Changing boundaries of geographic units

Applied work often uses data with a spatial component

- counties
- commuting zones
- congressional districts
- ...

Either you want to

- have consistent spatial units over time (e.g. changing county boundaries across Census years)
- combine data from different spatial units (e.g. construct variables for congressional districts using county-level information)

Area-based harmonization methods

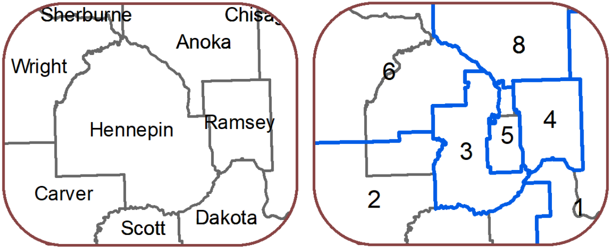
- Suppose you want to get county information (e.g. total income) for congressional districts (CD); use stock variables, not shares, you can build shares and percentages **after** the boundary harmonization

Area-based harmonization methods

- Suppose you want to get county information (e.g. total income) for congressional districts (CD); use stock variables, not shares, you can build shares and percentages **after** the boundary harmonization
- Hornbeck (2010) pioneered the following procedure
  1. Using ArcGIS (or other mapping software), intersect the CD boundaries with all counties or parts of counties which fall into a given CD
  2. Take total income in each county and sum it within the CD
     - counties solely lying in that CD receive a weight of 1
     - counties that partially lie in that CD receive a weight of $\frac{a}{A}$, where a is the area of the county that lies in the CD and A is the total area of the county

Area-based harmonization methods

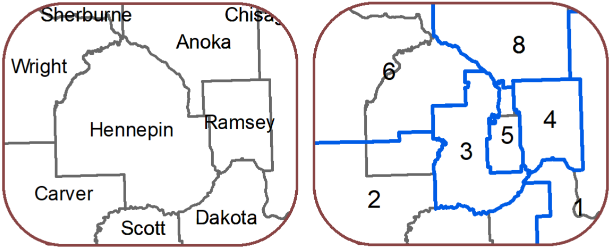
- Suppose you want to get county information (e.g. total income) for congressional districts (CD); use stock variables, not shares, you can build shares and percentages **after** the boundary harmonization
- Hornbeck (2010) pioneered the following procedure
  1. Using ArcGIS (or other mapping software), intersect the CD boundaries with all counties or parts of counties which fall into a given CD
  2. Take total income in each county and sum it within the CD
     - counties solely lying in that CD receive a weight of 1
     - counties that partially lie in that CD receive a weight of $\frac{a}{A}$, where a is the area of the county that lies in the CD and A is the total area of the county
- The last step assumes that total income is uniformly distributed across the county area
  - this will not hold in the presence of **urbanization/agglomeration** economies to the extent that origin and reference unit boundaries intersect



a) county boundaries

b) congressional districts

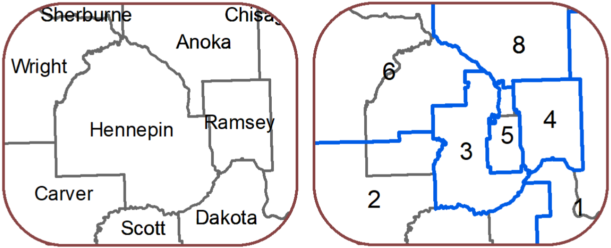
- CD3: $\frac{\text{Area}_{\text{CD3}}}{\text{Area}_{\text{Hennepin}}} \times (\text{total income}_{\text{Hennepin}})$ since CD3 is entirely in Hennepin county



a) county boundaries

b) congressional districts

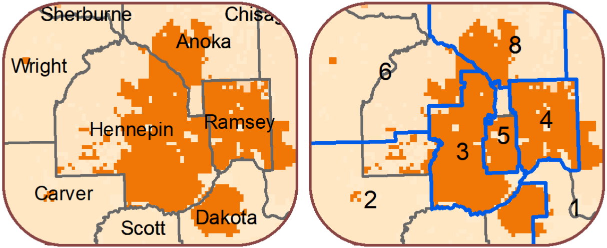
- CD3: $\frac{\text{Area}_{\text{CD3}}}{\text{Area}_{\text{Hennepin}}} \times (\text{total income}_{\text{Hennepin}})$ since CD3 is entirely in Hennepin county
- CD5: $\frac{\text{Area}_{\text{CD5 in Hennepin}}}{\text{Area}_{\text{Hennepin}}} \times (\text{total income}_{\text{Hennepin}}) + \frac{\text{Area}_{\text{CD5 in Anoka}}}{\text{Area}_{\text{Anoka}}} \times (\text{total income}_{\text{Anoka}})$, since CD5 covers parts of both Hennepin and Anoka



a) county boundaries

b) congressional districts

- CD3: $\frac{\text{Area}_{\text{CD3}}}{\text{Area}_{\text{Hennepin}}} \times (\text{total income}_{\text{Hennepin}})$ since CD3 is entirely in Hennepin county
- CD5: $\frac{\text{Area}_{\text{CD5 in Hennepin}}}{\text{Area}_{\text{Hennepin}}} \times (\text{total income}_{\text{Hennepin}}) + \frac{\text{Area}_{\text{CD5 in Anoka}}}{\text{Area}_{\text{Anoka}}} \times (\text{total income}_{\text{Anoka}})$, since CD5 covers parts of both Hennepin and Anoka
- Total income in CD 2/4/6/8 are the weighted sums of total income from counties and county parts belonging to these CDs.



a) county boundaries

b) congressional districts

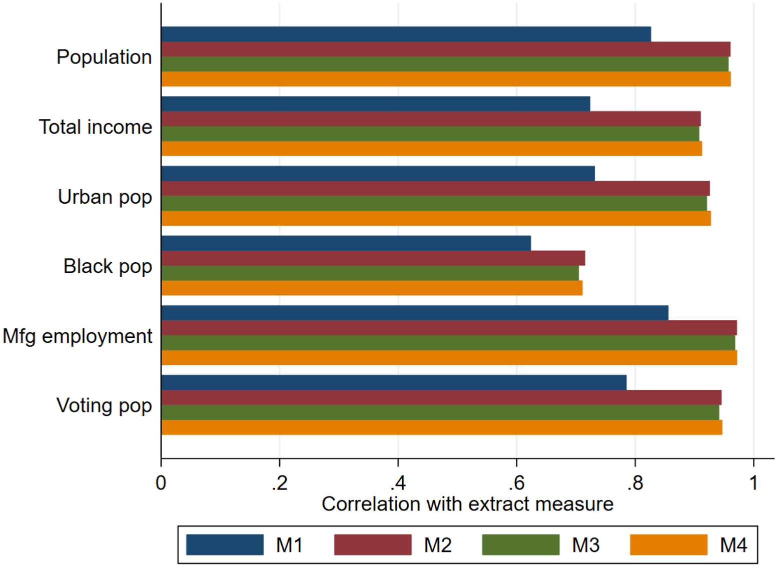
- But new population data by Fang and Jawitz (2018) shows that population is not uniformly distributed within counties
  - they develop different spatial models for 1×1km grid cell population distributions from 1790 to 2010
  - yellow and orange are approximately above/below mid-range population (for simplicity)

New population-based crosswalks

- Using this new population data, Ferrara, Testa, and Zhou (2021) provide a complementary approach to Hornbeck (2010) with new crosswalks for
  - counties-to-counties over time (Census decades)
  - Counties-to-congressional districts
- from 1790 to 2020 using population-based weights instead of area- based weights. The crosswalks can be downloaded at https://doi.org/10.3886/E150101
- How much does adjusting for population distribution matter?

Comparison with actual CD-level data

- Ferrara, Testa and Zhou (2021) take official Census data for CDs which exist for 1960, 1970, 1980 and 1980
- They generate CD-level variables from county-level data using
  - area-based weights (model M1)
  - population-based weights (models M2-4)
  - (slight differences between M2-4 in how the sub-county population distribution is generated; in practice it makes little difference)
- And then they compare how closely the different crosswalks can replicate the official data (next slide)
  - relative to M1, M2-4 have a **20% higher correlation** with the official CD data


Population
Total income
Urban pop
Black pop
Mfg employment
Voting pop
0
.2
.4
.6
.8
1
Correlation with extract measure
M1
M2
M3
M4

Assumptions and caveats

- Of course, there are still some assumptions and caveats
  - Fang and Jawitz (2018) still rely on **uniformity within defined urban and rural areas** (though this is relaxed further in M4)
  - earlier Census years increasingly rely on mathematical assumptions about urban extents over official Census data from the year 2000; this may not be unreasonable given that city population growth tends to follow certain power distributions (see Chen, 2015), but it is important to pick a reference year that pre-dates any "treatment"
  - population-based approaches **work well for stock variables that are correlated with population**, like income, number of workers, etc. If a variable is inversely correlated with population, say air quality, then population-based weights can be worse than area-based ones
    - Solution: turn variable into something that correlates positively with population (e.g. use air pollution instead of air quality)

How to apply the crosswalks

- Suppose you want to harmonize county level data on the stock of Mexican-born population in 1960 to the 88th CDs
  1. Get the county-level data for 1960 for the total number of persons born in Mexico. It is critical to harmonize only county-level stock variables for weights to be appropriate
  2. Given some set of county identifiers (e.g. FIPS or NHGIS codes), merge the 1960 county file with the 1960 to 88th Congress crosswalk file
  3. Multiply the number of persons born in Mexico with the provided weights
  4. Finally, collapse (i.e. sum) the weighted counts for each variable by CD identifiers. Round or mark as missing any cell as needed. The unit of observation is now the CD level

References

- Chen, Y. "The distance-decay function of geographical gravity model: Power law or exponential law?", Chaos, Solutions & Fractals, Vol. 77, pp. 174-189.
- Fang, Y. and Jawitz, J.W. (2018) "High-resolution reconstruction of the United States human population distribution, 1790 to 2010", Scientific Data, Vol. 5, `https://doi.org/10.1038/sdata.2018.67`.
- Ferrara, A., Testa, P., 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.
- Hornbeck, R. (2010) "Barbed Wire: Property Rights and Agricultural Development", Quarterly Journal of Economics, Vol. 125(2), pp. 767-810.