A REGRESSION APPROACH TO ESTIMATING THE AVERAGE NUMBER OF PERSONS PER HOUSEHOLD*

STANLEY K. SMITH, JUNE NOGLE, AND SCOTT CODY

In the housing unit method, population is calculated as the number of households times the average number of persons per household (PPH), plus the population residing in group quarters facilities. Estimates of households and the group quarters population can be derived directly from concurrent data series, but estimates of PPH have traditionally been based on previous values or estimates for larger areas. In our study, we developed several regression models in which PPH estimates were based on symptomatic indicators of PPH change. We tested these estimates using county-level data in four states and found them to be more precise and less biased than estimates based on more commonly used methods.

Small-area population estimates are used for a wide variety of purposes. Businesses use them to develop profiles of customers, to identify market clusters, and to determine optimal site locations. State and local governments use them to establish political boundaries, to monitor the impact of public policies, and to estimate the need for schools, roads, parks, public transportation, and fire protection. Research analysts use them to study urban sprawl, environmental conditions, and social trends. Such estimates are used as denominators for calculating many types of rates and to determine the allocation of billions of dollars in public funds each year. Clearly, the production of accurate and timely population estimates is of tremendous importance for many purposes.

The housing unit (HU) method is the most commonly used method for making small-area population estimates in the United States (U.S. Bureau of the Census 1990). It can be used for many types of geographic areas, from states down to counties, cities, school districts, ZIP-code areas, census tracts, and individual blocks. It can accommodate a number of different data sources and can be applied in many different ways. Perhaps most important, it has often been found to produce reasonably accurate population estimates (e.g., Hodges and Healy 1984; Lowe, Myers, and Weisser 1984; Smith 1986; Smith and Mandell 1984). Given these attributes, it is not surprising that the U.S. Census Bureau recently adopted the HU method as its sole methodology for producing subcounty population estimates.

The HU method is based on the assumption that almost everyone lives in some type of housing structure. In this method, population can be estimated as

$$P_t = (H_t \times PPH_t) + GQ_t,$$

where P_t is the population at time t, H_t is the number of occupied housing units (i.e., households) at time t, PPH_t is the average number of persons per household at time t,

*Stanley K. Smith, June Nogle, and Scott Cody, Bureau of Economic and Business Research, University of Florida. Address correspondence to Stanley K. Smith, Bureau of Economic and Business Research, 221 Matherly Hall, University of Florida, Gainesville, FL 32611-7145; E-mail: sksmith@ufl.edu. The authors thank the U.S. Census Bureau for financial support in the early stages of this project and the state demographers of Illinois, Texas, and Washington for providing historical data. They also thank the editor and referees for many helpful comments and suggestions.

and GQ_t is the group quarters population at time t. Estimates of the group quarters population typically include persons without permanent living quarters (e.g., the homeless population).

Households can be estimated using building permits, electricity customers, property tax records, water-meter records, and other data that reflect changes in the housing stock (e.g., Brown 1999; Smith and Cody 1994; Starsinic and Zitter 1968; U.S. Bureau of the Census 1998). The group quarters population can be estimated from public records or data provided by the administrators of group quarters facilities (e.g., Smith and Cody 1994). For PPH, however, data that accurately reflect changes since the most recent census are not readily available. As a result, most applications of the HU method have held past values constant or have extrapolated historical trends (e.g., Starsinic and Zitter 1968; U.S. Bureau of the Census 1998). Some analysts have refined these approaches by tying PPH changes in small areas to changes in larger areas for which current estimates are available (e.g., Smith 1986; State of New Jersey 1984) or by adjusting for changes in the mix of housing units (e.g., Findley 1979; Smith and Lewis 1980). These approaches work well when PPH values remain constant or follow stable trends, but produce inaccurate estimates when PPH values or trends are changing rapidly.

We hypothesize that PPH estimates can be improved if they incorporate place-specific data that are more recent than the previous census. One way to incorporate such data is to conduct sample surveys or to interview local experts (e.g., Roe, Carlson, and Swanson 1992). This approach is useful when estimates are needed only for a small number of places, but is costly and time-consuming when estimates must be prepared for a large number of places. Another approach is to develop regression models in which changes in PPH are related to changes in symptomatic indicators of household size. Several models have been suggested (e.g., Comprehensive Planning Organization of the San Diego Region 1974; Voss and Krebs 1979), but none has been widely used or thoroughly evaluated. We believe the regression approach is potentially useful because it incorporates postcensal data and can be applied to a large number of places at relatively low cost.

In our research, we constructed several regression models for estimating PPH and tested them using county-level data from four states. We focused on counties because they exhibit substantial variations in PPH levels and trends and consistent data covering several periods are readily available. Furthermore, the HU method is often used for county-level estimates: Shahidullah's (1999) survey of state demographic agencies reported that of 15 states that produced county population estimates, 11 used the HU method alone or in combination with other methods. It is likely that many of the lessons learned from constructing and evaluating county models will be relevant for subcounty models as well.

We conducted our empirical analyses using data from Florida, Illinois, Texas, and Washington. These states have 67, 102, 254, and 39 counties, respectively, providing a total sample size of 462. They represent different geographic regions and reflect a great deal of diversity in terms of age structure, racial and ethnic makeup, household composition, and other factors affecting PPH. Each has an active demographic agency that was able to provide the necessary input data; in each state, this agency is a member of the Federal-State Cooperative Program for Population Estimates (FSCPE). We believe that these counties provide an excellent sample for constructing and testing regression models.

We start by describing national PPH trends and comparing PPH values and trends for the states and counties in our sample. Then, we describe and estimate several regression models and use the results to develop county PPH estimates for 1980, 1990, and 2000. We compare these estimates with census data for the same years, evaluate the results for each model, and compare them with results produced by more commonly used methods. We investigate several extensions to our basic approach and suggest areas for which further research is needed. We close by drawing several conclusions regarding the value of using regression models to estimate PPH.[1]

HISTORICAL TRENDS AND GEOGRAPHIC DIFFERENCES

The average household size in the United States fell tremendously during the past two centuries, from 5.8 in 1790 to 4.8 in 1900 and 2.6 in 2000 (Kobrin 1976; U.S. Census Bureau 2001). This decline was largely the result of falling fertility and mortality rates and a declining propensity for people to live in extended families (Kobrin 1976). All regions of the country have experienced substantial declines in household size, but current values vary considerably from place to place.

In our sample, Texas had the largest PPH in each census year from 1970 to 2000, Florida had the smallest, and Illinois and Washington fell somewhere in between. Decade changes were similar in all four states, with PPH values falling by 11%–12% during the 1970s and 3%–4% during the 1980s. State PPH values changed little during the 1990s, falling slightly in Illinois, rising slightly in Texas, and remaining virtually unchanged in Florida and Washington.

Mean PPH values for counties also declined in all four states between 1970 and 2000 (see Table 1). Differences in county PPH values have generally become smaller over time, but substantial differences remain. For any given state, the largest county PPH values are often 30%, 40%, or 50% higher than the smallest values. In Texas, the largest values are more than 75% higher than the smallest values. During any given decade, PPH values may rise in some counties and fall in others. Even during the 1990s—when state and national PPH values changed little—PPH trends varied considerably from one county to another.

Given these differences, county PPH estimates that are based solely on previous values, historical trends, or changes at the state or national level will often be inaccurate. Can regression models be developed that provide more accurate PPH estimates than the simple methods that have traditionally been used?

CONSTRUCTING REGRESSION MODELS

Socioeconomic theory, demographic dynamics, and the availability of relevant data guide the choice of models and variables used for constructing population estimates (McKibben and Swanson 1997). For the present analysis, we wanted independent variables that reflected PPH changes over time and were available annually for counties. A number of factors may be expected to influence PPH trends, such as changes in fertility rates, marital status, living arrangements, age structure, racial/ethnic composition, foreign immigration, and housing characteristics (e.g., Findley 1979; Gober 1990; Kobrin 1976; Myers and Doyle 1990). Unfortunately, data for many of these factors were not available for counties for the years covered by this study. We explored several possibilities and—on the basis of their availability and our preliminary empirical analyses—chose the following independent variables: (1) births per household (Births); (2) school enrollees in grades

[1] Some applications of the HU method have distinguished among different types of housing structures (e.g., single family, multifamily, and mobile home). Theoretically, this distinction may improve population estimates because different types of housing structures often have different growth rates, occupancy rates, and PPH values. The data needed to apply the HU method by type of structure, however, are often unavailable (e.g., when households are estimated from electricity or telephone customers) or unreliable (e.g., data on mobile homes). Furthermore, the improvements in accuracy that may result from differentiating by type of structure have not been well established empirically. The U.S. Census Bureau, many business enterprises (e.g., Claritas), and most state demographers do *not* differentiate by type of structure when they apply the HU method; in the present analysis, neither do we. Whether models differentiated by type of structure may improve PPH (and population) estimates is a matter for future research.

Table 1. County PPH Characteristics by State: 1970–2000

	Mean Value			
State	1970	1980	1990	2000
Florida	3.05	2.70	2.54	2.48
Illinois	2.98	2.68	2.55	2.48
Texas	3.08	2.80	2.71	2.64
Washington	2.99	2.65	2.56	2.55

	Range of Values			
State	1970	1980	1990	2000
Florida	2.35 to 3.73	2.25 to 3.25	2.18 to 3.00	2.13 to 3.09
Illinois	2.62 to 3.56	2.42 to 3.08	2.30 to 2.98	2.21 to 2.97
Texas	2.43 to 4.39	2.23 to 4.05	2.15 to 3.90	2.13 to 3.75
Washington	2.52 to 3.36	2.29 to 2.92	2.25 to 3.03	2.16 to 3.26

	Mean Percentage Change		
State	1970–1980	1980–1990	1990–2000
Florida	–11.31	–5.62	–2.45
Illinois	–9.94	–5.01	–2.72
Texas	–8.67	–3.40	–2.20
Washington	–11.38	–3.31	–0.59

	Range of Percentage Changes		
State	1970–1980	1980–1990	1990–2000
Florida	–20.0 to –4.6	–11.5 to 4.7	–9.2 to 4.5
Illinois	–18.0 to –4.6	–8.3 to –0.7	–7.3 to 2.4
Texas	–21.0 to 0.8	–12.5 to 4.2	–15.3 to 4.8
Washington	–14.9 to –4.4	–11.1 to 4.5	–7.8 to 7.6

K–12 per household (School); and (3) Medicare enrollees age 65 and older per household (Medicare).

To account for differences in the sizes of county populations, we expressed all three independent variables as ratios by dividing the number of births, school enrollees, and Medicare enrollees, respectively, by the number of households in each county. Births per household and school enrollees per household were expected to have a positive impact on PPH because they reflect the presence of infants and children, respectively. Medicare enrollees per household were expected to have a negative impact on PPH because they reflect the presence of older people, who often live in one- or two-person households.[2]

[2] We considered several other independent variables in our preliminary analyses. Nonwhite births as a proportion of total births and nonwhite deaths as a proportion of total deaths were used as measures of the racial composition of the population. Increases in those variables were expected to have a positive impact on PPH because nonwhite households are larger than white households, on average. However, both variables were found to be statistically insignificant in most regressions and contributed little to the explanatory power of the equations. We also evaluated single-family units as a proportion of the total housing units as a measure of the

We collected county-level data for 1970, 1980, and 1990 for each variable. Data on births and school enrollments were provided by the FSCPE agencies in each state. Medicare data were provided by the Centers for Medicare and Medicaid Services, a branch of the U.S. Department of Health and Human Services. Data on PPH and households were derived from decennial census counts. Using these data, we estimated four ordinary least squares (OLS) regression models:

Model 1 (Basic): $PPH_t = \beta 1 + \beta 2\ (\text{Births}_t) + \beta 3\ (\text{School}_t) + \beta 4\ (\text{Medicare}_t)$.

Model 2 (Ratio): $(PPH_t\ /\ PPH^*_t) = \beta 1 + \beta 2\ (\text{Births}_t\ /\ \text{Births}^*_t) + \beta 3\ (\text{School}_t\ /\ \text{School}^*_t) + \beta 4\ (\text{Medicare}_t\ /\ \text{Medicare}^*_t)$.

Model 3 (Change): $(PPH_t - PPH_c) = \beta 1 + \beta 2\ (\text{Births}_t - \text{Births}_c) + \beta 3\ (\text{School}_t - \text{School}_c) + \beta 4\ (\text{Medicare}_t - \text{Medicare}_c) + \beta 5\ (PPH_c)$.

Model 4 (Ratio Change): $[(PPH_t\ /\ PPH^*_t) - (PPH_c\ /\ PPH^*_c)] = \beta 1 + \beta 2\ [(\text{Births}_t\ /\ \text{Births}^*_t) - (\text{Births}_c\ /\ \text{Births}^*_c)] + \beta 3\ [(\text{School}_t\ /\ \text{School}^*_t) - (\text{School}_c\ /\ \text{School}^*_c)] + \beta 4\ [(\text{Medicare}_t\ /\ \text{Medicare}^*_t) - (\text{Medicare}_c\ /\ \text{Medicare}^*_c)] + \beta 5\ (PPH_c)$.

Model 1 (Basic) is the simplest model, using county-level data from a single point in time (*t*). Model 2 (Ratio) also refers to time *t*, but expresses all variables as ratios of county to state values (state values are identified by an asterisk). Model 2 is a variant of the ratio-correlation method, with variables defined as ratios for one point in time, rather than as ratios of proportions from two points in time (e.g., Crosetti and Schmitt 1956; Namboodiri 1972).

Models 3 and 4 express all variables as changes over time ($t - c$, where c is the year of the previous decennial census). The variables in Model 3 are based solely on county-level data, whereas the variables in Model 4 are based on ratios of county to state data. In addition to the independent variables included in Models 1 and 2, Models 3 and 4 include the PPH value from the previous census as a predictor of the current PPH. This variable is expected to have a negative impact on changes in PPH. When overall PPH levels are falling, for example, declines are expected to be greater for counties with high PPH values than for counties with low PPH values (Smith and Lewis 1980). Model 4 is a variant of the difference-correlation method of population estimation (e.g., O'Hare 1976; Schmitt and Grier 1966).

We began our empirical analysis by combining counties from all four states into a single sample. We produced separate regressions for 1970, 1980, and 1990 for Models 1 and 2 and for 1970–1980 and 1980–1990 for Models 3 and 4. The results are shown in Table 2. We used the same variable names throughout the table, but the reader is reminded that the variables are defined somewhat differently in each model.

For Models 1 and 2, all three independent variables had the expected signs and were statistically significant for all three years. The independent variables explained a relatively high proportion of the variation in the dependent variable in both models (especially Model 2). The regression coefficients were more consistent from one period to another for Model 2 than for any of the other models.

composition of the housing stock. This variable was expected to have a positive impact on PPH because single-family units have more residents than do multifamily units or mobile homes, on average. In some regressions, this variable had the expected sign and was statistically significant; in others, it had the wrong sign or was statistically insignificant. Because none of these variables had much impact on PPH estimates, we excluded them from further analyses.

Table 2. OLS Regression Coefficients for the Full Sample: 1970–1990

	Model 1 (Basic)			Model 2 (Ratio)		
Variable	1970	1980	1990	1970	1980	1990
Births per Household	10.199*	11.192*	12.213*	0.124*	0.086*	0.125*
School Enrollees per Household	0.359*	0.512*	0.473*	0.180*	0.180*	0.154*
Medicare Enrollees per Household	–0.487*	–0.164*	–0.288*	–0.035*	–0.021*	–0.021*
Intercept	2.477*	2.072*	2.077*	0.727*	0.751*	0.738*
Adjusted R^2	0.756	0.744	0.763	0.843	0.821	0.836

	Model 3 (Change)		Model 4 (Ratio Change)	
Variable	1970–1980	1980–1990	1970–1980	1980–1990
Births per Household	1.851*	1.544*	0.029*	0.040*
School Enrollees per Household	0.225*	0.934*	0.070*	0.132*
Medicare Enrollees per Household	–0.621*	–0.219*	–0.047*	–0.022*
Persons per Household (Lagged)	–0.162*	–0.033*	–0.041*	–0.006
Intercept	0.241*	0.014	0.138*	0.011
Adjusted R^2	0.559	0.490	0.415	0.470

*Significant at $p < .01$.

Models 3 and 4 show the regression results when the variables are expressed as changes over time. The three independent variables common to all four models had the expected signs and were statistically significant in every instance. The lagged PPH variable had the expected negative sign and was statistically significant in three out of four instances. Adjusted R^2 values were considerably lower for Models 3 and 4 than for Models 1 and 2; apparently, it is more difficult to explain variations in changes in PPH over time than variations at a given time.

These results support the hypotheses regarding the effects of the independent variables on county PPH values: increases in births per household and school enrollees per household tend to raise PPH values, and increases in Medicare enrollees per household tend to lower PPH values. Other things being equal, the higher the PPH value in a particular year, the greater the decline during the following decade. How can these results be used to construct PPH estimates?

CONSTRUCTING PPH ESTIMATES

One straightforward procedure for estimating PPH is to apply the regression coefficients from one period to values of the independent variables in a later period. Using Model 1, for example, we could multiply the coefficients shown in Table 2 for 1990 by the 2000 values of the independent variables for a particular county, and sum the resulting products (plus the intercept) to provide an estimate of that county's PPH in 2000. The same steps could be followed for each of the other models.

Following this procedure, we constructed PPH estimates for each county on the basis of each of the four regression models. For Models 1 and 2, we constructed estimates for 1980, 1990, and 2000. For Models 3 and 4, we constructed estimates only for 1990 and 2000 because the 1960–1970 data we needed to construct 1980 estimates were not available.

The 2000 estimates required several adjustments to the input data. When postcensal estimates are produced by businesses and governmental agencies, data for the reference year are often unavailable because of time lags in their collection, tabulation, and release. In these circumstances, estimates must be based on extrapolations of recent trends. In this study, data from the sources described earlier for births, school enrollees, and Medicare enrollees were available only through 1998. We dealt with this problem by extrapolating 1990–1998 trends forward to 2000 for each of these variables.[3]

Values of the independent variables in 2000 required estimates of the number of households for each county in 2000. For Florida, we used the household estimates produced by the state FSCPE agency. For the other three states, we developed 2000 estimates by applying 1990 occupancy rates to the 1998 HU estimates produced by the U.S. Census Bureau and extrapolating the results forward to 2000. Models 2 and 4 also required state-level estimates of PPH in 2000. For Florida, we used the estimate produced by the state FSCPE agency. For the other three states, we used 1998 estimates from the Current Population Survey (CPS), extrapolated forward to 2000. These adjustments add a degree of uncertainty, but reflect conditions that must be faced when regression models are actually used for postcensal PPH estimates.

In addition, we made PPH estimates using an equally weighted average of estimates from the individual regression models (Average). Averages capture more information than can be incorporated in a single model and reduce the chances of making large errors; they often have been found to produce more accurate estimates and projections than do individual methods (e.g., Ahlburg 1999; Sanderson 1999; Smith and Mandell 1984). For 1980, Average was based on estimates from Models 1 and 2; for 1990 and 2000, it was based on estimates from all four models.

To provide a standard of comparison, we made PPH estimates using three simple but widely used methods. Model A uses each county's PPH in the most recent census as an estimate of its PPH during the following decade. Model B uses the percentage change in each county's PPH during the previous decade as an estimate of its change during the following decade. Model C uses the percentage change in a state's PPH since the most recent census as an estimate of PPH change for each of that state's counties. We refer to these as *traditional* methods because they are the ones that have been most commonly used for making postcensal PPH estimates.

EVALUATING PPH ESTIMATES

Errors for each model were calculated by comparing PPH estimates with the actual PPH values reported in the decennial census. We used two error measures to evaluate the estimates. The mean absolute percentage error (MAPE) is the average error when the direction of error is ignored. This is a measure of precision, or how close the estimates were to census values, regardless of whether they were too high or too low. The mean algebraic percentage error (MALPE) is the average error when the direction of error is included. This is a measure of bias, or the tendency of estimates to be too high or too low. These measures have been widely used for evaluating the precision and bias of population estimates and projections (e.g., Ahlburg 1999; Keilman 1999; O'Hare 1976; Smith 1986; Swanson and Tedrow 1984).

[3] Alternatively, the regression models could be run using lagged data for the independent variables (Swanson and Beck 1994). For example, 1988 values for the independent variables could be used to estimate PPH in 1990. PPH values for 2000 then could be based directly on the regression coefficients and 1998 data for the independent variables.

Table 3. Mean Absolute Percentage Errors (MAPEs) and Mean Algebraic Percentage Errors (MALPEs) for County PPH Estimates: 1980, 1990, and 2000

	1980		1990		2000	
	MAPE	MALPE	MAPE	MALPE	MAPE	MALPE
Regression Models						
1 (Basic)	7.0	6.5	3.6	0.9	3.6	–0.2
2 (Ratio)	2.9	–0.2	2.7	0.4	3.9	2.0
3 (Change)	—	—	4.6	–4.5	2.2	–1.8
4 (Ratio change)	—	—	3.5	3.2	2.5	1.8
Average	4.2	3.1	1.9	0.0	2.2	0.4
Traditional Models[a]						
A (No Change)	10.7	10.7	4.5	4.3	3.0	2.3
B (Extrapolated county percentage change)	—	—	7.2	–6.9	3.3	–2.1
C (State percentage change)	3.0	–1.7	2.2	0.8	3.8	3.5
Regression Models Including IRS Variables						
1 (Basic plus IRS)			2.5	–1.4	2.6	0.6
2 (Ratio plus IRS)			2.5	0.8	3.1	1.9
3 (Change plus IRS)			—	—	1.9	–1.4
4 (Ratio change pus IRS)			—	—	2.1	1.3
Average (including IRS)			2.2	–0.3	1.8	0.6

[a] Model A: $PPH_t = PPH_c$; Model B: $PPH_t = PPH_c + [(PPH_c - PPH_{c-1}) / PPH_{c-1}] PPH_c$; and Model C: $PPH_t = PPH_c(\text{State } PPH_t / \text{State } PPH_c)$.

The errors are summarized in Table 3. Several patterns stand out. First, MAPEs for the four regression models generally fell within a relatively small range. Except for Model 1 in 1980, MAPEs for all models and all years ranged only from 2.2% to 4.6%. There was no clear trend regarding changes in precision over time. For some models, MAPEs declined from one estimation year to the next; for others, they rose or remained constant. No single model provided the most precise estimate in every year.

Second, there was little indication of any systematic bias in the regression models. MALPEs for individual models were positive for some years and negative for others. For 1980, MALPEs for individual models ranged from –0.2% to 6.5%; for 1990, from –4.5% to 3.2%; and for 2000, from –1.8% to 2.0%. We believe that some models and periods will display an upward bias and others will display a downward bias, but there is no inherent tendency for regression-based PPH estimates to be either too high or too low.[4]

Third, the Average method generally produced better results than did the individual regression models. In 1980, Average was based on only two models; it performed considerably better than Model 1 but not as well as Model 2. In 1990 and 2000, Average was based on all four models. In 1990, both the MAPE and the absolute value of the MALPE were smaller—sometimes much smaller—for Average than for any of the individual mod-

[4] We also evaluated MAPEs and MALPEs for counties grouped by size and growth rate. Differences in population size and growth rate had no consistent impact on either the precision or bias of regression-based PPH estimates.

els. In 2000, the MAPE and absolute value of the MALPE were smaller for Average than for three of the four models. We believe that regression models may be more useful for estimating PPH when used in combination with each other than when used individually.

Fourth, the regression models were generally more precise and less biased than the traditional estimation methods. Errors were especially large for Model A in 1980 and Model B in 1990. It is not surprising that Model A—based on the assumption that PPH would not change—had a positive MALPE in all three years. Model B had a negative MALPE in both 1990 and 2000. Model C produced reasonably precise, unbiased estimates in 1980 and 1990, but had the second largest MAPE and the largest MALPE (in absolute terms) of all the models in 2000.[5]

Finally, the regression models produced substantially fewer large errors (i.e., 5% or more) than did the traditional models (not shown here). In 1990, Model A had more than seven times as many large errors as Average, Model B had more than 12 times as many, and Model C had 28% more. In 2000, Model A had almost twice as many large errors than Average, Model B had almost three times as many, and Model C had more than three times as many. The ability to reduce the number of large errors may be the greatest advantage of regression models over traditional methods for estimating PPH. We return to this point later in the article.

EXTENDING THE ANALYSIS

Incorporating Data From the Internal Revenue Service

The birth, school enrollment, and Medicare data used in this study performed well, but other variables may also be useful. One with particular promise is the average number of exemptions per federal income tax return (IRS). Because large households generally claim more exemptions than do small households, adding this variable to the regression models may raise their explanatory power and improve their estimation performance.

To test this hypothesis, we added IRS as an independent variable in Models 1–4 (results not shown here). Adding this variable substantially raised adjusted R^2 values in both 1980 and 1990 (1970 IRS data were not available). The IRS coefficient had the expected positive sign and was statistically significant in every equation. Adding IRS affected the coefficients of the other variables, but did not change their signs and generally did not alter their levels of significance.[6]

Using these new regression coefficients and data for 1990 and 1998—along with the 2000 adjustment procedures described earlier—we constructed PPH estimates using Models 1 and 2 for 1990 and all four models for 2000. The resulting estimation errors are shown in the bottom panel of Table 3. Adding the IRS variable consistently improved the precision of the PPH estimates, but had no consistent impact on bias.

We believe that IRS data are excellent indicators of changes in PPH. Their most important attribute for estimation purposes, however, may be their geographic availability. IRS data are tabulated not only for counties but for cities and ZIP-code areas as well.

[5] The accuracy of Model C is somewhat overstated for 1980 and 1990 because census data were used to calculate decade changes in state-level PPH. When this model is actually used for postcensal estimates, census data will not be available, and changes in state-level PPH will have to be based on survey data or some other type of estimate.

[6] The IRS data were provided by the U.S. Census Bureau in two separate data sets: one for 1980–1990 and the other for 1990–1998. However, the 1990 numbers in the first set differed from the 1990 numbers in the second set because of changes in the procedures used to assign addresses to specific geographic locations. Differences were usually less than 2% but were greater than 5% in some counties. We dealt with this problem by using 1980–1990 data to estimate the regression coefficients and 1990–1998 data to develop PPH estimates for 1990 and 2000. Given the similarities of the two data sets, we doubt that this adjustment had much impact on the empirical results.

Data for other indicators of PPH change, on the other hand, are frequently unavailable (or unreliable) below the county level. Therefore, IRS data may be particularly useful for constructing PPH estimates for subcounty areas. Although administrative and processing changes have introduced some inconsistencies into the time series, the use of IRS data for small-area PPH estimates is a topic that merits further research.

Developing State-Specific Models

In the analyses thus far, all counties were combined into a single sample. However, data for the independent variables may not be perfectly consistent from one state to another because of differences in accounting procedures or record-keeping practices. In addition, Florida, Illinois, Texas, and Washington most likely differ from each other in ways that affect PPH but are not accounted for in the regression models. We tested for these effects by adding dummy variables for Florida, Illinois, and Washington to each regression model and each period (not shown here). The dummy variables were statistically significant in most instances, and their inclusion often substantially raised the equation's explanatory power.

Given these results, it is possible that regression models that are estimated separately for each state will provide more accurate PPH estimates than will models that are based on the entire sample of 462 counties. To test this possibility, we ran Models 1–4 separately for the counties in each state. We do not show the results here, but we can report that the coefficients had the expected signs and were statistically significant in most instances. Adjusted R^2 values were frequently higher for the state-specific models than for the models covering the sample as a whole.

We then made two sets of county PPH estimates for each state, one based on state-specific regression coefficients and the other based on coefficients for the sample as a whole. The results of this exercise were mixed. In some instances, errors that were based on state-specific models were smaller than errors that were based on the entire sample, and in other instances they were larger. Table 4 shows the results for Average from these two sets of estimates.

For Florida, using a state-specific model reduced MAPEs for both years. For Illinois, using a state-specific model had no impact on the MAPE for 1990 but raised it slightly for 2000. For Texas, using a state-specific model had no impact on the MAPE for either year. For Washington, using a state-specific model raised the MAPE for 1990 but reduced it for 2000. Using state-specific models had no consistent impact on MALPEs, sometimes raising them, sometimes lowering them, and sometimes changing their signs.

These results do *not* show that PPH estimates that are based on state-specific models are uniformly better than estimates that are based on regressions covering the entire sample. This finding may be particularly useful when county estimates must be made for the entire nation (or for a large number of states) because it implies that it is not necessary to construct separate models for each state. It is possible, however, that state-specific models would produce better results if they were customized in some way (e.g., using different variables or sets of models for each state). Again, further research is needed before firm conclusions can be drawn.

Constructing State PPH Estimates

The results summarized in Table 4 show little evidence of bias, given that MALPEs were close to zero for all states in both years. This finding implies that state PPH estimates that are based on county regression models are likely to be relatively accurate. To test this possibility, we made 1990 and 2000 PPH estimates for each state using a weighted average of county PPH values, with the weights determined by each county's share of state households (not shown here). Estimates based on the state-specific regression models were compared with state PPH estimates derived from the CPS and census values for each year.

Table 4. **Mean Absolute Percentage Errors (MAPEs) and Mean Algebraic Percentage Errors (MALPEs) for County PPH Estimates (Average), Using Total Sample and State-Specific Regression Results**

		1990 MAPE		1990 MALPE	
	Number of Counties	Total Sample	State-Specific	Total Sample	State-Specific
Florida	67	2.5	1.9	2.2	1.3
Illinois	102	1.5	1.5	0.4	0.1
Texas	254	2.0	2.0	–0.8	0.6
Washington	39	1.6	2.0	0.8	–1.8
Average[a]	462	1.9	1.9	0.0	0.4
		2000 MAPE		2000 MALPE	
	Number of Counties	Total Sample	State-Specific	Total Sample	State-Specific
Florida	67	2.0	1.8	1.3	–0.5
Illinois	102	1.8	2.0	0.2	–0.1
Texas	254	2.3	2.3	0.0	0.4
Washington	39	2.6	2.2	2.4	1.7
Average[a]	462	2.2	2.2	0.4	0.3

[a]Average weighted by each state's share of the number of counties in the entire sample.

State PPH estimates derived from county PPH estimates were very close to census values. In 1990, no state had an error larger than 1.2%; in 2000, no state had an error larger than 2.4%. In both years, estimates were higher than census values in two states and lower in two states. Despite widely diverging trends at the county level, state estimates that were based on county estimates were precise and displayed no consistent bias.

The regression-based estimates were frequently closer to census counts than were those that were derived from the CPS. The 1990 regression-based estimates were more precise than were the CPS estimates in two states, the CPS estimate was more precise in one state, and the two had the same error in one state. For 2000, the regression-based estimates were more precise in three of the four states. In addition, the regression-based estimates had an equal number of positive and negative errors in both years, whereas the CPS estimates were below census counts in every instance. These results provide further evidence of the potential usefulness of regression-based PPH estimates.

Evaluating the Impact of PPH Errors on Population Estimation Errors

How large are PPH errors compared with errors for the other components of the HU method? A study of 1980 municipal estimates in New Jersey reported MAPEs of 4.4% for PPH, 6.1% for households, and 7.4% for total population (State of New Jersey 1984). A study of 1980 municipal estimates in Washington reported MAPEs of 4.4% for PPH, 4.9% for housing units, and 6.0% for total population (Lowe et al. 1984). A study of 1990 subcounty estimates in Florida reported MAPEs of 5.0% for PPH, 11.2% for households, and 11.9% for total population; for county estimates, MAPEs were 2.3%, 5.1%, and 5.4%, respectively (Smith and Cody 1994). All three studies found large errors for estimates of the group quarters population, but concluded that those errors contributed little to overall estimation error because—in most places—the group quarters population accounts for a small proportion of the total population.

For subcounty areas, PPH errors in all three studies fell between 4% and 5%; for the single set of county estimates, the PPH error fell within the range of errors shown in Table 3. What impact do these errors have on errors for population estimates? Given the interactive nature of the components of the HU method, it is impossible to answer this question definitively. Suppose, for example, that the household estimate for a particular county were 5% above the "true" number of households. The population estimate for that county would be more accurate if the PPH estimate were 5% *below* its true value than if it were perfectly accurate. On the other hand, if the PPH estimate were 5% *above* its true value, it would add to (and magnify) the 5% positive error in the household estimate. The impact of PPH errors on population estimation errors thus depends on the errors in the household estimates. If PPH and household errors have the same sign, they reinforce each other; if they have opposite signs, they offset each other.[7]

One way to control for interaction effects is to evaluate hypothetical population estimates in which two of the components (e.g., households and the group quarters population) are based on known census values and the third (e.g., PPH) is based on an estimate. Lowe et al. (1984) reported that errors in PPH estimates would have produced a MAPE of 4.3% for 1980 municipal population estimates in Washington, even if the other components had been estimated perfectly. In a similar analysis of 1990 Florida estimates, Smith and Cody (1994) reported that errors in PPH estimates would have produced MAPEs of 2.2% and 5.5%, respectively, for county and subcounty population estimates, even if the other components had been estimated perfectly. In both studies, MAPEs for the hypothetical population estimates were similar to MAPEs for the PPH estimates. The same result was found for measures of bias.

We believe that errors for PPH estimates represent approximate upper limits on the potential contribution of PPH errors to population estimation errors. For most sets of estimates, the contribution of PPH errors to population estimation errors will be less than that suggested by these upper limits because PPH errors and household errors frequently offset each other (to some extent). The effects of offsetting errors can be seen in the studies for New Jersey, Washington, and Florida cited earlier: in all three studies, the MAPEs for population estimates were less than the sum of the MAPEs for the PPH and household estimates (or, for Washington, the PPH and housing unit estimates). Reductions in PPH errors, then, will generally have a less-than-proportional impact on population estimation errors.

CONCLUSION

Regression methods have been used for estimates of the total population for many years (e.g., Crosetti and Schmitt 1956; O'Hare 1976; Swanson and Tedrow 1984), but have seldom been used for PPH estimates. In this article, we described several regression models and discussed their performance in estimating PPH for counties in Florida, Illinois, Texas, and Washington. On the basis of this analysis, we believe that regression models are capable of producing PPH estimates that are more precise and less biased than those produced by more commonly used methods.

It is unlikely, however, that a single model will provide the best PPH estimates under all circumstances. In this study, for example, Model 3 had the largest MAPE of any regression model in 1990 but the smallest in 2000, whereas Model 2 had the smallest MAPE in 1990 and the largest in 2000 (see Table 3). Because the performance of individual models varies over time—and it is impossible to know in advance which model will perform better

[7] This analysis ignores the potential impact of errors in group quarters estimates on population estimation errors. In most instances, this impact is small.

Table 5. Mean Absolute Percentage Errors (MAPEs) and Mean Algebraic Percentage Errors (MALPEs) for PPH Estimates for Counties in Which PPH Changed by 5% or More: 1990 and 2000

	1990		2000	
Models	MAPE	MALPE	MAPE	MALPE
Regression Models				
1 (Basic)	2.3	–1.7	5.0	0.5
2 (Ratio)	2.9	–2.6	5.8	2.1
3 (Change)	3.3	–3.2	2.7	–1.3
4 (Ratio change)	4.5	4.5	2.9	2.0
Average	1.3	–0.8	2.8	0.8
Traditional Models[a]				
A (No change)	7.2	7.2	7.3	6.4
B (Extrapolated county percentage change)	5.9	–5.2	3.7	1.7
C (State percentage change)	3.5	3.5	8.1	7.8
Regression Models Including IRS Variables				
1 (Basic plus IRS)	2.4	–0.7	3.4	0.2
2 (Ratio plus IRS)	2.5	1.1	4.3	1.4
3 (Change plus IRS)	—	—	2.6	–1.4
4 (Ratio change plus IRS)	—	—	2.6	1.1
Average (including IRS)	2.2	0.2	2.1	0.3

[a]Model A: $PPH_t = PPH_c$; Model B: $PPH_t = PPH_c + [(PPH_c - PPH_{c-1}) / PPH_{c-1}]\ PPH_c$; and Model C: $PPH_t = PPH_c(\text{State } PPH_t / \text{State } PPH_c)$.

for any given set of estimates—we favor the use of an average based on several models. Not only does an average reduce the odds of making large errors, it has often been found to produce smaller errors than *any* of the individual estimates (or forecasts) making up the average (e.g., Ahlburg 1999; Sanderson 1999; Smith and Mandell 1984).

The greatest advantage of regression models over traditional methods may be their ability to perform well in places where PPH is changing rapidly or following unusual trends. Regression-based methods explicitly account for postcensal changes in counties' demographic characteristics. Traditional methods, on the other hand, rely on previous census data or postcensal changes at the state or national level. Consequently, traditional methods are likely to perform well in counties where PPH is changing slowly or following stable trends, but poorly in counties where PPH is changing rapidly or following unusual trends.

To test this hypothesis, we calculated MAPEs and MALPEs for counties in which PPH changed by 5% or more during the decade preceding the estimation year.[8] The results for 1990 and 2000—the only years for which all four regression models could be applied—are shown in Table 5. In every instance, the Average estimates were more

[8] In our sample of 462 counties, PPH changed by 5% or more in 424 counties during the 1970s, in 174 counties during the 1980s, and in 60 counties during the 1990s.

precise and less biased than were the estimates produced by traditional methods. In most instances, MAPEs were two, three, or four times larger for the traditional methods than for Average; differences in MALPEs were even greater. The most important benefit of regression-based PPH estimates, then, may be their ability to reduce the large errors often produced by traditional methods in places that are undergoing substantial changes in demographic composition.

The regression models discussed in this article performed well, but many questions remain to be answered. What additional models and variables should be tested? What statistical adjustments may improve the results? Do some models perform better under some circumstances than others? Do some variables affect PPH differently in some places than in others? Can the same models be used everywhere, or may customized models improve PPH estimates for some places? What type of average provides the best estimates (e.g., simple, weighted, or trimmed)? Given current data limitations, can useful PPH estimates be developed for subcounty areas? Future research will undoubtedly lead to further improvements in the performance of regression-based PPH estimation models.[9]

REFERENCES

Ahlburg, D. 1999. "Using Economic Information and Combining to Improve Forecast Accuracy in Demography." Unpublished paper, Industrial Relations Center, University of Minnesota.

Brown, W. 1999. "Use of Property Tax Records and Household Composition Matrices to Improve the Housing Unit Method for Small Area Population Estimates." Paper presented at the Population Estimates Methods Conference, U.S. Bureau of the Census, Suitland, MD.

Comprehensive Planning Organization of the San Diego Region. 1974. *A Model for Estimating Household Size in the San Diego Region*. San Diego: Author.

Crosetti, A.H. and R.C. Schmitt. 1956. "A Method of Estimating the Intercensal Population of Counties." *Journal of the American Statistical Association* 51:587–90.

Findley, S.E. 1979. "Nonlinear Estimation of Household Size: The Minnesota Housing Unit Method." Technical Memorandum 79-1. St. Paul: Minnesota State Planning Agency.

Gober, P. 1990. "The Urban Demographic Landscape: A Geographic Perspective." Pp. 232–48 in *Housing Demography,* edited by D. Myers. Madison: University of Wisconsin Press.

Hodges, K. and M.K. Healy. 1984. "A Micro Application of a Modified Housing Unit Model for Tract Level Population and Household Estimates." Paper presented at the annual meeting of the Population Association of America, Minneapolis, MN, May 3–5.

Keilman, N. 1999. "How Accurate Are the United Nations World Population Projections?" Pp. 15–41 in *Frontiers of Population Forecasting,* edited by W. Lutz, J. Vaupel, and D. Ahlburg. New York: The Population Council.

Kobrin, F.E. 1976. "The Fall in Household Size and the Rise of the Primary Individual in the United States." *Demography* 13:127–38.

Loftin, C. and S.K. Ward. 1983. "A Spatial Autocorrelation Model of the Effects of Population Density on Fertility." *American Sociological Review* 48:121–28.

[9] Some analysts have noted that the basic assumptions of OLS regression analysis may be violated when data are geographically referenced (e.g., Loftin and Ward 1983; Mencken and Barnett 1999; Voss, Hammer, and Long 2001). When error terms are spatially autocorrelated, OLS estimates will be unbiased but inefficient. Accounting for spatial autocorrelation is a complex undertaking because there are numerous ways to measure the proximity of one geographic area to another, to model the autocorrelation structure, and to conduct statistical tests. Because the primary objective of our research was to determine whether regression models can be developed that produce better PPH estimates than the methods currently used—*not* to draw statistical inferences regarding the determinants of PPH—we did not adjust for the potential impact of spatial autocorrelation. One avenue for future research is to investigate whether testing (and perhaps adjusting) for spatial autocorrelation may improve regression-based PPH estimates.

Lowe, T.J., W.R. Myers, and L.M. Weisser. 1984. "A Special Consideration in Improving Housing Unit Estimates: The Interaction Effect." Paper presented at the annual meeting of the Population Association of America, Minneapolis, MN, May 3–5.

McKibben, J.N. and D.A. Swanson. 1997. "Linking Substance and Practice: A Case Study of the Relationship Between Socio-Economic Structure and Population Estimation." *Journal of Economic and Social Measurement* 23:135–47.

Mencken, F.C. and C. Barnett. 1999. "Murder, Nonnegligent Manslaughter, and Spatial Autocorrelation in Mid-South Counties." *Journal of Quantitative Criminology* 15:407–22.

Myers, D. and A. Doyle. 1990. "Age-Specific Population-per-Household Ratios: Linking Population Age Structure with Housing Characteristics." Pp. 109–30 in *Housing Demography,* edited by D. Myers. Madison: University of Wisconsin Press.

Namboodiri, N.K. 1972. "On the Ratio-Correlation and Related Methods of Subnational Population Estimation." *Demography* 9:443–53.

O'Hare, W. 1976. "Report on a Multiple Regression Method for Making Population Estimates." *Demography* 13:369–79.

Roe, L.K., J.F. Carlson, and D.A. Swanson. 1992. "A Variation of the Housing Unit Method for Estimating the Population of Small, Rural Areas: A Case Study of the Local Expert Procedure." *Survey Methodology* 18:155–63.

Sanderson, W. 1999. "Knowledge Can Improve Forecasts: A Review of Selected Socioeconomic Population Projection Models." Pp. 88–117 in *Frontiers of Population Forecasting,* edited by W. Lutz, J. Vaupel, and D. Ahlburg. New York: The Population Council.

Schmitt, R. and J. Grier. 1966. "A Method of Estimating the Population of Minor Civil Divisions." *Rural Sociology* 31:355–61.

Shahidullah, M. 1999. Unpublished survey of state members of the Federal-State Cooperative Program for Population Estimates. Springfield: Illinois Center for Health Statistics.

Smith, S.K. 1986. "A Review and Evaluation of the Housing Unit Method of Population Estimation." *Journal of the American Statistical Association* 81:287–96.

Smith, S.K. and S. Cody. 1994. "Evaluating the Housing Unit Method: A Case Study of 1990 Population Estimates in Florida." *Journal of the American Planning Association* 60:209–21.

Smith, S.K. and B. Lewis. 1980. "Some New Techniques for Applying the Housing Unit Method of Local Population Estimation." *Demography* 17:323–39.

Smith, S.K. and M. Mandell. 1984. "A Comparison of Population Estimation Methods: Housing Unit Versus Component II, Ratio Correlation, and Administrative Records." *Journal of the American Statistical Association* 79:282–89.

Starsinic, D.E. and M. Zitter. 1968. "Accuracy of the Housing Unit Method in Preparing Population Estimates for Cities." *Demography* 5:475–84.

State of New Jersey. 1984. "An Evaluation of Population Estimating Techniques in New Jersey." Office of Demographic and Economic Analysis, 1980 Test of Methods Report No. 3. Trenton, NJ: Department of Labor.

Swanson, D.A. and D. Beck. 1994. "A New Short-Term County Population Projection Method." *Journal of Economic and Social Measurement* 20:25–50.

Swanson, D.A. and L.M. Tedrow. 1984. "Improving the Measurement of Temporal Change in Regression Models Used for County Population Estimates." *Demography* 21:373–81.

U.S. Bureau of the Census. 1990. "State and Local Agencies Preparing Population and Housing Estimates." *Current Population Reports*, Series P-25, No. 1063. Washington, DC: U.S. Government Printing Office.

———. 1998. "Subcounty Population Estimates Methodology." Available on-line at www.census.gov/population/methods/e98scdoc.txt

U.S. Census Bureau. 2001. "Profile of General Demographic Characteristics." Table DP-1 in *2000 Census of Population and Housing*. Washington DC: U.S. Government Printing Office.

Voss, P., R. Hammer, and D. Long. 2001. "Demography and Regression Analysis: What Demographers Can Learn From Quantitative Geographers." Paper presented at the annual meeting of the

Southern Demographic Association, Miami Beach, FL, October 11–13.

Voss, P. and H. Krebs. 1979. *The Use of Federal Revenue Data for Improving Current Estimates of Average Household Size for Minor Civil Divisions: An Evaluation*. Technical Series 70-5. Madison: Applied Population Laboratory, University of Wisconsin–Madison.