New Area- and Population-based Geographic Crosswalks for U.S. Counties and Congressional Districts, 1790-2020

Readme File

Andreas Ferrara[1], Patrick A. Testa[2], and Liyang Zhou[3]

October 2022

Crosswalks

Contains the area- and population-based crosswalks as .csv files.

- **CountyToCounty**: contains the crosswalks to harmonize county boundaries over time relative to a chosen base year, with area-based weights (M1) as well as five population-based weights (M2–M6). Each subdirectory (1790 to 2020) has the crosswalks for the year in the folder name. Inside the directory, the first year is the year you want to harmonize and the second year is the base year (e.g. crosswalk_1830_1790.csv).

- **CountyToCD-xxx**: contains the six crosswalk weights for mapping counties to the congressional district level. There are three different directories that provide crosswalks based on different treatment of the time dimension:

  - **DOA**: the Census year of apportionment associated with a given Congress (e.g. counties from the 1800 Census are associated with the CD boundaries from the Congresses between 1803 to 1812).

  - **DOC**: the Census decade shared with the starting year of a given Congress (e.g. counties from the 1800 Census are associated with the CD boundaries from the Congresses between 1801 to 1810).

  - **Nearest**: the nearest Census year associated with the starting year of a given Congress (e.g. counties from the 1800 Census are associated with CD boundaries from the Congresses between 1795 to 1804).

---

[1] University of Pittsburgh, Department of Economics. Email: a.ferrara@pitt.edu

[2] Tulane University, Department of Economics. Email: ptesta@tulane.edu

[3] University of Pittsburgh, Department of Economics. Email: LIZ113@pitt.edu

- **Coordinates**: contains latitude and longitude of county and CD centroids to facilitate the estimation of spatial autocorrelation corrected standard errors, distances, etc.

Weighting models

Crosswalk files include six weighting models:

- **M1**: an area-based model, equivalent in construction to existing area-based crosswalks.

- **M2**: a population-based model, with county area divided into urban and rural areas, based on historical population estimates from Fang and Jawitz (2018).

- **M3**: a population-based model, with county area divided into urban and rural areas after excluding non-inhabitable areas, based on historical population estimates from Fang and Jawitz (2018).

- **M4**: a population-based model, with county area divided into urban and rural areas after excluding non-inhabitable areas, with additional weighting for topographic suitability, based on historical population estimates from Fang and Jawitz (2018).

- **M5**: a population-based model, with built-up settlement areas indicated in space (1810-2020 only), based on Leyk et al (2020).

- **M6**: a population-based model, with built-up property counts indicated in space (1810-2020 only), based on Leyk et al (2020).

Teaching material

Contains a set of basic Power Point (.ppt) and Latex (.tex) files to incorporate in classes that wish to teach students how to crosswalk county-level data over time or to different congressional districts using the crosswalks produced by Ferrara, Testa, and Zhou (2021). The relevant figures are included as .png files and a compiled version of the .tex file is included as .pdf file.

Replication-files

Contains the zip file to replicate results from Lee, Moretti, and Butler (2004) with both area- and population-based crosswalks as in Ferrara et al. (2021). The zip file contains

- **Code**

  - gen_chars.do – produces the Lee et al. (2004) data using the different types of crosswalks

- tables.do – produces the balancing tests in Lee et al. (2004) using the different types of crosswalks

- generate_model_correlation_figure_LMB.do – produces the correlation figure between the crosswalked county-to-CD data (using the different crosswalks) and the official CD data by the Census used by Lee et al. (2004). This is Figure 3 in Ferrara et al. (2021).

- **Data**

  - County: this folder contains data from the Census as well as the County and City Data Books used in Lee et al. (2004).

  - Generated: this folder contains the generated CD-level data using the crosswalks provided by Ferrara et al. (2021) as well as the generated data to produce Table 2 in Lee et al. (2004).

Cite

Ferrara, A., Testa, P.A., and Zhou, L. (2021) "New Area- and Population-based Geographic Crosswalks for U.S. Counties and Congressional Districts, 1790-2020", CAGE Working Paper No. 588 [link. Also available at SSRN]

Version history:

V4: added models M5 and M6 to each county-to-county and county-to-congressional district crosswalk file (for census years 1810−2020).

V3: i) added county- and CD-centroid latitude and longitude data files to facilitate computation of spatial standard errors, distance metrics, etc., ii) re-packaged data files into zip files to adhere to OpenICPSR file upload limits, iii) all crosswalks are now provided as both .csv and Stata files.

V2: updated teaching materials to incorporate the reference to the CAGE working paper.