**Protocol and Data Documentation: Commuting Zones Crosswalk**

Document initially created on 12/10/2019, most recent edit/addition on: 12/10/2019

Documentation author: Heather Rollins

**Contents of Document:**

1. Document Overview
2. Data Objectives
3. Raw data sources and data availability
   1. Crosswalk data
   2. State and county FIPS code and name data
4. Methods
   1. Data downloads
   2. Loading in the Data to R, selecting columns
      1. Crosswalk Data
      2. State and County FIPS Codes Data
   3. Cleaning County FIPS Codes data
   4. Building the Clean Crosswalk
   5. Additional counties
   6. Setting the CZ name
5. Final derived datasets
6. Quality Issues and comments

**Part 1: Document Overview**

This document is associated with the code and clean data sets in the folder:

\\drexel\departments\SOPH\Shared\UHC\Projects\Bilal\_DP5\Data\Crosswalks\US\

The specific clean data files in this folder associated with this document are:

…\Clean\county\_FIPS\_CZ\_crosswalk.csv

…\Clean\county\_FIPS\_codes\_names\_2016.csv

…\Clean\state\_FIPS\_codes\_names\_2016.csv

The R code file associated with the creation of the crosswalks is:

\\drexel\departments\SOPH\Shared\UHC\Projects\Bilal\_DP5\Data\Crosswalks\US\Data\cz\_fips\_cw\_code.R

The purpose of this document is to:

* Explain and fully document the process of obtaining, cleaning, and manipulating data which was done to produce the final clean data sets
* Describe the final clean data sets
* Discuss any issues encountered with the data and/or data cleaning process

Important abbreviations/acronyms in this document:

|  |  |
| --- | --- |
| Acronym | Full name |
| CZ | Commuting Zone |
| CBSA | Core-Based Statistical Area |
| FIPS | Federal Information Processing Standards |
| CW | Crosswalk |

**Part 2: Data Objectives**

The goal was to create a clean crosswalk of county FIPS code to Commuting Zone (CZ) using the 2010 CZ definitions. This process involved cleaning and simplifying the 2010 county-CZ-CBSA crosswalk already available. Additionally, state and county FIPS codes and names files were cleaned and saved, and these files were used to also add state and county names to the FIPS-to-CZ crosswalk.

**Part 3: Raw Data Sources & Data Availability**

**Crosswalk Data**

The County-to-CZ crosswalk is available at the website: <https://sites.psu.edu/psucz/data/>

The link used to download the crosswalk is: <https://sites.psu.edu/psucz/files/2018/09/counties10-zqvz0r.csv>

The data is available as a CSV file. The data has the following columns:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Description | Variable type | Values |
| FIPS | 5-digit combined state and county FIPS code | Number | 10001, 10003, 10005, … |
| OUT10 | CZ code assigned to this county | Number | 1 – 625 |
| Pop10 | Population count of this county | Number | 82 – 9818605 |
| CBSA10 | CBSA/MMSA code | Number | 10100 – 49820 |
| CBSAName | Name assigned to the CBSA code | Character | Example: “Aberdeen, SD” … |
| …(others ignored) |  |  |  |

**State and County FIPS Code and Name Data**

State and county FIPS code and name data are available from the following website:

<https://www.census.gov/geographies/reference-files/2016/demo/popest/2016-fips.html>

The link used to download the county FIPS code and name data is:

<https://www2.census.gov/programs-surveys/popest/geographies/2016/all-geocodes-v2016.xlsx>

The link used to download the state FIPS code and name data is:

<https://www2.census.gov/programs-surveys/popest/geographies/2016/state-geocodes-v2016.xls>

The county FIPS data is available as an excel file. Rows 1-3 rows are a header, row 4 is blank, and row 5 has column names. The remainder of the rows, starting with row 6, has the data of interest.

The county FIPS data has the following columns:

| Column Name | Description | Variable type | Values |
| --- | --- | --- | --- |
| Summary Level | 3-digit code indicating the level of information in the row. | Numeric value as text (character) | 010, 040, 050, 061, …  Leading zeros added to value. |
| State Code (FIPS) | 2-digit state FIPS code | Numeric value as text  (character) | 00, 01, 02, …  Leading zeros added to value. |
| County Code (FIPS) | 3-digit county FIPS code  *(county, borough, census area, municipality, parish)*  Unique set of codes per state.  County codes can be duplicated for county subdivisions, places, and consolidated cities. | Numeric value as text  (character) | 000, 001, 003, …  Leading zeros added to value. |
| County Subdivision Code (FIPS) | 5-digit county subdivision FIPS code  *(city, town, township, village, borough, municipality, reservation, location, grant, purchase, etc.)*  Unique set of codes per state/county | Numeric value as text  (character) | 00000, 00070, 00100, 00108, 00116, …  Leading zeros added to value. |
| Place code (FIPS) | 5-digit place FIPS code  *(city, town, village, borough, municipality, metro/ metropolitan/ consolidated/ unified government, etc.)*  Unique set of codes per state/county | Numeric value as text  (character) | 00000, 00065,00100, 00113, 00116, …  Leading zeros added to value. |
| Consolidated City Code (FIPS) | 5-digit consolidated city FIPS code.  *(city, metro/ metropolitan/ consolidated/ unified government, etc.)* | Numeric value as text  (character) | 00000, 00065, 00100, 00113, 00116, …  Leading zeros added to value. |
| Area Name (including legal/statistical area description) | Name of the county, or other area, if the row does not specify a unique county | Character/String | Examples:  “United States” “Alabama” “Autauga County” “Abbeville city” “Addison town” “Aleutians East Borough” “Aleutians West Census Area” “Anchorage Municipality” |

The state FIPS data is available as an excel file. Rows 1-4 rows are a header, row 5 is blank, and row 6 has column names. The remainder of the rows, starting with row 7, has the data of interest.

There are 4 “header” type rows within the data for the region of the US. Under these 4 headers, there are a further 9 header rows for the division of the region. Under these divisions, there are rows for the states.

The county FIPS data has the following columns:

| Column Name | Description | Variable type | Values |
| --- | --- | --- | --- |
| Region | 1-digit code indicating the region of the USA | Numeric value as text (character) | 1 = Northeast Region (incl. divisions 1-2)  2 = Midwest Region (incl. divisions 3-4)  3 = South Region (incl. divisions 5-7)  4 = West Region (incl. divisions 8-9) |
| Division | 1-digit number indicating the division, which is a group of states within a region.  The header row for a division and the states within that division all have the same division value | Numeric value as text (character) | 0 = header rows for regions  1 = New England Division  2 = Middle Atlantic Division  3 = East North Central Division  4 = West North Central Division  5 = South Atlantic Division  6 = East South Central Division  7 = West South Central Division  8 = Mountain Division  9 = Pacific Division |
| State Code (FIPS) | 2-digit state FIPS code | Numeric value as text  (character) | 00, 01, 02, …  00 = row for a region or a division  01-56 = state FIPS codes  Leading zeros added to value. |
| Name | Name of the state, region, or division | Character/String | Examples:  “Northeast Region” “New England Division” “Connecticut” Maine”, … |

**Part 4: Methods**

The code written to download and clean the data is saved in the R code file:

…\Bilal\_DP5\Data\Crosswalks\US\Data\cz\_fips\_cw\_code.R

**Data Downloads**

All data was downloaded using the R code in the file cz\_fips\_cw\_code.R. The data was downloaded from the links specified in *Part 3: Raw Data Sources & Data Availability*. This was done using the R function:

download.file(url, destfile).

The data was saved directly to the folder path:

[\\drexel\departments\SOPH\Shared\UHC\Projects\Bilal\_DP5\Data\Crosswalks\US\Data\](file:///\\drexel\departments\SOPH\Shared\UHC\Projects\Bilal_DP5\Data\Crosswalks\US\Data\)

The files of the original data are:

CZ crosswalk: …\Bilal\_DP5\Data\Crosswalks\US\Data\psu\_edu\_cz\_county\_cw\_2010.csv

County FIPS codes: …\Bilal\_DP5\Data\Crosswalks\US\Data\ censusgov\_state\_geocodes\_2016.xls

State FIPS codes: …\Bilal\_DP5\Data\Crosswalks\US\Data\censusgov\_county\_geocodes\_2016.xlsx

**Loading in the Data to R, selecting columns**

After saving, the data was read into R.

**Crosswalk Data**

The crosswalk existed as a CSV file, which can be easily read into R using:

read.csv(file, stringsAsFactors = F)

Using the package data.table, the csv was converted to a data.table. By doing this, renaming the columns of a data.table becomes very easy, using the function:

setnames(x, old, new)

Using this function, the variables “OUT10” and “FIPS” were renamed to “cz” and “fips”, respectively.

Then the data was subset to only those two columns (for CZ code and county FIPS code), plus the column for the 2010 population (“Pop10”).

**State and County FIPS Codes Data**

Since the state and county FIPS codes files are excel files (.xls and .xlsx respectively), we can use the package readxl and the function read\_xls() to read in the files. In the function, we can specify the column variable types (“text”) and the number of rows to skip (5 rows for state data, 4 rows for county data).

The state and county data were merged into one data set by the "State.Code..FIPS." column in the county data and the "State..FIPS." column in the state data. A new variable “fips” was created by concatenating the 2-digit state code and the 3-digit county code. The data was subset to the columns for state code, county code, 5-digit FIPS, the area name (from the county data) and the name from the state data. Columns were also renamed.

**Cleaning County FIPS Codes data**

Rows where either the state code was “00” or the county code was “000” were removed.

The data contained rows with duplicated county FIPS codes. We needed to eliminate rows where the county was duplicated and the name specified an area other than a county (e.g. city, town, village,…).

For each state, we compared the number of data rows, the number of unique 5-digit county FIPS codes, and the number of unique county codes where the county names contained the word “county”. The states were found to have count mismatches are:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| State FIPS | State Name | N data rows | N unique County FIPS | N unique county FIPS with “County” in name | T/F: N row matches N unique FIPS? | T/F: N unique FIPS matches N Unique FIPS with “County” in name? |
| 09 | Connecticut | 177 | 8 | 8 | FALSE | TRUE |
| 17 | Illinois | 1536 | 102 | 102 | FALSE | TRUE |
| 18 | Indiana | 1102 | 92 | 92 | FALSE | TRUE |
| 20 | Kansas | 1581 | 105 | 105 | FALSE | TRUE |
| 23 | Maine | 549 | 16 | 16 | FALSE | TRUE |
| 25 | Massachusetts | 365 | 14 | 14 | FALSE | TRUE |
| 26 | Michigan | 1621 | 83 | 83 | FALSE | TRUE |
| 27 | Minnesota | 2761 | 87 | 87 | FALSE | TRUE |
| 29 | Missouri | 427 | 115 | 114 | FALSE | FALSE |
| 31 | Nebraska | 533 | 93 | 93 | FALSE | TRUE |
| 33 | New Hampshire | 269 | 10 | 10 | FALSE | TRUE |
| 34 | New Jersey | 586 | 21 | 21 | FALSE | TRUE |
| 36 | New York | 1071 | 62 | 62 | FALSE | TRUE |
| 38 | North Dakota | 1783 | 53 | 53 | FALSE | TRUE |
| 39 | Ohio | 1689 | 88 | 88 | FALSE | TRUE |
| 42 | Pennsylvania | 2639 | 67 | 67 | FALSE | TRUE |
| 44 | Rhode Island | 44 | 5 | 5 | FALSE | TRUE |
| 46 | South Dakota | 1328 | 66 | 66 | FALSE | TRUE |
| 50 | Vermont | 269 | 14 | 14 | FALSE | TRUE |
| 55 | Wisconsin | 1986 | 72 | 72 | FALSE | TRUE |

Missouri had one county that did not have any rows with the word “county” in the area name. This data was for St. Louis City Missouri (FIPS 29510). There was only row for this FIPS code. There was also a county named St. Louis County Missouri (FIPS 29189) in the data. Both FIPS codes appear in the County-to-CZ crosswalk, so it was decided that it was important to not remove St. Louis City Missouri (FIPS 29510) from the data.

Besides this one data row (St. Louis City Missouri, FIPS 29510), for these 20 states with duplicates, rows not containing the word “county” in the name would be removed.

This cleaned set of county and state FIPS codes and names was saved. Additionally, a data set of the unique state FIPS codes and names was extracted from this data set and then saved.

**Building the Clean Crosswalk**

A new variable for state FIPS code was added to the crosswalk by dividing the 5-digit FIPS code by 1000 and rounding down using the R function floor().

In all 3 data sets (CZ crosswalk, state FIPS codes & names data, county FIPS codes & names data), all numeric-type variables were converted from character to numeric.

After determining that the state FIPS code 72 (Puerto Rico) was not present in the CZ crosswalk data, this state code was removed from both the state FIPS codes & name and the county FIPS codes and names data sets.

The data set with state FIPS code and state name was merged first to the FIPS-to-CZ crosswalk (matched by the 2-digit state FIPS code) to be able to have at least a state name for all rows. This was done in case a specific row in the CZ-to-FIPS crosswalk had no match in the county FIPS codes & names data set, which is a situation that would leave us with no county name information. Then, the state FIPS code & names data was merged to the crosswalk, matched by the 5-digit county FIPS code.

**Additional Counties**

From work done previously to clean and manipulate the IRS migration data sets, we knew that there were many counties that shifted boundaries and/or changed names/FIPS codes over time.

For more information on county changes, please see the following documents:

\\drexel\departments\SOPH\Shared\UHC\Projects\Bilal\_DP5\Data\Environment\US\Migration\Clean\IRS Migration Protocol and Data Documentation.docx

\\drexel\departments\SOPH\Shared\UHC\Projects\Bilal\_DP5\Data\Environment\US\Migration\Documentation\Cross References \_ Federal Communications Commission.pdf

\\drexel\departments\SOPH\Shared\UHC\Projects\Bilal\_DP5\Data\Environment\US\Migration\Documentation\county\_geography-\_changes2015.pdf

\\drexel\departments\SOPH\Shared\UHC\Projects\Bilal\_DP5\Data\Environment\US\Migration\Documentation\FIPS\_County\_Code\_Changes.pdf

The county changes to be mindful of are summarized below.

* 12025 & 12086
  + (1997) Dade County, FL 12025 renamed to Miami-Dade, FL 12086
* 51780 & 51083
  + (1995) South Boston, VA 51780 merges into Halifax County, VA 51083
* 51560 & 51005
  + (2001) Clifton Forge, VA 51560 merges into Alleghany, VA 51005
* 51515 & 51019
  + (2013) Bedford City, VA 51515 merges into Bedford County, VA 51019
* 46113 & 46102
  + (2015) Shannon County, SD 46113 renamed to Oglala Lakota County, SD 46102
* 02158 & 02270
  + (2015) Wade Hampton Census Area, AK 02270 renamed to Kusilvak Census Area, AK 02158
* Other Alaska Census areas/borough which experience changes:
  + (...-1992): 02231
  + (1992-...): 02282
  + (1992-2007): 02232
  + (...-2008): 02280 02201
  + (2007-...): 02230 02105
  + (2008-...): 02275 02198
  + Exists all years: 02130

The issues in the data and the manual edits performed in the data for each county are summarized in the table below:

| County FIPS & name | CZ | Row in data | Name in data | CZ in data | Get CZ from | Edits required? (Summary) |
| --- | --- | --- | --- | --- | --- | --- |
| 12025, Dade County, FL | 98 | **NO** | No | No | 12086 | New row with all info |
| 12086, Miami-Dade County, FL | 98 | Yes | Yes | Yes | - | None |
| 51780, South Boston, VA | 589 | **NO** | No | No | 51083 | New row with all info |
| 51083, Halifax County, VA | 589 | Yes | Yes | Yes | - | None |
| 51560, Clifton Forge, VA | 583 | **NO** | No | No | 51005 | New row with all info |
| 51005, Alleghany, VA | 583 | Yes | Yes | Yes | - | None |
| 51515, Bedford City, VA | 585 | Yes | No | Yes | - | County name/info |
| 51019, Bedford County, VA | 585 | Yes | Yes | Yes | - | None |
| 46113, Shannon County, SD | 377 | Yes | No | Yes | - | County name/info |
| 46102, Oglala Lakota County, SD | 377 | Yes | Yes | No | 46113 | Add CZ, match 46113 |
| 02270, Wade Hampton Census Area, AK | 30 | Yes | No | Yes | - | County name/info |
| 02158, Kusilvak Census Area, AK | 30 | Yes | Yes | No | 02270 | Add CZ, match 02270 |
| 02231, Skagway-Yakutat-Angoon Census Area, AK | None | **NO** | No | No | - | New row with all info |
| 02282, Yakutat Borough, AK | 23 | Yes | Yes | Yes | - | None |
| 02232, Skagway-Hoonah-Angoon Census Area, AK | None | **NO** | No | No | - | New row with all info |
| 02280, Wrangell-Petersburg Census Area, AK | 28 | **NO** | No | No | - | New row with all info |
| 02201, Prince of Wales-Outer Ketchikan Census Area, AK | 24 | **NO** | No | No | - | New row with all info |
| 02230, Skagway Municipality, AK | 18 | Yes | Yes | Yes | - | None |
| 02105, Hoonah-Angoon Census Area, AK | 23 | Yes | Yes | Yes | - | None |
| 02275, Wrangell Borough, AK | 28 | Yes | Yes | Yes | - | None |
| 02198, Prince of Wales-Hyder Census Area, AK | 24 | Yes | Yes | Yes | - | None |
| 02130, Ketchikan Gateway Borough, AK | 24 | Yes | Yes | Yes | - | None |

**Setting the CZ Name**

Since the 2010 CZs no not have names, we chose to name each CZ by the county with the largest population within each CZ.

For example, CZ 1 contains 9 counties, and the county with the largest 2010 population with CZ is 01101 Montgomery County, AL, with a population of 229363. Therefore, CZ 1 was assigned the name Montgomery County, Alabama.

**Part 5: Final datasets**

The final CZ-to-county FIPS clean data set is saved at the following location:

\\drexel\departments\SOPH\Shared\UHC\Projects\Bilal\_DP5\Data\Crosswalks\US\Clean\county\_FIPS\_CZ\_crosswalk.csv

This data contains columns for: CZ code, 5-digit county FIPS code, 3-digit county FIPS code, county name, 2-digit state FIPS code, state name, 2010 population, and CZ name. The data includes rows for every county which existed between 1990 and 2019.

Additionally, basic data sets for state name & 2-digit FIPS code and county name & FIPS code (both for 2016) were also saved at the following locations:

\\drexel\departments\SOPH\Shared\UHC\Projects\Bilal\_DP5\Data\Crosswalks\US\Clean\state\_FIPS\_codes\_names\_2016.csv

\\drexel\departments\SOPH\Shared\UHC\Projects\Bilal\_DP5\Data\Crosswalks\US\Clean\county\_FIPS\_codes\_names\_2016.csv

The state FIPS/name data set contains columns for: state 2-digit FIPS code, state name.

The county FIPS/name data set contains columns for: state 2-digit FIPS code, county 3-digit FIPS code, 5-digit FIPS code (combined state and county), county name, state name.

**Part 6: Quality Issues and comments**

There are several counties in the crosswalk data which did not exist in 2010. All counties which existed between 1990 and 2019 were included in the crosswalk. Some of these counties were added manually according to existing information.

Counties which did not have a 2010 CZ assignment (because they did not exist in 2010) were assigned to the CZ which would be the best match, according to the CZ assignment of 2010 counties which overlapped the same geographic area of the non-2010 county.

There are two counties with no matching CZ, because they each overlap with multiple counties with disagreeing CZ assignments. These counties are 02232 Skagway-Hoonah-Angoon Census Area and 02231 Skagway-Yakutat-Angoon Census Area.