SPATIAL ANALYSIS AND THE USE OF CENSUS DATA

David W. Wong
The University of Connecticut

ABSTRACT. Discrete choice models and logit type models have been used extensively in economics in the past two decades. However, one of the major practical constraints of using them in geography is the availability of appropriate data sets. Usually, information about the characteristics of choice alternatives and the socio- economic attributes of the individuals have to be collected through expensive and labor intensive interviews or surveys. On the other hand, census data have been widely used in socio-economic research and urban analysis, but not in relatively advanced spatial modeling because the summary table format (U.S. Census 1983a) presenting census data cannot facilitate this type of analysis. For example, only the total number of people in each category of census variables in given areas (e.g. tracts or block groups) is available. This paper suggests the use of a relatively old technique, which carries more than a dozen names such as biproportional fitting, cross-Fratar method, and mostellerizing, etc., may help us to "recreate" the "original" census information. The numbers of people or housing unit in categories cross-classified by census variables can be estimated by this method. Thus, advanced spatial modeling such as discrete choice modeling may be performed while some common mistakes pertaining to statistical analysis in geographical research can be avoided.

In the first section, I will discuss the general data requirements to perform spatial statistical analysis or modeling. I will identify the limitations of some popular data sets which fail to fulfill the above requirements, and the useful information captured by them. Then I will briefly discuss the iterative biproportional fitting technique focusing on its theoretical underpinning and procedure. In the third section, I propose how this technique can be used to "recreate" individual level census data. Some problems and limitations of the technique are enumerated in the fourth section. In addition, directions for future investigation are suggested.

Data Requirements for Spatial Analysis

If we survey the empirical studies in urban analysis for the past two decades, it is very obvious that the literature is dominated by statistical analysis. A very popular approach, which is inherited from the urban ecological tradition (Berry and Kasarda 1977), is to treat each census area as an observation and perform statistical testings such as regression analysis. Examples of such analysis include Nelson's (1988) research on urban revival and numerous studies she cited. However, the recent resurgence of studies in the modifiable areal unit problem (MAUP)(Openshaw 1984) have proven that it is no longer legitimate to treat areal units such as census tracts as observations in statistical analysis. It is demonstrated that using different spatial resolutions will yield dramatically different results (Fotheringham and Wong 1990).

If we put urban analysis into the larger context of human behavioral research, it is not difficult to reason that the subject of analysis should be an individual rather than an areal unit. Studies on the modifiable areal unit problem also implied that studies using the smallest areal unit or an individual will give us the most conservative results. Thus we should avoid the ecological fallacy in regression analysis and should explore the availability of individual level data or spatial data having a higher resolution than the census tract data.

Census tract data are commonly used in urban analysis because they capture very valuable and rich information. The accuracy is relatively high for such a large scale survey. However, the major drawbacks of using them, as I mentioned before, are the

problems related to ecological fallacy and the MAUP, and individual level data should be used instead. However, census data possess information extremely valuable to geographers. Most census data report the variables or attributes of the subjects on the census areal unit basis. The census areal unit, which can be census tract or census block group, has a geographical reference. In other words, census data provide some crude geographical information of the subjects under investigation, despite the fact that the spatial resolution and precision are not extremely high.

Another set of data, which has attracted attention recently and is used in some migration studies, is the Public-Use Microdata Samples (PUMS). The PUMS consists of individual housing and person records which provide census variables information (U.S. Census 1983b). Currently, the PUMS come in two set of samples: 1% and 5% of the total U.S. population. Using the PUMS in analysis, one can avoid the ecological fallacy and the criticisms evolved from the MAUP research. But the major difficulty in using the PUMS in geographical research is that only very crude locational information on the individual is available. The most detailed geographical reference for a person or a housing unit is whether it is located inside or outside the central city of an SMSA. This type of locational information is of very low utility in spatial analysis.

In short, the format of the PUMS is very suitable for spatial analysis but they do not possess useful locational information. The census data capture the locational information but do not have the appropriate structure. The procedure of iterative proportional fitting (IPF), which will be discussed in the next section, is a potential technique to combine the locational information in the census data with the PUMS data structure.

Iterative Proportional Fitting

The IPF method is also known as the Cross-Fratar procedure (Upton 1985) or the Furness growth-factor method (Macgill 1977) in transportation engineering and as the RAS method or biproportional fitting method in economics. In geography, the IPF technique has not drawn much attention until quite recently even though its theoretical underpinning is embedded in spatial interaction modeling. Upton (1985) elaborates the theoretical link between spatial interaction models and the IPF technique.

In the context of estimating migration flow, Plane (1982) employed the minimum information principle discussed by Snickars and Weibull (1977) to formulate a set of spatial interaction models. The principle is to minimize the information objective gain function:

$$I(\mathbf{P}, \mathbf{Q}) = \Sigma_i \Sigma_{i \neq j} p_{ij} \; ((\ln p_{ij}/q_{ij}) - 1) \qquad (1)$$

where p_{ij} is the predicted or estimated probability that any individual in region i will migrate to region j, q_{ij} is the *a priori* estimate of the probability, and $\mathbf{P}$, $\mathbf{Q}$ are the matrices of the corresponding probabilities. The minimization procedure is subject to the constraints imposed on the marginal probabilities and is used to estimate a set of probabilities (p_{ij}' s so that they are the least distinguishable from the original set of probabilities (q_{ij}' s .

As elaborated by Macgill (1977), this minimum information principle has been utilized, albeit implicitly, in the biproportional fitting procedure. The results obtained from the iterative marginal fitting procedure should satisfy the minimum information principle. That is, the resultant table is least distinguishable from the original one.

Mosteller (1968) proposed that the IPF procedure can be used to 'scale down' a contingency table so that the marginal sums are ones, and this standardized table can preserve the interaction effect among variables. To explain the IPF procedure, I will follow Mosteller's suggestion of using it to standardize a contingency table.

Original table

		Actual	Required
100	10	110	1
5	2	7	1
105	12	117	
1	1		

(a)

1st Iteration

		A	R
.909	.091	1.00	1.00
.714	.286	1.00	1.00
1.623	.377		
1.000	1.000		

(b)

2nd Iteration

		Actual	Required
.560	.241	.801	1
.440	.759	1.199	1
1.000	1.000		
1	1		

(c)

After 4 cycles

		A	R
.667	.333	1.00	1.00
.333	.667	1.00	1.00
1.000	1.000		
1.000	1.000		

(d)

Figure 1 The Procedure of Iterative Proportional Fitting

Figure 1 illustrates the procedure in detail. Figure 1a is the original structure of the contingency table (which is similar to **Q** in Equation 1). The value in each cell is the number of observations in each cross-classified category. The first step is to adjust the table in a row-wise manner. The values in Figure 1b are obtained by dividing the original frequencies in each cell by the actual row sums and multiplying by the required row sums. For example, the cell (1,1) is 0.909 = (100/110)(1). By applying this formula to each cell, all row sums equal 1, but the actual column sums are not. The next step is to adjust the table in a column-wise manner and Figure 1c is the resultant table after the third step. The value in each cell is obtained by dividing the value in each cell in Figure 1b by the actual column sums and multiplying by the required column sums (0.56 = (0.909/ 1.623)(1). Now the columns sum to 1, but by adjusting the column sums, the row sums deviate from 1. The next step is to adjust the table again so that the row sums are 1. The iterative procedure should continue until the actual row and column sums are the same as the required value (in this case 1, Figure 1d). This standardization procedure proposed by Mosteller can 'scale-down' a contingency table, but it can also be used to 'scale-up' or inflate a table so that the column and row sums match some pre-set values.

From census data, we can obtain the number of observations in each category of a census variable in each census areal unit. For example, we can easily find out the number of renters and owners in each census block group. Number of observations in each income category in each block group is also available. The values of these two

variables (tenancy and income group) can be regarded as the required marginal sums or the pre-set column and row totals for the IPF procedure.

From the PUMS, we can cross tabulate all observations by the above two variables to obtain the numbers of people in the whole SMSA fall into the categories defined by the variables. This cross-tabulated table is served as the original table for the IPF procedure. Performing the iterative fitting procedure, the original matrix is 'scaled down' so that its column and row sums match the required numbers extracted from the census block group data. Thus the resultant table shows the number of people who own houses and fall into a particular income group.

The above is the simplest case using the IPF procedure. It involves the fitting of only two dimensions. In reality, we often want to know the number of observations cross-classified by more than two variables. Evans and Kirby (1974) demonstrated that when the procedure is used in multidimensional situations, the estimates will converge and yield unique solutions. Thus the IPF procedure has the potential to apply to real world situations.

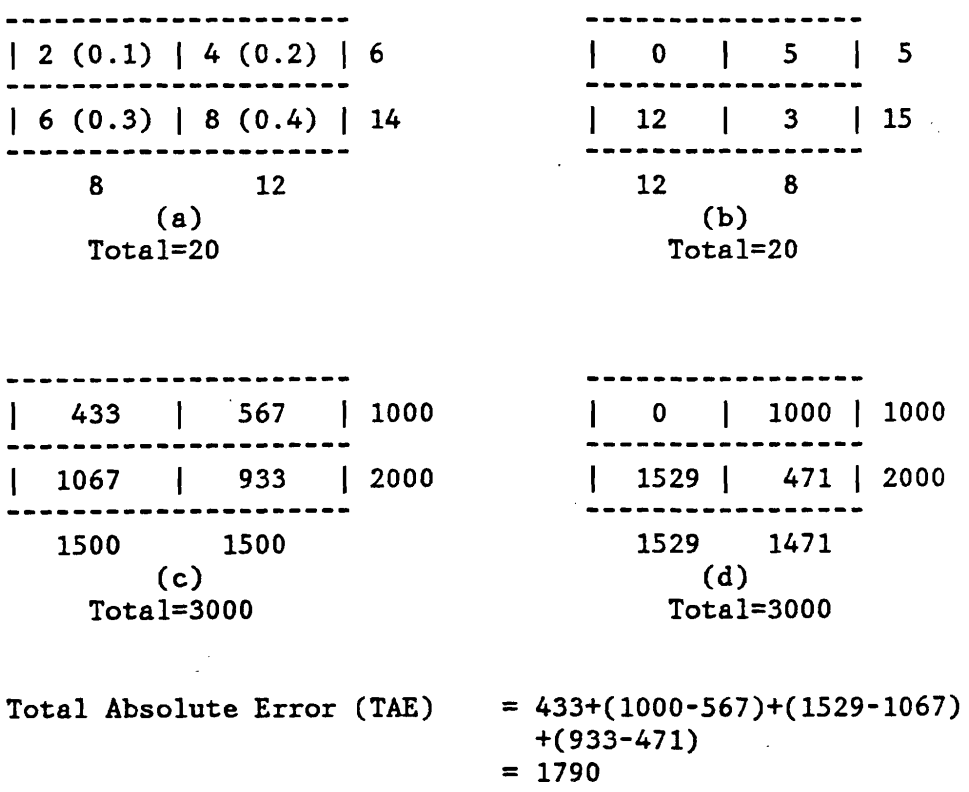
Limitations and Reliability

The IPF procedure is evolved from techniques of analyzing contingency tables such as log-linear modeling (Fingleton, 1981). Thus its usefulness is intimately related to the inherited limitations of log-linear modeling. From the above illustration, we can see that the IPF procedure can provide estimate for cells in the contingency table of a census areal unit. However, what we can extract from the contingency table is a categorical scale of the census variables and in no way can we estimate the variables in ratio or interval scale using this estimation scheme. Even if the estimate for each cell is accurate, the individual information is only available in categorical scale. Only a limited number of modeling techniques or statistical procedures can be used to analyze this type of data. In other words, the IPF procedure can generate estimates only for categorical data modeling. More powerful statistical procedure cannot be applied.

In the particular example applying the IPF procedure in this paper, contingency tables are created from the PUMS to be inflated or deflated. However, as I discussed earlier, the PUMS offer very crude locational information and it is not possible to create original contingency tables of finer scale beyond the city-suburb division. As a result, one original table is for all areal units inside the city and the another is for units outside the city. Adopting the same original table for all the census areal units within the city or outside the city will introduce a spatial smoothing effect. That is, variations among census units are smoothened or reduced. Only if other individual level data sets are available and offer higher spatial resolution than the PUMS can the smoothing effect can be eliminated.

Samples of 1% and 5% are available from the PUMS. The 5% sample is usually large enough for most studies. However, each PUMS file can be regarded as one of the many possible samples of the population. Thus, the impacts of sampling error on the accuracy of the estimates should not be ignored. The results of a simple simulation experiment indicate that the sample error in creating the original contingency table can cause significant errors in the estimates.

Figure 2a is a table which can be regarded as the true distribution of any two variables. The probabilities of any observation being classified as that category are in parentheses. Based upon this probability distribution, 20 observations are allocated to the four categories, and 35 random allocations are made. A wide range of 2-by-2 tables are generated. The difference between the outcomes of various allocations and the table in 2a is attributable to random error. Figure 2b is the 'original' table of one of the extreme



Total Absolute Error (TAE) = 433+(1000-567)+(1529-1067)
+(933-471)
= 1790

Figure 2. Effects of Random Error on Iterative Proportion Fitting Procedure

---

cases. If both tables 2a and 2b are inflated by the same magnitude (from 20 total observations to 3000 observations) subject to the same sets of marginal sum constraints (1000, 2000, 1500 and 1500), two very dissimilar matrices can be obtained (Figure 2c and 2d). Total absolute error (TAE), which is an index indicating the deviation of a matrix from another (Upton 1985), is also calculated and it is quite large.

This experiment illustrates another limitation of the IPF procedure. Since one of the cells in the **Q** is 0, no observation in the random sample falls into this cross-classified category. As a result, the inflated population matrix (Figure 2d) cannot satisfy all the marginal sum constraints and the results will not converge.

In the future, several directions should be pursued before we can apply the IPF procedure to estimate individual level data for applied research. The smoothing effect is a major limitation of the IPF technique to offer us accurate estimates. The most direct way to tackle this issue is not to apply the same original matrix for a larger number of areas. Theoretically, if we have sample for each areal units, then each area will have its original tables and spatial smoothing will not occur. However, it is very costly to collect a sample for each areal unit. If we have to use the PUMS, the smoothing effect seems to be unavoidable. But we should still investigate to what extent and of what magnitude the smoothing effect exists and how likely it will create detrimental effects to further applications of the estimated results.

The ultimate concern of using this estimation technique is whether it can give us relatively accurate estimates. The understanding of the random error effect seems to be the key to this issue. Intuitively, increasing sample size will reduce random error. Then the question is whether the 1% or the 5% sample is enough for this type of estimation. I suspect that the accuracy or the reliability of the estimates is also related to the size of

the table and the way we reduce interval and ratio data into categorical data during the cross-tabulation process of creating the original table. We can use the equal interval categorization scheme or we can define the categories in such a way that almost the same numbers of observation are found in each category. The latter categorization scheme will be very unlikely to create tables with zero counts in any cells and this scheme seems to be more preferable for the IPF procedure than the former scheme, even though the former scheme is the most common one. All the above issues can be investigated through simulation experiments using either real world data or artificial data, and the results should inform us of under what conditions the IPF procedure can yield the 'best' estimation results.

References

Berry, B. and J. Kasarda (1977). Contemporary Urban Ecology (New York: Macmillan).

Evans, S. and H. Kirby (1974). "A Three-dimensional Furness Procedure for Calibrating Gravity Models" Transportation Research 8:105-22.

Fienberg, S. (1970) "An Iterative Procedure for Estimation in Contingency Tables" The Annals of Mathematical Statistics 41: 907-917.

Fingleton, B. (1981). "Log-linear Model, Mostellerizing and Forecasting" Area 13:123-129.

Fotheringham, A. and D. Wong (1990). "The Modifiable Areal Unit Problem in Multivariate Statistical Analysis" Environment and Planning A (forthcoming).

Macgill, S. (1977). "Theoretical Properties of Biproportional Matrix Adjustments" Environment and Planning A 9:687-701.

Mosteller, F. (1968). "Association and Estimation in Contingency Tables" Journal of the American Statistical Association 63(321): 1-28.

Nelson, K. (1988). Gentrification and Distressed Cities (Madison: The University of Wisconsin Press).

Openshaw, S. (1984). The Modifiable Areal Unit Problem, CATMOG No. 38, (London: Geo Books).

Plane, D (19820. "A Information Theoretic Approach to the Estimation of Migration Flow" Journal of Regional Science 22: 441-456

Snickars, F. and J. Weibull (1977). "A Minimum Information Principle" Regional Science and Urban Economics 7: 137-168.

Upton, G. (1985). "Modelling Cross-tabulated Regional Data" in Measuring the Unmeasurable, P. Nijkamp, H. Leitner and N. Wrigley, eds. (Amsterdam: Martinus Nijhoff Publishers).

U.S. Bureau of the Census (1983a). 1980 Census of Population and Housing, Census Tract, Buffalo, NY SMSA (Washington: U.S. Government Printing Office).

U.S. Bureau of the Census (1983b). 1980 Census of Population and Housing, 1980: Public-Use Microdata Samples Technical Documentation (Washington: U.S. Government Printing Office).