instrumentation and post-coding methods, only low rates of place-of-work address coding resulted.)

The only convenient source of comparable employment distributions across cities that we have been able to find is the travel-to-work supplement of the annual housing surveys (AHS) of 1975 and 1976. The tract-to-tract files available from these surveys provide a set of comparable home-to-work commuting data for 41 SMSAs. On these files, home-to-work traffic volumes are cross-tabulated by census tract of employment and tract of residence. By aggregating trips across tracts of origin for each tract of destination we can obtain an estimate of the total daytime employment in each tract.

Unfortunately, the AHS tract-to-tract files allow us to break down these tract-level employment totals along only two dimensions: (a) mode of travel to work and (b) time of departure from home. We know nothing about occupations or industry types for the employment of each tract. We know only how many people came to work there, how they got to work, and what time they left home. Despite these limitations, this information appears useful for our purposes.

Fortunately, the Bureau used 1970 census geographic definitions for all of the AHS waves. This means that the 41 AHS SMSA data files can be matched directly with 1970 census files. SMSA and tract codes are identical across the two data sources. In choosing to work with 1970 population and 1975-1976 employment distributions we simply make the assumption that population and employment patterns in the AHS SMSAs did not shift radically during the early 1970s.

To compute pattern distances between population and employment distributions we need geographic coordinates for the centroids of all census tracts in all SMSAs. It is possible to obtain such coordinates from the tract boundary files created by the Bureau, again in conjunction with the urban atlas project. We used the Bureau's EASYCORD program to compute the areas and centroids for all tracts in the 35 SMSAs common to both the AHS and the tract boundary files.

Data that describe levels of gasoline consumption across cities are available in a restricted form from the census of retail trade, which is conducted by the Bureau every five years. These data are reported in machine-readable format at both county and SMSA levels. Of course, SMSA definitions change frequently as additional counties are added to each, and so the SMSA retail trade data are not comparable over time. Since, however, SMSAs outside of New England are proper collections of counties and county equivalents, it is possible for most of the U.S. to aggregate county-level retail trade data to the level of 1970 SMSAs. We have done this for the data of the 1972 and 1977 retail trade censuses. The 1967 data have already been aggregated and reported in machine-readable format by 1970 SMSAs by the Bureau in the 1972 County and City Data Book (CCDB) files. Due to data reporting formats and SMSA definition changes in the 1970s, only these 1967 retail trade data of the 1972 CCDB are available to describe 1970 New England SMSAs. The 1972 CCDB files also provide a convenient machine-readable source of other SMSA-level 1970 census variables, such as total population, density, percentage urban, median income, and transit ridership, which we also want to include in our analysis.

For research purposes, one problem with all of the machine-readable retail trade files is that they report only total dollar sales of gasoline service stations. In printed reports, additional data are given for total gallons of gasoline by selected counties and for some censuses by selected SMSAs. Other printed materials permit factoring of service station sales by states into component trade items. These figures suggest that gasoline sales make up about 70 percent of all station sales across all states.

At this point we have not resolved to our complete satisfaction all data problems in this area. To date, we have contented ourselves with the use of the dollar sales data as a proxy for gasoline consumption. To improve marginally the accuracy of this variable, we have divided it by a weighted, state-level estimate of the per gallon price of gasoline in each SMSA. We hope in future months to be able to key other data from the retail trade census printed reports and to further improve the accuracy of this proxy.

## RESULTS

Table 1 presents some preliminary results of our study. Pattern distances computed between total tract-level population and employment distributions are reported for 23 AHS SMSAs. The square root of pattern distance has been listed, because this measure translates directly into units of miles and therefore more strongly correlates with actual travel distances. Table 1 also gives for all SMSAs the proxy for 1972 gasoline consumption (i.e., total gasoline service station sales in 1972 divided by average 1972 retail prices and 1970 population) along with several other 1970 census variables. SMSAs are ranked in order of increasing pattern distances between population and employment distributions.

Figures 1 and 2, which sketch tract-level distributions of daytime employment for the Milwaukee and Houston SMSAs, illustrate the principal dimension of the data of Table 1. Relative to Houston, Milwaukee is much more spatially compact, and thus distances

Table 1. Root pattern distance between population and employment distributions and other data items for selected SMSA.

| SMSA | Population in 1970 (000s) | Density (persons/mile$^2$) | Percentage Workers Commuting by Transit | Gasoline Use [total station sales in 1972/(price x 1970 population)] | Root Pattern Distance Between Population and Employment (miles) |
|---|---|---|---|---|---|
| Milwaukee, WI | 1403 | 964 | 12.0 | 406 | 2.10 |
| Allentown, PA-NJ | 543 | 501 | 3.2 | 417 | 2.12 |
| Columbus, OH | 916 | 613 | 8.1 | 445 | 2.18 |
| Miami, FL | 1267 | 621 | 9.1 | 449 | 2.26 |
| Omaha, NE-IA | 540 | 353 | 7.2 | 463 | 2.27 |
| Buffalo, NY | 1349 | 848 | 10.4 | 320 | 2.33 |
| Cleveland, OH | 2064 | 1359 | 13.3 | 412 | 2.46 |
| Denver, CO | 1227 | 336 | 4.4 | 443 | 2.56 |
| Louisville, KY-IN | 826 | 911 | 6.7 | 442 | 2.68 |
| St. Louis, MO-IL | 2363 | 574 | 8.1 | 460 | 2.81 |
| Oklahoma City, OK | 640 | 300 | 1.6 | 435 | 2.84 |
| New Orleans, LA | 1045 | 530 | 20.4 | 387 | 2.93 |
| Atlanta, GA | 1390 | 805 | 9.4 | 576 | 2.93 |
| Grand Rapids, MI | 539 | 380 | 2.2 | 526 | 2.98 |
| San Diego, CA | 1357 | 319 | 4.3 | 431 | 3.12 |
| Raleigh, NC | 228 | 267 | 3.8 | 455 | 3.26 |
| Portland, OR-WA | 1009 | 276 | 6.0 | 449 | 3.32 |
| Kansas City, MO-KS | 1253 | 454 | 5.5 | 531 | 3.37 |
| Indianapolis, IN | 1109 | 361 | 5.8 | 556 | 3.54 |
| Sacramento, CA | 800 | 234 | 2.3 | 500 | 3.59 |
| Houston, TX | 1984 | 316 | 5.4 | 479 | 3.74 |
| Cincinnati, OH-KY-IN | 1384 | 644 | 8.3 | 456 | 3.94 |
| Seattle-Everett, WA | 1421 | 337 | 7.1 | 477 | 4.01 |

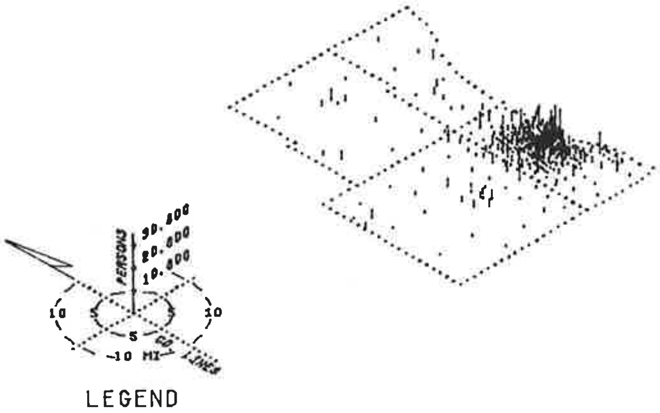
between population and employment distributions are smaller and per capita gasoline consumption is less. Note also the difference in transit use between the two cities.

Table 2 presents the correlations that exist among the variables of Table 1. The pattern of correlations suggests that some collinearity exists among the variables of pattern distance, gasoline consumption, density, and transit use. The data indicate that SMSAs that have low per capita gasoline consumption tend to be SMSAs that have large populations, high densities, larger percentages of workers who commute by transit, and smaller distances between population and employment distributions. None of this is particularly surprising.

What is surprising is that, of all variables, the measure of pattern distance has the strongest correlation (r = 0.49) with the proxy for gasoline consumption. Furthermore, a multiple regression of gasoline consumption on all of the other variables of Table 1 indicates that the pattern distance variable remains the most significant predictor of gasoline consumption even after the variables population size, density, and transit use are entered into the regression. Thus, pattern distances between population and employment distributions appear quite useful for analysis of urban form-efficiency relationships.

Figures 3-6 assist our understanding of the value of the pattern distance measure. From the data of Table 1, we might expect that gasoline consumption would be higher in Columbus than in Cincinnati. The density of Cincinnati is slightly greater than that of Columbus. Transit use in Cincinnati is slightly greater than transit use in Columbus. However, the pattern distance computed between population and employment for Cincinnati is almost twice that determined for Columbus. To an extent, this difference between the spatial structures of the two cities is visible in Figures 3-6. Can this be the explanation for the slightly higher rate of gasoline consumption in Cincinnati?

Figure 1. Milwaukee daytime employment by tracts, 1976.



SUMMARY DISCUSSION

Of course, the relation between gasoline consumption and urban structure across U.S. SMSAs is much more complex than the few data above could describe. In more comprehensive analyses, we are also examining the significance of other predictors of SMSA gasoline use such as through traffic, tourism, climate, local automobile fleet efficiencies, and economic conditions. At this point, our analysis is incomplete. However, we can report that a preliminary reading of all data available indicates that the pattern distance measure will survive the challenges of alternative explanations.

More detailed descriptions of employment distributions across SMSAs would be nice. For example, a breakdown of employment by at least broad occupational or industrial classes would be useful. Such data could be used as a surrogate for land use. We could begin to explore the relative efficiencies of alternative land use structures and mixtures. There is some hope that data such as these will be available from the 1980 census as a result of the place-of-work address coding now under way at the Bureau for the transportation planning community.

Since the currently available AHS employment data are cross-tabulated by mode of travel and the 1970 census data provide numerous tract-level population distributions, we are computing pattern distances between employment by two modes of travel (driving alone and public transportation) and residential distributions for two broad classes of workers, (white-collar and blue-collar workers). This may allow us to say something about the relative effects on SMSA gasoline consumption of the residential distribution of these two occupational groups with respect to general employment opportunities and tran-

Figure 2. Houston daytime employment by tracts, 1976.

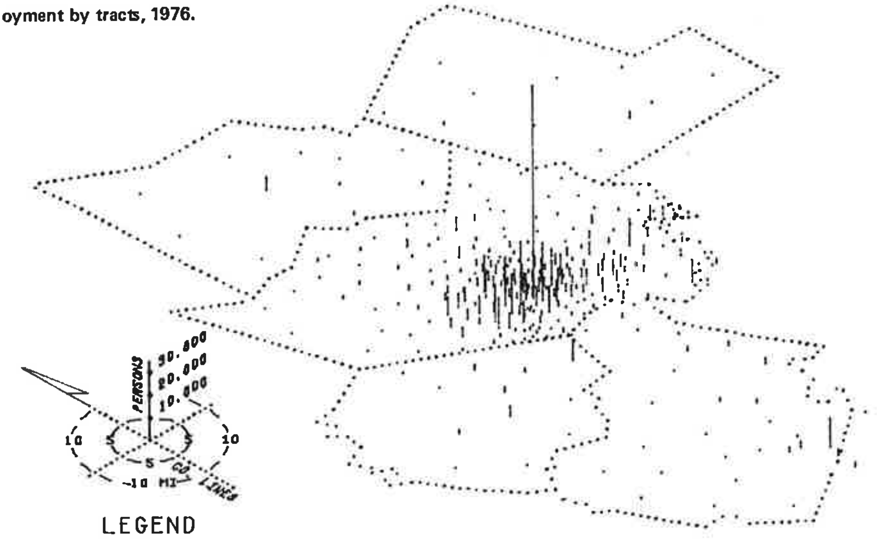


Table 2. Correlations between SMSA variables of Table 1.

| Variable | Population | Density (persons/mile$^2$) | Commuting by Transit | Gasoline Use | Pattern Distance |
|---|---|---|---|---|---|
| Population | 1.00 | 0.43[a] | 0.40[b] | -0.06 | 0.12 |
| Density | | 1.00 | 0.60[c] | -0.33 | 0.48[a] |
| Transit use | | | 1.00 | -0.40[b] | -0.25 |
| Gasoline use | | | | 1.00 | 0.49[a] |
| Pattern distance | | | | | 1.00 |

Note: All variables are defined in Table 1.

[a] Statistically significant at 0.05 level.

[b] Statistically significant at 0.1 level.

[c] Statistically significant at 0.005 level.

Figure 3. Columbus daytime employment by tracts, 1976.

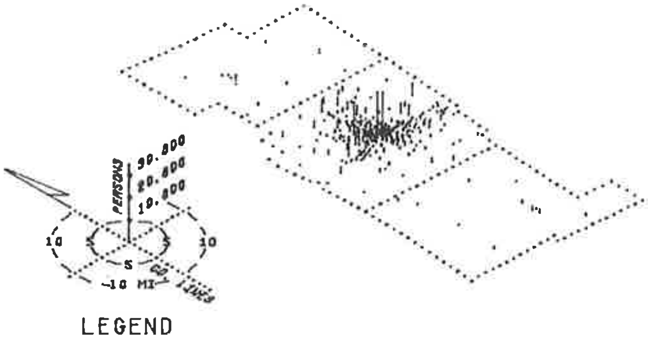


Figure 4. Columbus nighttime population by tracts, 1970.

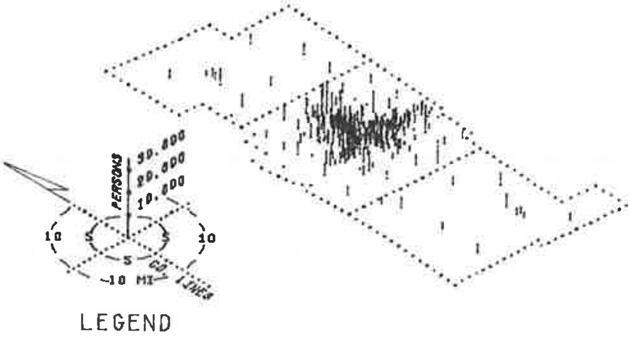


Figure 5. Cincinnati daytime employment by tracts, 1976.

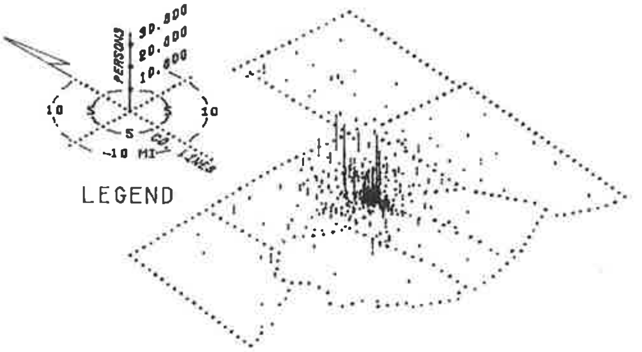
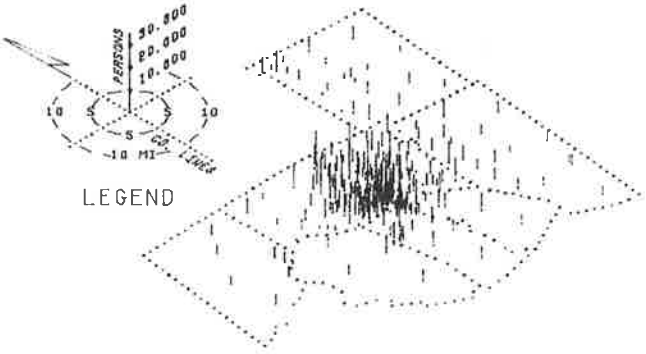


Figure 6. Cincinnati nighttime population by tracts, 1970.



sit services. By examining the distance between the residences of these two occupation groups, we may find that the social clustering of occupation groups in space (i.e., the pattern distance between white collars and blue collars) proves just as significant a predictor of gasoline consumption across cities as the distance between homes and work places. In a similar manner, we are investigating the extent to which racial residential segregation influences gasoline consumption across SMSAs.

In any case, analyses suggest that the pattern distance measure will prove advantageous to our present study of the relation between urban spatial structure and gasoline efficiency. For reasons mentioned earlier, we suspect that the concept can be shown to be useful in other areas of urban research as well. The method appears to work much in the manner of graphics, allowing us to sift through large quantities of spatial data to determine those spatial relations that make a difference.

Techniques of spatial analysis have the advantage over graphics that they tie directly to methods of statistical hypothesis testing. Occasionally, they have the disadvantage of being somewhat more difficult to compute. The data base requirements of the two methods are for the most part identical and modest. Since the most severe problem of urban research is often the insufficiency of information with respect to the questions that we must try to answer, it seems appropriate to devote resources to the development of such purely descriptive methods that offer the potential of insight despite the incompleteness of our data.

## ACKNOWLEDGMENT

This work was supported at the Research Triangle Institute by the Research and Special Programs Administration, U.S. Department of Transportation. The views expressed in this report are mine and do not necessarily represent the views of the Research Triangle Institute or the U.S. Department of Transportation. An earlier version of this paper was presented at Harvard Computer Graphics Week in Cambridge, Massachusetts, July 27-31, 1981. I wish to thank the Harvard Laboratory for Computer Graphics and Spatial Analysis for permission to reprint this paper.

## REFERENCES

1. D.S. Neft. Statistical Analysis for Areal Distributions. Regional Science Research Institute, Philadelphia, Monograph Series, No. 2, 1966.
2. O.D. Duncan, R.P. Cuzzort, and B. Duncan. Statistical Geography: Problems in Analyzing Areal Data. Free Press of Glencoe, Glencoe, IL, 1961.

3. R.M. Ray, III. An Entropy Maximization Approach to the Description of Urban Spatial Organization. Department of City and Regional Planning, Univ. of North Carolina at Chapel Hill, Chapel Hill, Ph.D. dissertation, May 1977.
4. G.G. Dantzig. Linear Programming and Extensions. Princeton Univ. Press, Princeton, NJ, 1963.
5. R. Dorfman, R.A. Samuelson, and R.M. Solow. Linear Programming and Economic Analysis. McGraw-Hill, New York, 1958.
6. R. Bachi. Statistical Analysis of Geographical Series. Bulletin de l'Institut International de Statistique, Vol. 36, 1957, pp. 229-240.