Correlates of Day-Hiking Travel: The Effects of Aggregation

Daniel R. Fesemaier, Michael F. Goodchild, and Stanley R. Lieber

ABSTRACT: *The general distance decay model has been applied in numerous studies to model outdoor recreation participation. Using fourteen socioeconomic characteristics of the household, this study identified the correlates of the "resistance" to travel parameter of the distance decay model for three levels of generalization: City, County, and Planning Region. The effects of aggregation are compared with the results of Avery and Clark, and related to the spatial arrangement of the predictor variables.*

KEYWORDS: *Distance decay, travel pattern, scale change, autocorrelation, predictability.*

AUTHORS: Daniel R. Fesenmaier and Michael F. Goodchild *are with the Department of Geography, University of Western Ontario;* Stanley R. Lieber *is with the Department of Geography, Southern Illinois University, Carbondale.*

A version of this paper was initially prepared for the Annual Meeting of the Association of American Geographers, East Lakes Division, September 1978. Acknowledgement is given to the Illinois Department of Conservation for data which was collected through a Statewide Comprehensive Outdoor Recreation Grant, and analyzed in this paper.

Journal of Leisure Research, *Volume 12, Number 3, pp. 213-228.*

Introduction

Outdoor recreation forecasting methods developed since the early 1960's have been based upon association between patterns of participation and demographic characteristics such as age, sex, available leisure time, and income of the family. At the same time the "distance decay" model, which includes measures of availability and "resistance" to travel, has been utilized to model recreation trip patterns. Initially, these models were applied at a macro-scale where factors which generally determine outdoor recreation participation were identified. However, with an increasing tendency to regard the individual as the primary unit of observation, geographers found that new regularities exist and that the macro-level no longer provides accurate information at the micro-scale. This paper, therefore, attempts to examine one aspect of the "distance decay" model which may be dependent upon level of aggregation, namely the "resistance" to travel parameter and its correlates. Given that this model may be useful in predicting participation rates as well as travel characteristics, an approach is taken such that various levels of aggregation are generated and analyzed. In essence, then, our principal tasks are:

1. To identify "resistance" to travel coefficients for each study unit at each level of aggregation, and
2. To identify the correlates of these "resistance" parameters for each aggregate level.

Outdoor Recreation Choice Model

For decades researchers in human behavior have attempted to identify the important causes of differing human behavior patterns (Ittleson 1973; Saarinen 1976). A model that these researchers and others have developed identified, albeit simplistically, the relationship which exists among individuals, their environment, and their available opportunities. Figure 1 presents a simplified diagram of the individual choice process and identifies three sets of important factors: 1) individual characteristics, 2) social relationships and social structure, and 3) available opportunitiies. The individual characteristics of the household such as age, income, education, and family composition explain differences in participation among households. For example, one might expect that an older household not having young children might take extensive vacations whereas younger families may be more restricted.

Changes in social structure and the social institutions which establish societal norms may also affect the amount of leisure time. The trend towards early retirement in the United States, for example, has resulted in a net increase in leisure time available to Americans while cultural institutions such as Boy Scouts, and Campfire Girls, and 4-H Clubs have promoted self-reliance and romantic values regarding the outdoors and the environment.

Although "potential" households' demands in any behavioral setting may be represented by the desires and wishes of decision makers, actual decisions can only be expressed in the context of available opportunities. Participation, therefore, is a measure of the tradeoffs between "potential" demand as dictated by the characteristics of the individual and surrounding social structure and the

constraints of availability. For example, suppose one hiking area is very large and attractive but far from a decision maker's home while another is much smaller but closer to his residence. This individual must evaluate the advantages to be obtained against the extra cost and effort incurred in traveling to the distant site.

The Distance Decay Model

The descriptive model of Figure 1 has been fomulated into many mathematical models. One such model is the so-called "distance decay" model. In its simplistic form (Equation 1) two parameters are employed to model the "behavioral" and "accessibility" components of demand.

$$I_{ijk} = K_j e^{b_j d_{ik}} \qquad \text{Equation 1}$$

where:

I_{ijk} = the number of times a member of household (i) in region (j) visited place (k) during the study period.

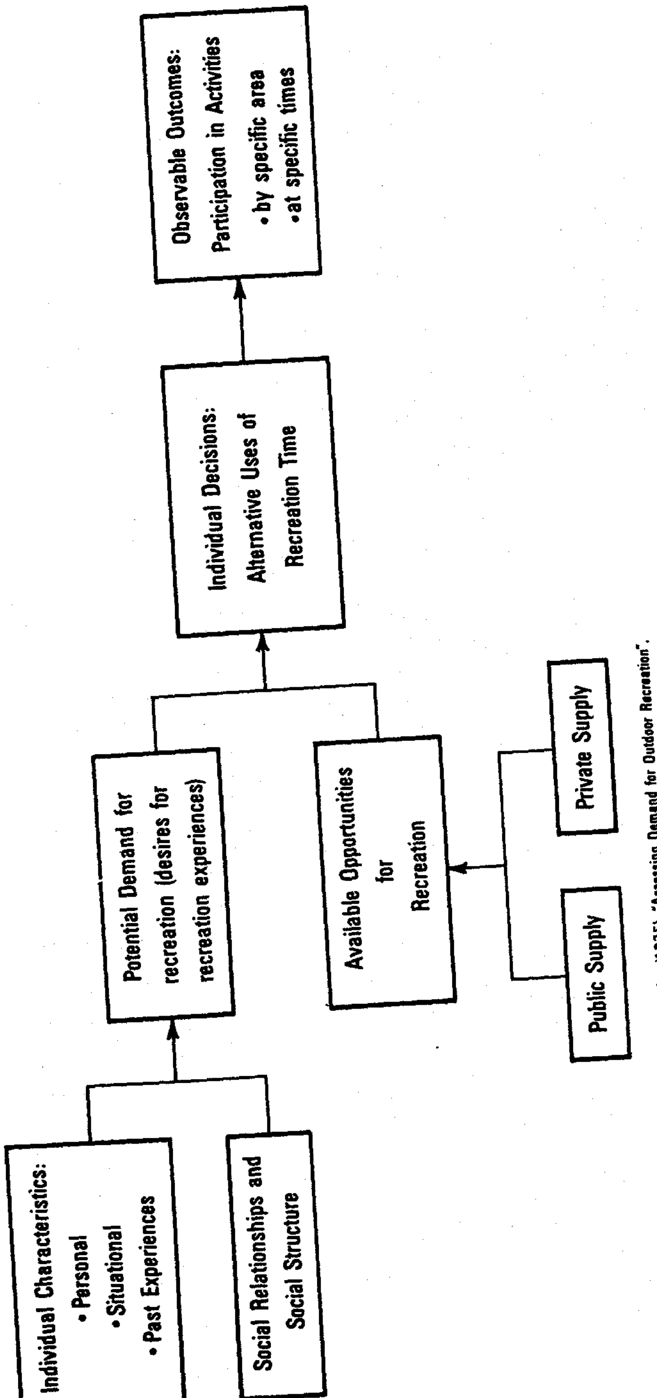
K_j = a constant for region (j).

b_j = a constant expressing the importance of distance to the residents of region (j).

d_{ik} = the distance members of household (i) must travel to place (k).

e = 2.718

Participation is believed to decline exponentially as a function of the distance needed to travel to the recreation site; (bj), then, is expected to be negative and determines the rate at which participation declines with. If, for example, (bj) is small and negative, distance has little influence on participation whereas if (bj) is negative and large, participation declines sharply with distance (see Figure 2). That is, "all things being equal," the farther a park is located away from a population the fewer will be its visitors. With this condition the (bj) parameter has been considered to be a measure of individuals' response to distance and is determined by their behavioral setting i.e., individual and social characteristics (Cesario, 1975). If however, the available facilities are not uniformly distributed the (bj) may also reflect availability, the limiting factor of potential demand considered in Figure 1. The (Kj) parameter, however, has commonly been considered to describe the propensity to interact, in other words, the tendency to participate within a given system. Since in this study (Kj) is estimated for each region (j), (Kj) may also reflect certain socioeconomic as well as availability characteristics within region (j) that determine individuals'



Source: U.S. Bureau of the Interior, Bureau of Outdoor Recreation (1975), "Assessing Demand for Outdoor Recreation", U.S. Government Printing Office, Washington, D.C.

Figure 1. Model of individual recreation choice.

proclivity to recreate. Although (Kj) and (bj) may be described by similar soicoeconomic or availability characteristics, each parameter provides a numerical description of separate tendencies of the population; that is, (Kj) describes "propensity" whereas (bj) describes "reaction to distance," given a particular behavioral and locational setting. The primary purpose of this study, then, is to investigate the extent to which the (bj) values for each region can be predicted from information about that region, including the socioeconomic characteristics of its residents. In other words, we attempt to identify the socioeconomic correlates of "resistance to travel" at varying levels of scale.

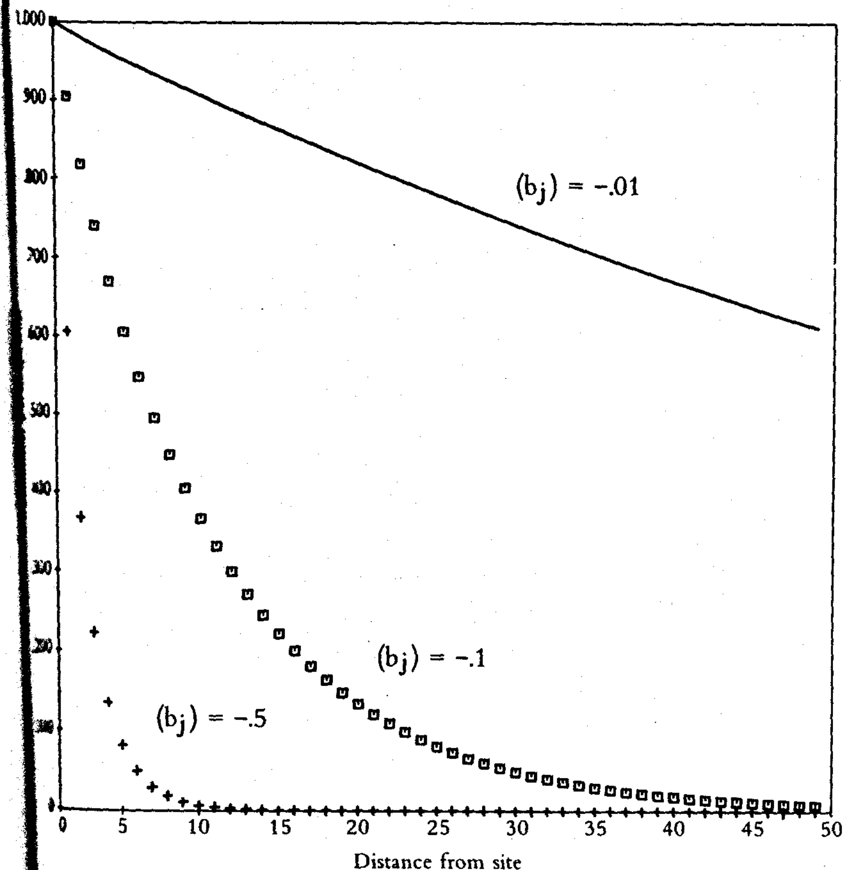


Figure 2. Impedance of distance.

A Case Study: Day-hiking in Illinois

During the winter of 1977, a telephone based recreation participation survey of 9,238 households was conducted for the Illinois Department of Conservation (1977). Random digit dialing techniques were used to ensure that the sample was representative of the general Illinois population. Twenty-three different outdoor recreation activities formed the basis of the survey; for each activity in which the household participated the names of places of the facilities visited and the length of stay were obtained for up to five of the most frequented recreation sites of each activity. In addition, detailed demographic and socioeconomic data, including a complete household description, were obtained.

During the survey it became apparent that day-hiking was the most popular activity of outdoor recreation within the state. It also is an activity which requires relatively little experience and is easily undertaken and thus, allows the decision maker a wide variety of potential locations to choose from. In essence, day-hiking presents the least number of barriers (social, economic, or availability) to participation and for this reason was chosen for further investigation.

Computational Model

Referring to Equation 1, for this study Iijk was estimated from the number of days in which each household participated in day-hiking at location (k) in an entire year. dik was estimated in two ways, from the *number of miles* and the *time in minutes* household (i) needed to travel to recreation site (k). Equation 1 was linearized to

$$\ln I_{ijk} = \ln K_j + b_j d_{ik}$$

where ln denotes the logarithm to base 2.718, or "natural" logarithm, and therefore, (bj) can be estimated using simple linear regression.

Equation 2 was calibrated at three levels of aggregation which are commonly used for planning purposes: 1) by the city in which the individual household is located, 2) by the county in which the individual household is located, and 3) by the Governor's Planning Region in which the individual is located. Each individual surveyed was initially assigned to one of 355 cities included in the study; however, only 90 of these cities included sufficient information such that the operational distance decay model could be calculated. Similarly, of the 102 counties within Illinois, only 50 counties contained sufficient information; and, of the 9 Governor's Planning Regions, 6 were included. Utilizing those survey responses of households that resided in each of the specific "regions" (city, county, or Planning Region), a (bj) parameter was calculated. Thus, 90 (bj) estimates were made for the cities, while 50 and 6 were generated for the counties and Planning Regions, respectively.

For each of the three aggregate levels, fourteen socioeconomic variables included in the survey were aggregated and expressed as the arithmetic mean. They are identified as follows:

ADDYRS	=	the number of years living at a particular address
AVAGE	=	the average age of the family
CHILD02	=	the number of children in the household between 0 and 2 years of age
CHILD612	=	the number of children in the household between 6 and 12 years of age
CHILD317	=	the number of children in the household between 13 and 17 years of age
CHTWSIZ	=	the size of town the interviewee lived in as a child
INCOME	=	the amount of income the household generates in one year
IN-SCHOOL	=	the number of children attending primary or secondary school
PCMALE	=	the percentage of males in the family
TWNYRS	=	the number of years the interviewee has resided within the current town or city
VACDAY	=	the number of vacation days the household head receives in a year
WKHRS	=	the number of hours the household head works in a week

Using the estimated (bj) values and the aggregated socioeconomic variables, step-wise multiple regression was run for each aggregation level to identify the correlates of (bj). Table 1 identifies the five predictors which explain the greatest amount of variation for each of the regression analyses.

Results

Step 1: (bj) calibration

As previously mentioned, the parameter (bj) connotes certain behaviroral and/or locational implications; hence, the sign and magnitude of its estimates reflect the importance of distance to recreation participation. In this analysis, the estimates of (bj) from distance in miles at the city level ranged from -1.74 to +.77 having an average of -.023 and being strongly negatively skewed. This negative mean is in conformity with the general hypothesis that visitation is a decreasing function of distance; that is, with all things being equal, the farther the recreation site is located away from the individual, the fewer times he will visit. The positive values also observed, however, indicate that "all things are not equal" and therefore distance may not be a "negative" factor in some cases. Consider potential reasons for differences in the sign; for example, in cities such as Chicago and St. Louis few "natural" areas exist which allow extensive hiking and the ones that are present are commonly located in the periphery of the city. This inaccessibility, then, may allow the recreator to emphasize different aspects of the trip. Perhaps a store will be visited along the way, or the "aesthetics" of the drive may determine where he will recreate.

Figures 3 and 4 display the general pattern of (bj) throughout Illinois. There appear to be three classis of (bj)'s: 1) the strong positive (bj) [indicated by the peaks]. They appear to be located in large cities such as Chicago, Champaign-Urbana, and Decatur and may reflect the lack of nearby facilities; 2) weak positive and negative (bj) [the level areas]. Such values may reflect an "indifference" to distance whereby the trip to the recreation area may be only one aspect of the day's outing. For example, along with hiking, the family may go shopping or visiting in a nearby town. This reaction to distance is characteristic of most rural areas of the state; and, 3) the strong negative (bj). This last area is identified by the pits and is typified by abundant nearby facilities located in southern Illinois.

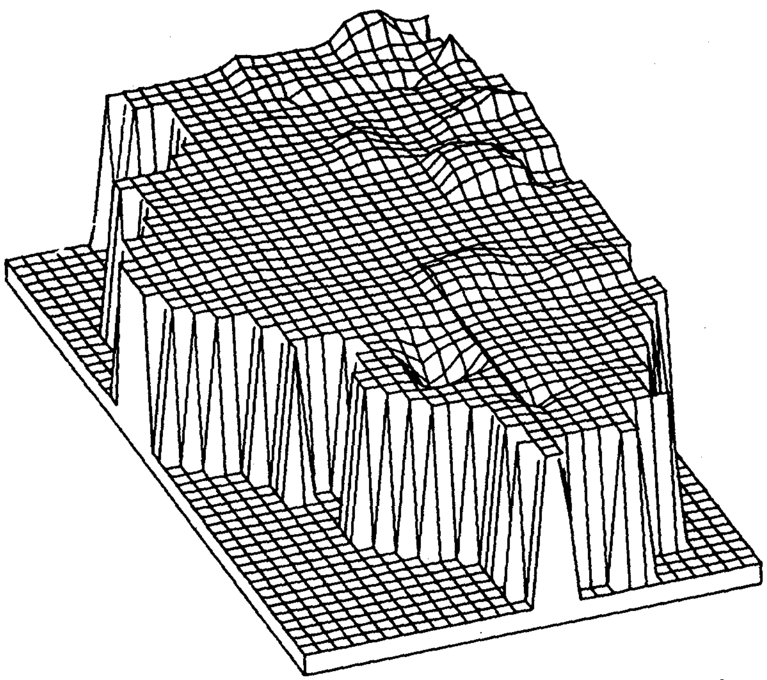


Figure 3. Distribution of (b_j) estimates in Illinois, distance measured in miles.

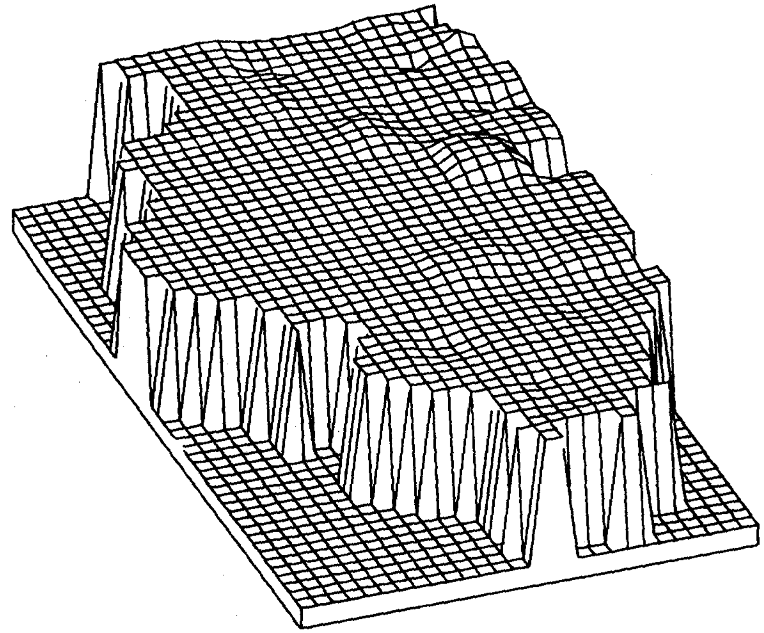


Figure 4. Distribution of (b_j) estimates in Illinois, distance expressed as travel time in minutes.

Table 1 shows a great deal of difference in the results for the three levels of aggregation. One must ask, then, how scale change affects outdoor recreation forecasting estimates. About this the geographical literature is far from clear. Many studies have attempted to define the applicability of a particular "distance decay" model at a given level of aggregation; however, none have attempted to determine the applicability over varying levels of aggregation. In the sociological and economic literature, however, the problem of aggregation has been extensively investigated. Instigated by Robinson's work "Ecological Correlations and the Behavior of Individuals" (Robinson 1950), Blalock (1964), Grunfeld and Griliches (1960), and Clark and Avery (1976), among others, delved into the problems associated with inference at multiple levels of aggregation. Robinson argued that it may be incorrect to make inferences at varying levels implying that regularities exist at each level and that a macro-level model cannot provide accurate information at the micro-scale. Contrary to this viewpoint is Blalock's; he argued that systematic grouping to maximize variation in the independent variables will yield an unbiased slope coefficient at the macro-level, although correlation coefficients will be affected. Grunfeld and Griliches, seeking to explain in mathematical terms the effects of aggregation wrote:

TABLE 1

The Correlates of (b_j)

1a. Distance measured in miles

	City (N=90)		County (N=50)		Governor's Planning Region (N=6)	
Overall r =	.508		.611		1.000	
Significance (α level)	.000		.003		.063	
Variables:	Beta:					
	TWNYRS	.315	TWNYRS	.268	TWNYRS	-.821
	INSCHOOL	.229	WKHRS	.170	VACDAY	-.743
	CHILD1317	.251	INCOME	-.232	AVAGE	-.526
	AVAGE	.225	AVAGE	.257	CHTWSIZ	.101
	ADDYRS	-.211	CHILD1317	.230		

1b. Distance expressed as time in minutes

	City (N=90)		County (N=50)		Governor's Planning Region (N=6)	
Overall r =	.244		.563		1.000	
Significance (α level)	.487		.025		.009	
Variables:	Beta:					
	ADDYRS	.162	TWNYRS	.297	TWNYRS	-1.000
	CHILD612	-.179	WKHRS	.118	INSCHOOL	-.826
	AVAGE	-.140	INCOME	-.156	ADDYRS	-.461
	CHILD1317	.133	CHILD02	-.116	INCOME	.264
	PCMALE	-.092	INSCHOOL	.164		

The fact that the aggregate r^2 is usually higher than the micro r^2s is due mainly to what may be best called a 'grouping' or 'synchronization' effect. It is the result of the empirical fact that most of the groupings that are likely to be used are such that aggregation will increase the variance of the denominator of r^2 relative to its numerator. The synchronization effect can be expressed as follows: The higher the correlation between the independent variables of different individuals or behavior units, ceteris paribus, the higher the r^2 of the aggregate equation relative to the r^2s of the micro equations. (Grunfeld and Griliches, 1960, p. 4)

To identify more specifically what is meant by "synchronization" these authors consider a simple bivariate example which allows identification of the relevant properties of aggregation. In summary, the ratio between the aggregate r_a^2 and the micro r_m^2 can be written as follows:

$$\frac{r_a^2}{r_m^2} \doteq (b^2S_{xm}^2 + S_{um}^2) / [b^2S_{xm}^2 + S_{um}^2 (\frac{1+(k-1)p_u}{1+(k-1)p_x})]$$

where:
- S_{xm}^2 = the variance of x_m for k individuals
- S_{um}^2 = the variance of u_m (random error) for k individuals
- b = the regression coefficient (slope) for all k individuals
- p_x = the intercorrelation among individuals of the independent variable for all pairs of individuals
- p_u = the intercorrelation of the disturbance (residuals) for all pairs of individuals

From this equation we see that the relationship between the aggregate r_a^2 and the micro r_m^2 depends upon the relative sizes of p_x and p_u.

With this ratio in mind one can now understand the effects of aggregation. The aggregation process generates "mean" values and, as aggregation occurs, these values approximate the overall average more closely. In effect, the correlation within each independent variable (p_x) increases. Additionally, as aggregation occurs, the correlation within residuals (p_u) decreases. This is caused by the increasing randomness in residuals; that is, initially the model is partially misspecified (indicated by the low r^2) and therefore the residuals are not completely random. However, as the averaging of independent variables improves model fit, the residuals become increasingly random. This increase in r^2, then, is not just a reflection of model fit but also a reflection of the distribution in residual error.

Blalock (1964) used Robinson's notion of "ecological" regression to investigate the effects of various aggregation procedures and found that the severity of bias varies with grouping procedure. He indicated that most aggregation is a composite of two different types of partitioning:

1) Cases can be grouped according to their scores on the x independent variable, or,

2) cases can be grouped according to scores on the y dependent variable.

In both methods subgroups would be grouped together, irrespective of spatial relationships, to maximize differences between each aggregate group. A third type of aggregation was also identified; that is, the subunits could be grouped in some way unrelated to their x or y values, for example, proximity. This form of aggregation, however, was considered to be an aspect of either of the previous two in that proximity aggregation would tend, in some cases, to maximize variation in x while in other circumstances maximizing variation in y. Using these three aggregation methods and a simple random aggregation method, Blalock demonstrated that the values of the correlation coefficient (r_{xy}) and regression coefficient (b_{xy}) are stable under random aggregation; that (r_{xy}) and (b_{xy}) are unstable when aggregating by y; that (r_{xy}) is unstable but (b_{xy}) is stable when aggregating by x; and, that proximity aggregating produces unstable (r_{xy}) and (b_{xy}) that moderate in severity depending upon the relative degree of change in x or y. Hannan (1971) has suggested that the degree of bias in the regression coefficient (for proximity aggregation) is contingent upon the nature of the data which is being analyzed. He stated that it is an empirical question whether proximity grouping will maximize variation in x or y. In a recent paper supporting Hannan's contention, Clark and Avery (1976) argued that the spatial distribution of values, the spatial autocorrelation of the x and y variables within the study region, determines the extent to which spatial aggregation biases the regression or correlation coefficient. That is, if both variables are *not* spatially correlated, then spatial aggregation may provide unbiased estimates of (r_{xy}) or (b_{xy}); however, if either variable is spatially correlated, even spuriously, then the regression variables will be systematically grouped and therefore will bias the results. Other conditions, however, have been identified that also provide unbiased results. These are:

1) When the variances of macro-level and micro-level variables are equal (Hannan and Burstein 1974); and,

2) When the group mean of the independent variable (x) has no effect on y with x controlled (Firebaugh 1978).

Under these limited conditions, then, unbiased estimates of (r_{xy}) and (b_{xy}) may result.

Figures 5 and 6 display the spatial arrangement of two socioeconomic characteristics (AVAGE and WKHRS) included in this study; in addition, the mean and standard deviation statistics are included to identify the homogenizing effect of aggregation. One can see that there is a high degree of variation initially and that aggregation into larger units decreases differences among groups as evidenced by the decrease in standard deviation size. These findings agree with the model developed by Grunfeld and Griliches. That is, at the micro-

Figure 5.
Distribution of AVAGE values in Illinois.

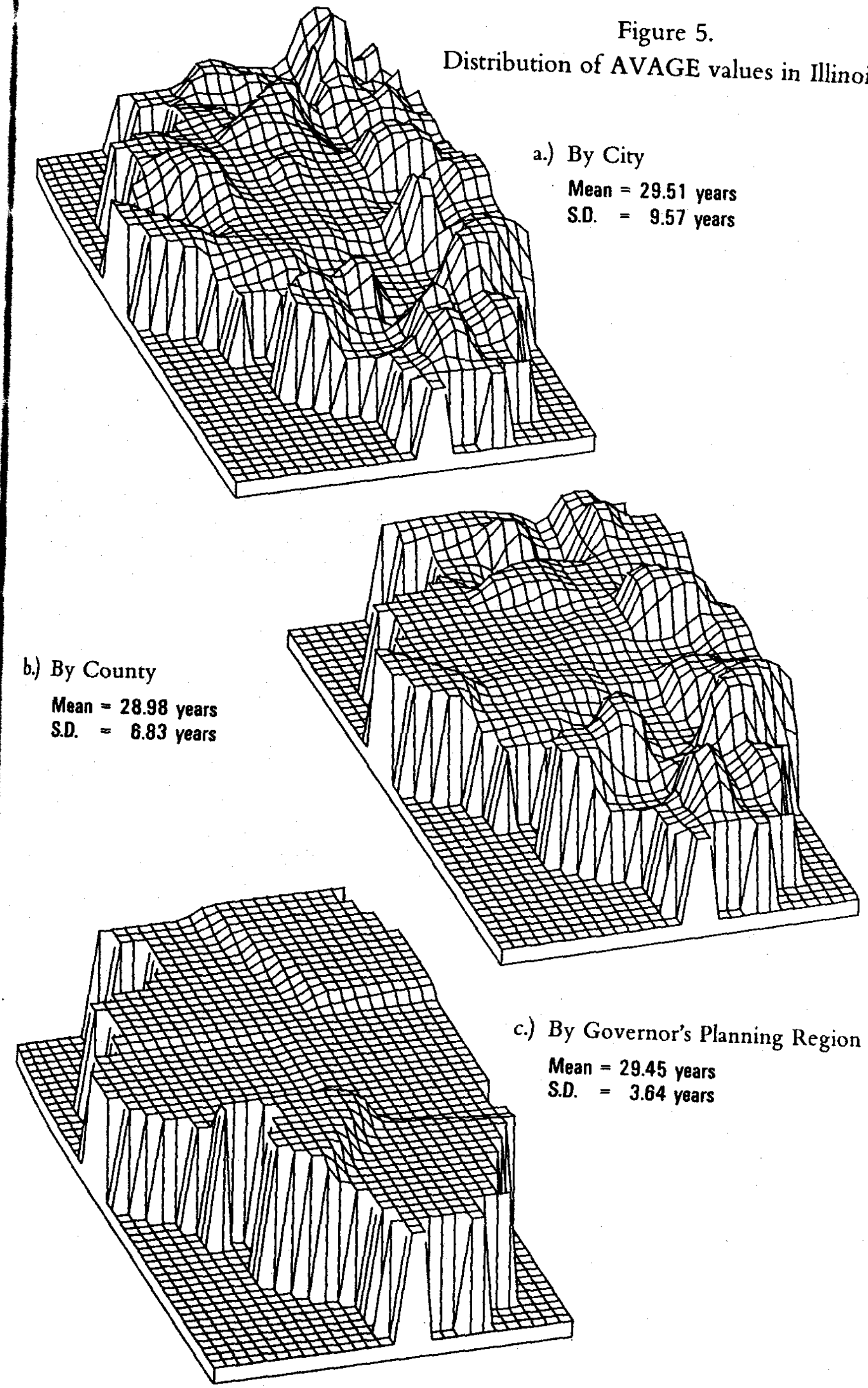
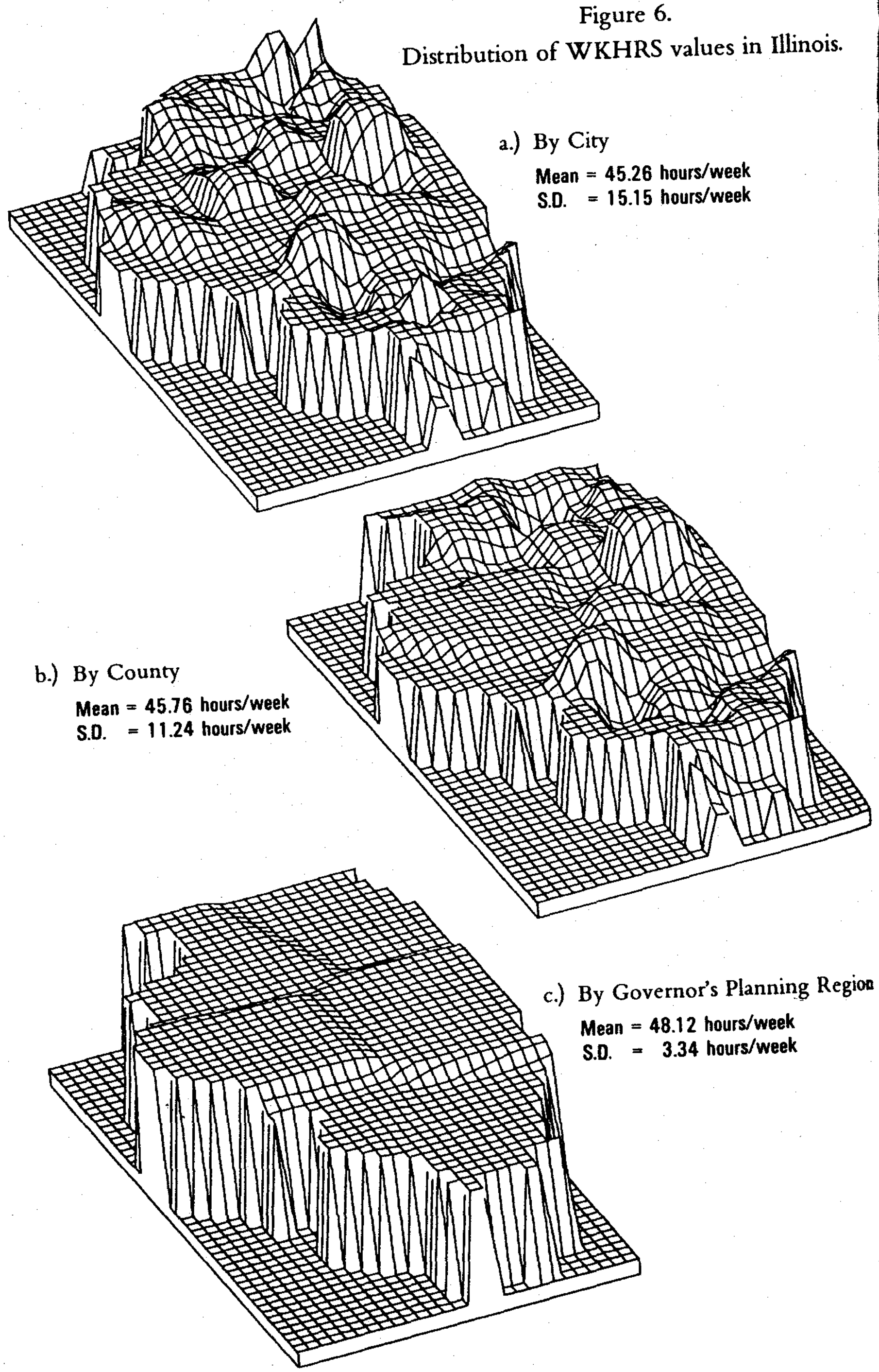


Figure 6.
Distribution of WKHRS values in Illinois.



a.) By City
Mean = 45.26 hours/week
S.D. = 15.15 hours/week

b.) By County
Mean = 45.76 hours/week
S.D. = 11.24 hours/week

c.) By Governor's Planning Region
Mean = 48.12 hours/week
S.D. = 3.34 hours/week

level the criterion variables are spatially correlated and when regressed, result in a high correlation among residuals; however, as spatial aggregation occurs, "average" groups are formed which result in low autocorrelation among variables; i.e., residuals become increasingly random. In addition, one can see that the variation within variables declines differentially (AVAGE declines 28.6 percent and 46.7 percent while WKHRS declines 28.8 percent and 70.2 percent when counties and planning regions are formed, respectively) and that the powers of the predictor variables vary substantially. As shown in Table 1a INSCHOOL is the second best predictor at the city level but provides little "explanation" at any other aggregate level. Similarly in Table 1b TWNYRS does not appear among the best five predictors at the city level but correlates perfectly at the planning region. These results, then, suggest that the study variables exhibit spatial correlation such that any spatial aggregation of this particular data set will produce biased results.

Conclusions

The spatial aggregation of data may significantly affect the use of regression in identifying the correlates of recreation travel behavior. For the specific set of aggregated data in this study, the correlation substantially increased while the relative importance of the independent variables varied widely. As suggested by Clark and Avery, it appears that the spatial structure of data and the level of aggregation determines, at least partially, model fit and the relative importance of predictor variables. It is also clear, however, that grouping spatially may not always result in bias; that is, if neither x nor y are directly or indirectly spatially correlated, or if the conditions identified by Hannan and Burstein (1974) or Firebaugh (1978) are met, then spatial aggregation will produce no bias.

The implications for any statistical analysis of outdoor recreation information, therefore, are as follows: 1) At any one scale there is a tradeoff between or predictability of the model. Although the city model may have been sensitive to the individual behavior pattern, the lack of accuracy in this model allows little confidence in its use as a planning tool. On the other hand, the macro-level model appears insensitive to the individual but provides great accuracy in identifying factors which generally affect travel patterns; and, 2) The conclusions of any study must be treated with caution. Even when two studies are conducted at similar levels, because the spatial arrangement of predictor variables from aggregation may not be identical or because the conditions identified by Hannan and Burstein or Firebaugh cannot be met, neither similar coefficients nor correlates may be obtained from a "comparative" study. The researcher faced with a problem of making inference with or comparisons of grouped data, then, must first identify the criteria by which the data were originally aggregated, and second, assess the relationship between the grouping criteria and the variables included in the study. Only after these steps have been taken can the researcher decide whether any meaningful conclusions can be made.

References

Blalock, Hubert M. 1964. *Causal inferences in non-experimental research.* Chapel [illegible] University of North Carolina Press.

Cesario, Frank J. 1975. A new method for analyzing outdoor recreation trip data. *Jo[illegible] of Leisure Research* 7(3): 200-215.

Clark, W. A. V and Avery, Karen L. 1976. The effects of aggregation in statistical ana[illegible] *Geographical Analysis* 8(4): 428-438.

Firebaugh, Glenn. 1978. A rule for inferring individual-level relationships from aggr[illegible] data. *American Sociological Review* 43(August): 557-572.

Grunfeld, Yehude and Griliches, Zvi. 1960. Is aggregation necessarily bad?. *Revi[illegible] Economics and Statistics* 42(February): 1-13.

Hannan, M. T. 1971. *Aggregation and Disaggregation in Sociology.* Lexington, M[illegible] Heath.

Hannan, M. T. and Burnstein, Leigh. 1974. Estimation from grouped observation. *A[illegible] ican Sociological Review* 39(June): 374-392.

Ittleson, William H. 1973. *Environment and cognition.* New York: Seminar Press.

Lieber, Stanley. 1979. "Recreation Behavior in Illinois." Carbondale, Illinois: Depar[illegible] of Geography, Southern Illinois University, typewritten.

Robinson, William S. 1950. Ecological Correlations and the behavior of individuals. *A[illegible] ican Sociological Review* 15(June): 351-357.

Saarinen, Thomas F. 1976. *Environmental planning: perception and behavior.* Bo[illegible] Houghton-Mifflin Co.

U. S. Department of the Interior, Bureau of Outdoor Recreation. 1975. *Assessing de[illegible] for outdoor recreation.* Washington, D. C.: U. S. Government Printing Office.