Year-to-Year Correlation in National Health Interview Survey Estimates

Chris Moriarity[1], Van L. Parsons[2]
[1,2]National Center for Health Statistics, Centers for Disease Control and Prevention
3311 Toledo Rd, Hyattsville, MD 20782

Abstract

The National Health Interview Survey (NHIS) is a multi-purpose health survey conducted by the National Center for Health Statistics (NCHS). Public-use microdata files and complex sample design variance estimation structures are available at the NCHS Internet web site. In 2007, NCHS expanded the variance estimation structures to cover all available years of public-use microdata files, which enabled researchers to compute appropriate variance estimates when conducting analyses of NHIS annual data pooled across years. We present research findings about the level of correlation in NHIS estimates from year-to-year, which provide insights in areas such as: 1) the amount of error in variance estimates for pooled data within a sample design period if the annual data are incorrectly assumed to be uncorrelated; 2) the trend in correlation over a sample design period.

Key Words: Complex sample; Variance estimation; Pooling survey data

1. Introduction

The National Health Interview Survey (NHIS) is the principal source of information on the health of the civilian noninstitutionalized population of the U.S. The NHIS sponsor is the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention. The NHIS is a continuous survey that has been in operation since 1957. In the current NHIS sample design, implemented in 2006, when full funding is available for the survey, we anticipate obtaining completed interviews at approximately 35,000 living quarters (households and noninstitutional group quarters such as college dormitories) each year. All eligible (i.e., civilian) persons at a sampled address are included in the NHIS interview, yielding a sample of approximately 87,500 persons each year in the current sample design period when full funding is available. Each interview is conducted via a personal visit to the living quarters by an employee of the U.S. Bureau of the Census, which is the data collection agent for the NHIS.

The NHIS sample consists of clusters of living quarters chosen within a first-stage sample of U.S. counties. This sampling method is used to control the costs related to personal visit interviewing. The cost of conducting personal visit interviews in a simple random sample of U.S. living quarters would be prohibitive, due to the amount of travel that would be required.

The current sample design of the NHIS was implemented in 2006, based on Census 2000 information. The current sample design is very similar to the previous sample design, which was in effect from 1995 to 2005, and was based on 1990 Census information. The sample design period before that spanned 1985 to 1994, and was based on 1980 Census information. The other previous sample design periods for which microdata files are available are: July 1, 1962 (beginning of fiscal year 1963) to 1972, and 1973 to 1984. Although the NHIS began in July 1957, microdata files were not retained prior to fiscal year 1963, so it is not possible to assess the year-to-year correlation in NHIS estimates using microdata prior to fiscal year 1963.

Prior to 1968, the NHIS operated on a fiscal year basis (July 1 to June 30), and the oldest existing NHIS microdata files are fiscal year files. For example, fiscal year 1963, the first year for which NHIS microdata files are available, corresponds to the period July 1, 1962 to June 30, 1963. In 1968, the NHIS transitioned to a calendar year basis, and both 1968 fiscal year and 1968 calendar year microdata files were created. After 1968, all full-year NHIS microdata files are calendar year files.

We discuss below why there is year-to-year correlation in NHIS estimates. We then describe the standardized variance estimation structures that now exist for 1963-2007 NHIS, and present some analyses that estimate year-to-year correlations and evaluate the performance of the standardized structures.

2. Sample Design Feature That Introduces Year-to-Year Correlation in NHIS Estimates

The first-stage sample of U.S. counties that are selected at the beginning of a sample design period remains the same for the sample design period. Thus, each annual sample resides in the same geographic areas. Although no address is selected for interview more than once during a sample design period, the addresses selected for the annual sample in a given U.S. county in one year usually are geographically proximate to the addresses selected in the county the previous year. Unless the two annual samples are drawn from an area that is very heterogeneous, households with similar characteristics are likely to be included in the two annual samples. Treating two annual samples within a sample design period as statistically independent for variance estimation will omit the year-to-year correlation between the samples, leading to biased variance estimates.

3. Variance Estimation for the NHIS Public-Use Microdata Files

NCHS provides variance estimation guidance and structures online at the NHIS Methods webpage:

http://www.cdc.gov/nchs/about/major/nhis/methods.htm

NCHS did a large-scale expansion of the guidance and structures in 2007, including guidance/structures for analysis of pooled data. For example, the 1997-2005 guidance contains a section "Merging Data Files and Combining Years of Data in the NHIS" that provides an in-depth discussion of recommended methods for analysis of pooled data. We applied these recommended methods in our analyses.

Briefly, the standardized structures are based on previously-developed structures. In some time periods, modifications and/or extensions of earlier structures were implemented. For example, we created the 1963-1972 standardized structure in 2007 by extending a previous 1969-1972 structure backwards in time; we also made a modification to the previous 1969 structure to make it legitimate for pooled analyses.

All of the structures consist of Pseudo-Strata, containing at least two Pseudo-PSUs per Pseudo-Stratum. The presence of two or more Pseudo-PSUs per Pseudo-Stratum and the use of a "with replacement" sampling assumption assure that contemporary complex sample design variance estimation software (e.g., SUDAAN, Stata, SPSS, SAS survey procedures, R) is capable of producing standard error estimates from NHIS data, including subgroup analyses.

Within a sample design period, pooled analyses should treat the data years being pooled as dependent, because the annual samples were drawn from the same geographic areas, as described above. Sample cases from a given geographic area, sampled in different years, should be assigned to the same Pseudo-Stratum for variance estimation. The standardized structures assure this. When a pooled analysis crosses a sample design change boundary (e.g., 2005-2006), the annual samples from the different designs should be treated as independent. Some modification of the Pseudo-Strata variables is needed to assure distinct Pseudo-Strata values in different sample design periods; an algorithm for achieving this appears below. For a pooled analysis that is a combination of "within" and "across" sample designs (e.g., 2004-2006), the years within a sample design (e.g., 2004-2005) should be treated as dependent, and then the chunks of "within" data (e.g., 2004-2005, 2006) should be treated as independent from each other.

From 1997 to the present, the NHIS public-use files contain Pseudo-Stratum and Pseudo-PSU codes that can be used directly in variance estimation software. From 1980 to 1996, the NHIS public-use files contain sufficient information to create Pseudo-Stratum and Pseudo-PSU codes. For public-use files prior to 1985, auxiliary data files are available at the NHIS Methods webpage. Note that for the 1980-1984 public-use files, the variance estimation information on the files is equivalent to the content of the auxiliary files. However, the auxiliary files contain a Pseudo-Stratum variable, which the user would have to create if working only with the information in the public-use files. When NCHS released the oldest NHIS microdata files (fiscal year 1963 to calendar year 1968) for the first time in 2008, auxiliary data files for variance estimation were provided at the NHIS Methods webpage at the same time the microdata files were released.

A summary of the standardized structures follows:

Table 1: NHIS public-use microdata file standardized variance estimation structures

Data Years	Structure
1958[a]-1962[a]	N/A (no microdata)
1963[a]-1972	143 strata (142 strata for 1968, 1969), 2 PSUs per stratum
1973-1984	149 strata, 2 PSUs per stratum
1985-1994	62 strata, 4 PSUs per stratum (3 PSUs per stratum in 1985, 2 PSUs per stratum in 1986)
1995-1996	99 strata; 2-4 PSUs per stratum in 1995, 2-3 PSUs per stratum in 1996
1997-2005	339 strata, 2 PSUs per stratum
2006-end of sample design period that began in 2006	300 strata, 2 PSUs per stratum

[a] Fiscal year (e.g., fiscal year 1958 was July 1, 1957 to June 30, 1958)

The range 1995-2005 is one sample design period, but there are different public-use variance estimation structures for 1995-1996 and 1997-2005. Due to confidentiality concerns, a number of changes were made in the NHIS public-use files for 1997, including a new public use file variance estimation structure. Although the 1996 and 1997 samples are correlated, it is not possible to estimate the correlation using the public use file variance estimation structures that are available.

Note that in the above table, the number of strata never exceeds 999. Also, the Pseudo-Stratum values in the NHIS public-use files for more recent years and auxiliary files for earlier years never exceed 999. Thus, an algorithm that guarantees distinct Pseudo-Strata values for different sample design periods in all pooled analyses of NHIS data from fiscal year 1963 to the present is as follows:

1. For the period fiscal year 1963-calendar year 1972, add 1000 to the Pseudo-Strata values.
2. For the period 1973-1984, add 2000 to the Pseudo-Strata values.
3. For the period 1985-1994, add 3000 to the Pseudo-Strata values.
4. For the period 1995-1996, add 4000 to the Pseudo-Strata values.
5. For the period 1997-2005, add 5000 to the Pseudo-Strata values.
6. For the period 2006-end of current sample design period, add 6000 to the Pseudo-Strata values.

Provided that no future structure contains more than 999 Pseudo-Strata, this algorithm can be continued indefinitely in the obvious way.

A pooled analysis that encompasses all of the NHIS person files between 1963 and 2007 consists of over 5 million records, with 1,411 degrees of freedom (number of PSUs - number of strata) when Pseudo-Strata values are altered as described above.

4. Results

4.1 Error in assuming that correlated NHIS data are uncorrelated

A simple example follows that illustrates the magnitude of error that can occur by incorrectly assuming that two years of data within a sample design period are uncorrelated. In this example, 1995 and 1996 data are pooled, weights are divided by 2 (because 2 years of data are being pooled), and the data are used to estimate the total U.S. population:

Estimate: ~263,081,000
Correct standard error estimate, accounting for correlation: ~8,208,000
Incorrect standard error estimate, not accounting for correlation: ~5,506,000 (33% underestimate)

The substantial positive correlation for the two years of data is omitted from the standard error estimate when the two years of data are treated as independent.

When 2006 and 2007 data are pooled, or 1987 and 1988 data are pooled, the incorrect standard error estimate is a 26% underestimate. If the 1995 and 1996 data are pooled, and weights adjusted proportional to sample size (1995 sample size is 102,467, 1996 sample size is 63,402), the correct standard error estimate, accounting for correlation, is ~6,515,000, and the incorrect standard estimate is then a 15% underestimate.

It is not difficult to perform an analysis of two data years within a sample design period with the erroneous assumption that the two data years are uncorrelated. In contrast, as shown in Table 1 above, the structures are different enough in different periods that it is unlikely that a data user would inadvertently analyze two data years as correlated when they should not be, so we do not examine this type of error here.

4.2 Trend in correlation over a sample design period - Census Region population estimates

Gonzalez, et al. (2001) presented a model for approximating the relative standard error of a prevalence estimator based on "n" years of pooled data, where for simplicity the year-to-year correlation was assumed to be constant in the model. Gonzalez, et al. noted, however, that the year-to-year correlation is likely to decrease within a sample design period as time increases.

The standardized variance structures permit a direct examination of the trend in correlation over a sample design period by aggregating weighted sums at the Pseudo-PSU level for two different data years, concatenating the data into one file at the Pseudo-PSU level for analysis, generating covariance matrices at the Pseudo-Stratum level (we used the PROC CORR module in SAS), summing the matrices across Pseudo-Strata to obtain an overall covariance matrix, which we then transformed into a correlation matrix. Although this is an operation that can be carried out for any two years within a sample design period, the focus of our research was to select a year early in a given sample design period and then compute correlations between that year and all later years in the sample design period, in order to assess the change in correlation as the gap between years increased. We carried this analysis out for the entire 1963-2007 period, within each group of years shown in Table 1 above, for estimates of the total U.S. (civilian noninstitutional) population in each of the four Census Regions (Northeast, Midwest, South, West). Total population is a generic statistic that is known to have increased at a fairly steady rate throughout the 1963-2007 period. *A priori*, we anticipated seeing high positive correlations in adjacent years, positive correlation throughout a sample design period, and a decline in correlation as the gap between years increased.

For 2006 data pooled with 2007 data, the correlations range between 0.79 and 0.87. More specifically, the estimated correlation between the 2006 population estimate for the Northeast Census Region and the 2007 population estimate for the Northeast Census Region is 0.79, the estimated correlation between the 2006 population estimate for the West Census Region and the 2007 population estimate for the West Census Region is 0.87, and the correlations between the 2006 and 2007 population estimates for the other two Census Regions are between 0.79 and 0.87. No trend analysis can be done for the current sample design period until more data years are available.

For 1997 data pooled with 1998 data, the correlations range between 0.47 and 0.62, declining to a range between 0.27 and 0.39 for 1997 data pooled with 2005 data. All correlation estimates for 1997 data pooled with single years of data in the 1998-2005 interval are positive.

We observed several negative correlation estimates in 1995 data pooled with 1996 data, which suggests that the variance estimation structure for this period is not as robust as for 1997-2005 and 2006-2007. The correlations range between -0.22 and 0.34. The 1996 NHIS sample size is reduced because approximately 3/8 of the 1996 sample was used for research for the new NHIS questionnaire that was implemented in 1997. The number of Pseudo-PSUs is 393 in 1995, 295 in 1996. The initial concatenation used the 1995 Pseudo-PSUs. This concatenation included all of the 1995 data, with zero contribution from 1996 for 98 of the Pseudo-PSU level sums. A second concatenation of the 1995 and 1996 data, using only the Pseudo-PSUs present in 1996 (which excludes some of the 1995 data), gives a different correlation estimate outcome with a range of -0.03 to 0.63.

There were large reductions in the NHIS sample in 1985 and 1986 due to budget shortfalls, so we examined 1985-1986 separately from 1987-1994. For 1987 data pooled with 1988 data, the correlations range from 0.71 to 0.91, declining to a range between 0.41 and 0.85 for 1987 data pooled with 1994 data. All correlation estimates for 1987 data pooled with single years of data in the 1988-1994 interval are positive. For 1985 data pooled with 1986 data, one correlation estimate is negative. The correlations range between -0.22 and 0.48. The number of Pseudo-PSUs is 186 in 1985, 124 in 1986. The initial concatenation used the 1985 Pseudo-PSUs. This concatenation included all of the 1985 data, with zero contribution from 1986 for 1/3 of the Pseudo-PSU level sums. A second concatenation of the 1985 and 1986 data, using only the Pseudo-PSUs present in 1986 (which excludes some of the 1985 data), gives a different correlation estimate outcome with a range of 0.21 to 0.51.

For 1973 data pooled with 1974 data, one correlation estimate is negative. For 1973 data pooled with other years between 1975 and 1984, there is considerable year-to-year variability, with several other negative correlation estimates occurring.

For 1963 data pooled with 1964 data, the correlations range from 0.29 to 0.84. For 1963 data pooled with other years between 1965 and 1972, there are two correlation estimates for the West Census Region that are negative. Note that 1968 was the year when NHIS made the transition from fiscal year to calendar year; for our analysis of data in the 1963-1972 period, we used calendar year data for 1968. Correlation estimates for 1963 data pooled with 1972 range from 0.06 to 0.63. For the unique case of 1968 fiscal year data pooled with 1968 calendar year data, where there is a six month overlap in the data files (67,608 records that are common to both person files), the correlations range from 0.45 to 0.84.

Graphs of the correlations for the 1963-1972 period, the 1973-1984 period, the 1987-1994 period, and the 1997-2005 period appear as Figures 1-4 at the end of this paper. In Figure 1, the abbreviation "63-64" on the horizontal axis denotes correlations between 1963 and 1964 data (one year gap), the abbreviation "63-65" denotes correlations between 1963 and 1965 data (two year gap), etc., with the same convention used for Figures 2-4. The greater variability of the correlations for the earlier periods is apparent at a glance. The level of correlations for the 1987-1994 period is higher than the 1997-2005 period.

Note that year-to-year correlation is a function of the survey design, the weighting methodology used to modify sampling weights, the functional form of the statistic being estimated, and of course, how the three just-mentioned components interact with the specific characteristic targeted for measurement over time. For this paper we are focusing on the basic estimator of total and the impact of the year-to-year design clustering on the correlation of survey total estimator over time. Historically, the public NHIS variance structures have promoted linearization techniques, but these techniques are not amenable to incorporating the poststratification weighting adjustments into the variance estimator. Our non-explicit use of the poststratification in variance estimation, i.e., treating the final weight as an inverse of probability selection weight, should provide correlations of totals somewhat consistent with those produced by "pure" sampling weight strategies. Furthermore, in the interests of confidentiality, and providing consistent design structures for each NHIS cycle, the public-use design structures are coarsened versions of the true structures. Variances and correlations produced by the public-use design structures may be less stable than those produced by use of the in-house design structures, particularly for estimates based on small sample sizes or on samples clustered in only a few Pseudo-PSUs.

5. Conclusions, Summary

As expected, there is considerable year-to-year correlation in NHIS estimates. There does appear to be a downward trend in correlation as the gap between years increases within a sample design period, but it is clear that the correlation is not ignorable even when pooling data from opposite ends of a sample design period.

More recent variance estimation structures appear to perform better than earlier ones for pooled analyses, particularly the 1973-1984 structure.

Our approach of aggregating data to the Pseudo-PSU level and then concatenating across years at the Pseudo-PSU level generally worked well for producing correlation estimates, except for 1986 and 1996 when sample reductions eliminated a large number of Pseudo-PSUs. It may be difficult or impossible to reliably estimate correlations in this type of situation. When some type of estimate must be obtained, we recommend the one obtained from the smaller group of Pseudo-PSUs that contain data from both years.

When pooling data years with a large sample size disparity (e.g., including 1986 and/or 1996), using pooled weights adjusted proportional to sample sizes rather then the number of years being pooled is an option worth considering, leading to lower standard error estimates.

References

Botman S, Moore T, Moriarity C, and Parsons V. Design and Estimation for the National Health Interview Survey, 1995-2004. Vital Health Stat 2(130). 2000.

Gonzalez JF, Jones C, Moriarity C, and Parsons V. Approximation of Relative Standard Errors of Multi-Year Estimators in the National Health Interview Survey, ASA Proceedings of the Joint Statistical Meetings, 2001.

Figure 1

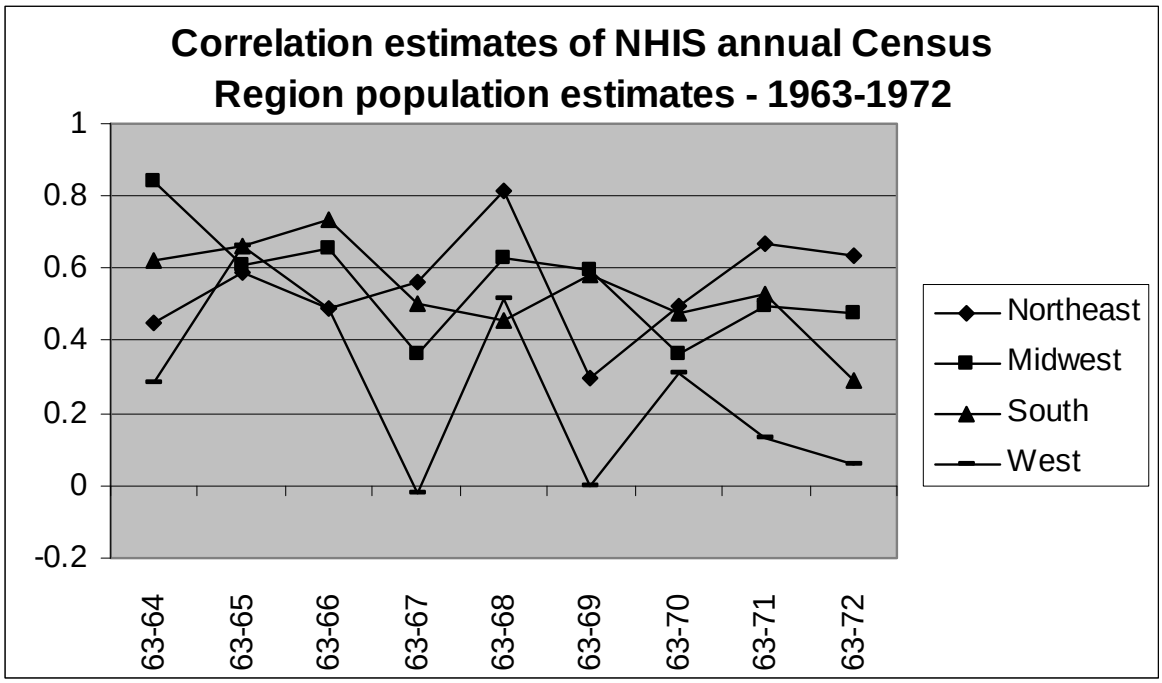

Figure 2

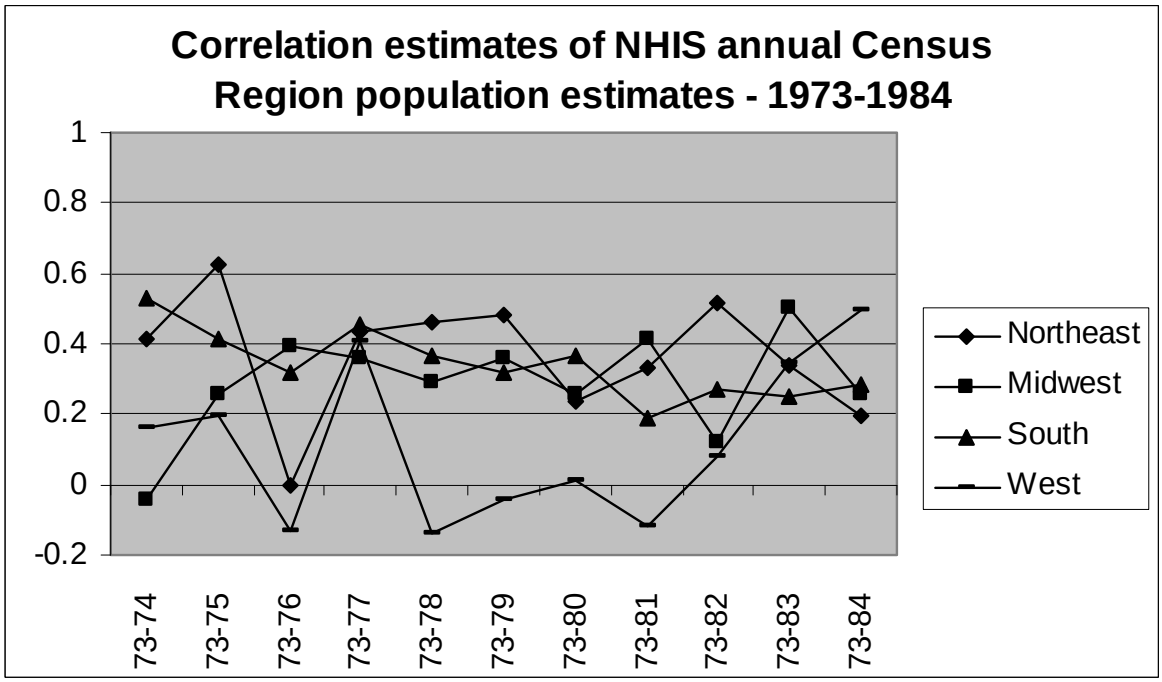

Figure 3

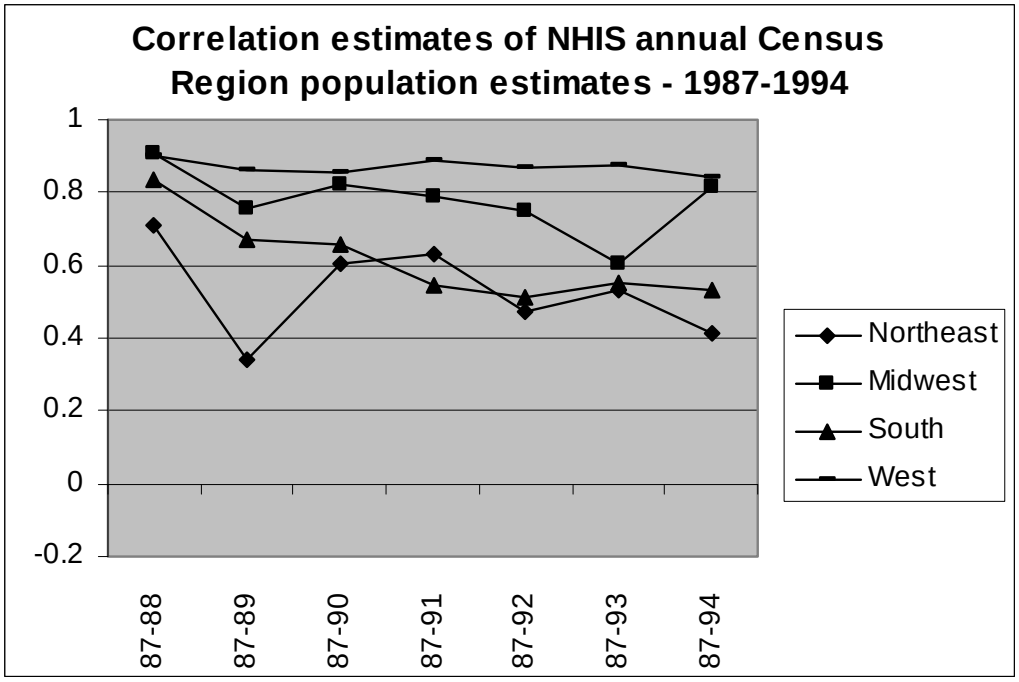

Figure 4