DIR 75-1
SP-9

August
1975

RURAL DEVELOPMENT POTENTIALS:
A MULTIVARIATE APPROACH

Lynn Reinschmiedt

Lonnie L. Jones

Technical Article No. 11952 of the Texas Agricultural
Experiment Station, Texas A&M University.

Paper presented at AAEA annual meeting, Columbus, Aug. 10-13, 1975.

Staff papers are made available for distribution without formal review by the Department of Agricultural Economics.

ABSTRACT:

Session Number XXXIII Session Title Rural Policy and Recreational Development

Rural Development Potentials: A Multivariate Approach

Lynn Reinschmiedt and Lonnie L. Jones, Texas A&M University

Limited funds available for developing rural industrial or recreation activities demand efficient use. Principal component analysis is demonstrated in selecting significant variables to be used in discriminant analysis. Resulting discriminant coefficients indicate the relative degree of potential each variable contributes toward classifying counties as industrial or recreation oriented.

RURAL DEVELOPMENT POTENTIALS:
A MULTIVARIATE APPROACH

by

Lynn Reinschmiedt and Lonnie L. Jones*

In recent years considerable attention has been devoted to the so-called "rural development problem," which broadly includes problems of low income, a high incidence of poverty, inadequate or expensive community services, net out-migration and high dependency rates in rural areas (Tweeten, p. 43). Numerous programs have been enacted over the years directed toward alleviating the "rural problem," the most recent of which is the Rural Development Act of 1972. The act interprets the main objective of rural development as "to encourage and speed economic growth in rural areas, to provide for jobs to improve the quality of rural life, and to do so on a self-earned, self-sustaining basis" (Tyner, p. 36).

In the past as well as the present the rural development approach on a national scale has been characterized by a lack of funding scattered over numerous projects, agencies, and various other groups and bodies. Hence not all communities who need funds will be able to acquire assistance and the approaches taken may produce results inferior to that possible (Tyner, p. 37).

Research Assistant and Associate Professor, respectively, Depart-ment of Agricultural Economics, Texas A&M University.

The authors are grateful to W. B. Smith, Ron Lacewell, and R. J. Freund for helpful comments.

Technical Article No. 11952 of the Texas Agricultural Experiment Station.

In discussing the focus rural development research should take Tyner indicates that "the most clear-cut dimension of the rural development problem is assisting the decision-maker who has the capability through his own initiative and resources to do something about his problems" (Tyner, p. 37). The purpose of this paper is to describe a procedure that will aid community leaders and agency personnel in selecting the direction of focus that has highest pay off from rural development activities.

Rural development actitivies are predominantly focused in two areas--industrial and recreational development. However, many rural areas have little choice as to which effort will yield results from concentration of their development efforts, because of limited physical characteristics of the county.

A previous study has shown discriminant analysis to be useful in determining a county's potential for recreation or industrial development (Bromley). Bromley divided counties in Wisconsin into two potential groups--recreation or industrial. Discriminant function coefficients were obtained, and used in conjunction with the relevant data on other counties in question to classify them into one of the two groups. Such a procedure allows the decision maker to observe factors and distinguish the relative importance of particular factors to potential development patterns for the county.

Framework of Analysis

The study area for the present analysis included 53 non-SMSA counties in eastern Texas (Figure 1). East Texas has undergone significant industrial and recreational development in recent years due

EAST TEXAS SMSA's AND STUDY REGION

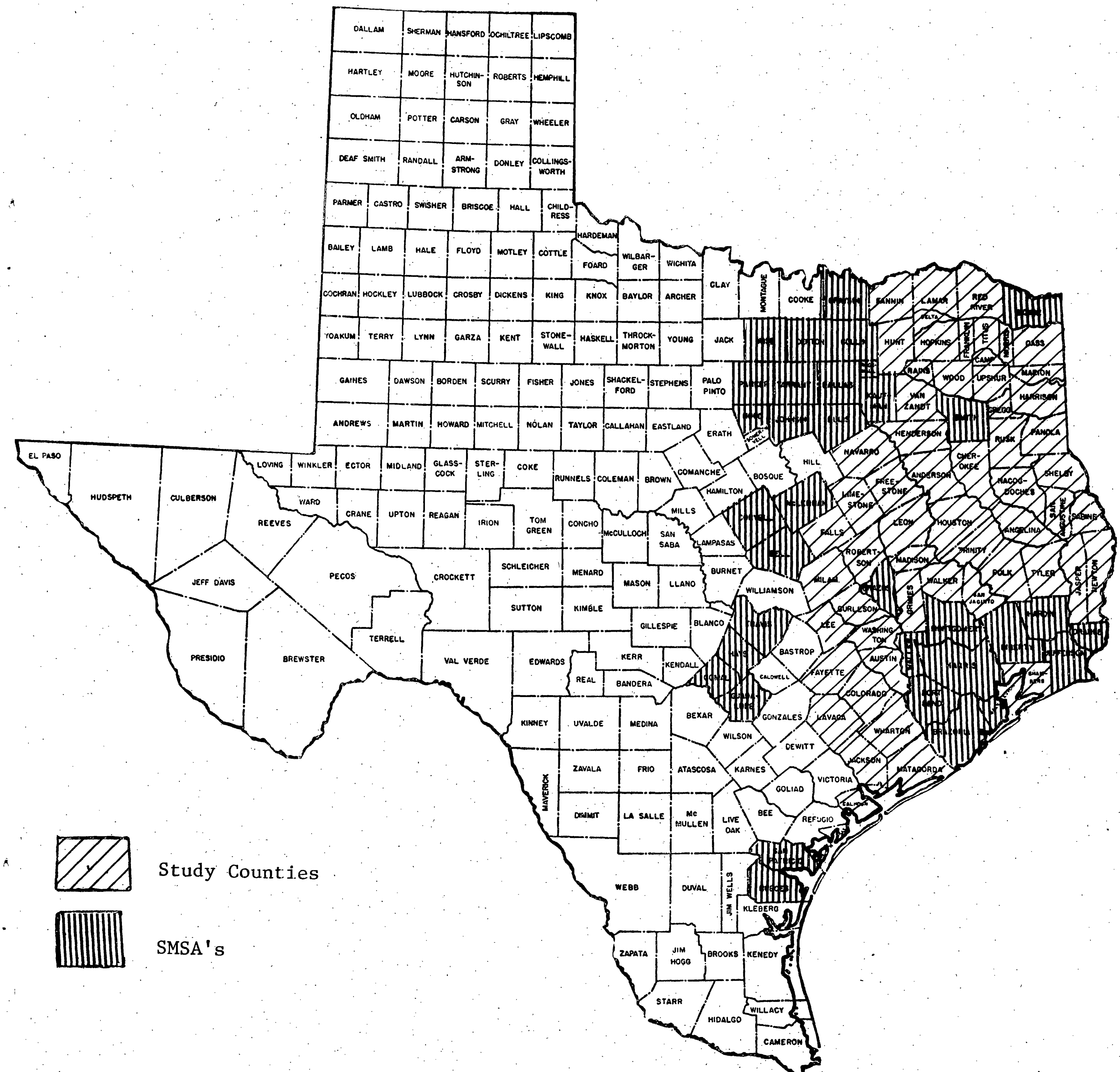

Figure 1.

to several factors. The area is close to several large markets (Dallas-Fort Worth, Tyler, Waco and Houston, along with several other smaller SMSA's) that serve as outlets for industrial goods. Metropolitan areas also provide an abundant supply of recreation seekers. When coupled with the agreeable climate, wooded terrain and water resources of East Texas, recreation plays an important part in the local economy. Agricultural activity is predominantly a small farm type with varying levels of cultivation and livestock enterprises.

The 53 counties were ranked with respect to two criteria: employment in manufacturing, and employment in the recreation and services industry. Two mutually exclusive groups were formed for the discrimination analysis by selecting the top counties in each group. Eleven counties were placed into the industry group and twelve in the recreation group. Table 1 contains 29 variables selected to reflect distinguishing characteristics of each group. Variables X_{13}, X_{23}, and X_{27} were dropped from the analysis due to numerous missing values, leaving 26 variables.[1]

Analysis

Multivariate Analysis of Variance

Prior to discriminant analysis, it is necessary that a significant overall difference indeed exist between the assigned industrial and recreation oriented counties. A multivariate analysis of variance (MANOVA) was performed on the two groups to test the

Table 1. County characteristics and variable name.

Variable Name	Description
X_1	Highway distance to nearest SMSA
X_2	Highway distance to second nearest SMSA
X_3	Highway distance to nearest interstate highway
X_4	Total annual rainfall
X_5	Average number days between killing frost
X_6	County economic index
X_7	Total county retail sales in 1970 ($1000)
X_8	County agricultural employment in 1970
X_9	County service employment in 1970
X_{10}	County manufacturing employment in 1970
X_{11}	County wholesale trade employment in 1970
X_{12}	County entertainment and recreation employment in 1970
X_{13}	County forest resources
X_{14}	County value added in manufacturing per capita in 1970
X_{15}	County agricultural income (1970)
X_{16}	Highway distance to nearest public junior college
X_{17}	Highway distance to nearest public university with graduate degree program
X_{18}	Highway distance to nearest city with commercial airport
X_{19}	Total number of interstate common carriers licensed to serve city
X_{20}	Highway distance to nearest deep water port
X_{21}	Highway distance to nearest city having general hospital
X_{22}	Number beds in general hospital per 1000 in 1970 (county)
X_{23}	Highway distance from city to nearest reservior with 5000 sq. feet
X_{24}	Number acres in county parks per 1000 people in 1970
X_{25}	Highway distance to nearest state park
X_{26}	Rated capacity of city H_2O system in excess of peak daily demand
X_{27}	Rated capacity of city sewage system in excess of peak daily demand
X_{28}	Number acres owned by industrial foundation
X_{29}	County tax rates (1970

hypothesis of no overall group effect. This hypothesis was as follows:

$$H_0: \begin{bmatrix} \pi_{1\ 1} \\ \vdots \\ \pi_{1\ 26} \end{bmatrix} = \begin{bmatrix} \pi_{2\ 1} \\ \vdots \\ \pi_{2\ 26} \end{bmatrix} \quad (1)$$

which states that the treatment effect vectors are equal. This hypothesis was rejected at the 99% level of significance with an F-statistic of 4.61. Given this information it was concluded that a significant group effect was present between the industrial and recreation groups of counties.

Principal Components Analysis

The second step utilized principal components to reduce the number of variables. Principal components is a technique whereby most of the variation in a multivariate system can be summarized in fewer variables (Morrison, D.F., p. 288). The technique was applied to the original 26 variables and eigenvalues were calculated to measure the amount of variation accounted for by each component of variation. Dividing the eigenvalues by the number of variables and multiplying by 100 determines the percent of total variance explained (Rummel, p. 467). A list of the eight eigenvalues that were greater than one and their cumulative percentage values of the total variation explained by that component are listed in Table 2. The first principal component had an eigenvalue of 5.86 and accounted for 23% of the total variation in the data system. The eight components cumulatively explained 85% of the total variation.

The rotated factor matrix was utilized to reduce the number of variables. Each factor from the principal components analysis contains values (loadings) for each variable. The loadings are weighted

Table 2. Eigenvalues and cumulative percentage of eigenvalues.*

Eigenvalue	Cumulative percentage
5.86	.23
3.59	.37
3.27	.50
2.61	.61
1.97	.69
1.79	.76
1.32	.81
1.031	.85

*SAS only prints out a factor matrix for eigenvalues above 1, unless otherwise specified.

proportionally to its involvement in a component; the more involved a variable the higher the weight and a weight near zero is assigned to variables unrelated to a given component.

The loadings with the highest absolute values were selected from each factor for further analysis. Variable selection by this procedure is somewhat subjective in that the researcher decides on some minimum "cut-off" value for accepting variables, or he simply chooses the loading which seems to be high relative to the remaining values in that component. In this instance, the loadings were chosen on the basis of the relative weighting in each factor. A list of the variables selected is shown in Table 3.

In the first component variables X_6, X_7, X_9 and X_{11} were highly correlated. All these variables can be interpreted to represent an economic activity index of some measure. According to Morrison, if the variables in question are highly correlated, they are measuring about the same thing. The coefficients on these variables will be unstable and difficult to interpret. Variable X_6, the county economic index, was chosen to represent this group. The last variable in this component X_{26}, a community services variable, was somewhat difficult to relate to this pattern of variables. However, it was not excluded at this phase of the analysis.

The second component reflects an agricultural oriented pattern. One possible interpretation of the signs of these coefficients is that increases in variables X_5, X_8, and X_{15} are associated with increased agriculturally based economic activity, which would usually

Table 3. Variables selected from principal component analysis

Component	Variable	Loading	Description
1	X_6	.88	County economic index
	X_7	.97	Total county retail sales
	X_9	.96	Total county service employment
	X_{11}	.95	Total county wholesale trade employment
	X_{26}	.84	Rated excess capacity in excess of city water system
2	X_4	.67	Total average annual rainfall
	X_5	-.78	Average no. days between killing frosts
	X_8	-.86	County agriculture employment
	X_{15}	-.89	County agriculture income
3	X_1	.82	Highway distance to nearest SMSA
	X_2	.94	Highway distance to second nearest SMSA
	X_3	.79	Highway distance to nearest interestate highway
4	X_{17}	-.64	Highway distance to nearest university with graduate degree program
	X_{29}	.75	County tax rate
5	X_{19}	-.81	Total interstate common carriers
	X_{25}	-.80	Highway distance to nearest state park
6	*		
7	X_{22}	.89	Number beds in general hospital
8	X_{14}	.85	County value added per capita by manufacturing.

*The eigenvalue for component 6 was 1.79 (Table 2). None of the variables in component 6 could be separated from the group by significantly high loadings.

be associated with a lower level of recreational or industrial related activities. Given this reasoning one would expect rainfall to have a negative sign also, which is not the case. Again a large degree of correlation exists between the county agricultural income and county agricultural employment. County agricultural income was dropped from this group.

The third component describes a distance pattern. These coefficients may be interpreted to mean that there is a positive relationship between distance as related to recreation and industrial activity. That is, an increase in distance would increase recreation or industrial activity, which is the opposite relationship usually hypothesized. A recent paper, however, reported that projects located in or near major cities (250,000 population or more) were less successful in creating jobs than projects located further from major urban centers (Barrows and Bromley, p. 50). A similar relationship may exist within the East Texas area with respect to a county's proximity to an SMSA. There was a reasonably high degree of correlation between (X_1 and X_2) and (X_2 and X_3), thus X_2 was dropped from the analysis.

The remainder of components do not render any readily recognizable descriptive patterns. All of the remaining variables listed in these latter components were used for further analysis.

Discriminant Analysis: Multiple Regression Approach

The objective of the discriminant analysis was to classify objects by a set of independent variables into one of two or more mutually exclusive categories. In the present two group case, the discrimination procedure can be reduced to a multiple regression analysis. The linear classification procedure lets each individual's discriminant score Z_i be a linear function of the independent variables:

$$Z_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_{ni} \qquad (2)$$

where

X_{ji} = the ith county's value of the jth descriptive characteristic

b_j = the discriminant coefficient for the jth variable

Z_i = the ith individual's discriminant score

Z_c = the critical value for the discriminant score

The individual's discriminant score is assigned as follows:

$$\text{Group 1} = -\left(\frac{n_2}{n_1 + n_2}\right) \qquad (3a)$$

$$\text{Group 2} = \left(\frac{n_1}{n_1 + n_2}\right) \qquad (3b)$$

where

n_1 = number of counties in group 1 (manufacturing)

n_2 = number of counties in group 2 (recreation)

A linear classification procedure as outlined in equation 3 is optimal if the variance of the independent variables in Group 1 are the same as correlations in Group 2. That is, the covariance matrices are equal (Morrison, D. G., pp. 156-57).

The advantage of using the regression approach to discrimination is that a set of discriminant coefficients (B values) are produced as a by product. Certain other procedures (such as SAS) are essentially classification schemes that do not provide these coefficients. The discriminant coefficients are probably more important than the discriminant score for this particular analysis for recommendation purposes than the actual classification of counties.

The variable selection procedures of SAS (STEPWISE, FORWARD, BACKWORD, MAXR, MINR) were utilized to further reduce the number of variables in the discriminant analysis and to provide the discriminant coefficients. An eight variable model determined by the Maximum R^2 Improvement Procedure is presented for explanation purposes in Table 4.

Given these coefficients the counties can be classified into one of the two groups. As a tentative check two counties, one from each group, were checked to see if the groupings were within reason. Both counties, Angelina and Falls were classified into their appropriate groupings.

As mentioned earlier, the eight standardized regression coefficients listed in Table 4 may be more useful to the researcher or decision maker than the actual classification score. A shortcoming of discriminant analysis is that classifying counties into one group or another has relatively few policy implications other than simply getting an inventory type classification. Community leaders and rural development specialists are concerned with what can be done to improve the well-being of communities with lagging economic development. The regression coefficients enable the researcher to determine the relative influence certain variables exert on enhancing industrial or recreational development. For example, regional development specialists who decide where funds provided by federal agency are to be spent could utilize the information in Table 4. Funds may be better utilized in counties which have large negative valued coefficients if the goal is industrial development.

Table 4. Best eight variable model maximizing R^2 from variables selected by principal component analysis

Variable	Description	Standardized Coefficient
X_4	Total average annual rainfall	-.37[a] (.0175)[c]
X_{14}	Co. value added per capita (manufacturing)	-.27[b] (.00003)
X_{17}	Hwy. distance to nearest university with graduate degree	-.20 (.0024)
X_5	Average no. days between frosts	.33[a] (.0042)
X_8	County agri. employment	.17 (.00008)
X_{19}	Total interstate common carriers	-.11 (.0083)
X_1	Hwy. distance to nearest SMSA	-.16 (.0029)
X_{26}	Rated capacity of excess water system	-.12 (.0318)

[a]Significant at 90% level of significance

[b]Significant at 95% level of significance

[c]Standard errors

R^2 = .87, F ratio = 12.54. It should be noted that, since all possible equations were examined, the F statistic should be used only as a guide rather than a true indicator of statistical significance.

Specialists may give consideration to counties located near SMSA's and universities, each having values of -.16 and -.20, respectively. Results also indicate communities with water systems that are operating with excess capacity are conducive to industrial development. The remainder of these coefficients may be interpreted likewise. The variables which are not subject to control by community leadership also provide meaningful information. For instance, communities can be made more aware of the difficulties they will encounter if their county characteristics differ significantly from those defined as desirable.

For descriptive purposes the 13 variable model and its discriminant coefficients are presented in Table 5 to show all the variables selected by the principal component analysis.

Summary and Implications

Twenty-nine characteristics were collected on 53 non-SMSA counties in East Texas. Based upon manufacturing and recreation-entertainment employment figures, two mutually exclusive groups were chosen with 11 and 12 members, respectively. The purpose of the analysis was to categorize counties into two mutually exclusive groups based on their potential economic development pattern (industrial or recreation and agriculture) using discriminant analysis. Principal component analysis was used to reduce the number of variables in the problem. The most significant variables from each factor were utilized in the discriminant analysis. Each of the principal components was examined to determine if any meaningful interpretation could be designated to them. The first

Table 5. Best thirteen variable model maximizing R^2 from the variable selected by principal component analysis.

Variable	Description	Standardized Coefficient
X_4	Total average annual rainfall	-.32 (.0290)[a]
X_{14}	County value added/capita (manuf.)	-.24 (.00006)
X_{17}	Highway distance to nearest university with graduate degree program	-.19 (.0033)
X_5	Average number days between frosts	.36 (.0069)
X_8	County agriculture employment	.19 (.0001)
X_{19}	Total interstate common carriers	-.13 (.0123)
X_1	Hwy. distance to nearest SMSA	-.16 (.0050)
X_{26}	Excess capacity of water system	-.13 (.0833)
X_3	Hwy. distance to nearest interstate hwy.	-.08 (.0034)
X_{25}	Hwy. distance to nearest state park	.06 (.0055)
X_{22}	Number of beds in general hospital	-.02 (.0203)
X_6	County economic index	-.02 (.6869)
X_{29}	County tax rates	.003 (.5265)

R^2 = .88

F ratio = 6.21. See note at bottom of Table 5 with respect to F ratio.

[a]Standard errors

three principal components reflected patterns of economic activity, agricultural activity and a distance relationship, respectively.

Results of the principal components analysis were utilized in the multiple regression discriminate analysis. As a by-product of the multiple regression approach to discrimination, a set of discriminant coefficients were obtained. Significant variables identified in this procedure provide policy information that may be utilized for directing economic development funds and efforts into programs with relatively high expected returns. Proximity to SMSA's, proximity to universities, transportation services and community public services were associated closely with potential for industrial development in the 23 counties examined.

Certain limitations to the discriminant analysis should be noted. The procedure is somewhat restricted in that it describes past relationships, hence future classification of excluded counties may be tenuous. Also discriminant coefficients may not reflect cause and effect, as is the case of any statistical procedure. For example, a variable indicating a significant influence on industrial activity may be a result of industrial activity rather than a cause of increased activity. Nevertheless, the significant statistical associations identified by the discriminant analysis should be preferable to "catch all" approaches so often used in development recommendations. Since the variables are initially selected on a basis of expected economic influence, the significant variables identified are expected to have a close economic, as well as statistical, association with development.

A larger study area would be desirable. If the results of this preliminary study merit further consideration it is hoped that data can be collected on all non-SMSA counties in Texas. It is felt that this would provide a better perspective of the recreation-industry development problems. Another possible avenue of research is discrimination on the basis of three groups--recreation, agriculture and industry--or possibly a better distinction could be drawn between agriculture and industry rather than recreation and industry. The variables selected in this analysis lean toward the agriculture-industry distinction.

A more appropriate selection of variables would reflect per capita information rather than the gross figures used in the present analysis. The problem discussed in this paper appears to lend itself to a more thorough and expanded approach than the present analysis has undertaken.

Footnotes

[1]These variables would be expected to be significant in effecting the results of the analysis. Further pursuit of this study should benefit by obtaining estimates of these missing values by exhausting other data sources or by some missing data techniques.

References

Barrows, Richard L. and Daniel W. Bromley, "Employment of the Economic Development Administration's Public Works Program," American Journal of Agricultural Economics, 57 (1975): 46-54.

Bromley, Daniel W., "The Use of Discriminant Analysis in Selecting Rural Development Strategies," American Journal of Agricultural Economics, 53 (1971): 319-322.

Frank, Ronald E., William F. Massey and Donald G. Morrison, "Bias in Multiple Discriminant Analysis," Journal of Marketing Research, 2 (August 1965): 250-58.

Kerns, Waldon R., "Application of Multivariate Analysis to Determine Influence of Resource Base on Changes in Economic Activity (1960-1970)," Southern Journal of Agricultural Economics, 5 (December 1973): 79-87.

Morrison, Donald F., Multivariate Statistical Methods, McGraw-Hill, New York: 1967.

Morrison, Donald G., "On the Interpretation of Discriminant Analysis, Journal of Marketing Research, VI (May 1969): 156-63.

Rummel, R. J., "Understanding Factor Analysis," Journal of Conflict Resolution, XI: 444-80.

Tweeten, Luther, "The Need for the Systems Approach to Rural Development Research," Southern Journal of Agricultural Economics, 6 (July 1974): 43-52.

Tyner, Fred H., "Rural Development Research Under Scrutiny," Southern Journal of Agricultural Economics, 6 (July 1974): 35-42.