TREND SURFACE ANALYSIS AS A TOOL FOR HANDLING SPATIALLY-DEFINED NON-COINCIDENT DATA[1]

David L. Sjoquist, Georgia State University
Larry D. Schroeder, Georgia State University
Frank P. Jozsa, Jr., Georgia State University

Smoothing of economic time series data has a long history within economics. It is not uncommon to find economic practitioners estimating quarterly data from observed annual data or even estimating annual data from decenial Census data. An analogous data problem exists when analyzing social behavior of groups within a particular geographically-defined area. It is common for different (social) data sets to be defined for alternative, non-coincident geographical areas, e.g. Census tracts and voting precincts. Aggregation requires loss of information and degrees of freedom. Further, if no spatial boundaries are coincident, a smoothing or interpolation method must be used to make the data sets compatible.

Several analytical methods exist for smoothing areal data including Fourier analysis, filtering theory, or a combination of Fourier and gravity analysis.[2] However, in this paper we concentrate on a single regression technique for analyzing spatial data -- trend surface analysis (TSA) and explore its use in the preparation of data. Although emphasis is placed on TSA, the explicit objective of this paper is to compare the accuracy of several alternative methods in predicting spatially-defined, unobserved data.

After reviewing the underpinnings of TSA, we explain, in Section 2, two simple prediction methods which are used as a basis of comparison in the empirical portion of the paper. In Section 3 the prediction experiment is described with the final section containing the experimental results and the conclusions reached.

Section 1 - Trend Surface Analysis

The concept of TSA may be most easily explained in terms of a sample of data observed in either a random or regular spatial pattern. Assuming that some variable of interest, z, is measured at each of a number of geographical points, then each such point can be assigned a unique (x,y) coordinate relative to some common origin.

It is possible that the points, rather than being discrete points from a random sample, are summarizations of the level of the variable in a well-defined area immediately surrounding the (x,y) position on the map. One can thus draw a three dimensional solid with the height of the solid representing the aggregate or summary measure of a variable z within the prescribed area. For most physical and much social data it is also reasonable to suspect that the level of z for points near the boundary of any one areal unit would be, in part, associated with the level of z in the adjacent unit. Thus, one would suspect that one solid would blend into the surrounding solids so that the surface over the entire space is relatively smooth. Summarizing the surface over the space is, then, the goal of any surface analysis, where the method used in TSA is least squares regression.

Of course, least squares techniques assume specification of some form of the regression function. Unfortunately there is no theoretical justification for any particular functional form; therefore, a simple and pragmatic choice, followed throughout the remainder of the analysis, is to limit the analysis to four functional forms which creat 1st through 4th degree polynomial surfaces.

For fairly small areas and many variables it is reasonable to supect that the values of z in adjacent areas are positively correlated; however, this need not be the case. If there is no observed correlation in contiguous areas, it is reasonable to suspect that regression techniques using polynomial surfaces will not explain much of the geographical variation in z. Thus, before turning to alternative prediction methods, it is appropriate to consider one measure of the strength of areal association of a variable -- the "contiguity ratio", c, developed by Geary (1954)[3] with

$$c = \left[\frac{(n-1)}{2K_1}\right] \left[\frac{\Sigma'(z_t - z'_t)^2}{\Sigma_t (z_t - \bar{z})^2}\right]$$

where n = total number of areal units
t = any one unit
z = the variable being analyzed
$K_t = k_t$ with k_t the number of connections of contiguous units associated with unit t
Σ_t = sum over all units
Σ' = sum over contiguous units

If there is no areal contiguity, the value of c will be approximately equal to one. Geary presents both a randomization and probability distribution approach to the use of the ratio for hypothesis testing.

Section 2 - Alternative Prediction Algorithms

We have selected three alternatives with which we compare TSA -- "nearest neighbor prediction", "gravity prediction", and "modified TSA". The first two methods are naive but objective in the sense that they follow some preset computation algorithm. The last alternative is a by-product of TSA but requires a certain amount of subjective judgment.

The nearest neighbor prediction approach simply uses the value of z for the geographically (linear) closest observed point to the point to be estimated. The gravity prediction model uses a weighted combination of the values of z at the

closest (measured as linear distance) four points to the predicted point and weights them by the inverse of the distance squared.

As with any regression analysis, residuals occur in TSA which can then be examined to determine if particular geographical areas are associated with positive or negative residuals. The final alternative, modified TSA, uses this information plus any subjective information available to the investigator (for example, that a particular sub-area has characteristics which differentiate it from the surrounding areas) and allows the investigator to segment the original investigation area into subares, fit new TSA surfaces to these subareas, and then predict values of z within each subarea. Of course, if the residuals exhibit little or no contiguous covariance, as measured by the ratio c defined in Section 1, the modified TSA approach to prediction is unlikely to add to the predictive power of TSA.[4]

Section 3 - Prediction Experiment

Given the basic TSA prediction method and the three alternatives outlined above, we now describe the steps followed in comparing the predictive accuracy of the various methods. The data set used is the 1970 Census of Housing for Fulton and DeKalb Counties in Georgia. The procedure to compare the predictive ability of the four methods is to select variables available at both the Census tract and Census block level and use the tract data as control points (i.e., as if these were the only data available) to predict levels of the variables for a sample of Census blocks using each of the prediction methods. The predicted values for the block level are then compared to the actual block values.

For the experiment, three different variables are used -- mean housing value, percent of persons under 18 years of age, and percent of the population which is black.

Given thse variables, the following procedure is used. From the approximately 8,000 Census blocks in the two-county area we select a random sample of 400. On a map we "eye-ball" the geographical center of each tract and the 400 blocks. Then using a cartographic digitizer we assign to each of these centers an x,y coordinate relative to a common origin.

For each Census tract the information regarding the level of the three variables listed above are extracted from the information in the 1970 Census of Housing, "Block Statistics." For those tracts for which information was not provided we use the mean of the variable in question for all tracts which are contiguous to the tract with the missing data. For the selected block the same three variables are coded; however, in this case missing data were simply excluded from the prediction error computation.

Each of the four alternative prediction methods is then used to predict values for the sampling of Census blocks. From these predictions and the observed block-level values three error measures are determined -- the sum of squared errors, the simple correlation coefficients between the predicted and actual values of the variables, and the mean percentage error where the percentage error for any observation is determined as (error/actual value).

Section 4 - Results and Conclusions

The results of the empirical tests described above are not overly encouraging to the social scientists hoping to use TSA as a prediction technique for imputing values to non-observed spatial variables. We will first consider the TSA results for the entire sample region, then compare the prediction results with those from the two naive techniques. The final portion of the section contains results on two segmentations of the orginal data space.

The first variable to be predicted is the mean housing value of a Census block. This variable would, a priori, seem to be a likely candidate for TSA under a behavioral hypothesis that persons choose to live near persons with similar housing demands and thus differences in values should vary gradually over space. (Of course, it is for this very same reason that the two naive methods may also predict quite accurately.)

The upper panel of Table 1 contains the results for the mean housing value variable for the entire two-county area. At the lower portion of that panel is the value of the contiguity coefficient, .206. The value of 1-c, .794, can be interpreted as an areal correlation coefficient, which, using the standard normal test cited above, indicates that the null hypothesis of no areal association of housing values can be rejected at less than the .001 level of significance. (The value of the standard normal deviate is 14.5. Note also that each of the contiguity coefficients reported below are highly significant.)

The total sum of squared variation about the mean for the housing value variable is shown in the lower portion of the panel with the coefficients of determination for each of the four surfaces shown in column (1). Although the first degree surface does not explain even one-quarter of the total variation, the higher order surfaces have a much higher explanatory power.

The predictive power of the surfaces are shown in columns (2)-(4) of the Table based on 302 housing value levels in the sampled Census blocks. One sees from the error measures that the second degree surface does the best predicting for this sample and this particular variable, in the sense that it produces the lowest sum of squared errors and has the highest correlation between actual and predicted values of housing values.

The two lower panels of Table 1 contain summary information on the other two variables of interest. For the age variable we find a con-

tiguity coefficient of .136 indicating an even higher areal association for this variable than for housing values. The coefficients of determination for the age variable are lower than for comparable surfaces on the housing value variable, a result not unexpected given the seemingly more random nature of this variable. For this variable, too, the second degree surface does, by far, the best job of predicting. In fact, the third degree surface results in a negative correlation between predicted and observed values of the 387 block values predicted.

The results for the racial composition variable are not encouraging either. Although, as would be expected, there is a very significant contiguity ratio, the prediction error measures do not indicate unqualified success in predicting racial composition of Census blocks on the basis of TSA using Census tract data. These results are likely affected by locational housing patterns in Atlanta where 20.1% of the Census tracts are more than 90% black while 57.5% are less than 10% black.

Before turning to the results for the alternative prediction methods, we report on a secondary finding from the TSA regression analysis of the two-county area. As other authors have noted,[5] if the original variable under investigation shows a high areal association as measured by the contiguity ratio (as each of the variables studies here do), one may investigate the contiguity ratio of the residuals from the regression to ascertain how well the regression has explained the purely areal variations of the variable.

Unfortunately, for our data, one must conclude that the spatial relationships hypothesized do not reflect well the relationships which exist. For, as is shown in Table 2, the residuals from the twelve regressions still exhibit exceedingly high (and statistically significant) contiguity effects. Thus, we might conclude that either the functional forms chosen are inadequate or that the entire model used is incorrect. For example, perhaps other explanatory variables in addition to spatial location are necessary for improved explanatory powers. However, since the purpose of this paper is simply to compare the predictive power of TSA with several naive models, we turn now to these prediction results.

Shown in columns (1)-(3) of Table 3 are the results of predicting the values of the three variables using the nearest neighbor prediction technique. As shown there, for each of the variables and for each of the error measures, this naive prediction method performed better than the TSA approach. As might be expected the correlation between predicted and actual was especially high for both the mean housing value in a Census block and the percentage of blacks living in the block. The correlation was somewhat lower for the more randomly distributed age variable.

Interestingly, as is shown in columns (4)-(6) of the table, the gravity model of prediction did not do much better than the nearest neighbor approach. In fact for housing values, prediction results, as shown by the correlation coefficient, were better using the more naive model. This is likely due to more rapid changing of mean housing values over space than changes in the racial and age compositions of the population.

For the third alternative prediction method--modified TSA-- we use two types of subjective judgments to predict block-level variables. In the first of these, purely subjective judgments about the socio-economic composition of the two-county area was used to segment the area into three subareas which we call south, central and north to refer to the approximate relative locations of the three areas. Upon this subjective segmentation, TSA surfaces were determined for each. The R^2 and correlation results are shown in Table 4 for each of the three variables and subareas. In no instance did the segmentation produce better predictions than the naive models. In some cases the R^2 statistics were higher and prediction errors lower than for the unsegmented TSA results, in other cases poorer results were obtained. This indicates that segmentation would require a variable-by-variable approach since a segmentation which might be reasonable for one variable may be entirely different for other variables.

In the second approach to modified TSA we take advantage of the capability of TSA regression programs to map the surfaces as well as residuals from the regressions.[6] It is the capability of mapping residuals which is of primary use for the modified TSA prediction technique. By studying the residuals from the original regressions it is possible to combine this objective information with a certain amount of subjective judgment in spatially segmenting the data.

We performed this operation on the racial composition variable using the results of the residual maps from the original first through fourth degree surfaces. The results of this technique are shown in Table 5. Once again the results are mixed when compared with the original surfaces; however, prediction errors are still greater for the modified TSA approach than for the naive methods.

To summarize, we have reviewed and used trend surface analysis regression techniques for predicting values of socio-economic variables and have found that, although the technique provides an objective approach to summarizing the spatial distribution of variables, it does not perform as well as alternative, simpler approaches to prediction. Included in these alternative methods have been a nearest-neighbor approach, a gravity model based on the four nearest observed points, and a modified TSA method.

We, therefore, conclude that when faced with the problem of two data sets with non-coincident boundaries, alternatives to TSA are likely to be preferred in readying the two data sets for joint analysis. The social scientist may have to give up degrees of freedom and aggregate to common boundaries or may have to use other aggregation methods.

FOOTNOTES

1/ We wish to thank Truman Hartshorn for his help and the Bureau of Business and Economic Research at Georgia State University for financial assistance.

2/ For descriptions or applications of the other techniques mentioned, see Harbaugh and Preston (1966), Tobler (1969), and Hawkes (1973).

3/ An expanded discussion of the ratio is given in Duncan (1961).

4/ The modified TSA is then simply a method for further utilization of residuals, a technique discussed by Thomas (1968). Note, it is this aspect of the problem which essentially required that the regression program employed contain mapping provisions.

5/ For example, Geary (1954) and Hawkes (1973).

6/ Note that, except for this feature, the preceding results for TSA could have been generated using any ordinary regression program. The program which we used was written by O'Leary, et al. (1964) at the University of Kansas and was adapted for use on a Univac Spectra 7 computer.

REFERENCES

Duncan, O. D., R. P. Cuzzort and B. Duncan (1961), Statistical Geography (The Free Press).

Geary, R. C. (1954), "The Contiguity Ratio and Statistical Mapping," The Incorporated Statistician, Vol. 5, pp. 114-141. Reprinted in B. J. L. Berry and D. F. Marble, Spatial Analysis: A Reader in Statistical Geography (Prentice-Hall, 1968.)

Harbaugh, J. W. and F. W. Preston (1966), "Fourier Series Analysis in Geology," Short Course and Symposium on Computer Applications in Mining and Exploration (School of Mines, U. of Arizona). Reprinted in B. J. L. Berry and D. F. Marble, Spatial Analysis: A Reader in Statistical Geography (Prentice-Hall, 1968).

Hawkes, R. K. (1973), "Spatial Patterning of Urban Population Characteristics," American Journal of Sociology, Vol. 78 (March), pp. 1216-1235.

O'Leary, M., R. H. Lippert and O. T. Spitz (1964), "Fortran IV and Map Program for Computation and Plotting of Trend Surfaces for Degrees 1 Through 6," University of Kansas.

Tobler, W. R. (1969), "Geographical Filters and Their Inverses," Geographical Analysis, Vol. 1 (July), pp. 234-253.

Thomas, E. N. (1968), "Maps of Residuals from Regression," in B. J. L. Berry and D. F. Marble, Spatial Analysis: A Reader in Statistical Geography (Prentice-Hall), pp. 326-352.

U.S. Bureau of the Census (1971), Census of Housing: 1970 Block Statistics, Final Report HC(3)-56 (Government Printing Office).

Table 1

TSA REGRESSION AND PREDICTION RESULTS

Mean Housing Value

Degree Surface	(1) R^2	(2) Sum Squared Errors	(3) Correlation	(4) Mean Proportion Error[a/]
1st	.238	.363 E11	.530	.062
2nd	.432	.308 E11	.602	.074
3rd	.485	.634 E11	.586	-.342
4th	.615	.665 E15	.462	57.018
c = .206	no. blocks for prediction = 302		sum squared variation = .160 E11	

% Less Than 18

	(1)	(2)	(3)	(4)
1st	.004	.577 E5	.032	.100 E5
2nd	.157	.508 E5	.374	.987 E4
3rd	.254	.325 E8	-.150	-.591 E5
4th	.307	.191 E9	.056	.121 E6
c = .136	no. blocks for prediction = 387		sum squared variation = .174 E5	

% Black

	(1)	(2)	(3)	(4)
1st	.101	.465 E6	.265	.206 E6
2nd	.233	.385 E6	.402	.105 E6
3rd	.256	.176 E9	.300	-.510 E7
4th	.378	.537 E11	-.257	.912 E8
c = .124	no. blocks for prediction = 357		sum squared variation = .292 E6	

[a/] Proportion Error computed as $\frac{\text{Predicted - Actual}}{\text{Actual}}$; If zero, Actual set = .0001.

Table 2
CONTIGUITY COEFFICIENTS ON RESIDUALS

Degree Surface	Housing Value		% Less Than 18		% Black	
	c	1-c	c	1-c	c	1-c
1st	.067	.933	.094	.906	.073	.927
2nd	.087	.913	.108	.892	.086	.914
3rd	.095	.905	.120	.880	.088	.912
4th	.113	.887	.128	.872	.102	.898

Table 3

PREDICTION RESULTS USING NAIVE METHODS

Variable	Nearest Neighbor (1) Sum Squared Errors	(2) Correlation	(3) Mean Proportion Error	Gravity Model (4) Sum Squared Error	(5) Correlation	(6) Mean Proportion Error
Housing Value No. used = 302	.134 E11	.845	.052 E0	.141 E11	.838	.043 E0
% $\leq$ 18 No. used = 387	.499 E5	.476	.747 E5	.473 E5	.487	.743 E4
% Black No. used = 357	.161 E6	.821	.547 E5	.146 E6	.834	.676 E5

Table 4
THREE-WAY SEGMENTATION USING TSA

Degree Surface	South R^2	South r	Central R^2	Central r	North R^2	North r
			Housing Value			
1st	.304	.042	.246	.277	.070	.544
2nd	.509	.450	.299	.302	.444	.610
3rd	.565	.083	.548	.484	.639	.580
4th	.627	.080	.589	.279	.902	-.310
Total Variation	.369 E10		.440 E10		.842 E9	
			% Less Than 18			
1st	.077	-.157	.057	.101	.594	.442
2nd	.319	.054	.250	.240	.663	.310
3rd	.414	.184	.452	.074	.804	.259
4th	.505	.170	.522	-.148	.829	.470
Total Variation	.132 E5		.114 E4		.275 E4	
			% Black			
1st	.198	.315	.242	.420	.031	.040
2nd	.459	.215	.371	.483	.046	.015
3rd	.610	-.257	.478	.392	.099	.040
4th	.620	.031	.495	-.023	.118	-.030
Total Variation	.199 E5		.205 E6		.109 E4	

Table 5
SEGMENTATION USING RESIDUALS
FROM % BLACK SURFACE

Surface	South R^2	South r	Central R^2	Central r	North R^2	North r
1st	.024	.081	.116	.319	.037	-.074
2nd	.098	.093	.354	.401	.072	-.186
3rd	.153	-.033	.405	-.287	.264	-.182
4th	.243	.098	.471	-.277	.398	.188
Total Variation	.788 E4		.159 E6		.668 E4	