SOCIAL IMPACT DATA COMMONS

Supporting Local Decision-Making through the Aggregation of ACS Demographic Estimates within Locally-Relevant Geographies

Aaron Schroeder & Edward Wu
Social Decision Analytics Division (SDAD) of the UVA Biocomplexity Institute

2023 ACS DATA USERS CONFERENCE

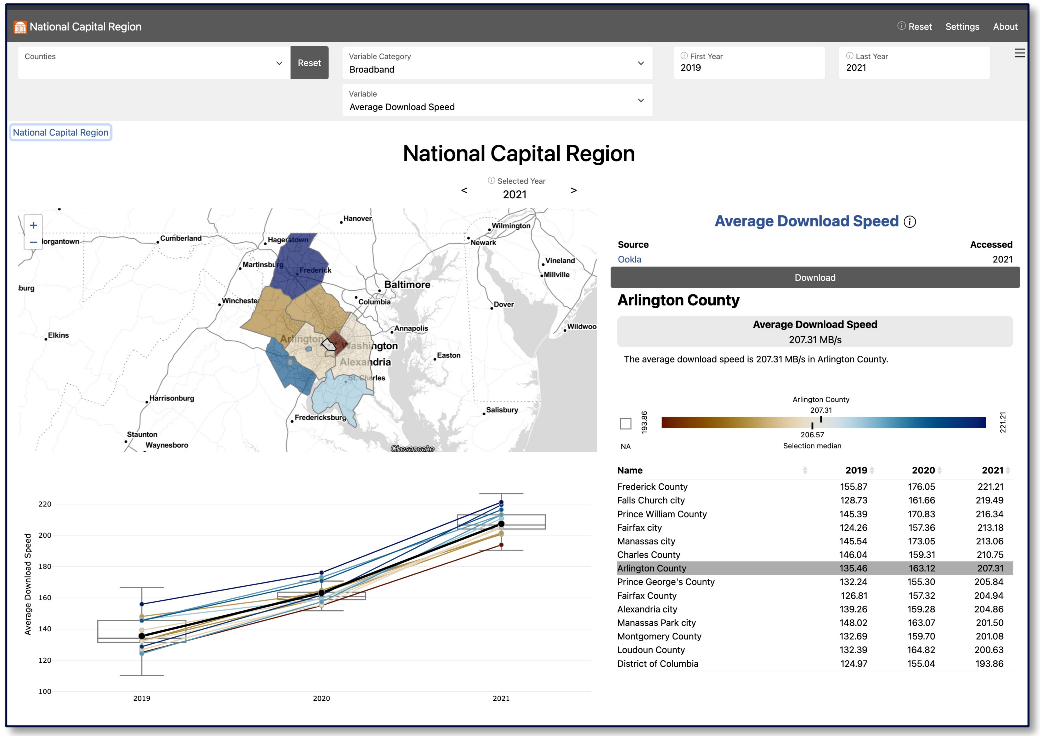
UVA | BIOCOMPLEXITY INSTITUTE

Center for Inclusive Growth

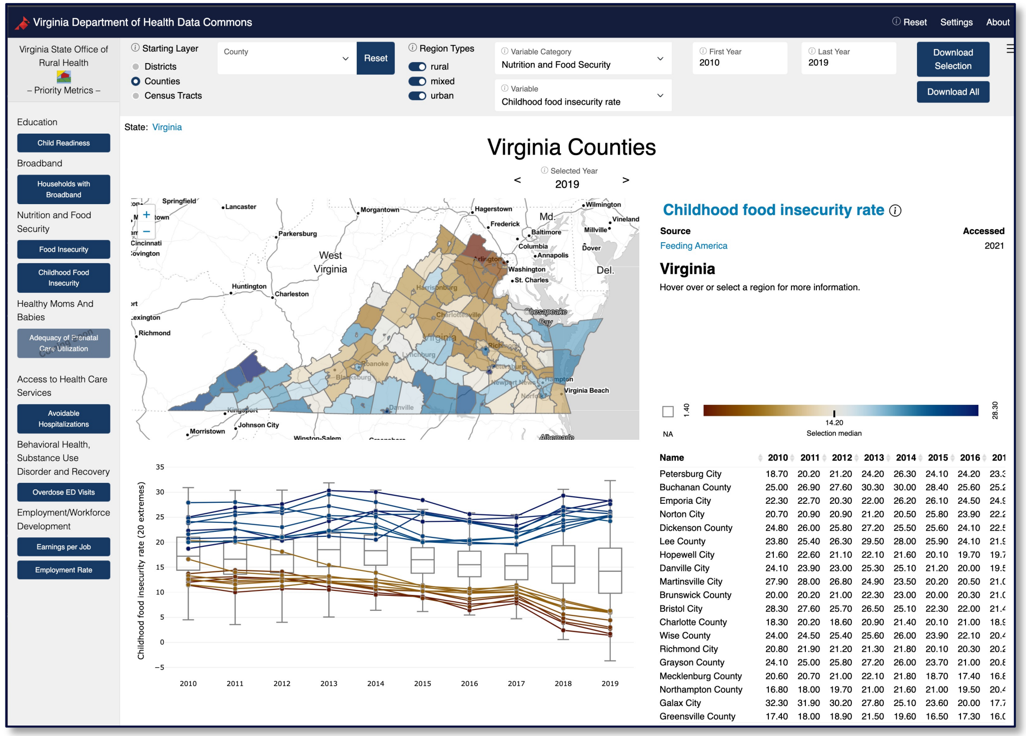
Two Data Commons Projects

1. Social Impact Data Commons to Inform Equitable Growth (Mastercard Center for Inclusive Growth) – National Capital Region
2. Data Commons to Support Department of Health Strategic Plans (Virginia Department of Health) – State of Virginia

Both apply methods for population redistribution



https://uva-bi-sdad.github.io/capital_region



https://uva-bi-sdad.github.io/vdh_rural_health_site

SOCIAL IMPACT
DATA COMMONS

Estimating Quantities in Useful Geographies

Translation of Census Demographics to New Geographies

- Providing local governments with estimates for quantities of interest in useful geographical areas can be helpful for crafting policy
- There can be a mismatch between the geographies of available data and geographies of interest
- Ex: American Community Survey (ACS) estimates are at the block group level, but Arlington County is interested in civic associations
- Many others, e.g., Metro Corridors, Business Districts

Census Block Groups

Waverly Hills Civic Association

Estimating Quantities in Useful Geographies

Translation of Census Demographics to New Geographies

- We combine information from data sets with varying degrees of granularity (Arlington County parcel data, ACS block group data, and ACS microdata)
- We obtain demographic estimates at the parcel/household level and aggregate these estimates up to the geography of interest

SOCIAL IMPACT
DATA COMMONS

Data Sources

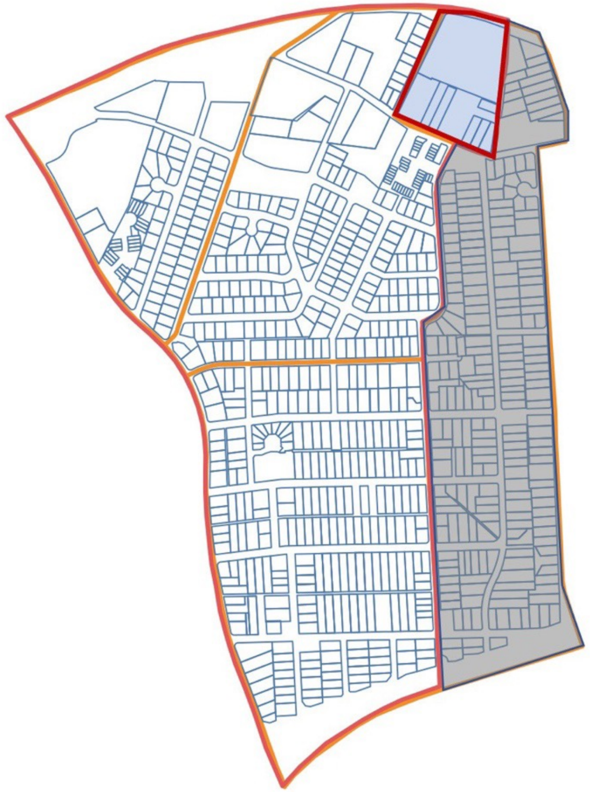
The three data sources currently using

- Arlington (local) Parcel Data
  - Approximately 34,000 parcels in Arlington
  - Variables include “House Type” (e.g., Single Family Detached, Apartments) and Own/Rent status
  - Units are measured at the parcel-level
- ACS Block Group Data
  - Approximately 200 block groups in Arlington County
  - Units are block groups, and estimates for household- and individual-level variables are provided at the block group level (e.g., number of households in the block group with less than $10,000 yearly income, number of residents in the block group with a master’s degree)
  - Household-level variables include income, rent/own, housing type
  - Individual-level variables include sex, race, educational attainment, age
- ACS PUMS Data
  - Approximately 11,000 individuals in 2 PUMAs (Public Use Microdata Areas) in Arlington
  - Each PUMA contains multiple block groups
  - Household-level variables include income, rent/own, housing type
  - Individual-level variables include sex, race, educational attainment, age

Estimating Quantities in Useful Geographies

Translation of Census Demographics to New Geographies

Census Block Groups

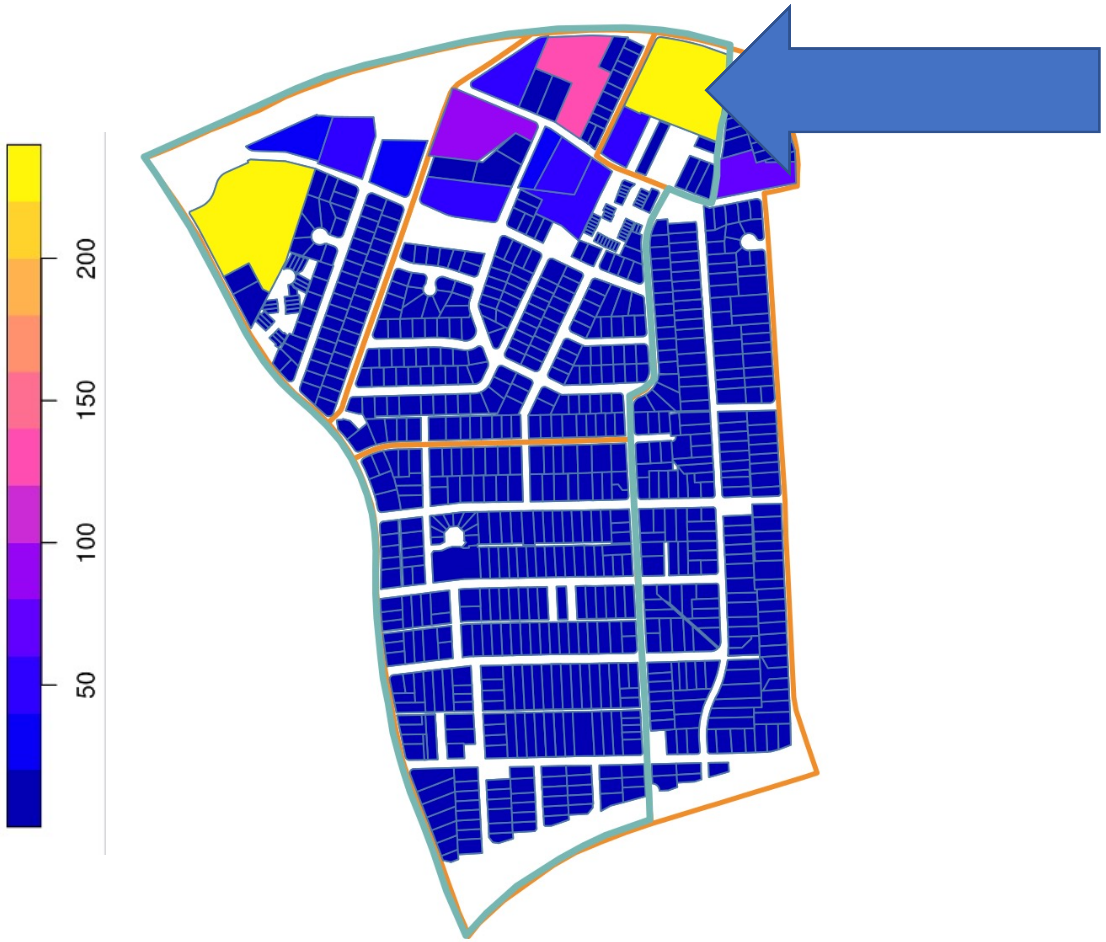


Waverly Hills Civic Association

SOCIAL IMPACT DATA COMMONS

Estimating Quantities in Useful Geographies

Translation of Census Demographics to New Geographies



Partial Overlap

- Leave it Off? (One popular site uses a 50% Approach)
- Some others use a Percentage Overlap\centroid approach?
- Both potentially significantly reduce the populations of most interest!
- Our first approach – Allocate demographics to parcels according to number of housing units

SOCIAL IMPACT DATA COMMONS

Approach 1

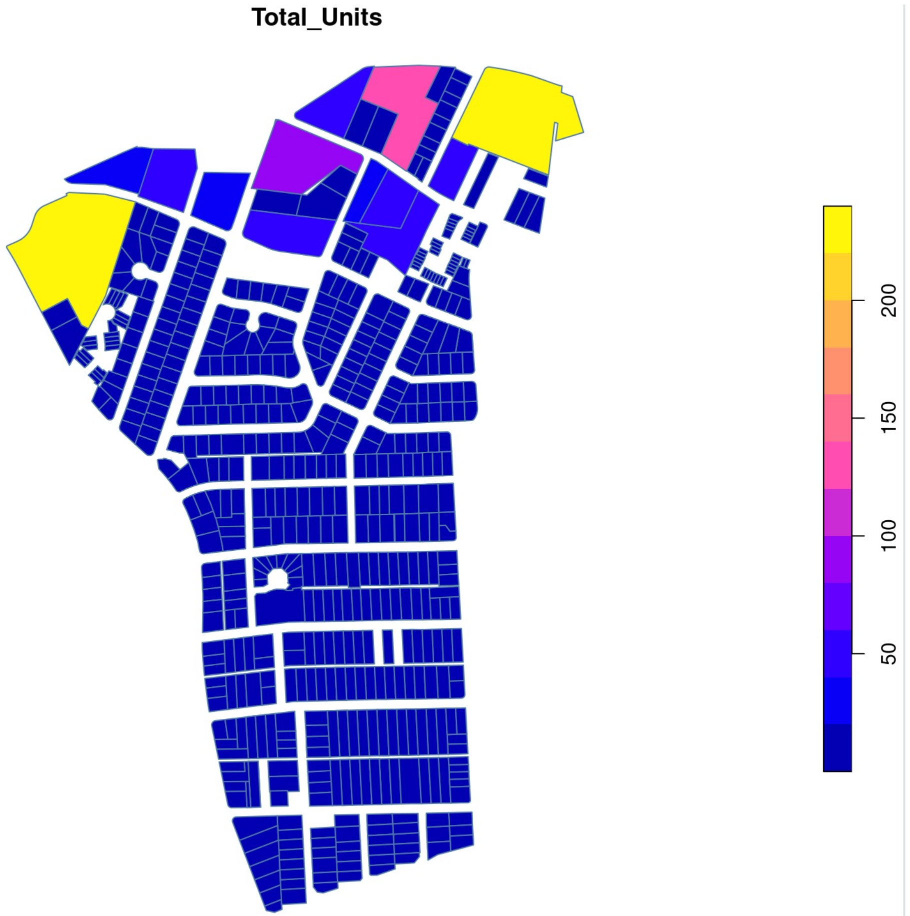
Proportional Distribution of ACS Estimates

- For each block group, we have an ACS estimate for the number of each race residing in that block group
- We distribute the ACS estimates to each parcel weighted by the number housing units on the parcel

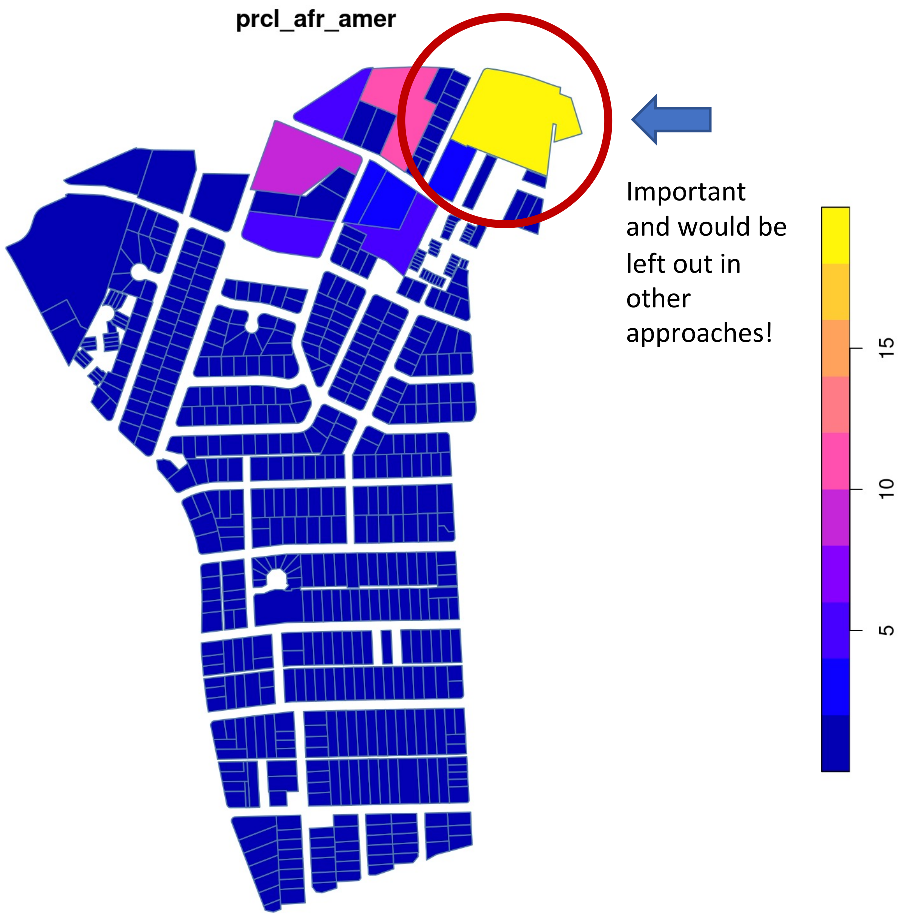
- Suppose block group 1 has 750 white residents and 250 housing units
- Each housing unit will be estimated to have 750/250 = **3** white residents
- Therefore, a parcel with 10 housing units will be estimated to have **30** white residents

Estimating Quantities in Useful Geographies

Translation of Census Demographics to New Geographies



Waverly Hills Parcel Units

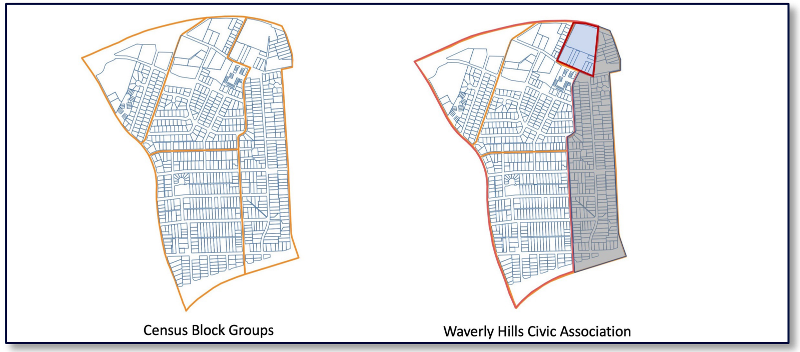


Waverly Hills Estimated Demographic per Parcel

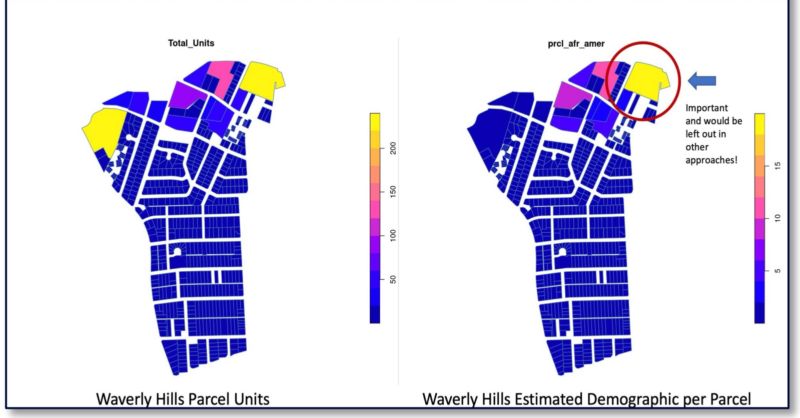
Estimating Quantities in Useful Geographies

Creating data and metrics in geographies that matter locally

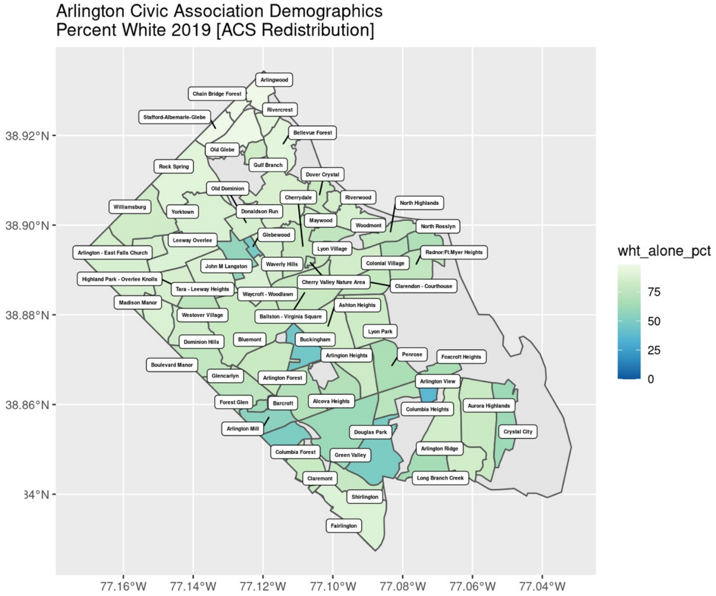
Translation of Census Demographics to New Geographies

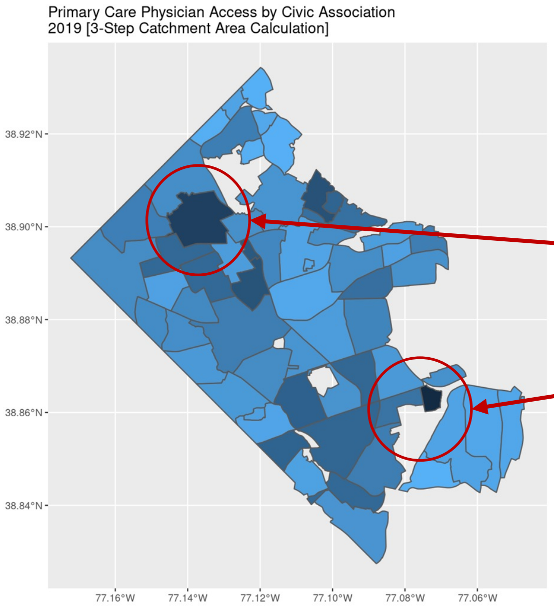


Reviewing multiple methods – creating new approaches



Enabling Analysis by Neighborhood, rather than Census Geographies





Need Civic Association Household Income Variable as Well (Coming Soon!)

SOCIAL IMPACT DATA COMMONS

Approach 2

Enhancement using Average Household Size

- Different types of housing have a different number of people on average
- In the Arlington PUMAs, single family houses average 2.7 residents and other homes (apartments, condos) average 1.7
- Approach 2 distributes ACS estimates to parcels weighting by both number of housing units and average household size

Approach 2

Example

- Suppose block group 1 has 750 white residents and 250 housing units, of which 150 are single family residences and 100 are apartments
- 528.3 white residents are estimated to live in single family residences:

$$\frac{2.7\times150}{2.7\times150+1.7\times100}\times750=528.3$$

- Each single family residence is estimated to have 528.3/150 = **3.5** white residents
- Each apartment is estimated to have 221.7/100 = **2.2** white residents
- A parcel with 10 apartments is estimated to have **22** white residents

Approach 3

Incorporating Covariates

- Approach 1 does not account for covariate information (such as house type and own/rent status)
- Approach 2 accounts for differences in household size among single family vs. other homes
- We might expect the distribution of race to differ for single-family detached homes vs. apartments, or for rented vs. owned homes

Approach 3

Incorporating Covariates using Raking

- ACS block group data include the marginal distribution of race, own/rent status, and house type
- The distribution of race conditional on house type and rent/own status cannot be estimated directly from the block group data
- Joint distributions *can* be estimated using PUMS data, but PUMAs are much larger than (not representative of) block groups

Solution: use Raking/Iterative Proportional Fitting to re-weight PUMS data to match block group marginals and estimate conditional distribution of race using re-weighted data

Approach 3

Example

Step 1: Calculate individual and household targets using Parcel data and ACS Block Group data

Block Group Targets
- 80% of residents are white
- 50% of households rent

Step 2: Use raking to re-weight PUMS data

Original PUMS Data
Individual and Household Level Data

Rake to match household targets →

Raked PUMS, where
- 50% of households rent

Rake to match individual targets →

Raked PUMS, where
- 80% of residents are white
- ~50% of households rent

Approach 3

Example

Block Group Targets
- 80% of residents are white
- 50% of households rent

Step 3: Use re-weighted PUMS data to obtain estimates of interest

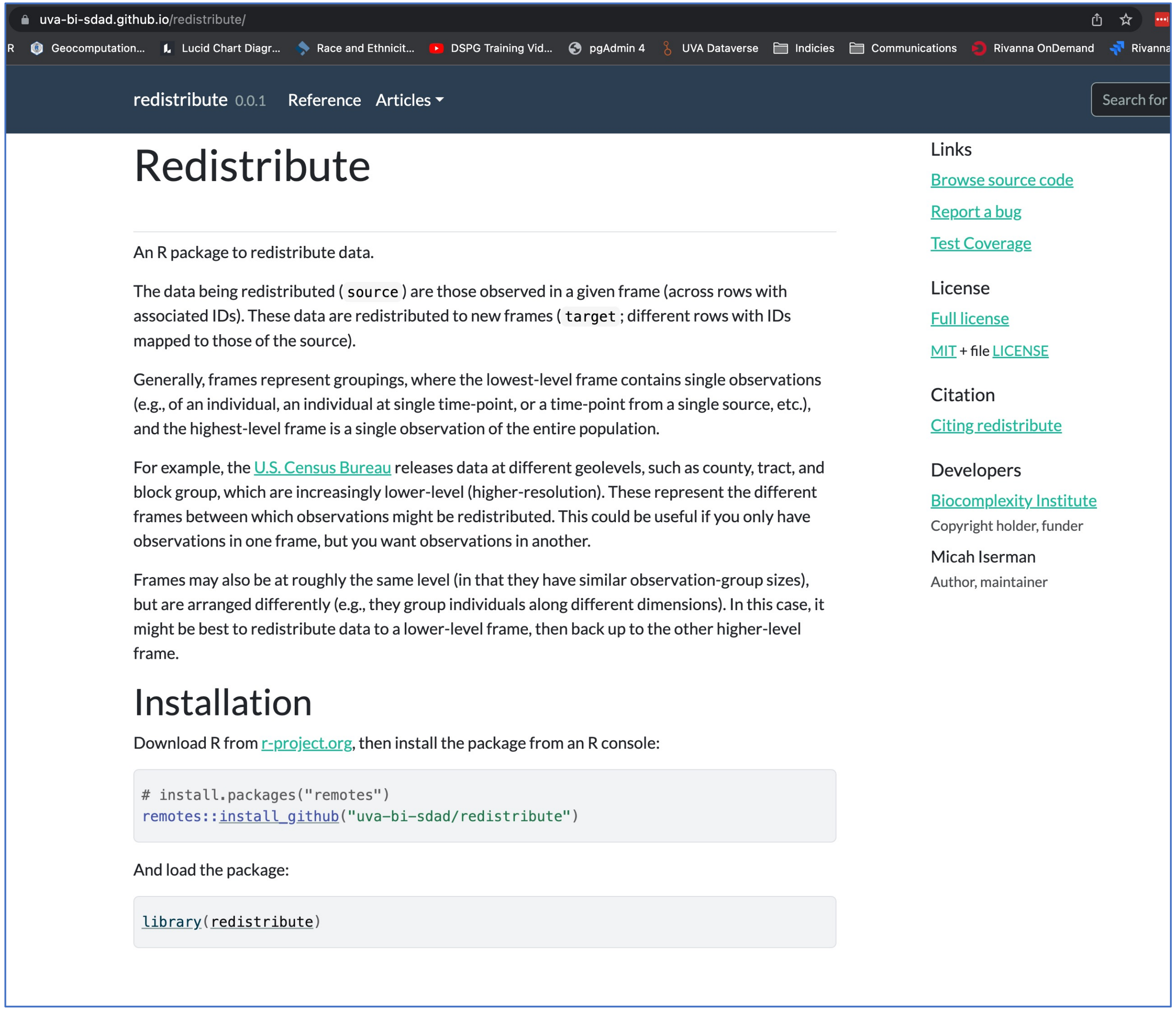
Raked PUMS, where
- 80% of residents are white
- ~50% of households rent

→

Among owners of single family detached homes in the block group:
88% are estimated to be White, 7% African American, 1% American Indian, and 4% Asian

R Package

Redistribute



uva-bi-sdad.github.io/redistribute/

redistribute 0.0.1 Reference Articles

Redistribute

An R package to redistribute data.

The data being redistributed (`source`) are those observed in a given frame (across rows with associated IDs). These data are redistributed to new frames (`target` ; different rows with IDs mapped to those of the source).

Generally, frames represent groupings, where the lowest-level frame contains single observations (e.g., of an individual, an individual at single time-point, or a time-point from a single source, etc.), and the highest-level frame is a single observation of the entire population.

For example, the U.S. Census Bureau releases data at different geolevels, such as county, tract, and block group, which are increasingly lower-level (higher-resolution). These represent the different frames between which observations might be redistributed. This could be useful if you only have observations in one frame, but you want observations in another.

Frames may also be at roughly the same level (in that they have similar observation-group sizes), but are arranged differently (e.g., they group individuals along different dimensions). In this case, it might be best to redistribute data to a lower-level frame, then back up to the other higher-level frame.

Installation

Download R from r-project.org, then install the package from an R console:

```
# install.packages("remotes")
remotes::install_github("uva-bi-sdad/redistribute")
```

And load the package:

```
library(redistribute)
```

Links

Browse source code

Report a bug

Test Coverage

License

Full license

MIT + file LICENSE

Citation

Citing redistribute

Developers

Biocomplexity Institute

Copyright holder, funder

Micah Iserman

Author, maintainer

Next Steps

- Model-based estimation approach is in progress
- Implementation of methods in an R package
- Validation using simulation studies and other approaches

- THANK YOU!

SOCIAL IMPACT DATA COMMONS