Calculating Residential Segregation Indices in A Reproducible Pipeline

Boyi Guo

Department of Biostatistics
University of Alabama at Birmingham

Last Updated 2022/02/17

Outline

Outline

- Motivation
- Pipeline Demonstration
- Discussion

Motivation

Why reproducible pipeline?

- Growing number of data requests since the publication of Cummings et al. (2021)
- Up the research reproducibility game
  - Growing emphasis in biomedical science research (Heil et al. 2021), and public health research (Peng and Hicks 2021)
  - See my previous talk *Reproducible Data Analysis Workflow* for easy starts

Pipeline Demonstration

Preparation

- Software & Package Installation
  - Git: https://git-scm.com/book/en/v2/Getting-Started-Installing-Git
  - R (recommend 4.0+, minimial 3.6+) & RStudio: https://www.rstudio.com/products/rstudio/download/
  - R package `renv`: https://rstudio.github.io/renv/index.html

Download Remote GitHub Repository

- Download the remote repository via https://github.com/boyiguo1/Tutorial-Residential_Segregation_Score
  - No GitHub account required
  - Download ZIP, de-compress and open the R project, i.e. `*.Rproj` file
  - [Advanced approach:] Create new project with version control
- Install the R packages with `renv`

```
renv::restore()
```

Set up your census API key

- Acquire your census api key string via https://api.census.gov/data/key_signup.html
- Replace your census API key in `_targets.R`
  - Search the file with the keyword “TODO:”

Run the pipeline

- To run the pipeline `tar_make()`
- To fetch a target object: `tar_load(object)`, e.g. `tar_load(rs_indices)` for the calcualted indices
- Other Utility
  - Pipeline progress or modification since last run

    ```
    tar_visnetwork()
    ```

  - Check *Addins* in the tool bar

There are many other fantastic functions from the R package `targets`. Please see https://books.ropensci.org/targets/walkthrough.htmls.

Switching between examples

- RStudio graphic user interface: View -> Show Git -> Dropdown list. [TODO: insert a screen shot here]
- Command line: `git checkout ChangeToBranchName`

Customization

- Understand the file system
  - `_targets.R`: the master file containing all steps of analysis
    - Similar to a normal R script file except that the assignment of objects follows a new syntax
    - `tar_target(name, command)` translate to `name <- command`
    - Use global search () to find all places needs customization
  - Self-defined functions are located in the folder `R`
    - You can use these functions to write your own pipeline to calculate remaining indcies introduced in Massey and Denton (1988)

Preliminary Findings & Remarks

Recap of Massey and Denton (1988)

- Surveyed 20 indices describing 5 dimensions of residential segregation
- Validated the segregation indices with *US metropolitan areas data* via factor analysis
- Suggested one index for each of the five dimensions

  "This interpretation [that researchers had on the five-dimenional indices as segregation] is an abstraction of empirical reality, not reality itself."

Dimensions of Residentital Segregation

- *Evenness*: spatial distribution of different groups among *units* in a metropolitan area
- *Exposure*: possibility of interaction between minority and majority group members
- Concentration: relative amount of physical space occupied by a minority group in the metropolitan area
- Centralization: how a group spatially located near the center of an urban area
- Clustering: which areal *units* inhabited by minority members adjoin one another, or cluster, in space

Indices Implemented in the Pipeline

- Dissimilarity index for Evenness: the percentage of population would have change residence to have the same percentage overall
  - 0.0 (complete integration) to 1.0 (complete segregation)
- Interaction index for Exposure: probability that a minority person shares a unit area with a majority person
  - 0.0 (complete segregation) to 1.0 (complete integration)
- Isolation index for Exposure: probability that a minority person shares a unit area with a minority person
  - 0.0 (complete integration) to 1.0 (Complete segregation)

Remarks (I)

- How to choose the areal unit?
  "We chose census tracts for the simple reason that more racial and ethnic data are available for them than for other geographic units."

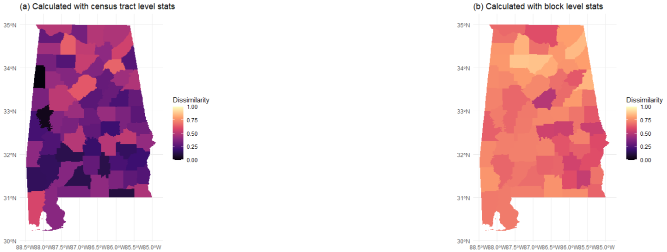
- The indices are not well-defined when the area contain neither majority or minority.

$$\sum_{i=1}^{n}[(\frac{x_i}{X})(\frac{y_i}{t_i})]$$

  - This is more likely to happen within smaller area unit, e.g. at census tract level in Arizona

Remarks (II)

- Measurement Consistency



(a) Calculated with census tract level stats

(b) Calculated with block level stats

Figure 1: 2010 Alabama Dissimilarity Index at county level calculated with census tract level statistics *(a)* and block level statistics *(b)*

References

References I

Heil, Benjamin J, Michael M Hoffman, Florian Markowetz, Su-in Lee, Casey S Greene, and Stephanie C Hicks. 2021. "Reproducibility standards for machine learning in the life sciences." *Nature Methods*, August. https://doi.org/10.1038/s41592-021-01256-7.

Massey, Douglas S, and Nancy A Denton. 1988. "The Dimensions of Residential Segregation." *Social Forces* 67 (2): 281–315.

Peng, Roger D., and Stephanie C. Hicks. 2021. "Reproducible Research: A Retrospective." *Annual Review of Public Health* 42 (1): 79–93. https://doi.org/10.1146/annurev-publhealth-012420-105110.