Reproducible Research

Jared Bennett

22 September, 2019

Reproducible Research

Today, we will do a case study of a publicly-available urban study.
You can find a copy of their paper in this directory.
Their code and documentation can be found here.

Summary of Analysis

Data was collected, processed, and merged using disparate files available from the US Census. In particular, the authors had to combine location data using shape files and distilled demographic data from question P22 in the census, which breaks down households by family/non-family and the age of the head of household. The data were merged using census-block identification numbers.

In the Cohort Location Model, the authors argue that changes in households' housing careers (i.e. housing consumption, residential mobility and location choices) across the life span happen in a continuum and therefore they hypothesise that such changes could create spatial sorting effects within metropolitan areas.

They test their model on the 50 largest US metropolitan regions (core based statistical areas or CBSAs in the paper), deriving household age and location from the 2010 Census. Distances were standardized against the farthest area from each city center $\frac{\text{distance of block}}{\max(\text{distance of block})}$.

As overall household counts are not evenly distributed by the age of the householder, simple counts of households at each location will not result in fair comparisons. Therefore, to evaluate the location choices of a given age of a household, they computed location quotients for each age group at each 1/100 of the standardised distance.

- HH_{ij}: Number of households labelled at age-group i and distance j from the city-center
- $HH_{.j}$: Number of households labelled at distance j from the city-center
- $HH_{i.}$: Number of households labelled at age-group i
- $HH_{..}$: Total number of households (in that city)

Location Quotient:

$$LQ = \frac{\left(\frac{HH_{ij}}{HH_{.j}}\right)}{\left(\frac{HH_{i.}}{HH_{..}}\right)}$$

Note: A LQ of 1 denotes that a given age cohort is represented at a given distance in the same proportion as that age cohort is represented in the entire metropolitan area. Location quotients (LQs) less (greater) than 1 indicate under(over)-representation in a given area or at a given distance.

They compare distance from city-center against location quotient, fitting the relationship using LOESS ("locally-smoothed line of best fit").

Questions

1) From the information above and the documentation provided, can you quickly identify where in the code (if present at all) the authors:
    a. Collected their data
    b. Cleaned/processed their data
    c. Calculated various statistics
    d. Produced the plots in their paper
2) In terms of coding what elements of their project do you like?
   Consider: Documentation, comments, organization, naming, workflow, data provenance, etc.
3) Similarly, list out things that you think could be improved.
4) Without examining the code itself, can you quickly discern the purpose of each file?
5) Without examining the code itself, can you quickly tell what each block of code does?
6) Are there exceptions to your answers above?
7) In what way do the authors document their workflow? Do you think this method is effective?

Gitignore

Take a look at the gitignore file. This is a good way to exclude certain files from being documented by Git. At the very least, you'll want to add large data sets and auxiliary files to gitignore.