Volume Title: Residential Location and Urban Housing Markets

Volume Author/Editor: Gregory K. Ingram

Volume Publisher: NBER

Volume URL: http://www.nber.org/books/ingr77-1

Publication Date: 1977

Chapter Title: Census Data and Housing Analysis: Old Data Sources and New Applications

Chapter Author: William C. Apgar, Jr.

Chapter URL: http://www.nber.org/chapters/c4310

Chapter pages in book: (p. 139 - 180)

Chapter Five

Census Data and Housing Analysis: Old Data Sources and New Applications

William C. Apgar, Jr.

INTRODUCTION

The supply of current, accurate, and detailed information on individual metropolitan areas has been increasing rapidly in recent years. While numerous agencies are responsible for this fortunate development, the Bureau of the Census has led the way with its expanded program of data collection and release, especially the Small Area Data Program of the Decennial Census of Housing and Population. The decennial census collects information on numerous housing attributes as well as the exact location of each residential dwelling unit. In addition, the census also gathers information on the economic and demographic characteristics of the occupants of each dwelling unit, including detailed place-of-work information for each employed member of the household over the age of fourteen.

The tremendous value of currently available Census Bureau data is evidenced by the vast quantities of social science research utilizing this important national resource. The Bureau of the Census has closely monitored the data needs of social science research and has responded to those needs by developing a series of sophisticated data

Note: This study is based on research funded by the Department of Housing and Urban Development under contract H-1843 to the urban studies group of the National Bureau of Economic Research. The author wishes to acknowledge the helpful comments of his colleagues in the study group, and especially John F. Kain and Gregory K. Ingram, who made extensive comments on an earlier draft.

summaries. For the 1960 census, in a major departure from previous practice, individual interview schedules were made available for a 1/1,000 sample of the population, identified by region of residence and city size. For the 1970 census, this program was expanded substantially with the release of a variety of 1 percent Public Use samples, which identified areas as small as individual counties. In addition, the Bureau of the Census has greatly expanded its program of data release for small areas and has made available a series of machine-readable data files which provide summary statistics and cross tabulations for a number of different levels of spatial aggregation, including census tracts, blocks, and minor civil divisions.

Further advantages of using Census Bureau information are its low cost and the extent and uniformity of its coverage. Few data sources can compete with available census data in these areas. To design and execute a special-purpose survey of housing consumption is a costly and time-consuming enterprise. Even when suitable data have been collected for other purposes, they are usually difficult and time consuming to use. More important, special-purpose data sources often provide information on only part of the urban housing market. For example, samples of sale prices of owner-occupied dwelling units, used in a number of recent studies of urban housing markets, exclude both renter and owner-occupied multiple units. Of equal importance is the difficulty or even impossibility of replicating analyses based on these highly specialized data sources. The use of nonuniform sources of data produces confusing and often unintelligible differences in results that may be specific to the location of the study, to the techniques of analysis, or to the data used.

By contrast, Census data are collected in a uniform manner for all owner- and renter-occupied and vacant housing units in the United States. This massive coverage permits the release of detailed complete-count housing information for states, counties, and large metropolitan areas as well as statistically reliable summary data for areas as small as individual city blocks. As a result, models estimated with Census data for one metropolitan area easily can be replicated for other metropolitan areas, and often can be replicated for spatial configurations ranging from individual census tracts or minor civil divisions to counties or entire states. Such flexibility is the unique strength of the decennial census.

Despite these numerous advantages, little systematic attention has been given to the efficient utilization of Census data for the analysis of urban housing markets. In part, this results from the exacting data needs of urban analysis. Recent theoretical and empirical work on urban housing markets has illustrated the need for large samples of

data on the physical characteristics, location, and price of residential dwelling units. If housing consumption is best described in terms of a large number of diverse attributes, as this research suggests, models of urban spatial structure must address the possible impact of the interaction of structure attributes, neighborhood, and location on both housing supply and demand. Since neighborhood attributes are likely to vary over space, and since the linkages between spatially separated housing submarkets are likely to be quite subtle, many issues cannot be resolved unless the tests are conducted with large samples containing microspatial detail.

At first glance, Census data seem ill-suited for such detailed microspatial analysis. Despite numerous requests, the Bureau of the Census refuses to release sample data identified by small areas for fear that such a procedure could result in the exact identification of the responses of an individual household or otherwise undermine the confidentiality of the Census program. While it is true that guidelines followed by the Census in its publication program to insure the confidentiality of individual responses make it more difficult to use its data, I will demonstrate that researchers have seriously underestimated the potential value of existing Census data for urban analysis. In fact, currently available Census products include all the summary statistics required to estimate a wide range of spatially detailed housing market models.

The estimation of empirical models from summary statistics is hardly a new idea, but the implications of this concept for Census data use have been generally overlooked. In practice, most social science empirical research utilizes samples of observations on a set of variables. Since any distribution or interaction present in the data could be generated from this raw sample data, the minimum information actually required for the estimation of a model is of little practical concern. In actual practice, few empirical models use all or even a large fraction of the information available in such samples. Typically, the methods employed to estimate these models use aggregate or summary statistics obtained from the raw data. Depending on the exact properties of the estimating technique used, alternative sets of summary statistics are calculated as intermediate steps in the estimation of the parameters of the model. In ordinary least squares models, for example, it is common to ignore many three-way or higher-order interactions present in the raw data. The majority of ordinary least squares models utilize only the simple pairwise correlations between the variables in the equation.

In the analysis of Census data, recognition of the minimum information required for the estimation of an empirical model is of

tremendous importance. Since both cost and considerations of confidentiality limit the amount and form of Census data, the efficient use of available data requires a precise statement of the model to be estimated and an exact enumeration of the required summary statistics. Needed summary data not contained in any single Census release often can be assembled by combining information from different published or machine readable sources. Additional summary statistics can be obtained directly from the Bureau of the Census through its program of special tabulations. In either case, the vast potential of Census data should be investigated before an empirical analysis is abandoned entirely or the analysis is recast to fit the specific nature of a non-Census data source.

In an effort to demonstrate the potential usefulness of Census housing data, I present an ordinary least squares model of housing price variation. The example was chosen for several reasons. First, the estimation of a single-equation model of the variation of housing prices over structure and neighborhood characteristics is a standard exercise in urban analysis. The example demonstrates that the replication of many spatially detailed models of price variation can be achieved without resort to special housing surveys. Second, the example presents a clear, yet simple illustration of the use of a number of Census data sources in the estimation of a single econometric model. Finally, an ordinary least squares model was chosen to demonstrate the potential usefulness of the release of raw product moments or simple correlation matrices of variables for blocks, tracts, or minor civil divisions. Such correlation matrices could vastly improve the quality of small-area data without further expanding the massive set of small-area cross tabulations already available.

Following this introduction, I summarize several recent empirical studies of urban housing markets concerned with the analysis of the spatial variation of housing prices and with the testing of alternative theories of housing market segmentation. Rather than emphasize the detailed and often conflicting findings of these studies, I concentrate on the difficulties inherent in applying available non-Census data sources to such an analysis.

I then present an ordinary least squares model of housing price variation estimated with Census data for the Pittsburgh SMSA. The empirical results both demonstrate the richness of the technique and provide a limited test of the extent of housing market segmentation in the Pittsburgh SMSA. A more general discussion of the estimation of alternative housing market models using Census data and a broad overview of the use of summary statistics in other forms of discrete

multivariate analysis follow. Finally, I make suggestions for the future release of Census data and offer a few concluding comments.

ECONOMIC ANALYSIS OF URBAN HOUSING MARKETS: AN OVERVIEW

In recent years, members of the urban studies group at the NBER have published a series of econometric analyses of urban housing markets (see Kain and Quigley 1975 for an excellent overview). Each of these studies employed a relatively large body of home interview data, which permitted extensive microeconomic analysis of urban housing markets in St. Louis, San Francisco, and Pittsburgh (Kain and Quigley 1975, Quigley 1972, and Straszheim 1975).

While the details of these studies differ, they share a common core of theory and method and their findings are broadly consistent. In each of these studies it is documented that individual households demand specific housing attributes or bundles of attributes. Some of these attributes are produced by individual housing suppliers, using land, durable capital goods, and less durable operating inputs. Other housing services, including neighborhood amenities and disamenities and public goods, are selected simultaneously with the choice of a dwelling unit. The production of these elements of housing consumption, however, are beyond the control of any single housing supplier, but rather depend on the collective action of large numbers of housing suppliers, demanders, and public officials.

Researchers have responded to the great complexity of housing markets in a variety of ways. The most common response is to assume it away. If housing markets are in long-run equilibrium, it is possible to ignore many aspects of heterogeneity in housing outputs and treat housing services as a single homogeneous commodity. As Olsen (1969, p. 614) contends: "In long run competitive equilibrium, only one price per unit applies to all units of housing stock and another price to all units of housing service regardless of the size of the package in which these goods come."

Given the empirical findings of the NBER econometric studies it seems unlikely that the treatment of housing output as a single homogeneous commodity selling in a single unified housing market is tenable. It is equally unlikely that at any instance housing markets are at or near long-run equilibrium or that prices for comparable components of housing service are uniform throughout a metropolitan area. Rather, as Straszheim (1975, p. 22) observes: "Heterogeneity in the existing stock, other differences in neighborhood desirability and the existence of discrimination imply that the urban

housing market is, in fact, a set of compartmentalized and unique submarkets delineated by housing type and location. Consequently, a great many markets must be considered, with complex interrelationships over time and space."

Straszheim was fortunate in having access to a data base with sufficient spatial and structural detail to test the implications of this theory of market segmentation and disequilibrium. Indeed, one distinguishing feature of the NBER econometric studies is their use of large samples of home interview data on households and dwelling units. These large samples permit the highly disaggregated micro analysis of housing prices and housing demand required to test hypotheses concerning a heterogeneous housing stock. It is of major importance that in these household surveys, information was collected on individual dwelling units, their location, and their occupants. Finally, these home interview data are especially useful because they cover both renter- and owner-occupied dwellings located throughout a metropolitan area.

Typically, the home interview data used by these NBER studies were obtained originally for other purposes. Each survey is unique unto itself, and each presents its own set of strengths and weaknesses. While these data have supported an impressive array of housing market studies, their unique features make generalizations across urban areas difficult.

In addition to NBER-sponsored activities, related analyses of housing markets have been presented by several other researchers, using relatively large samples of sales data generated by local property tax assessors, metropolitan mortgage bureaus, and local realtor groups (see Peterson 1974), and by the Federal Housing Administration (FHA) and other federal agencies.[1] These samples are often quite large and provide information on housing characteristics, location, and price. Property tax assessment data seem quite promising. Very often, for each parcel in the taxing jurisdiction, the assessor maintains a file on the characteristics of the lot and structure. For single-family homes, this is often accompanied by recent sales price and building permit information. Typically, all or part of this information is either on the public record or available to researchers subject only to appropriate assurances of confidentiality (see Peterson et al. 1973, Chap. 8). Given the improved quality of assessment techniques and the growing use of computerized retrieval systems, assessment data are likely to become an increasingly important source of housing market information.

This improved data on the sales price of owner-occupied housing are of considerable interest to both prospective home buyers and

public officials. Since home ownership represents a major source of wealthholding for middle-income families, evaluation of the impact of alternative government actions on the purchase price of single-family homes is an important public issue.

Tests of many theories of urban spatial structure, however, require data on rents as opposed to values. Knowledge of the relationship between employment accessibility and land values gives only approximate information on the relationship between accessibility and land rents. Even in a simple monocentric model of urban spatial structure, current land rents do not necessarily bear a simple relationship to current land values. The former depends on current transportation costs, population characteristics, and subjective evaluations of time, while the latter depends on these plus a market evaluation of their likely changes over time. As a result, sales data must be analyzed with great care. This point, however, is overlooked by most analysts of sales data, who readily compare values to rents, using simple real estate rules of thumb.

One common approach is to assume that the imputed rents for owner-occupied dwelling units are 1 percent of the total value of the property. For example, Polinsky and Rubinfeld present an elegant theoretical model of the benefits of environmental improvements; they then proceed to an empirical test of their theory which uses both renter- and owner-occupied dwelling units. To convert the price information they have for these two groups into comparable units, they multiply monthly rents by a factor of 100 (Polinsky and Rubinfeld 1975).

Similar rough approximations appear throughout the literature on housing demand and housing price formation. Recently, A. Thomas King (1972, 1973) developed a model of housing demand based on a household's maximization of a branched utility function. King formulated the theoretical approach in terms of income and housing rents, but he tested the theory using the sales prices and attributes of a sample of single-family homes located in the New Haven area. He recognized this discrepancy; yet, he concluded it was a trivial matter to convert values into rents (King 1972, p. 19, especially footnote 9).

While this assumption is convenient, it is unlikely that the treatment of the multiperiod investment aspects of housing purchases is as simple as the analyses by King and by Polinsky and Rubinfeld suggest. If housing markets are in long-run static equilibrium, the market price per unit of time for any piece of housing capital will be equal to the long-run supply price of capital. This in turn depends on the purchase price of capital, the rate of depreciation, and rate of interest. If alternative housing investments are

equally risky and have the same rate of depreciation and the same construction costs per unit, then in equilibrium, the value of any piece of housing capital will be a constant multiple of the annual rents generated by that capital stock.

There is considerable evidence, however, that the ratio of housing value to rent is not constant over structural type or location. Structures located in blighted areas often sell for only three or four times their annual rental receipts, while buildings in more desirable neighborhoods often sell for seven or eight times annual rents. Even within the same neighborhood type, the relationship between value and rent often differs by structural type (Peterson et al. 1973).

These problems make sales data particularly inappropriate for analysis of market segmentation. For any particular attribute, there could exist a well-defined pattern of spatial variation in current rents, reflecting excess supplies in some areas and excess demands in others. Yet depending on the extent to which these excess supplies and demands are expected to persist, the market value for the particular attribute could exhibit differing degrees of variation from its long-run supply price.

In addition to the problems of interpreting market value information, sales data also suffer from lack of uniform coverage. The typical analysis of housing sales data covers only single-family, owner-occupied dwelling units. In most areas these units are newer and more highly suburbanized than the rental housing or owner-occupied multiple stock. Thus, much of the great diversity of housing-structure types is ignored in analyses which concentrate on single-family, owner-occupied units.

Ann B. Schnare and Raymond Struyk (1974), for example, have recently completed a series of analyses using a sample of 2,200 single-family, owner-occupied houses located in thirteen suburban communities in the Boston SMSA. Using the standard regression approach, they attempted to explain sales price as a function of structure, neighborhood, and locational attributes. They concluded that there was no significant spatial variation in the sales price of individual housing attributes, and that substitution on the part of housing consumers and housing demanders in this submarket was adequate to prevent "widespread and pervasive market segmentation" (Schnare and Struyk 1974, p. 40). This could well be the case. It is impossible to determine, of course, whether their results are an indication that the market values have already discounted existing differentials in current rents, or whether, in fact, no current rent differentials exist within this relatively homogeneous subsample of dwelling units.

Despite the difficulty of analyzing housing markets, many researchers are attracted by the apparent simplicity of single-equation models of price variation and long-run equilibrium models of housing supply and demand. While model simplicity is often a virtue, the value of ignoring many crucial short-run-disequilibrium aspects of urban housing markets is less than obvious. Schnare and Struyk suggest that only simple models can be calibrated with available data and that data are not available to support empirical models of housing markets that incorporate the notions of housing market segmentation outlined above. They conclude (p. 2) that "the possibility of distinct market segmentation poses a real threat to the viability of statistical analyses of housing prices."

Their view of the possibilities for empirical analyses of housing markets seems overly pessimistic. As I illustrate in the next section, currently available Census Bureau data will permit many interesting tests to be made of housing market segmentation and the spatial variation of housing prices. Furthermore, these examples give only an inkling of the many analyses that can be conducted with Census data. While it is tempting to abstract from real-world complexities and present a simplified theory consistent with the most readily available data, it is often more rewarding to probe for ways to expand the capabilities of existing data sources for testing more realistic theories.

THE USE OF CENSUS SUMMARY STATISTICS IN AN ORDINARY LEAST SQUARES MODEL OF GROSS MONTHLY RENT

The census Public Use Sample Program has stimulated a great deal of empirical research based on Census data.[2] This program made available large samples of household interview data identified by subareas as small as individual counties or county groups with populations of 250,000 or more.[3] Unfortunately, the Bureau of the Census has concluded that confidentiality requirements prohibit the release of sample data identified by small areas of residence. Such microspatial detail would, of course, greatly expand the usefulness of the data for the analysis of many aspects of urban housing markets.

Fortunately, many forms of analysis do not require sample data. It is widely recognized that a raw product moment matrix provides the information needed for the estimation of the coefficients of an ordinary least squares model. Many available statistical packages provide the capability for processing a sample of observations, calculating the summary statistics needed to estimate the model, and

retaining this information for subsequent analysis. As a result, individual observations are not required to estimate a variety of ordinary least squares models. Instead, they can be estimated directly from Census data aggregated at different levels and obtained from a number of separate Census publications and computer tapes. For example, available Census tables permit estimation of housing market models that measure the effect of neighborhood and structural characteristics on the price of housing services.

This general proposition can be illustrated by Equation (5-1):

$$R_{ij} = \sum_{k=1}^{K} X_{ijk} A_k + \sum_{h=1}^{H} L_{ijh} B_h + e_{ij} \qquad (5\text{-}1)$$

which states that the rent of the ith dwelling unit in the jth subarea or neighborhood is a linear function of K attributes of the dwelling unit and its structure and H attributes of the neighborhood and location. The ordinary least squares (OLS) estimates of the coefficients require only information on the relationship between each of the variables taken two at a time. These moments can be obtained from data summarized at different levels of aggregation and spatial detail.

Consider, for example, the raw product moment between two of the structural variables. This calculation does not require information on the location of the dwelling unit, but only the joint distribution of the two variables over the entire study area. This can be obtained from samples of Census files containing no subarea information. In the case where the two attributes are defined in discrete terms, the moment can be obtained directly from a cross tabulation of the two variables aggregated to the areawide level. If the first variable is a dummy for the presence or absence of full plumbing and the second is a dummy for the presence or absence of central heating, then the raw product moment of the two variables is simply the count of dwelling units that have both attributes.

Most census variables are collected and presented in terms of a limited number of categories. Since many structural attributes can best be described in terms of the presence or absence of certain physical features, this categorization is adequate. In other instances, attributes are inherently continuous and their conversion to discrete categories results in a loss of information, particularly for open-ended categories. It should be observed that this shortcoming is not unique to census data and that the extensive pretesting of questionnaires by the Census enables it to use categories which minimize the loss of information.

For certain variables, continuous information is collected. Rent is one such variable. Even so, the continuous distribution of rent is not needed to estimate the linear rent equation. As long as all the structural variables are represented in discrete terms, then the raw moments involving those variables and rent can be exactly determined using the mean rent and the count of all rental dwelling units in the entire area for each level of the discrete structural variables.

An estimate of the variation of rent does require continuous information. While an approximation can be made by assigning values to each of a number of discrete intervals, this procedure produces an unknown loss of accuracy which affects measures of goodness of fit, including, of course, R^2. It should be observed, however, that since the variation of rent is not required for the calculation of the individual coefficients of an ordinary least squares model, the impact of the lack of continuous rent information in an ordinary least squares framework is greatly reduced.

Unlike structural attributes, locational detail is needed to obtain the product moments involving neighborhood attributes. If the neighborhood attribute is assumed to hold for the entire subarea, the only subareal structural information required is a summation of the attributes over all dwelling units in the subarea. The raw product moments involving both structural and neighborhood variables are weighted sums of the aggregate structural characteristics for each neighborhood, where the weights are the values of the neighborhood attribute in question.

Moments involving only neighborhood attributes can be obtained in a similar fashion. Since the neighborhood attributes are assumed constant within a subarea, the raw product moment between two neighborhood attributes can be exactly estimated by taking a weighted sum of the product of the two variables. In this instance the weights are the total number of observations in each subarea or neighborhood.

The locationally specific data required for neighborhood moment calculations can be obtained from a number of sources. For 1970, the Bureau of Census computer tapes contain the distributions of structural variables required to estimate rent functions for metropolitan areas at the census tract and block level. If parts of the study area are not tracted, the minor civil division or enumeration district can be used to form the subareas. The appropriate delineation of subarea depends on the nature of the neighborhood data used. These neighborhood variables can be obtained from land use planning studies, transportation surveys, and similar sources. If the data reveal that broad sections of the study area are highly homogeneous, census

tract aggregates could suffice as subareas. In other instances individual tracts or even block data might be needed to capture subtle spatial variations in neighborhood amenities.

The preceding discussion has been rather general and was intended to introduce the notion that the estimation of a linear rent equation does not necessarily require sample data with locational detail, but can be estimated with aggregate and subareal data of a special nature. In the next section I present an example to further illustrate this technique. I then discuss other similar models that can be estimated using aggregate census data.

A RENT EQUATION FOR THE PITTSBURGH SMSA

The statistical approach outlined in the preceding section can be applied to any urban place. To illustrate these techniques, I use Census tables for the Pittsburgh SMSA. Data published in the Metropolitan Housing Characteristics (MHC) series and available on the Fourth Count Summary tract tapes for Pittsburgh in 1970 were used to estimate Equation (5-2).

$$R_{ij} = A_0 + \sum_{k=1}^{9} X_{ijk} A_k + \sum_{h=1}^{6} L_{ijh} \tag{5-2}$$

where

j designates a census tract (702 in all) in the Pittsburgh SMSA.

i designates the individual occupied renter units in each tract.

X_1 to X_4 are dummy variables for structure. X is 1 if the dwelling unit is in a two-family house and 0 otherwise. The other three categories are 3-4 units, 5-9 units, and 10 or more units; single-family units are represented in the constant term.

X_5 to X_7 are age dummy variables. X_5 is 1 if the dwelling unit was built during 1950-1959 and 0 otherwise. The other two categories are: 1940-1949 and before 1940. Units built since 1959 are represented in the constant term.

X_8 is a dummy variable for plumbing. X_8 is 1 if the dwelling unit has only partial plumbing facilities and 0 otherwise.

X_9 is the number of rooms in the dwelling unit.

L_1 is a measure of accessibility for each tract in minutes of one-way travel time. It was calculated using a matrix of

zone-to-zone travel time and employment location and a standard exponential decay weighting function. The travel times are based on 1967 data. (See Ingram 1971, App. A, for a fuller discussion of this variable.)

L_2 is the average income of all families in the tract, in thousands of dollars.

L_3 is the black population as a percent of the total population in each tract.

L_4 is net residential density of each tract in units per residential acre.[4]

L_5 is a dummy for tract location. L_5 is 1 for tracts located in the City of Pittsburgh and 0 otherwise.

L_6 is a dummy for race of the head of household residing in the unit. L_6 is 1 for black-occupied units and 0 otherwise.

Data from Metropolitan Housing Characteristics (MHC) and Fourth Count do not permit an exact estimation of Equation (5-2). There are three minor sources of error. First, there are minor inconsistencies between the Fourth Count and MHC data. For five census tracts with fewer than fifty rental units, only the total number of renter-occupied dwelling units is provided. All the other variables for these tracts were suppressed. For several other tracts, some of the required variables were suppressed. In the Pittsburgh analysis, some or all of the tract level variables had to be imputed for approximately 0.3 percent of all rental dwelling units.

Second, the total number of renter-occupied dwelling units obtained from the Fourth Count and MHC differed by 0.5 percent. Since the two sources were released at different times, this discrepancy could reflect differences in error editing. In any event, it was a minor matter to adjust for these differences by scaling the tract level data to agree with the published aggregates.

The most difficult problem arises from the Census definition of the renter-occupied subsample. For the Pittsburgh SMSA, 5.3 percent of all renter households were enumerated as paying "no cash rent." Another 1.6 percent were units located on lots with ten or more acres. No rental information was collected for these units. Unfortunately, both groups are included in the cross tabulations of the structural attributes. To adjust for this problem, the product moments involving the rental information were scaled to reflect the differential coverage of the structural and rental data.

In Table 5-1, I present the means and standard deviations for gross monthly rent and fifteen explanatory variables for the entire Pittsburgh SMSA. In 1970, Pittsburgh had 245,085 renter-occupied

Table 5-1. Rent Regressions for Total SMSA Sample, Pittsburgh, 1970: Means and Coefficients of Individual Variables (number of observations = 49,017; figures in parentheses below means are standard deviations; figures below coefficients are standard errors)

	Mean	*Regression Coefficient*
Constant		−32.25
Structure-type dummies[a]		
2 units	0.21	−1.24
	(0.41)	(0.45)
3-4 units	0.15	0.18
	(0.36)	(0.52)
5-9 units	0.12	4.84
	(0.32)	(0.57)
10 or more units	0.19	14.52
	(0.39)	(0.56)
Year-built dummies[a]		
1950-1959	0.10	−29.07
	(0.30)	(0.67)
1940-1949	0.12	−41.68
	(0.33)	(0.65)
Before 1940	0.65	−47.69
	(0.47)	(0.54)
Plumbing dummy: partial plumbing[a]	0.07	−17.21
	(0.26)	(0.63)
No. of rooms per dwelling unit	4.12	12.97
	(1.47)	(0.13)
Accessibility (minutes of one-way travel time)	22.47	2.09
	(4.37)	(0.48)
Average tract income (thous. dol.)	9.20	8.04
	(2.65)	(0.74)
Tract percent black	11.59	0.01
	(24.36)	(0.01)
Net residential density (units per acre)	15.31	0.20
	(24.48)	(0.01)
Location dummy, Pittsburgh	0.36	1.21
	(0.48)	(0.46)
Race dummy, black[a]	0.13	−2.91
	(0.34)	(0.69)
Rent per unit	108.35	
	(52.25)	
R^2		0.557
Standard error		34.689
Number of observations		49.017

Source: Means are from U.S. Census of Population and Housing, 1970, *Metropolitan Housing Characteristics for the Pittsburgh SMSA*, and Fourth Count Summary Tapes for the Pittsburgh SMSA.

[a]See text for explanation of dummies.

dwelling units. Both the tract data and the SMSA level data used to estimate the model summed to this total. Data were collected from only a sample of the population except for a few variables, for which complete-count estimates were obtained by scaling a 20, 15, or 5 percent random stratified subsample. Since the census variables used in this model were all drawn from the 20 percent sample, it was assumed in calculating sample statistics and degrees of freedom that the actual sample for the entire SMSA contained only 49,017 renter-occupied dwelling units (e.g., 49,017 = 0.2 × 245,085).

With the exception of these minor problems of internal consistency, the estimates obtained are the same as those that could have been obtained using a sample of Census households identified by Census tract of residence. While Equation (5-2) could have been estimated for any metropolitan area, the Pittsburgh SMSA was used because data for several useful neighborhood characteristics such as measure of accessibility, tract net residential density, percent nonwhite, and mean tract family income were readily available. While there is no limit to the number of neighborhood variables that could have been used in the analysis, additional neighborhood variables would have turned out to be highly correlated with the four variables included, and would have needlessly complicated this illustration of technique. In addition, it would have been possible to use the moment matrices generated for this example and ridge regression techniques to investigate the effect that multicollinearity has on the stability of the coefficient estimates; but again, this would tend to confuse the central methodological thrust of the paper (see Hoerl and Kennard 1970).

Also shown in Table 5-1 are the estimated coefficients for Equation (5-2) and their standard errors for the entire Pittsburgh SMSA. Overall, the model performs reasonably well, explaining 55 percent of the variance, and each of the coefficients is highly significant. This second result is hardly surprising, however, given that the regression is based on nearly 50,000 observations.

The individual coefficients conform fairly well to expectations. Older dwelling units rent at a discount, as do units with only partial plumbing. Additional rooms rent for an extra $12.97 each. As compared to a single-family unit, large multiple units appear to rent at a premium. This finding in part reflects the fact that large multiples are newer on average than single-family renter units and provide services not measured by the structural variables included in the equation.

The neighborhood coefficients also have plausible value. Greater accessibility, central location, higher mean tract income, and higher density are all associated with higher rents. The percent black

coefficient is small and insignificant, but the dummy variable for black occupancy is significantly negative.

Perhaps more important than the test results of this exercise for the particular model is the illustration that a linear rent equation can be estimated by using census data obtained from two sources. The problems discussed previously, i.e., suppression of some tract data, release of MHC and Fourth Count data at different times, and apparent differences in error editing and in the treatment of no cash rent, could have created insurmountable problems. Fortunately, the inconsistencies that did exist were small. The two sources had virtually identical means and variances for the common structural characteristics, often differing in only the fourth or fifth significant digit. In addition, in those few instances where tract level cross tabulations were available, the moments available from tract data were virtually identical to those derived from published SMSA level data. Those differences that did appear seem to result from a difference of approximately 0.5 percent in the total number of renter-occupied dwelling units enumerated on the Fourth Count tapes compared to MHC. An adjustment for this discrepancy was made by scaling the tract level data by the appropriate ratio.

While the results shown in Table 5-1 are interesting in their own right, their more important value is that they illustrate the wide range of analyses possible using aggregate Census data. In the next section the analysis presented in the table is extended by presenting additional OLS estimates for several stratified samples.

TEST OF INTERACTION FOR THE OLS FRAMEWORK

One central empirical and theoretical issue in the analysis of housing markets is the extent to which particular housing attributes rent for equal amounts across space or across types of occupants. Tests of this market segmentation hypothesis have proliferated in recent years. In this section I illustrate how available Census data can be used to test that hypothesis in metropolitan housing markets.

In addition to the SMSA-wide data used in the previous section, the Metropolitan Housing Characteristic series also presents the same contingency tables for all households and for black-headed households for each city over 50,000 and the suburban ring. In the Pittsburgh SMSA, only the central city and the suburbs are identified, but in other SMSAs numerous non-central-city locations are identified as well. This further stratification permits estimation of Equation (5-2) for black- and white-occupied rental units located in the central city and the suburbs. Table 5-2 contains the means and

Table 5-2. Means and Standard Deviations of Structural and Neighborhood Variables for Renter-occupied Dwelling Units, by Race and Location, Pittsburgh, 1970 (figures in parentheses are standard deviations)

	White-occupied		*Black-occupied*	
	City	*Suburbs*	*City*	*Suburbs*
Structure-type dummy				
2 units	0.21	0.22	0.15	0.18
	(0.41)	(0.41)	(0.35)	(0.39)
3-4 units	0.17	0.14	0.16	0.14
	(0.37)	(0.35)	(0.37)	(0.34)
5-9 units	0.12	0.10	0.19	0.11
	(0.32)	(0.31)	(0.39)	(0.32)
10 or more units	0.29	0.15	0.17	0.16
	(0.45)	(0.36)	(0.38)	(0.36)
Year-built dummies				
1950-1959	0.07	0.12	0.07	0.11
	(0.26)	(0.32)	(0.25)	(0.31)
1940-1949	0.10	0.13	0.18	0.16
	(0.30)	(0.33)	(0.38)	(0.37)
Before 1940	0.74	0.61	0.64	0.64
	(0.44)	(0.49)	(0.48)	(0.48)
Plumbing dummy: partial plumbing	0.10	0.06	0.10	0.09
	(0.30)	(0.24)	(0.29)	(0.28)
No. of rooms per dwelling unit	3.72	4.29	4.09	4.34
	(1.53)	(1.43)	(1.45)	(1.29)
Accessibility	26.08	20.38	26.36	20.34
	(1.13)	(4.15)	(1.33)	(3.81)
Average tract income (thous. dol.)	9.63	9.58	6.09	7.80
	(3.21)	(2.08)	(2.04)	(1.81)
Tract percent black	7.75	3.13	70.32	2.77
	(16.49)	(7.46)	(29.63)	(2.61)
Net residential density	27.60	8.59	23.58	13.94
(units per acre)	(39.45)	(10.98)	(14.75)	(19.17)
Rent per unit	116.47	108.69	90.69	89.19
	(60.91)	(50.18)	(35.83)	(33.40)
Number of observations	13,182	29,362	4,573	1,900

standard deviations for each of the variables. Again, the sample size has been adjusted to reflect the fact that the underlying data were based on only a 20 percent enumeration.

Tables 5-3, 5-4, and 5-5 contain estimated coefficients for the stratified subsamples. Table 5-3 contains estimates for the central-city and suburban subsamples, and in Tables 5-4 and 5-5, the sample is stratified by both race and location.

Table 5-3. Coefficients of Individual Variables in Rent Regression for City and Suburban Samples, Pittsburgh, 1970 (figures in parenthese are standard errors)

	City-Sample Coefficients	*Suburban-Sample Coefficients*
Structure-type dummies		
2 units	1.683	−1.961
	(0.916)	(0.511)
3-4 units	7.392	−3.522
	(0.982)	(0.594)
5-9 units	14.713	−0.879
	(1.050)	(0.662)
10 or more units	24.724	8.333
	(1.031)	(0.659)
Year-built dummies		
1950-1959	−27.570	−30.932
	(1.429)	(0.734)
1940-1949	−34.814	−45.254
	(1.286)	(0.726)
Before 1940	−40.188	−51.692
	(1.091)	(0.605)
Plumbing dummy: partial plumbing	−17.683	−16.695
	(1.060)	(0.791)
No. of rooms	13.341	12.909
	(0.234)	(0.148)
Accessibility	0.210	2.196
	(0.251)	(0.047)
Average tract income	7.950	8.229
	(0.116)	(0.110)
Tract percent black	−0.050	0.152
	(0.153)	(0.020)
Net residential density	0.174	0.339
	(0.009)	(0.018)
Location dummy, Pittsburgh	–	–
Race dummy, black	2.248	−7.695
	(1.106)	(0.889)
Constant	6.556	−32.098
R^2	0.569	0.617
Standard error	37.241	30.656
Number of observations	17,755	31,262

Table 5-4. Coefficients of Individual Variables in Rent Regression for White-occupied Dwelling Units, by Location, Pittsburgh, 1970 (figures in parentheses are standard errors)

	SMSA Coefficients	*City-Sample Coefficients*	*Suburban-Sample Coefficients*
Structure-type dummies			
2 units	−1.23	2.79	−1.90
	(0.49)	(1.08)	(0.53)
3-4 units	0.37	10.20	−3.20
	(0.56)	(1.18)	(0.61)
5-9 units	5.91	21.20	−0.16
	(0.62)	(1.33)	(0.68)
10 or more units	17.44	32.70	9.68
	(0.61)	(1.26)	(0.69)
Year-built dummies			
1950-1959	−30.95	−35.56	−31.20
	(0.72)	(1.71)	(0.75)
1940-1949	−45.92	−48.19	−46.65
	(0.71)	(1.62)	(0.75)
Before 1940	−52.51	−56.15	−53.12
	(0.58)	(1.35)	(0.62)
Plumbing dummy: partial plumbing	−18.79	−20.01	−17.22
	(0.70)	(1.26)	(0.83)
No. of rooms	13.45	14.68	13.19
	(0.14)	(0.28)	(0.15)
Accessibility	2.09	−0.37	2.21
	(0.04)	(0.31)	(0.04)
Average tract income	7.67	7.01	8.15
	(0.10)	(0.13)	(0.11)
Tract percent black	−0.11	−0.31	0.10
	(0.02)	(0.02)	(0.03)
Net residential density	0.20	0.15	0.37
	(0.01)	(0.01)	(0.02)
Location dummy, Pittsburgh	2.07		
	(0.49)		
Constant	−26.92	38.80	−32.32
R^2	0.615	0.602	0.629
Standard error	33.42	38.46	30.56
Number of observations	42,544	13,182	29,362

Table 5-5. Coefficients of Individual Variables in Rent Regression for Black-occupied Dwelling Units, by Location, Pittsburgh, 1970 (figures in parentheses are standard errors)

	SMSA Coefficients	*City-Sample Coefficients*	*Suburban-Sample Coefficients*
Structure-type dummies			
2 units	−0.29	0.09	−1.15
	(1.11)	(1.36)	(1.86)
3-4 units	−2.04	−0.09	−7.30
	(1.15)	(1.37)	(2.07)
5-9 units	−2.68	0.01	−7.76
	(1.14)	(1.32)	(2.31)
10 or more units	0.54	1.85	−9.24
	(1.20)	(1.39)	(2.34)
Year-built dummies			
1950-1959	−11.35	−12.29	−13.54
	(1.66)	(2.01)	(2.96)
1940-1949	−8.50	−6.70	−12.72
	(1.42)	(1.64)	(2.83)
Before 1940	−11.13	−8.03	−19.74
	(1.23)	(1.46)	(2.60)
Plumbing dummy: partial plumbing	−11.98	−10.89	−14.43
	(1.29)	(1.52)	(2.39)
No. of rooms	10.98	11.42	9.33
	(0.29)	(0.35)	(0.55)
Accessibility	1.05	−1.38	1.50
	(0.15)	(0.35)	(0.18)
Average tract income	6.39	6.63	4.44
	(0.23)	(0.26)	(0.53)
Tract percent black	0.01	−0.02	0.12
	(0.02)	(0.02)	(0.03)
Net residential density	0.25	0.33	0.12
	(0.02)	(0.03)	(0.04)
Location dummy, Pittsburgh	6.40	–	–
	(1.22)		
Constant		41.95	−0.56
R^2	0.380	0.428	0.325
Standard error	27.70	27.14	27.71
Number of observations	6,473	4,573	1,900

Again, most of the coefficients are highly significant. The major exceptions are structure-type dummy variables in the black-occupied equations. For the black-occupied city subsample, none of the dummy variables has coefficients that are twice their standard errors. In general, the black renter equations do not perform as well as those for white renters. The R^2's are only 0.43 and 0.33 for the black-occupied city and suburban subsamples compared to 0.60 and 0.63 for the comparable white subsamples.

Finally, the coefficients of the basic equation seem to vary from one sample to the next. The rent paid for each additional room ranges from $14.68 for white households living in the city to $9.33 for black households living in the suburbs. The discount for buildings built before 1940 is $56.15 for white-occupied units in the city. The discount for black-occupied units is only $8.03 for units located in the city and $19.74 for black-occupied suburban dwelling units.

F tests on the equality of coefficients, shown in the tabulation below, confirm the finding that the coefficients of individual attributes are statistically different for blacks and whites living in the central city and the suburbs:

	R^2	*Stand. Error*	*Degrees of Freedom*	*F Test*
Pooled model	.554	34.69	–	–
Stratified models				
Race	.608	32.73	15;48,987	404.7
Location	.597	33.18	15;48,987	302.5
Race and location	.615	32.50	14;48,961	491.6

The hypothesis that the coefficients of the four subsamples differ by only a location and a race dummy can be rejected at the 0.01 level of confidence. Similarly, attempts to pool the data by race or location were rejected at the same level of confidence.

The foregoing table also demonstrates that the stratifications significantly improve the explanatory power of the model. The pooled model explains 55 percent of the variance in monthly gross rent. One-way stratification by location increases the explanatory power of the model by five percentage points, and one-way stratification by race increases the explanatory power by six percentage

points. Two-way stratification increases the explained variance by nearly seven percentage points—to 62 percent. Similarly, stratification of the equation by location and race reduce the aggregate standard error by approximately 7 percent.

As was noted earlier, there has been considerable debate in recent literature over the empirical importance of alternative market segmentation hypotheses. The central concern is the extent to which a single hedonic index for housing consumption can be identified, or whether market segmentation undermines the usefulness of the hedonic approach or any other unidimensional description of housing services. The issues take on added significance when it is realized that the validation of several major policy planning models is heavily dependent on the assumption that a single hedonic index can be identified using a single metropolitanwide or cross-metropolitan model of observed market rents; for example, important aspects of the calibration of the Urban Institute Housing Market Simulation Model are based on the single-equation hedonic approach (see de Leeuw and Struyk 1975).

Given the limited number of variables included in the illustrations presented in this study, it would be inappropriate to assign too much significance to the reported test statistics. It is likely that many of the differences observed in the coefficients reflect measurement problems and the omission of relevant variables. The inclusion of a more detailed set of dummy variables to represent the age and number of units in the structure should reduce these measurement problems, as should the inclusion of other indicators of structural quality such as the presence of central air-conditioning, the type of heating system, the number of bathrooms, and the nature of the available kitchen equipment. In addition, improved specification variables that measure the quality of local services and the attractiveness of the neighborhood would undoubtedly help. Finally, additional stratifications and tests for interaction should be performed before any set of equations is given much credence.

With these caveats in mind, it does seem that the results presented in these illustrations are consistent with the hypothesis of housing market stratification. While this interpretation could well be altered in light of additional estimation, it should be stressed that unlike models estimated with single-source data bases, the models presented here can be replicated for any SMSA, major city, or group of counties in the United States. Moreover, unlike many alternative analyses of housing prices, which utilize samples of sales prices, the equations estimated in this study are based on a sample of renter-occupied dwelling units. This frees the analysis from complications

introduced by the consideration of the multiperiod investment aspects of the purchase of a dwelling unit. Finally, it is also possible both to expand the list of individual structural attributes and enrich the neighborhood definition and stratification schemes. As I illustrate in the next section, these refined models can also be estimated by using Census housing data.

ALTERNATIVE MODELS OF HOUSING PRICES

The 1970 Census survey contained thirty-five housing questions. Of these, fifteen were collected for all dwelling units in the United States. Other housing questions were obtained for samples of 20, 15, or 5 percent. In each instance, the results of the sample surveys were scaled to complete-count control totals.[5] In addition to the data for number of rooms, structural type, year built, and partial plumbing, which were used in the previous sections, information was also gathered on the number of bedrooms and bathrooms; the type of bathroom, kitchen, cooking, heating, and air-conditioning facilities available for each dwelling unit; and the source of water sewage disposal, number of stories, and presence of elevators in each structure or building.

Extension of the OLS models presented in the previous section requires aggregate cross tabulations, or two-dimensional contingency tables of the included variables, as well as one-way frequency distributions at the neighborhood level. All the needed one-way frequency distributions are available on the Fourth Count Summary Tapes for census tracts and minor civil divisions. The Bureau of the Census has also announced plans to release similar data for census blocks and enumeration districts (Census 1973c). The usefulness of this data in the linear rent model depends, of course, on the level of data suppression and the consistency of this new data source with previously released Census data.

Obtaining the required contingency tables for extension of the linear rent model is somewhat more difficult. In addition to data published in the Metropolitan Housing Characteristics series, many contingency tables can be found on the Sixth Count Summary Tapes, which provide summary housing data for states, SMSAs, metropolitan counties, nonmetropolitan counties of 50,000 inhabitants or more, cities of 50,000 inhabitants or more, and central cities. For each of these areas, the Sixth Count tapes provide 109,061 cells of housing market data in 348 tables (Census 1972). Of these, 164 tables involve one or more population items, while the other 184 involve only housing characteristics.

Using Sixth Count data, it is possible to further stratify the rent equation presented in the previous sections by counties or by cities of 50,000 or more. For large SMSAs, this would represent a significant improvement over the central-city and suburban stratifications used in the previous Pittsburgh examples. Because of the large number of tables and diverse levels of stratification, the Sixth Count Summary information for a single state may occupy as many as fourteen reels of computer tape. While the size of the Sixth Count files makes the extraction of the required tables somewhat difficult, the improvement in model specification should be well worth the effort.

Even though a large number of tables are on the Sixth Count Summary tapes, the required information for use of all thirty-five housing variables in a single linear rent equation is not on these tapes or in other published Census information. Such an equation would require the calculation of 630 raw product moments, and only a fraction of the needed contingency tables are contained on the Sixth Count tapes. Thus, for example, in addition to testing for the impact on rent of age, structural type, plumbing, and number of rooms, it would also be possible to test for the impact of either the type of heating or air-conditioning equipment. All the required cross tabulations are available on Sixth Count. The inclusion of both the heating and air-conditioning variables in a single equation, however, requires the joint distribution of these two variables, information not contained on those tapes.

The lack of Sixth Count information should pose no real problem to the specification of more complicated rent equations. The missing contingency tables or product moments could be obtained from a special tabulation of the Census data. This, of course, is a costly activity requiring careful planning and allowance for sufficient time to permit the bureau to program and process the initial request. Once this has been done, subsequent requests for similar data for other metropolitan areas should prove less difficult and less expensive to obtain.

Somewhat less precise estimates of the required moments can be obtained directly from the Public Use Sample tapes. These tapes contain a 1/100 sample of all Census questionnaires and include all the housing information collected. As was noted earlier, however, the Public Use Sample tapes do not present detailed spatial information. This is no obstacle to the estimation of linear rent equations, since calculation of the raw product moments involving only structural attributes does not require information on the location of the dwelling unit.

The main difficulty with using Public Use Sample data in the estimation of product moments results from possible sampling errors and inconsistencies between sample and complete-count information. For example, estimates of variance of each of two variables obtained from complete-count data and similar estimates of the covariance between these two variables obtained from sample data could well produce an estimate of simple correlation between the variables which falls outside the zero-one interval, a disconcerting prospect, indeed.

Yet a simple adjustment of the sample data will eliminate possible inconsistencies. Consider, for example, the estimation of the raw moment between type of heating equipment and type of air-conditioning equipment available in the dwelling unit. The joint distribution of these two variables obtained from the sample data can be scaled to satisfy the known complete-count one-way distribution of each variable. This procedure preserves the cross-product ratios present in the sample data while ensuring that the sample cross tabulation is consistent with the complete-count data.[6] Finally, it should be observed that scaling the data also improves the efficiency of the sample estimate of the raw product moment involving the two variables, since the procedure is equivalent to creating a random stratified sample of observations.

Further improvement in the efficiency of the sample estimates of the raw product moments can be achieved by scaling the sample data to satisfy multiple-dimension contingency tables. For example, each observation of the sample tapes could be classified by type of structure, heating equipment, and air-conditioning equipment. The observations could then be adjusted to satisfy both the complete-count joint distribution of structural type and type of heating equipment and the complete-count joint distribution of structural type and type of air-conditioning equipment. This scaling requires an iterative technique, since adjusting the data to satisfy the first joint distribution may produce results inconsistent with the second joint distribution. As was true of the simple two-dimensional problem, this three-way scaling problem is equivalent to the formation of a random stratified sample and should further improve the efficiency of the sample estimates of raw product moments or other statistics.

By following the above procedures it should be possible to convert Public Use Sample data into usable estimates of raw product moments not available from complete-count data. This permits the estimation of rent equations using all of the housing variables collected by the Census Bureau, while still considering the potential impact of neighborhood and locational characteristics. While it is

possible that sampling error present in the estimated product moments will result in biased estimates of the coefficients of the model, careful stratification and scaling of the sample data along the lines previously suggested should greatly reduce the remaining error. In general, the value of the Public Use Sample data for any type of analysis will be enhanced by scaling that data to match available complete-count, one-way, or joint frequency distributions. Indeed, the Public Use Sample of the U.S. Census is one of the few sources of sample data for which complete-count control totals for each variable are available.

As the above discussion illustrates, Census data can support a very rich ordinary least squares analysis of housing price variation and market segmentation. The estimation of the regression equation required only raw product moments for each pair of variables. Although it is possible to use available Census data to perform tests of market segmentation by race and a crude stratification by location, other dimensions or market segmentations are also of interest.

Assuming that Census tracts correspond roughly to homogeneous neighborhoods, it is possible to estimate individual equations for each tract. In addition to the one-way frequency distributions used in the earlier analysis, the Fourth Count Summary tapes also present a number of useful tract-specific cross tabulations. For each tract in the Pittsburgh SMSA, for example, information is available on the joint distribution of rent payments and structural type, rent payments and age of structure, and age of structure and structural type. By assigning values to each of the rent categories it is possible to calculate the raw product moments needed to estimate the coefficients of a model in which it is assumed rent is a linear function of structural type and age for each tract.

Lack of available cross tabulations severely limits the possibilities for tract-specific rent equations. In an earlier paper, Apgar and Kain presented a series of tract-level regression equations. Data considerations, however, limited the analysis to four housing variables, i.e., structural type, age, plumbing, and number of bedrooms (see Apgar and Kain 1972). Even at this reduced level of detail, several of the required raw product moments could not be obtained from the data. Rather, it was necessary to estimate values for these missing moments that were consistent with the known tract-level one-way distributions of the variables.[7]

Despite these difficulties, the tract-specific regressions for the Pittsburgh metropolitan area did reveal some significant differences in the variation of attribute prices over neighborhood and location. It

is of more importance to the current discussion that the analysis of Pittsburgh tract data demonstrated the ease of estimating regressions directly from raw product moments. Three separate functional forms of the basic equation were estimated for owners and renters stratified by race and located in each of the 702 Census tracts in the Pittsburgh area. In all, sufficient data were available for the estimation of nearly 4,000 separate equations. If software development costs are excluded, the average computer cost for estimating each equation and printing out the results were approximately four cents. It is unlikely that an analysis using raw sample data could have been completed at any lower cost.

It should be clear, then, that the ability to estimate an OLS model of housing price variation is limited only by the detail of available summary data. As a result, the release of cross-product matrices for housing variables aggregated by block, tract, or minor civil division would do much to offset the lack of spatially detailed Census sample data. These matrices would be the building blocks for numerous regression analyses based on Census data and would greatly improve the quality of Census housing analysis. Indeed, such a program for improved quality of small-area data would have significant value for other areas of social science research; and in the final section of this study I briefly consider some of these alternative applications.

CENSUS SUMMARY DATA AND DISCRETE MULTIVARIATE ANALYSIS

I have presented a detailed discussion of the estimation of a series of OLS models of housing prices. This emphasis on housing prices should not obscure the fact that the techniques presented are equally applicable to other areas of research. Nor should it be assumed that the approach is limited to OLS analysis. Indeed, given the current availability of Census cross tabulations, many types of discrete multivariate analysis can be conducted with Census summary data. In an effort to underscore the potential richness of census-based econometric analysis, I present some alternative uses of the summary data.

The importance of home ownership as a vehicle for wealth accumulation is well documented. Yet it is clear that black households are systematically excluded from the benefits of owner occupancy (Kain and Quigley 1972). Two explanations come quickly to mind. First, it is likely that racial discrimination limits black residential choice to portions of the housing stock which are inappropriate vehicles for home ownership. Second, it is likely that

even if black residential areas do expand to include suitable housing stock, racial discrimination by lending institutions could prevent black households from achieving owner status.

It is possible to evaluate these two explanations with a logit model of home ownership. Consider, for example, the basic logit model, which equates the log of the odds of home ownership with a linear function of race, family income, family type, and neighborhood type. Related models could include the interaction between family income, neighborhood, and race, or other higher-order interactions.

As was the case with the OLS models, tests for the presence of interaction in this spatially detailed logit model of home ownership could be conducted with sample data identified by small areas. Assuming that the census tracts correspond to a desired typology of neighborhoods, it is also possible to estimate these logit models using a more limited set of neighborhood-specific cross tabulations. In this problem the parameters of the simple linear logit model can be obtained by combining the four-way cross tabulation—race by family income by family type by neighborhood type—and each of the two-way cross tabulations involving tenure into a single estimate of the full five-dimensional array—race by tenure by family type by family income by neighborhood type. Using an iterative procedure, an estimate can be obtained of the unknown five-dimensional array that satisfied the known set of marginal distributions or interaction assumptions. In this instance, the assumption is that the five-way array can be fully described with one four-dimensional marginal summary and four two-dimensional marginal summaries. All other possible higher-order interactions are assumed to be absent. By using the values produced from this estimated array, coefficients can be constructed for the linear logit model described above. While the procedure is somewhat complicated to describe, it is in fact computationally quite simple. The approach follows directly from Leo Goodman's demonstration that certain hierarchical hypotheses concerning the interaction structure of a multivariate contingency table correspond to the logit analysis of a dichotomous variable (Goodman 1970).

Hypotheses concerning the presence of interaction terms in the logit model can be tested by fitting more complex models. The only matrix not present in the Census tract summary data is the five-way cross tabulation involving all of the variables in the problem. This precludes the possibility of fitting the so-called saturated model, i.e., the model that corresponds to the hypothesis that all possible interactions are represented in the data in a statistically significant manner. If it is correct to reject this hypothesis, then it is possible to

test for the presence of other interactions in the logit framework. The tests are based on the usual chi-square goodness-of-fit statistic and follow directly from Goodman's observations on the partitioning of hierarchical hypotheses on multidimensional contingency tables.

Since the preceding discussion was somewhat terse, the reader is referred directly to the growing literature on discrete multivariate analysis, for example, the excellent textbook by Bishop et al. (1974). It is hoped, however, that this presentation has been sufficiently detailed to establish one rather simple point. Just as was true with OLS models of housing markets, available small-area Census Bureau summary data can support detailed logit analyses. As the above discussion illustrates, use of Census data in a logit analysis of home ownership could be quite rewarding. It is likely that such analysis would lead to the rejection of the simple hypothesis that each of the variables affects the log of the odds of home ownership in a linearly additive fashion. More complex specifications could be tested, however—a process that would undoubtedly enhance our understanding of the relationship between race and home ownership.

Numerous other examples of discrete multivariate analysis based on Census summary data could be presented. Literally thousands of multiple-dimension contingency tables are presented in the First to the Sixth Count Summary Tapes for 1970. In addition, for both 1960 and 1970, a number of other cross tabulations have been released as a byproduct of the publication of the special Census Subject Reports. These present the most detailed information available from Census sources on such diverse topics as modal choice, journey to work, intrametropolitan mobility of households, and the occupational and geographical mobility of workers. Each of these separate Census products is a potentially rich source of the summary statistics needed to perform useful logit, OLS, or other discrete multivariate analyses of important issues.

CONCLUSION

Although confidentiality requirements may prohibit the Bureau of the Census from releasing sample data identified by detailed geographic area, many empirical tests of important hypotheses do not require information on individual households, but rather a much reduced set of summary statistics aggregated at different levels. Such exercises will succeed only if these summary statistics are internally consistent; yet as this study has illustrated, current procedures generate data with several minor inconsistencies. Correction of these problems would greatly increase the accuracy and ease of estimation of Census-based econometric models.

First, the Census should attempt to reconcile the differences between the alternative sources of housing data. As was noted earlier, the total number of renter-occupied dwelling units obtained from the Fourth Count and Metropolitan Housing Characteristics do not agree. Since these two sources were released at different times, this discrepancy could reflect differences in error editing. While the desire for the prompt release of Census data is understandable, the need for their internal consistency is an equally important objective. If errors are discovered during the early phases of census processing, subsequent release of the adjusted data would be of considerable importance.

Second, the Census policy of data suppression should be reassessed. There is no apparent advantage in defining a Census tract with so few households that summary statistics for all households are suppressed. These small tracts should be aggregated with larger ones so that tract-level summary statistics can be presented that are consistent with similar statistics published for an entire county or metropolitan area.

Third, the Bureau of the Census should reassess its treatment of households who pay "no cash rent" and rental units located on lots with ten or more units. Unfortunately, both groups are included in the cross tabulations of the structural attributes. Since both types reflect special situations, it would be good sense to exclude them from all tabulations of renter-occupied housing units. Separate tabulations for those two situations could be presented instead. The current procedure does not provide sufficient information for analysis of those special cases but does increase the difficulty of using all rental data.

In addition to reviewing the procedures used to ensure data consistency, the Bureau of the Census should also review its programs of release of summary statistics and sample data. As was noted, the Public Use Sample of the U.S. Census is one of the few sources of sample data for which complete-count control totals for each variable are available. In addition, the Sixth Count Summary tapes provide information required to scale the sample data to satisfy a series of multidimensional contingency tables. Even if the bureau is reluctant to release Public Use Sample data with additional small-area detail, it should consider expansion of the Fourth and Sixth Count Summary tape program. This would greatly improve the efficiency of the summary statistics generated with existing Census sample data.

While sample data are required for many types of analysis, I have outlined a generally overlooked use for complete-count aggregate Census summary statistics. Increased availability of Census summary

statistics for tract, minor civil division, or metropolitan levels would enhance use of the data for ordinary least squares, logit, or other discrete multivariate analysis. These additional summary statistics could be provided as new contingency tables, but the creation of product moment matrices of individual Census variables for a variety of levels of aggregation would be a more compact way of releasing similar information. Other summary statistics could be released for individual tracts or blocks. Such procedures would permit small-area analysis of Census data while preserving the confidentiality of individual responses.

Finally, it should be observed that the confidentiality requirements of the Bureau of the Census do not represent a needless obstacle to research but rather, an important aspect of the Census program which should be rationalized and improved. The success of the Census in gathering complete and accurate information on millions of households rests in part on its ability to make a meaningful pledge of confidentiality. In a time of growing skepticism concerning governmental invasion of privacy, these pledges assume even greater importance. The tremendous nonresponse rates in many non-Census special surveys underscore the importance of the confidentiality issue.

Given the difficulty and expense of collecting any household interview information, an attempt by the Bureau of the Census to improve the quality of available small-area summary data should be well worth the effort. For 1970, the major costs of data acquisition and processing have already been incurred. For a small additional expenditure, considerable benefits could be obtained by reformating existing small-area information and creating new small-area summary statistics. For future censuses, possible improvement in question design should be considered as well. This would be especially important for housing analysis, which presents many difficult measurement problems.

The Bureau of the Census is in a unique position to gather and disseminate small-area data. Researchers would do well to develop analytical methods compatible with the requirements of the agency. Substantial empirical research can be conducted without infringing on the privacy of individuals. The Bureau of the Census already provides many valuable summary statistics, and I hope it will continue to expand its program of data acquisition and release.

NOTES TO CHAPTER FIVE

1. For a discussion of sales data gathered by federal agencies, see Musgrave (1969).

2. See, for example, *Review of Public Data Use* (December 1972), which contains a series of articles about Public Use Sample research.

3. For a complete description of the Public Use Sample data and other Census products see Census (1973a).

4. These land use data were obtained from the Southwest Pennsylvania Regional Planning Commission.

5. For a discussion of sampling procedures, see Census (1973a).

6. For a discussion of the implications of scaling a multidimensional contingency table see Bishop (1969). Also see Goodman (1970).

7. See Apgar and Kain (1972, appendix) for a complete discussion of the procedure used to estimate the missing moments.

REFERENCES

Apgar, William C., Jr., and John F. Kain. 1972. "The Residential Price Geography of the Pittsburgh SMSA." Paper presented at winter meetings of Econometric Society of America. Toronto, Canada. December.

Bishop, Yvonne M. 1969. "Full Contingency Tables, Logits, and Split Contingency Tables." *Biometrics*, June.

Bishop, Yvonne M., et al. 1974. *Discrete Multivariate Analysis*. Cambridge, Mass.: M.I.T. Press.

Census. 1972. U.S. Bureau of the Census. Social and Economic Statistics Administration. *Data Access Description*. Computer Tape series, CT-7. August.

_____. 1973a. U.S. Bureau of the Census. *A Procedural History of the 1970 Census of Population and Housing.*

_____. 1973c. U.S. Bureau of the Census. *Small Area Data Notes*. September. mnmnmnmnmnm

de Leeuw, Frank, and Raymond J. Struyk. 1975. *The Web of Urban Housing*. Washington, D.C.: Urban Institute.

Goodman, Leo. 1970. "The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications." *Journal of the American Statistical Association*, March.

Hoerl, Arthur E., and Robert W. Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems" and "Ridge Regression: Applications to Nonorthogonal Problems." *Technometrics*, February.

Ingram, Gregory K. 1971. "A Simulation Model of an Urban Housing Market." Ph.D. dissertation, Harvard University.

Kain, John F., and John M. Quigley. 1972. "Housing Market Discrimination, Homeownership, and Savings Behavior." *American Economic Review*, June.

_____. 1975. *Housing Markets and Racial Discrimination: A Microeconomic Analysis*. Urban and Regional Studies 3. New York: National Bureau of Economic Research.

King, A. Thomas. 1972. "Land Values and the Demand for Housing." Ph.D. dissertation, Yale University.

_____. 1973. "Households in Housing Markets: The Demand for Housing Components." Working Paper. Processed. Bureau of Business and Economic Research, University of Maryland. March.

Musgrave, John C. 1969. "The Measurement of Price Changes in Construction." *Journal of the American Statistical Association*, September.

Olsen, Edgar. 1969. "A Competitive Theory of the Housing Market." *American Economic Review*, September.

Peterson, George E. 1974. "The Effect of Zoning Regulations on Suburban Property Values." Working Paper. Processed. Washington, D.C.: Urban Institute.

Peterson, George E., et al. 1973. *Property Taxes, Housing, and the Cities.* Lexington, Mass.: D.C. Heath and Company.

Polinsky, A. Mitchell, and Daniel L. Rubinfeld. 1975. "Property Values and the Benefits of Environmental Improvements: Theory and Measurement." Discussion Paper 104. Processed. Institute of Economic Research, Harvard University. March.

Quigley, John M. 1972. "Residential Location: Multiple Workplaces and a Heterogeneous Housing Stock." Ph.D. dissertation, Harvard University.

Review of Public Data Use; December 1972. Entire issue was devoted to Public Use samples.

Schnare, Ann B., and Raymond J. Struyk. 1974. "Segmentation in Urban Housing Markets." Paper presented before Committee on Urban Economics of Conference on Housing Research. Washington University, October 4-5.

Straszheim, Mahlon. 1975. *An Exonometric Analysis of the Urban Housing Market.* Urban and Regional Studies 2. New York: National Bureau of Economic Research.

Comments on Chapter Five

Eric A. Hanushek

Apgar's study has three distinct aspects. First, there is a methodological discussion about the best ways to use available data on urban housing. Second, there is an implicit research strategy for future analyses of urban structure. And, third, there is a set of implications for the Bureau of the Census pertaining to methods of providing "more information" without sacrificing guarantees of confidentiality.

The first two topics—methodology and research strategy—are best considered within the context of the actual empirical analysis presented by Apgar. This is an analysis of relative rental prices of housing units within a metropolitan area—in this case the Pittsburgh metropolitan area. The motivation for this analysis comes from three sources: First, the empirical analyses of housing prices which have been done in the past have generally used unique data sources and model specifications, making it difficult to compare the results across studies. Second, past studies (and projected future ones) have concentrated more on sales of owner-occupied units, and these units might not adequately reflect the total housing market. Third, if because of market segmentation different prices are in effect for similar housing services in different locations, large bodies of data are called for to sort out the "linkages between spatially separated housing submarkets."

The thrust of Apgar's study is that the Census provides a large and consistent body of micro data which can be used to learn more about

Note: In preparing these comments, I benefited from discussions with John Quigley.

urban form and the spatial aspects of urban housing markets. It is his contention that the richness of the published Census data has been largely overlooked and that the tabulations provide all the information needed to estimate models of housing prices for individual units and not just aggregations such as housing prices for Census tracts.

The starting point for Apgar's analysis is a presumption that the Census Bureau collects and reports in various forms all the data about attributes of housing structures that would be needed for a properly specified model of housing prices or rents. To summarize his methodology, let us then begin with the presentation of Census housing data and the information requirements for estimation of ordinary least squares coefficients. For each metropolitan area, the Census publishes a number of cross tabulations of aspects of housing units (such as structural type, dwelling unit age, rent, and value). These are published for an entire SMSA, for all cities over 50,000 in population, for the suburban ring, and by race of the occupant. Additionally, Census publications provide frequency counts of these attributes and others for individual Census tracts.

Two basic analytical schemes, then, can be followed to analyze prices of housing units. First, if we observe that the required cross-product information for least squares estimation where the independent variables are categorical is simply the information contained in the two-way cross tabulations, we see that a model based upon estimation from individual units can be developed for any geographical area (or set of households) for which a complete set of two-way tabulations is available. Alternatively, we could estimate a model of housing prices based upon aggregate Census tract data, that is, using Census tracts as the observational units. There are two problems with the first method: First, the Census does not publish all of the cross tabulations of variables that might be desired; and second, there are differences among geographic subareas, for example, in neighborhood, accessibility and public services, which we would like to include in a housing price model but for which we do not have the needed cross-product data. The second method—estimation based upon aggregate Census tract data—suffers from possible aggregation biases and losses of efficiency because none of the within-Census tract variation in housing attributes is used in estimation.[1]

Apgar proposes to combine these two methods into a mixed estimation technique. The essence of the technique is to estimate the parameters of housing prices partly on the basis of individual data for the whole area and partly on the basis of aggregate Census tract variables. (As a footnote, I would add that the tract variables do not

have to be provided by the Census but can be taken from other sources as long as the measurements are consistent with the tract boundaries.) Although it builds upon the fairly well known "pure" alternatives, this is an interesting methodological development. As Apgar points out, there is a considerable history of empirical analyses which appear to overlook such a mixed estimation strategy. Nevertheless, judgments on this technique must be based upon its empirical usefulness.

Apgar has provided us with an example of how this methodology can be applied by estimating a series of rental price models for the Pittsburgh metropolitan area. These are hedonic indexes of rental prices based upon structural type, age of the structure, plumbing, number of rooms, tract-specific accessibility, net density, income, and racial composition. Similar models are estimated for the entire SMSA, for the central city and suburban ring, and for blacks and whites in the central city and ring. He further suggests, on the basis of availability and consistency of the data and the costs of doing such estimation, that a profitable research strategy is to replicate this analysis for each of the other tracted SMSAs in the United States. My remarks relate directly to these conclusions.

To begin, let us consider what we have learned from the analysis of the Pittsburgh data. The basic hedonic approach has been followed in a wide variety of circumstances,[2] although none of those previous attempts has been based upon 50,000 observations. Several possible uses have motivated past applications and provide some justification for the enterprise. First, the hedonic index can be used to develop standardized bundles of housing services or different aspects of those services. Since housing services are multi-dimensional, including dwelling unit attributes, neighborhood attributes, and locational attributes such as accessibility and public services, it is necessary in many analyses of urban housing markets and urban form to standardize the bundle. However, we do not have good ways of treating multi-dimensional goods when we do not have any price information linking the individual attributes. In the hedonic approach, estimates are made of the price associated with particular underlying components of housing services. When the market is in equilibrium, these estimates can be interpreted as shadow prices for the individual attributes. Use of these indexes, or portions of them, facilitates analyses of the supply and demand of different quantities of housing services and the interrelationships of housing bundles with urban form. Second, hedonic indexes provide a framework within which it is possible to consider the effects of accessibility, various externalities in price determination, and racial discrimination. By

standardizing carefully for different attributes of dwelling units, we can then concentrate on these issues. Third, we can look at market segmentation where the term is meant to imply differences in the relative valuations of different underlying housing attributes for identifiable sub-markets such as geographical areas or racial groups. Apgar wishes to concentrate mainly upon this third use, but presumably this type of analysis of the Census would be applicable to the other purposes also.

Data from the decennial census for individual dwelling units provides not only a very large sample, which can be used for the estimation, but also the possibility of replication in other cities. These aspects are not without costs, however. We must trade these advantages against an equivalent investment in obtaining more detailed and richer data about individual aspects of structure, neighborhood, and location for one or more metropolitan areas. In using the Census data, there seem to be considerable costs in terms of the quality of the data. To begin with, the information available is very rudimentary. There are data for five structural types, four age groupings, a measure of completeness of the plumbing, and the number of rooms. This is certainly not a very complete description of an individual dwelling unit. Most important, there is no information about the quality of the unit (except, perhaps, for the plumbing variable). It is doubtful that we would want to rely upon this portion of the model to provide the foundation for a further or more detailed analysis of relationships involving structural attributes. However, it is important to note that *all of the potential gains* obtained from using this mixed aggregation method consist of efficiency gains in the estimation of the coefficients for the four structural attributes measured. Further, these efficiency gains (compared to estimation using Census tract variables) accrue only to the extent that the model is properly specified, that is, that it is linear in terms of those four attributes of the housing stock. Apgar suggests that more complete models than the ones he presents can be estimated. However, limitations arising from the presentation of data by the Census (i.e., missing cross tabulations of variables) imply that models cannot be much more extensive than those estimated.

Apgar's neighborhood measures are also very crude. They pertain only to the tract level, and this is probably a very poor level of aggregation for the purpose of uncovering neighborhood effects. In these measures, one again finds a lack of qualitative information except perhaps in the measure of accessibility, a variable not found in basic Census data. There is no information about public services. Thus, it appears that the Census provides a large amount of not too

detailed data, and that data base does not take us very far in addressing two of the possible uses of hedonic price indexes for housing, namely, providing a good method for standardizing housing bundles or allowing detailed analysis of neighborhoods or externalities. In passing, I note that on the question of racial discrimination, Apgar's results (in terms of the dummy variable for black households) indicate that whites are discriminated against in the SMSA as a whole and, when division by central city-suburb is made, whites are discriminated against in the suburbs. However, we are not likely to take these estimates of discrimination seriously in a model which so imperfectly measures the housing attributes. Apgar does not seem to take this seriously since he does not even mention it.

One area of investigation remains, and that one—the test of market segmentation—Apgar addresses most directly. While for many analyses these models may not be complete enough in terms of the individual components, they might still be useful in making some overall statements about market segmentation or about differences in relative prices over different submarkets. However, for those purposes, this type of research also seems to fall somewhat short of what might be desired. As noted before, the maintained hypothesis is that there are subtle interactions between structural attributes, neighborhood characteristics, and location. Apgar then notes that the large Census samples provide a data base that can be used to test such interactions. However, the only interactions between submarkets and various prices that he can analyze are those between central city and suburb and between racial groups similarly divided. Within each of these four cells, he can only look at linear models in which there is no interaction between housing characteristics and neighborhoods. He cannot, for example, determine if there is any interaction between neighborhood quality and housing size. If we had not become conditioned by the Census presentation of data, we would probably not define submarkets as simply central city and suburb. Note that this submarket definition is imposed on the mixed estimation strategy, but it is not mandatory in an estimation strategy based upon Census tract aggregates; tracts may be divided in a variety of ways to test market segmentation hypotheses with aggregate data. Apgar does find statistically significant differences between the individual models. But there is a question about how to interpret them, particularly since the measures of structure and neighborhood are poor and public services are not measured at all.

Finally, let us return to the overall issue of the implicit research strategy. In the end, Apgar cites the cheapness of the regression analyses (although that significantly understates the costs of replica-

tion) and implies that we should replicate this analysis for other SMSAs.[3] What would we learn from such an analysis? In all likelihood we would find that there were significant differences in the shadow prices of different attributes. Such cross-sectional differences are already known to exist (see Ball's work, cited earlier). Further, recent research by Alan Goodman indicates that the hedonic indexes for a single location show considerable change over even short periods of time.[4] For example, obtaining estimates of the effect of a race dummy variable in 150 cities probably would not increase our understanding of racial submarkets for housing services. This is at least my view because I am unwilling to accept the specification dictated by the housing attributes currently available from the Census—and especially those four structural attributes that are completely cross-tabulated.

Since hedonic indexes represent reduced form models of the housing market, we would be tempted to explain observed differences among areas in terms of a variety of structural differences in either specific supply or demand conditions. However, it would be difficult to do this in a systematic manner without knowledge of the underlying supply and demand relationships within the different housing markets. Therefore, it would seem better to concentrate attention on a single housing market, where we could hope to isolate either more information about structural aspects or more precise information about some aspect of the housing market.

A more appealing research design than this Census strategy would call for a more intensive analysis of a given city. If the money for this analysis were spent on a carefully designed survey that addressed, say, the issue of neighborhoods, or public services, or market segmentation, I would speculate that we could increase our knowledge of urban structure more than by following the Census route. The appeal of using specialized data samples in the past has been that they often contain particular information about one aspect of urban housing markets. Indeed, the analytical design of many such studies has revolved around capitalizing on one or two particularly rich features of a body of data. Certainly, it would be helpful to replicate some of these studies with consistent data from other sources. However, consistency is not an absolute virtue, particularly when consistency implies losing all of the richness of any given analysis and imposing a maintained hypothesis which we would not in general be willing to accept.

A key point is that we are very uncertain about the correct specification of the housing price model. Even with the use of specialized data, there are many uncertainties about the specification

of various aspects of the total housing bundle. Replication of the Census analysis, based upon its minimal number of measures of housing attributes, cannot give us much guidance in terms of appropriate specification. The goal of comparisons across different estimates of hedonic indexes has mainly been to learn about model specification, and that issue has little to do with the efficiency of estimation of a few structural parameters. In fact, the questions of micro data and efficiency have meaning only within the context of reasonably well-specified models. There seems to be little one can say about the virtue of "efficient" estimates of possibly highly biased coefficients.

It also seems important to note one other aspect of the research design in this area. Data availability is not the only problem. There are some serious conceptual problems that have not been adequately addressed. For example, neighborhood is a concept that has received considerable attention. Yet there are few discussions of how one should go about measuring neighborhoods. In another area, the appropriate measurement and treatment of public services is unclear. Should we measure levels of services? Value added in public provision of services? Value added adjusted for costs? Or should we assume that all differences in public services are reflected in price differences for given housing units? Apgar completely ignores the issue, but even those who have considered it do not seem to have made much progress. In short, our conceptual and measurement tools need some refining. Also, our methods of analysis need some further consideration. Should we continue to analyze the reduced form models implied by the hedonic indexes or should we move toward more structural demand and supply relationships?

These issues take us considerably beyond the scope of this particular paper. However, they do suggest that at our current state of knowledge a continuation of smaller-scale analyses that focus upon particular smaller issues may not be a wasteful strategy. Thus, in terms of methodology and research strategy, it appears that Apgar makes a very valid point that Census housing data have not been fully exploited by their users. However, the generalization of this point in terms of research strategy is a bit misleading. The prime shortcoming of current models of housing prices and urban structure does not seem to be efficiency of parametric estimates but reasonableness of model specification.

The final point of Apgar's study—implications for the Bureau of the Census—is very well taken. The presentation of published Census data is not the best possible. Inconsistencies in data and missing detail appear to be introduced unnecessarily. Through consideration

of research needs and uses of the Census data, improvements could be made at virtually no cost.

As preparations are made for the 1980 census, design considerations obtained from potential research uses should be brought in. An important suggestion that is developed by Apgar is that the Census could routinely present cross-product information. As a practical matter, this would not have to be available in hard copy; machine-readable information would be sufficient. This step would expand the value of the data without compromising the policy of confidentiality. (It would probably have the added advantage of forcing the Census Bureau to provide consistent data.) Nevertheless, perhaps a more important issue to take to the Census Bureau is a better understanding of the specific data that they should collect. In other words, even here the importance of model specification should not be underestimated.

NOTES TO COMMENTS ON CHAPTER FIVE

1. Note that the efficiency gains of using the micro data might be offset by biases that arise from incomplete data in the cross tabulations or by errors resulting from the use of two different published sources.

2. See Michael J. Ball, "Recent Empirical Works on the Determinants of Relative House Prices," *Urban Studies*, June 1973.

3. Apgar's current model includes measures of accessibility by tract and net residential density by tract. Obtaining similar information for other cities could be very expensive. Furthermore, this does not include any estimate of the costs that would be involved in sorting out the results for 150 cities.

4. Alan C. Goodman, "Neighborhood Effects, Hedonic Prices, and the Residential Housing Choice" (Ph D. dissertation, Yale University, 1976).