Spatial Data Documentation

Intergenerational Transfer of (IGT) Poverty Project

2015-2017

Prepared by

Purdue Center for Regional Development

Department of Agricultural Economics, Purdue University

Purdue University IGT Team Members:

- Michael Delgado, PhD[b]
- Raymond J. G. M. Florax, PhD[b] (in memoriam)
- Bo Beaulieu, PhD[a]
- Brigitte Waldorf, PhD[b]
- Michael Wilcox, PhD[a]
- Huan Li, PhD[b]
- Tim Smith[b]
- Andrey Zhalnin, PhD[a]
- Yong Jee Kim[b]
- Indraneel Kumar, PhD[a]

This report is prepared by Indraneel Kumar and Yong Jee Kim. The following team members contributed to the spatial harmonized databases: Indraneel Kumar, Yong Jee Kim, Andrey Zhalnin, Huan Li, and Tim Smith.

This research is partly financed through USDA grant 58-6000-5-0044. We acknowledge support of the Economic Research Service (ERS), USDA.

---

a. Purdue Center for Regional Development
b. Department of Agricultural Economics

Table of Contents

List of Tables

List of Figures

1. Spatial Data Documentation

This document presents the structure, development, and metadata for the spatial data, a component of the PSID-Geo developed for the Intergenerational Transfer (IGT) of Poverty Project during 2016-2017. The IGT research team required data at various hierarchies, in particular, individual and different spatial levels to test various hypotheses. For individual level, the research team has employed the Panel Study of Income Dynamics (PSID)[1] database available from University of Michigan. For spatial level, the research team has processed the data in-house at the Purdue Center for Regional Development (PCRD). In particular, harmonized geospatial databases at the county and tract levels for the entire U.S. have been developed for the five decades from 1970 to 2010.

We begin by explaining harmonization of spatial and tabular data, followed by a description of small-area geographies and their coverages available in the U.S. Census Bureau. We refer to select literature to explain the spatial harmonization process and validation of the results. It should be noted that this document is not intended to be a literature review of the spatial harmonization. We compare the Neighborhood Change Database (NCDB) and Longitudinal Tract Database (LTDB), two major crosswalks for harmonization available to researchers and practitioners in the U.S. We provide reasons for selecting the LTDB crosswalk for the IGT Poverty Project. The methodology for LTDB implementation is presented along with flow charts to explain the steps. The data team performed spatial harmonization for counties and census tracts for five consecutive decades, from 1970 to 2010. The remainder of the tracted areas (un-tracted county parts) were harmonized for 1970 and 1980. The chosen reference geography and year was the Decennial Census 2010. The attribute values from 1970, 1980, 1990, and 2000 were adjusted and imputed to the reference geography and year. In total, seventeen major socioeconomic and demographic domains for five decades were harmonized for the IGT research. This report concludes with definitions and limitations of those census domains and variables.

[1] http://psidonline.isr.umich.edu/.

1.1 What is data harmonization?

The harmonization of spatial and tabular data is comprised of processes that make the data comparable and compatible across categories, classification systems, temporal or diverse periods, and geographies. A data harmonization project could entail either one or all of the aspects dependent on the research purpose, data sources, characteristics, and availability of the metadata. Within the Big Data parlance, data harmonization is popularly described as compiling data from disparate sources and making it compatible for decision-making purposes, including forecasting and predictive analytics. The spatial harmonization, in particular, is focused on fixing the geography for one specific reference period and adjusting the attributes of same geographies from other periods to match the reference geography. This is applicable when changes in geographies occur over the decadal census. For example, same census tract would show considerable changes in geography between 1990, 2000, and 2010. Spatial harmonization process would use 2010 geography as a reference and adjust attribute values for 1990 and 2000 to account for differences in the geographical areas. Spatial harmonization enables space-time analysis, and in particular, improves the “metrics for change”. One such example is identifying persistent poverty areas at a sub-county geographic level, which entails analysis of three consecutive decadal censuses.

1.2 Decennial census small-area geography

The Purdue research team has focused on two census geographies, tracts and counties for the IGT Poverty Project. Both census tracts and counties are common geospatial units available from the U.S. Census Bureau at a sub-state level. Census block groups and census blocks are even smaller geographies than the census tracts and were introduced during the 1940 Decennial Census (U.S. Bureau of the Census 1994). However, census block was discontinued in the post 2010, 5-year American Community Survey (ACS). The earliest form of census tracts that could be traced to census surveys were enumerations districts delineated for 4,000 individuals as early as the 1880 Decennial Census (U.S. Bureau of the Census 1994). However, the “census tract” per se was conceptualized by Walter Laidlaw before 1910 in New York to delineate boundaries of the neighborhoods and collect data on the demography (U.S.

Census Bureau 1930). The IGT Poverty Project uses census tracts as the comparable boundary to delineate the neighborhoods. Earlier, tracts were delineated for populations of 3,000 to 8,000, which later was fixed at 4,000 individuals by the U.S. Census Bureau. In 1910, major cities in the Midwest and East Coast, such as Chicago, St. Louis, Cleveland, Boston, Pittsburgh, etc., were tracted for census purposes. Following the Decennial Census of 1910, geographical coverage of the tracts increased, and by 1990, the entire contiguous U.S. was tracted for the first time (Refer to Figure 1). Since there was an upper threshold for population contained within the tracts, as resident populations increased, the tract geography changed considerably between the two consecutive census periods.

Another challenge for data harmonization for the IGT Poverty Project was that only major urban areas in U.S. were tracted for the decades of 1970 and 1980. The research team developed processes to fill-in the remainder of the un-tracted areas in the contiguous U.S. On a similar vein, though the tracts for the entire U.S. became available in 1990, the numbers and configurations of tracts varied considerably from 61,000 polygons in 1990 to nearly 74,000 polygons in 2010. The Decennial Census data were available on the public domain; however, the boundaries of sub-county (tract) geographies were not consistent, compatible, or comparable across the decades. This entailed application of the spatial harmonization techniques to match the geographies and their attribute data across the study period. Refer to Figures 2, 3, and Table 1.

In comparison to census tracts, counties are relatively stable geographies. However, since 1970, around 140 significant changes were uncovered for county boundaries across the decades. These included new counties, such as Yellowstone in Montana or Broomfield in Colorado, which were carved out of the existing counties. In addition to major adjustments to area, names were also changed for the counties. For example, Oglala Lakota County, home of Lakota and Sioux Indian tribes in South Dakota, was formerly known as the Shannon County.

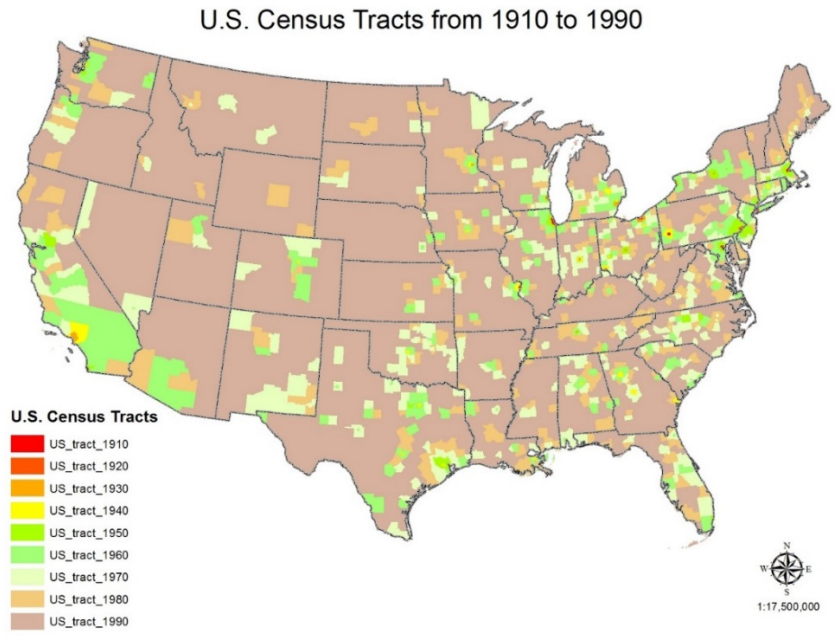


Figure 1: U.S. Census Bureau Tract Coverage from 1910 to 1990

Source: NHGIS, PCRD

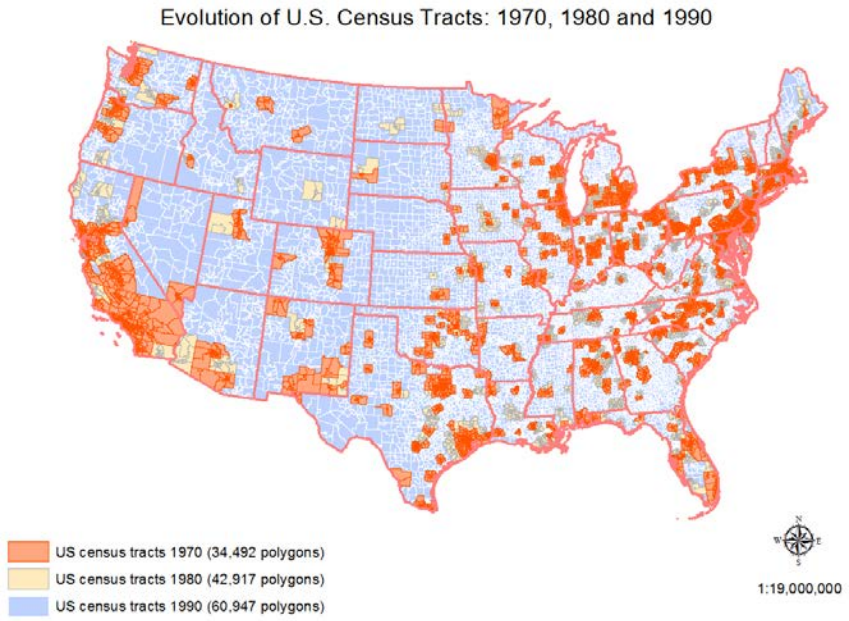


Figure 2: U.S. Census Bureau Tract Coverage from 1970, 1980 to 1990

Source: NHGIS, PCRD

Table 1: Number of Census Tracts in U.S.

Decennial Census	Total Tracts in U.S.	Tracts in Contiguous U.S.
1910	1,980	1,980
1920	5,746	5,746
1930	7,008	7,008
1940	7,563	7,446
1950	12,634	12,494
1960	23,096	22,982
1970	34,494	34,251
1980	46,197	45,923
1990	60,947	60,513
2000	65,312	64,868
2010	72,765	72,271

Note: In 2010, total number of tracts including Puerto Rico is 73,669.

Source: NHGIS, PCRD

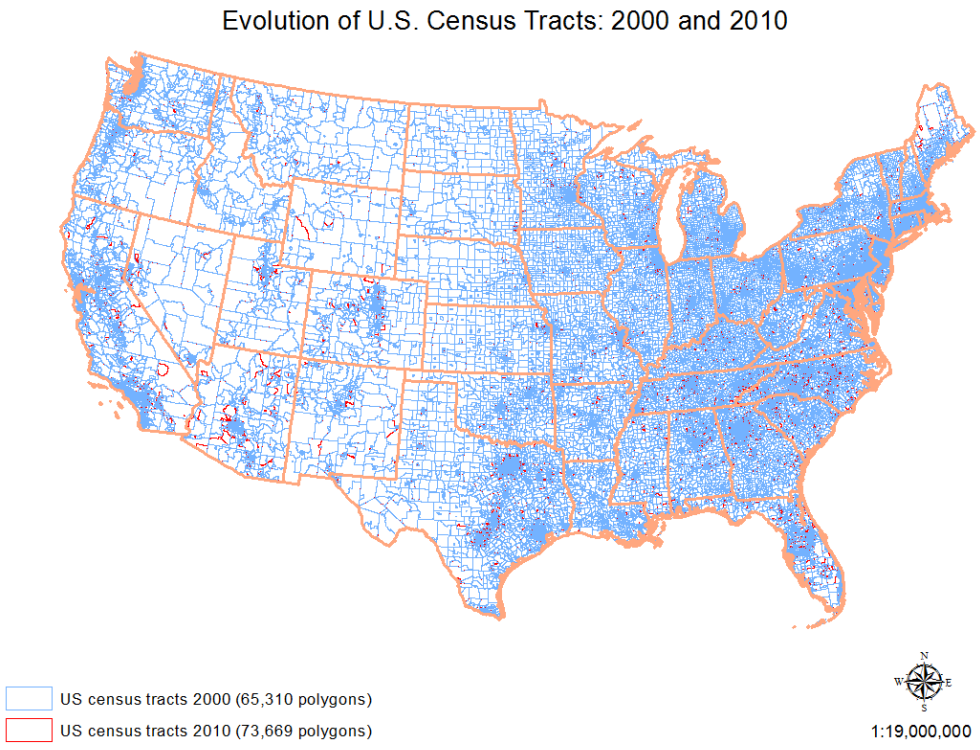


Figure 3: U.S. Census Bureau Tract Coverage from 2000 to 2010

Source: NHGIS, PCRD

2. Spatial Harmonization Process

The spatial harmonization process utilizes areal interpolation methods. This makes spatial harmonization different than the data harmonization processes that focus on attributes and definitions. Xie (2006) elicits that areal interpolation deals with "spatial relationships" and use either "areal weighting" or "distance-based" methods dependent on vector- or raster-based data outputs. The areal-weighting method employs ratios of geographic areas between two periods as weights. An important underlying assumption includes homogeneous densities during both the periods. The assumption for homogeneous density has been noted as a limitation for socioeconomic variables by previous researchers. Goodchild, Anselin and Deichmann (1993) and Xie (2006) mentioned that homogenous density assumption should be considered an "exception than the rule", as densities could vary between two censuses. One of the ways researchers have improved areal weighting is by incorporating ancillary information, such as land use/land cover (LULC) and transport network or street segments (Xie, 2006). The Neighborhood Change Database (NCDB)[2], a commercially available harmonized data for U.S., used street segments as ancillary information to develop part of the tract database (Cornelius and Tatian 2010). In addition to ancillary information, researchers have recommended finer spatial resolutions and iterative algorithms for interpolation. These spatial interpolation models are known as the Dasymetric Models, and are focused on estimating attributes at much finer resolution by utilizing ancillary data for street density, nighttime lights, land cover, slopes, and parcel data (Nagle, et al. 2014). For example, LandScan[3] is a Dasymetric model-based U.S. population data available at 100 by 100 meter grid through the Oak Ridge National Laboratory (ORNL). It was developed primarily to serve the Department of Homeland Security, however, it is now being utilized for planning and response for natural disasters, demand estimation for infrastructure, and central business district (CBD) planning.

It should be noted that areal weighting, interpolation, Dasymetric model, and other spatial harmonization techniques address the Modifiable Areal Unit Problem

[2] http://www.geolytics.com/USCensus,Neighborhood-Change-Database-1970-2000,Products.asp
[3] http://web.ornl.gov/sci/landscan/

(MAUP), but only to a certain extent. The census data collected at individual level, but presented at aggregated spatial levels, have MAUP because there are innumerable ways to aggregate and define spatial units such as a census tract (Openshaw 1983). Further, data aggregated by spatial units in a different way might give different statistical results, and techniques for spatial autocorrelation would not be able to address the MAUP completely (Openshaw 1983). A spatial pattern visible on a map is due, in part, to the underlying distribution of the variable as well as zoning or layout of the spatial units (Martin 1996; Lloyd 2014).

2.1 Validation of spatially-harmonized data

The outputs from spatial and data harmonization -- either through aggregation, collapsing, or imputation processes-- result into new data series. Researchers have cautioned to validate the harmonized data results, wherever feasible. For example, Noble, et al. (2011) suggested computing standard errors for the imputations, enabling users to conduct sensitivity analysis for the harmonized databases. Such analysis is only feasible if actual data values were known, such as comparison of the imputed and raw values through the Public Use Microdata Areas (PUMAs) as proposed for the ongoing Integrated Spatio-Temporal Aggregate Data Series (ISTADS) project (Noble, et al. 2011). A recent research comparing three harmonized databases from Neighborhood Change Database (NCDB), Longitudinal Tract Database (LTDB), and National Historical Geographic Information System (NHGIS) to actual change-values from census, reveals that several ancillay information and complex interpolation methods do not necessarily increase the accuracy of the output (Logan, Stults and Xu 2016). The researchers found that simple areal interpolation with an additional ancillary information, such as, water or land cover layer, could provide reasonable accuracy. Logan, Stults and Xu (2016) revealed that 2000 to 2010 harmonization for LTDB and NHGIS standardized geographies for tracts performed better than NCDB with lower range of error margins, when compared to the actual “tract population change file” prepared by the U.S. Census Bureau. The lack of adequate description of procedures in NCDB stalled the researchers effort to uncover soucres of diffrences from the census change file (Logan, Stults and Xu 2016). Our literature search reveals that at least four

spatially harmonized census data are either available or “under preparation” in U.S. These include ISTADS, NCDB, LTDB, and NHGIS standardized geographies, which cover different decades. The following section compares and contrast NCDB and LTDB.

2.2 Neighborhood Change Database vs. Longitudinal Tract Database

The Neighborhood Change Database (NCDB) and the Longitudinal Tract Database (LTDB) are two major datasets providing harmonized spatial data for sub-county geographies in the U.S. The NCDB was developed in early 1990s by the Urban Institute, with support from the Rockefeller Foundation to harmonize census tract data for 1970, 1980, and 1990 (Geolytics n.d.). Later, Geolytics added the 2000 and 2010 geographies and made adjustments to the previous harmonized databases, and transformed NCDB into a proprietary product. The NCDB provides tract level census data for 1970, 1980, 1990, and 2000 into normalized 2010 tract geographies. This includes variables from decennial census and long form survey including the 2010 census and 2006-2010 American Community Survey (ACS). In addition to tracts, NCDB provides census block groups and census blocks data for select decades. In comparison, LTDB was developed recently through research support from the Russell Sage Foundation and the Brown University. It provides tract level crosswalks to harmonize census data from 1970, 1980, 1990, and 2000 to the 2010 tract geographies.

The LTDB crosswalk files, which are weights for counties and tracts, are available in the public domain. This enables to spatially harmonize any tract-level variable from the previous four decennial censuses to the 2010 geography. It is maintained and provided by the Spatial Structures in Social Sciences (S4) research group at the Brown University (Brown University n.d.). The LTDB also provides a backward crosswalk to convert 2010 to 2000 geography. Appendix 2 compares key characteristics of the NCDB and LTDB. It is evident that processes for harmonizing previous censuses to 2010 census vary between NCDB and LTDB, which could result in different outcomes. According to Tatian (2003), the spatial harmonization in NCDB included a combination of methods, such as areal interpolation, ancillary data on streets, block level population, and even an earlier version of Census Bureau’s tract correspondence file known as the Under Class Data Base (UDB). The spatial harmonization in LTDB used a consistent method of areal interpolation and the ancillary

water layer (Logan, Xu and Stults 2014). As mentioned previously, a variety of ancillary information could not improve the accuracy of imputation in all cases. For the IGT Poverty project, the team decided to proceed with the LTDB because it is a freely available database with full documentation and crosswalk files.

3. Methodology for LTDB Implementation

The LTDB from Brown University provides individual crosswalk[4] files for 1970, 1980, 1990, and 2000 to harmonize to the 2010 tract geographies. Census tracts from a previous decade can change in the current decade because of consolidation, splitting, or minor changes of many-to-one, one-to-many, or many-to-many tracts (Tatian 2003 and Logan, Xu and Stults 2014). Post 1990, more than 60% of tracts remained unchanged with no discernible changes into the tract geographies between 2000 and 2010 (Logan et al. 2014). However, in 1970 and 1980, tracts only covered major urban areas and a large un-tracted region remained in the contiguous U.S. Also, the number of tracts changed considerably over the decades. For example, in 1970, there were around 34,250 tract polygons in contiguous U.S., which changed to 46,000 polygons in 1980 followed by 61,500 in 1990 and 65,000 in 2000, and finally around 72,300 tract polygons in 2010 (Refer to Table 1).

In developing the LTDB database, Brown University researchers assumed any change of less than 1% in a tract's geographical area was tantamount to "no-change" because U.S. Census Bureau made minor corrections to digitized boundaries and shape files through the decades (Logan, Xu and Stults 2014). The 1% limit has the benefit of eliminating any changes because of the non-topological geometry and characteristics of the shape files. The GIS files from previous decades are mostly available as shape files, which are easy to use, but comprised of non-topological vector GIS data (ESRI 1998). The LTDB crosswalks contained weights that were applied to the tract level data for the particular decade obtained from the NHGIS. The result was a database where 1970, 1980, 1990, and 2000 values were harmonized to 2010. Figure 4

[4] https://s4.ad.brown.edu/Projects/Diversity/Researcher/LTDB1.htm

shows a schematic of the tract data process including error redistribution and iteration. The sum total of harmonized tract data is the “control total” and should match to the sum total of the original data. In most cases, the difference was reduced to 0.04% or even less by the first iteration, and the second iteration made minor improvements.

Any GIS database development requires identifying or creating a unique identifier for the spatial unit. For harmonized tract database, we developed 11-digit tract ID from Census 2010 as the GEOID or the unique identifier for a census tract. For example, tract 06083002402, contains first two digit (06) as state ID, next three digits (083) as county ID, and next six digits (002402) as tract ID with the last two digits showing decimals of the actual tract ID. The decimals were created when a particular census tract from previous census was split into two or more census tracts so that the original census tract could be identified for temporal analysis (U.S. Bureau of the Census 1994). The LTDB provided crosswalk files in MS Access and CSV formats. For example, crosswalk for 1970 to 2010 would contain tract ID for 1970, corresponding tract ID for 2010, and respective weights. One-to-many transformation was common as one census tract in 1970 was split into 2, 3, 4 or even more census tracts in 2010. It is to be noted that weights were not distributed equally in 2010, but it was based on areal weighting and water layer to identify developed versus undevelopable parts of the census tract.

3.1 Remainder of county implementation

As noted previously, census tracts covered only major urban areas in U.S. during 1970 and 1980, and a significant area of the contiguous U.S. remained un-tracted. The IGT research required consistent geography for the contiguous U.S., since 1970. The research team, in particular, Professor Raymond Florax, came up with the idea to fill-in the gaps during 1970 and 1980, and develop a consistent set of geospatial data for the five decades. Spatial databases, such as census tracts and counties, are also known as lattice data, and any gaps or missing geographies would create an incomplete spatial system. This could have implications for spatial analysis, such as contiguity-based weight matrices with neighbor less spatial units. Unlike a grid, which is a regular lattice, counties and tracts form irregular lattices, and hence dependent on vertices and edges to define adjacency-based neighborhood structure. A complete lattice-based spatial system for contiguous U.S. would facilitate spatial analysis.

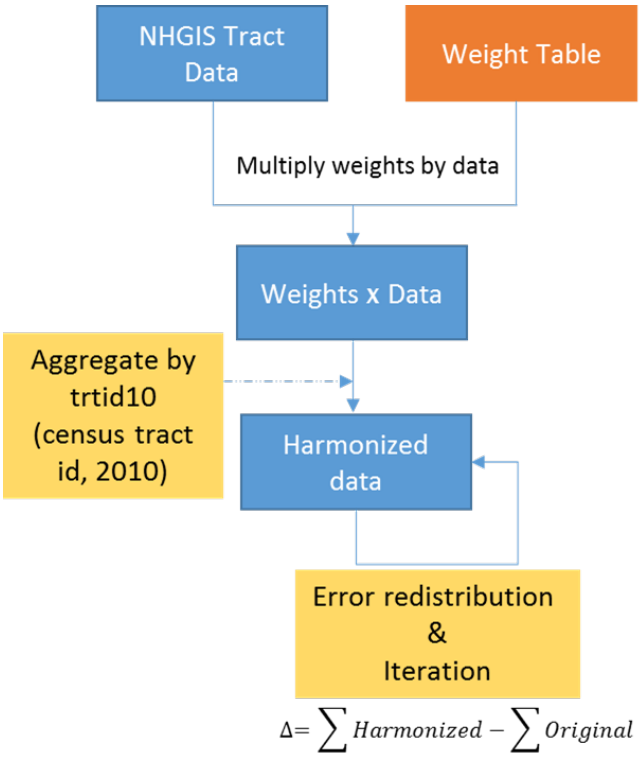


Figure 4: Census Tract LTDB Implementation from 1970 to 2010

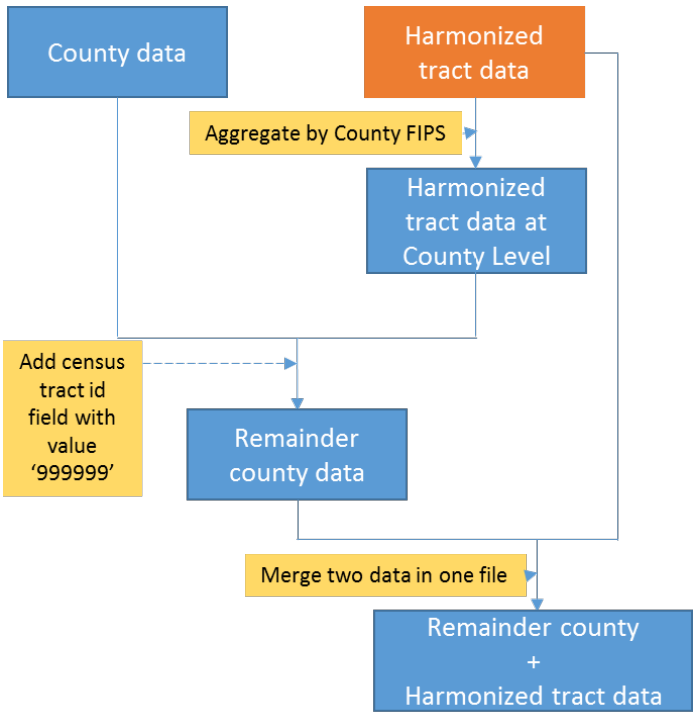


Figure 5: Remainder of County (Un-tracted) and Harmonized Tracts 1970 and 1980

Figure 5 includes the schematic for developing harmonized tracts for un-tracted, remainder of county areas for 1970 and 1980 censuses. The existing tracts in 1970 and 1980 were harmonized to 2010. Data for remainder of county (un-tracted portion) were estimated by deducting the sum of harmonized tracts data from the county total. The remainder of county area was filled-in with 2010 tract geographies, and the data were proportionately distributed based on geographical area. We also considered major water features in developing estimates for developable land area by using the land use/land cover information if available.

A step-wise description of remainder of county process is provided:

1. Data Preparation

- Harmonize 1970 and 1980 decennial censuses tract data by LTDB weights
  ⇨ This step would give 2010 harmonized 1970 and 1980 census tract data
- Harmonize 1970 and 1980 decennial censuses county data by county weights
  ⇨ This step would give 2010 harmonized 1970 and 1980 county data
- Subtract 2010 harmonized 1970 and 1980 census tract data from 2010 harmonized 1970 and 1980 county data
  ⇨ This step would give total remainder of county data

2. Weight (crosswalk) for Remainder of County

- Calculate areas of 2010 harmonized census tracts (1970 and 1980) after excluding 1975 water layer from GIRAS (Geographic Information Retrieval and Analysis System), USGS (United States Geological Survey)
  ⇨ This step would give areas of developable lands
- From LTDB weight file, extract tracted areas of 1970 and 1980 in terms of 2010 harmonized census tracts
- Calculate weights by dividing un-tracted area and sum of all un-tracted areas from 1970 and 1980 from the previous step

$$weight_i = \frac{area_i}{\sum_{k=1}^{n} area_k}$$

Where, $weight_i$ is weight of tract i, k is un-tracted area in the same county of i, and $area_i$ is area (developable land)[5] of tract i.

3.2 County crosswalk implementation

In addition to census tracts and un-tracted areas in 1970 and 1980, a spatially harmonized database was developed for counties from 1970 to 2010. We used the NHGIS decennial county boundary GIS files and developed weight (crosswalk) tables in-house to harmonize the county data. As discussed in a previous section of this report, counties are consistent and stable geographies. However, we discovered around 140 major changes in county boundaries over the five decades. The county crosswalk or weight tables have values as "1" if there were no changes in boundaries. A python code was developed to analyze the intersection of county boundaries between two decades. In general, changes were one-to-many from previous decades to 2010, and the python code estimated the geographical areas of the split and developed county weights based on proportionate areas.

Appendix 3 includes the entire schematic of harmonized data developed for counties, tracts, and un-tracted remainder of county areas. The schematic lays the processes, tabular data, weight tables, and various outputs. The original census data came from NHGIS, tract crosswalks came from LTDB, S4 at Brown University, county and remainder of county (un-tracted 1970 and 1980 areas) crosswalks were developed by PCRD at Purdue University. We developed harmonized data for five decades and later the research team used interpolation methods to develop the annual interpolated database. Since the PSID data are available annually, the interpolated county and tract data provided the corresponding spatial (neighborhood and county) information for the hierarchical model. Unlike PSID, which includes confidential information for individuals, the spatial database is based on publicly available data and crosswalks. However, the merged PSID-geo file becomes a confidential database because of location information of the surveyed individuals. The PSID-geo file is being used by the research team to estimate the econometric models.

[5] Developable land is estimated after considering the GIRAS data.

4. Census Variable Definitions

The IGT Project team identified socioeconomic variables from 17 major domains to harmonize data from 1970, 1980, 1990, and 2000 to 2010 tract geographies. Appendix 1 lists major domains and specific census datasets obtained from the NHGIS.

1. Dropout Prevalence

The dropout prevalence retained similar definitions over decades, albeit classifications changed during specific census periods. For example, the 1970 Decennial Census shows “not enrolled in school and not high school graduate” and “not enrolled in school and high school graduate”. Since all the five decennial censuses have similar definitions, dropout prevalence is comparable across the decades. However, the 1970 dropout prevalence requires caution if it is to be compared with other dropout prevalence data for two reasons. First, the universe of 1970 Decennial Census is different from other censuses. The 1970 Decennial Census used persons 16 to 21 years of age, whereas other censuses used 16 to 19 years of age. Second, 1970 Decennial Census does not mention number of persons who are enrolled in school. Although, we can use population between 16 to 21 years of age as a denominator, we do not know the exact denominator for dropout rate calculation. Therefore, 1970 dropout rate cannot be directly compared with other decennial censuses.

Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Dropout prevalence	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)

	Available	Available with caution	Unavailable

Categories	1970	1980	1990	2000	2010
Dropout rate					

2. a. Educational Attainment (Total)

The definition for educational attainment has changed since the 1990 Decennial Census. In 1970 and 1980, there were no specific questions asked about the level of education and degrees attained by the individuals. For educational attainment information of 1970 and 1980, the “years of school completed” is used to derive the

variables. We can know years spent for elementary, high school, and college education. However, years of school completed does not inform specifically about the type of college degree, such as bachelor, master, professional school, or a doctorate degree. From 1990 and onwards, census provided detailed educational attainment categories including separate enumeration for “some college” and “associates” degrees. The universe of educational attainment remained consistent through all the decennial censuses, which includes individuals of age 25 years and over.

Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Educational attainment	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF4, 2000_SF3a*	ACS(2006-2010)

Categories	1970	1980	1990	2000	2010
Lower than high school					
High school					
Some college or Associate’s degree					
Equal or above Bachelor’s degree					

2. b. Educational Attainment by Race

Educational attainment by race has similar issues as educational attainment of total population, which includes definitions for 1970 and 1980 Decennial Census. In addition to the definitions, educational attainment by race has one more challenge pertaining to the definition for Hispanics. Before 1990 Decennial Census, the definition of Hispanics was not universally established. The 1970 Decennial Census used Spanish American, whereas 1980 Decennial Census used population of Spanish origin as the universe. Therefore, users should be cautious when comparing educational attainment of Hispanics across the decadal censuses.

Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Educational attainment	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF4, 2000_SF3a*	ACS(2006-2010)

	Available	Available with caution	Unavailable

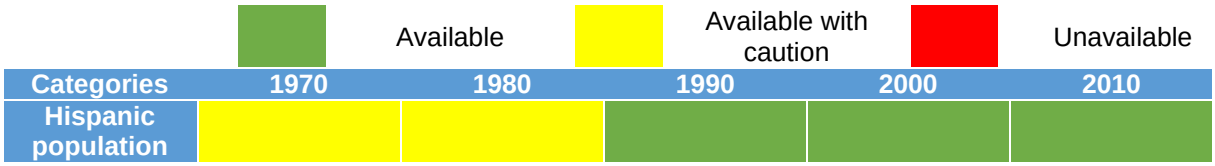
Categories		1970	1980	1990	2000	2010
White	Lower than high school					
	High school					
	Some college or Associate's degree					
	Equal or above Bachelor's degree					
Black	Lower than high school					
	High school					
	Some college or Associate's degree					
	Equal or above Bachelor's degree					
Hispanic	Lower than high school					
	High school					
	Some college or Associate's degree					
	Equal or above Bachelor's degree					

3. Ethnicity

The ethnicity (Hispanics) were not clearly defined in the census before 1990. Since there was no specific definition for Hispanics, 1970 Decennial Census provided Hispanic population information with four different categories: 1. Any of five Spanish

categories of the question on "origin or descent", 2. Puerto-Rican birth or parentage, 3. Spanish language, 4. Not of "Spanish language" but of Spanish surname (in 5 Southwestern states only). In the 1980 Decennial Census, information about ethnicity could be determined from questions related to Spanish origin. Therefore, ethnicity information should be compared cautiously between pre-1980 and post-1990 censuses, and users should be aware of the limitations of this variable.

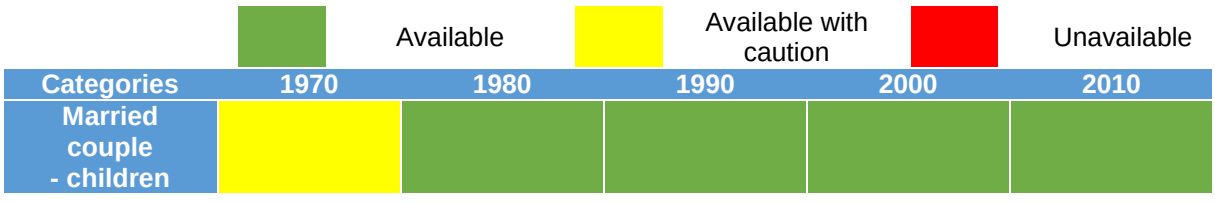
Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Ethnicity	1970_Cnt4Pb	1980_STF1	1990_STF1	2000_SF1a	2010_SF1a



4. Family Structure

Since 1980 Decennial Census, all the decennial censuses have retained similar variable definitions: “Married couple – children”, “Married couple – no children”, “Male householder – children”, “Male householder – no children”, “Female householder – children”, and “Female householder – no children”. Although the 1970 Decennial Census has a different variable definition, similar variables can be developed by aggregation. The 1970 Decennial Census has more specific variable definitions about owning children, like “no children”, “children under 6 years old”, and “children not under 6 years old”. If we aggregate “children under 6 years old” and “children not under 6 years old” as one variable, then it is equivalent to “own children”. The family structure definition is comparable except 1970 through all the decennial censuses.

Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Family structure	1970_Cnt4Pb	1980_STF3	1990_STF1	2000_SF1a	2010_SF1a



Married couple - no children					
Male householder - children					
Male householder - no children					
Female householder - children					
Female householder - no children					

5. Housing Age

Since 1970 Decennial Census, all censuses have retained similar variable definitions. Housing age has been classified usually based on 10-years interval, so it is possible to match the variable definitions by aggregation. However, there is a change of universe at the 1990 Decennial Census. During 1970 and 1980 decennial censuses, "year-round housing units" was used as the universe. In other words, seasonal and recreational vacant units were excluded. On the other hand, "total housing units" has been used as universe since the 1990 Decennial Census. Therefore, 1970 and 1980 decennial census housing age data should be compared to other decennial censuses with caution.[6]

1970 census	Other decennial censuses
Less than 10 years	Less than 10 years
10 – 19 years	10 – 19 years
20 – 29 years	20 – 29 years
More than or equal to 30 years	30 – 39 years More than or equal to 40 years

Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Housing Age	1970_Cnt4H	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)

[6] https://www.census.gov/hhes/www/housing/census/historic/units.html

	Available	Available with caution	Unavailable

Categories	1970	1980	1990	2000	2010
Less than 10 years					
10 – 19 years					
20 – 29 years					
More than or equal to 30 years					

6. Housing Value

Housing value includes the distribution of housing units across different value intervals. Although value intervals are not directly comparable because of inflation, this can be used to show trends of housing values across decades. If users need to compare housing values, they should be cautious that the universe of 1980 is different from other decennial censuses. The 1980 Decennial Census included housing values of owner-occupied non-condominium housing, whereas other census surveys included housing values of owner-occupied housing in general. Besides the change in universe, value intervals also changed through the decades. For example, the 1970 Decennial Census starts with “less than $5,000” as the first interval, and “$50,000 or more” as the last interval. On the other hand, 2010 Decennial Census starts with “less than $10,000” as the first interval, and “$1,000,000 or more” as the last interval. Therefore, users need be cautious about housing value intervals between different decennial censuses, not only because of inflation but also due to different value ranges.

Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Housing Value	1970_Cnt2	1980_STF1	1990_STF1	2000_SF3a	ACS(2006-2010)

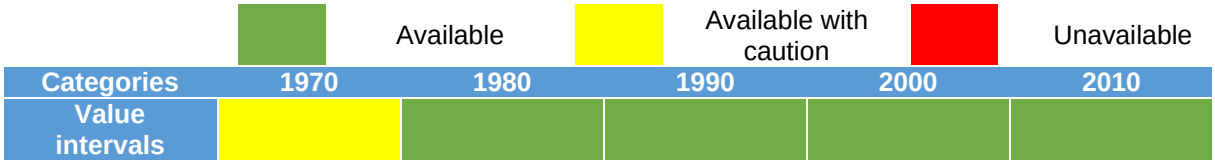
	Available	Available with caution	Unavailable

Categories	1970	1980	1990	2000	2010
Value intervals					

7. Income/Income by Race

Income data include number of households in different income intervals. Although, not directly comparable because of inflation, it can be used to show trends of household incomes. Users need to be aware that the 1970 Decennial Census has a slightly different universe. The 1970 Decennial Census has used “family” as its universe, while other decennial censuses have used “households” as their universe. In terms of definition, family needs at least two or more individuals to be established, whereas household needs at least one person. Besides changes to the universe, intervals for variable changed as well through the decades. For example, the 1970 Decennial Census starts “under $1,000” as the first interval, and “$50,000 and over” as the final interval. On the other hand, 2010 Decennial Census starts “under $10,000 as the first interval, and “$200,000 or more” as the final interval.

Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Income/ Income by race	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)



8. Industry

Decennial censuses provide the number of employed persons by major industry sectors. However, Census Bureau uses their own industry codes and definitions to group industry sectors and sub-sectors, which is different from the North American Industry Classification System (NAICS) approach. Additionally, NAICS codes are updated every 5 years as part of the Economic Census conducted in years ending with a 2 and 7. The Decennial Census industry codes are updated every 10-years. Users need to be cautious when comparing an industry sector from a specific decennial census to other censuses. They are urged to check the detailed definitions. Besides differences in the number of industry sectors, the 1970 Decennial Census has used

employed persons 14 years and over as the universe. The census from 1980 and onwards have used employed persons 16 years and over as the universe. It is important to pay careful attention to the differences in the universe when comparing industry data from 1970 to other decennial censuses.

Categories	1970	1980	1990	2000	2010
# of industry category	13	14	17	13	20

Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Industry	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)


Available

Available with caution

Unavailable

Categories	1970	1980	1990	2000	2010
Major industry groups					

9. Occupation

Decennial censuses provide the number of employed individuals in major occupation groups. For definitions of major occupation groups, Census Bureau has developed occupation codes that are different from Standard Occupational Classification (SOC) developed by the Bureau of Labor Statistics (BLS). Unlike, industry sector revisions that are undertaken every 5 years for NAICS, the SOC is revised once every 10 years. For example, the SOC was revised in 2000 and 2010, and the next revision is anticipated in 2018. In other words, a decennial census occupation group might not be directly comparable to data available as SOC through the BLS. Also, there are subtle changes in occupation definitions over different census periods. Therefore, users are cautioned to look into the occupation definitions when comparing data from different decennial censuses. Besides differences in definitions, similar to industry sectors, the 1970 Decennial Census had a different universe of employed persons 14 years and over, whereas other decennial censuses used employed persons 16 years and over as the universe.

Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Occupation	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF4	ACS(2006-2010)

	Available	Available with caution	Unavailable

Categories	1970	1980	1990	2000	2010
Major occupation groups					

10. Percent Driving

Percent driving includes percentages of car, truck, or van used as primary means of transportation to work. The decennial censuses retained similar definitions with slight changes in categories. Users need to be careful when comparing 1970 decennial data to other decennial censuses for two reasons. First, the 1970 Decennial Census has a different definition for driving. The 1970 census surveyed “Private auto, driver” and “Private auto, passenger” while other decennial censuses used “Car, truck, or van: Drove alone” and “Car, truck, or van: Carpooled”. So, there are some important variations in definition and classification. Second, the universe of 1970 is different from other decennial censuses. In 1970, the universe was defined as workers, regardless of their ages. On the other hand, from 1980 and onwards, the census used workers of age 16 years and over. Therefore, users need be cautious in comparing 1970 percent driving with other decades.

Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Driving	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)

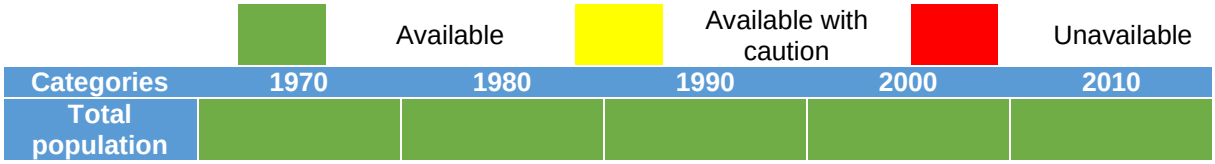
	Available	Available with caution	Unavailable

Categories	1970	1980	1990	2000	2010
Percent of driving to work					

11. Population

Population shows total number of resident individuals in the census tract or county regardless of socioeconomic characteristics. Population definitions and universe have remained consistent through all the decennial censuses.

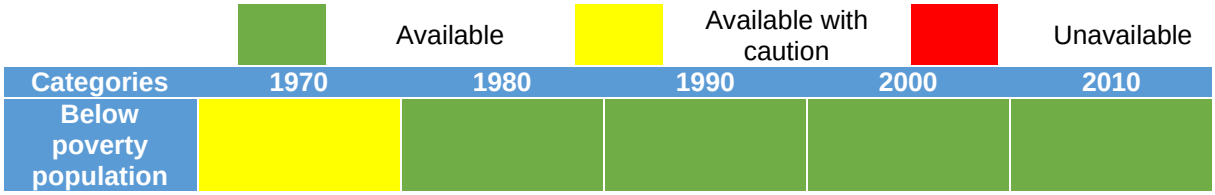
Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Population	1970_Cnt4H	1980_STF1	1990_STF1	2000_SF1a	2010_SF1a



12. Poverty

Poverty is defined as the population falling under the poverty threshold. From 1970 to 2010, the definition has not changed. However, the database contains two distinct poverty variables for 1970 Decennial Census based on individual and family. The individual poverty variable in 1970 Decennial Census included the population who were under the poverty threshold. We do not know the exact denominator (universe) for poverty rate calculation in 1970, however, total population has been used to estimate the poverty rate. For IGT purposes, the database also includes family under poverty threshold. From 1980 to 2010, the criteria for universes have remained the same.

Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Poverty	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)



13. Public Assistance

The specific questions for public assistance recipients started at the 1990 Decennial Census. Although 1970 and 1980 decennial censuses do not have questions asking public assistance recipients directly, they include income types from different sources. Thus, public assistance recipients can be estimated for 1970 and 1980.

Other censuses	1970 census	1980 census
With public assistance income	• Public assistance or welfare payments	• Public assistance income
No public assistance income	• Wage and salary • Nonfarm self-employment • Farm self-employment • Social security or Railroad retirement • Public assistance or welfare payments • All other income	• Wage or salary income • Nonfarm self-employment income • Farm self-employment income • Interest, dividend, or net rental income • Social Security income • All other income

Users comparing 1970 and 1980 public assistances with other decennial censuses need be aware of two issues. First, 1970 Decennial Census asked persons 14 years and over for their sources of income. So, when users need to calculate the percentage of public assistance recipients, the denominator should be from populations 14 years of age and over. In addition, 1970 Census used an individual as the unit, whereas other decennial censuses used households as the units. Therefore, 1970 data cannot be compared directly with other decennial censuses. Second, 1980 Census used households with incomes as their universe. Users need to be cautious about the denominator to calculate the percent of public assistance recipients for 1970 and 1980 decennial censuses.

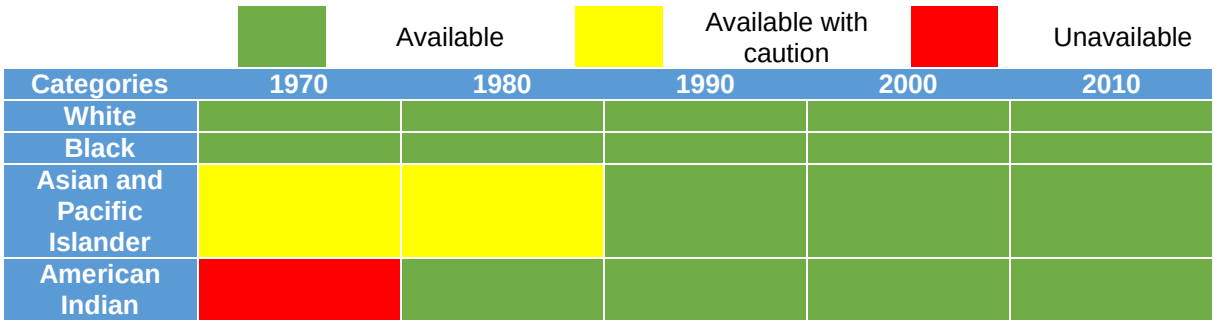
Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Public assistance	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)

Categories	1970	1980	1990	2000	2010
Percent of public assistance recipients					

14. Race

The five decennial censuses include White and Black populations albeit with some differences in other types of race categories. For example, the 1970 Decennial Census surveyed some "Asian and Pacific Islander" categories, like Japanese, Chinese, Filipino, Hawaiian, and Korean. Although the sum of all those categories could be a proxy for "Asian and Pacific Islander", it does not show the exact number of "Asian and Pacific Islander" population. This is same for the 1980 Decennial Census, which enumerated Japanese, Chinese, Filipino, Korea, Asian Indian, Vietnamese, Hawaiian, Guamanian, and Samoan. However, they did not include total Asian population. Other decennial censuses have a category for "Asian and Pacific Islander" or "Asian alone" and "Pacific Islander alone". Therefore, comparing 1970 and 1980 "Asian and Pacific Islander" population to other decennial censuses requires caution. For "American Indian", the 1970 Decennial Census did not survey American Indian population specifically. Except for 1970 Decennial Census, other censuses surveyed American Indian population. As per Lujan (1990), in 1970 Census, enumerators noted the race for households in the reservations based on observations. Whereas in 1980 Census, households or respondents did the self-selection for race (Lujan 1990). This resulted in such low values for American Indians that the IGT research team did not develop any maps for American Indians for 1970. Users need be cautious of these aspects in the census data for race.

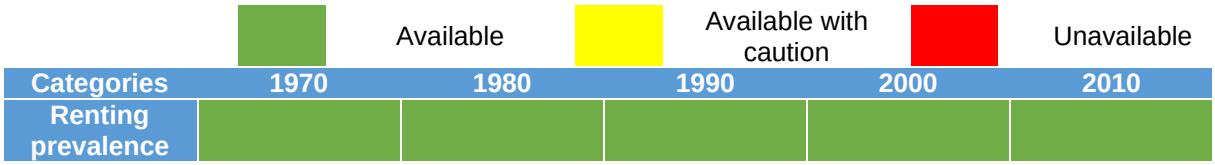
Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Race	1970_Cnt2	1980_STF1	1990_STF1	2000_SF1a	2010_SF1a



15. Renting Prevalence

Except for 1970 Decennial Census, all other decennial censuses provide distinct information for two types of "owner-occupied" and "renter-occupied" housing units. Although, the 1970 Decennial Census provides renting prevalence information in four different categories, they are aggregated to two groups similar to other censuses. For example, sum of "owned or being bought" and "cooperative or condominium units which are owned or being bought" can be considered as "owner occupied". In addition, the sum of "rented for cash rent" and "rented units occupied without payment of cash rent" can be considered as "renter occupied". Therefore, renting prevalence can be calculated for all decennial censuses. Users need be aware of the differences in definitions in the 1970 Census.

Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Renting prevalence	1970_Cnt2	1980_STF1	1990_STF1	2000_SF3a	ACS(2006-2010)



16. Travel Time

Travel time is defined as the time spent on travel to workplace, and it has been surveyed by the U.S. Census Bureau since 1980. Hence, travel time data for 1970 is not available. Although, the 1980 Decennial Census has different travel time intervals, other decennial censuses can be matched to the 1980 intervals by aggregation. In addition, the universes are the same for all decennial censuses. Therefore, travel time data are comparable for 1980, 1990, 2000 and 2010 censuses.

1980 decennial census	Other decennial censuses
Less than 5 min	Less than 5 min
5 – 9 min	5 – 9 min
10 – 14 min	10 – 14 min
15 – 19 min	15 – 19 min
20 – 29 min	20 – 24 min 25 – 29 min

30 – 44 min	30 – 34 min 35 – 39 min 40 – 44 min
45 – 59 min	45 – 59 min
60 or more min	60 – 89 min 90 or more min

Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Travel Time	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)

Available Available with caution Unavailable

Categories	1970	1980	1990	2000	2010
Less than 5 min					
5 – 9 min					
10 – 14 min					
15 – 19 min					
20 – 29 min					
30 – 44 min					
45 – 59 min					
60 or more min					

17. Unemployment Rate

The unemployment rate is defined as number of unemployed persons in the labor force. Although, there are differences in age classifications for labor force, all of the decennial censuses provide numbers of employed and unemployed individuals in the labor force. Therefore, the unemployment rates are comparable through all decennial censuses.

Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Unemployment rate	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF4	ACS(2006-2010)

Available Available with caution Unavailable

Categories	1970	1980	1990	2000	2010
Unemployment rate					

Bibliography

Brown University. *Census geography: Bridging data for census tracts across time.* n.d. https://s4.ad.brown.edu/projects/diversity/Researcher/Bridging.htm (accessed 10 16, 2016).

Cornelius, Craig, and Peter Tatian. *Appendix J: Description of Tract Remapping Methodology.* Documentation, Somerville: Geolytics, 2010.

ESRI. *ESRI Shapefile Technical Description.* Technical documentation, Redlands: Environmental Systems Research Institute, Inc., 1998.

Geolytics. *History of the Neighborhood Change Database (NCDB).* n.d. http://www.geolytics.com/USCensus,Neighborhood-Change-Database-1970-2000,Data,History,Products.asp (accessed 10 16, 2016).

Goodchild, Michael F, Luc Anselin, and U Deichmann. "A Framework for the Areal Interpolation of Socioeconomic Data." *Environment and Planning A*, 1993: 383-397.

Lloyd, Christopher. "Modifiable Areal Unit Problem." In *Exploring Spatial Scale in Geography*, by Christopher Lloyd, 29-44. John Wiley & Sons Ltd., 2014.

Logan, John R., Brian Stults, and Zengwang Xu. "Validating Population Estimates for Harmonized Census Tract Data, 2000-2010." *Annals of the American Association of Geographers* (Taylor & Francis) 106, no. 5 (2016): 1013-1029.

Logan, John R., Zengwang Xu, and Brian Stults. "Interpolating U.S. Decennial Census Tracts from as early as 1970-2000: A Longtudinal Tract Database." *The Professional Geographer*, 2014: 412-420.

Lujan, Carol. *As Simple As One, Two, Three: Census Underdocumentation Among the American Indians and Alaska Natives.* Research Group Staff Working Paper, U.S. Census Bureau, 1990.

Martin, David. *Geographic Information Systems: Socioeconomic Applications.* London: Routledge, 1996.

Nagle, Nicholas, Barbara Buttenfield, Stefan Leyk, and Seth Speilman. "Dasymetric Modeling and Uncertainty." *Annals of Association of American Geographers*, 2014: 80-95.

Noble, Petra, David Van Riper, Steven Ruggles, Jonathan Schroeder, and Monty Hindman. "Harmonizing Disparate Data across Time and Place: The Integrated Spatio-Temporal Aggregate Data Series." *Historical Methods* (Routledge) 44, no. 2 (2011): 79-85.

Openshaw, S. *The Modifiable Areal Unit Problem.* Concepts and Techniques in Modern Geography, Norwich: Geo Books, 1983.

U.S. Bureau of the Census. *Geographic Areas Reference Manual.* Manual, Department of Commerce, 1994.

U.S. Census Bureau. *Census Tracts in American Cities.* U.S. Census Bureau, 1930.

Xie, Zhixiao. "A Framework for Interpolating the Population Surface at the Residential-Housing-Unit Level." *GIScience & Remote Sensing*, 2006: 233-251.

Appendix 1: Census Variable Data Source: NHGIS[7] (National Historical Geographic Information System)

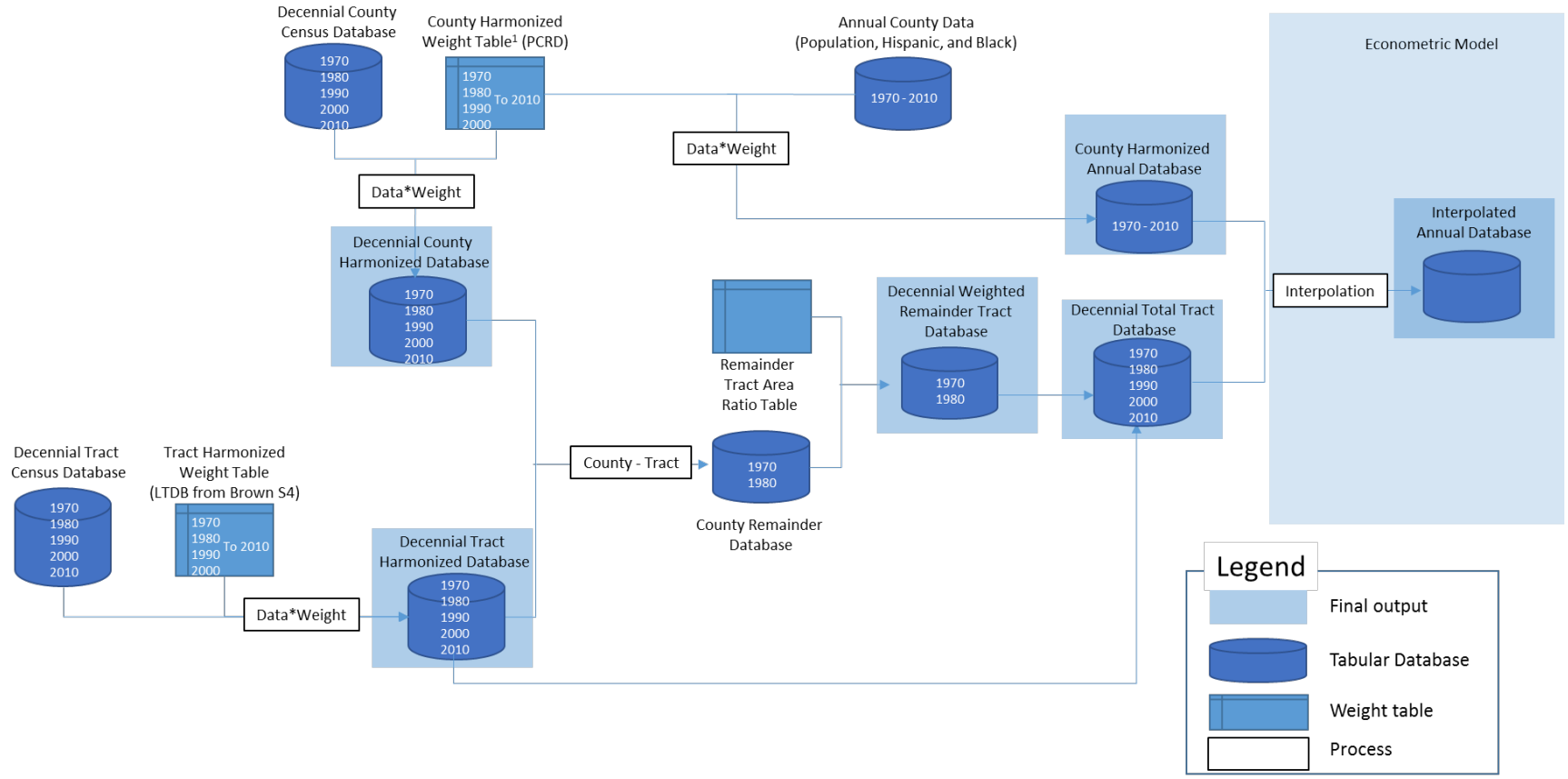
Categories	1970 dataset	1980 dataset	1990 dataset	2000 dataset	2010 dataset
Driving	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)
Dropout prevalence	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)
Educational attainment	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF4, 2000_SF3a*	ACS(2006-2010)
Ethnicity	1970_Cnt4Pb	1980_STF1	1990_STF1	2000_SF1a	2010_SF1a
Family structure	1970_Cnt4Pb	1980_STF3	1990_STF1	2000_SF1a	2010_SF1a
Housing Value	1970_Cnt2	1980_STF1	1990_STF1	2000_SF3a	ACS(2006-2010)
Housing Age	1970_Cnt4H	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)
Industry	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)
Income	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)
Income by race	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)
Occupation	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF4	ACS(2006-2010)
Population	1970_Cnt4H	1980_STF1	1990_STF1	2000_SF1a	2010_SF1a
Public assistance	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)
Poverty	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)
Race	1970_Cnt2	1980_STF1	1990_STF1	2000_SF1a	2010_SF1a
Renting Prevalence	1970_Cnt2	1980_STF1	1990_STF1	2000_SF3a	ACS(2006-2010)
Travel Time	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF3a	ACS(2006-2010)
Unemployment rate	1970_Cnt4Pb	1980_STF3	1990_STF3	2000_SF4	ACS(2006-2010)

[7] https://www.nhgis.org/

Appendix 2: Comparison of LTDB and NCDB

Criteria	LTDB	NCDB
Crosswalk (weight matrix or bridge file)	Available free-of-cost	Cost ($)
Crosswalk-weight calculation	2000 to 2010 based on a combination of population and area. 2000 to 2010 uses blocks to aggregate up to tracts	Similar procedure
	1970, 1980 and 1990 to 2010 based on areal weighting	1970-1980 Census Bureau's tract correspondence file for earlier version. 1980-1990 based on areal weighting. 1990-2000 based on areal and street weighting
	Ancillary water layer is used to identify locations with no land area	Ancillary Tiger street data was used to develop weights from 1990-2000
Documentation	All details available free-of-cost	Some documentation including user's guide, appendices and variables listing are available free-of-cost
Variables	Any variable from census can be harmonized using weights table including other sources	A large number of harmonized variables is available. NCDB provides variables and not the weight table
Year of introduction	2014	Early 2000s
Citation # by Google Scholar	114 (82 citations since 2015)	823 (173 citations since 2015)
Provided by	Brown University	Geolytics

Appendix 3: Tract, County, and Remainder of County Harmonized Database



[1] Weight table is calculated based on NHGIS decennial county boundary

Data and crosswalk sources:

a) National Historical Geographic Information System, Minnesota Population Center
b) Spatial Structures in the Social Sciences, Brown University
c) Purdue Center for Regional Development, Purdue University
d) Department of Agricultural Economics, Purdue University