CDC SVI calculation and validation

Heli Xu

2/8/23

Table of contents

[SVI calculation and validation 1](#_Toc126856659)

[PA ZCTA-level SVI 2018 2](#_Toc126856660)

[Add ZCTA information to CDC-released SVI 2](#_Toc126856661)

[Aggregating ct data to ZCTA level 3](#_Toc126856662)

[Correlation: aggregated CDC data vs. calculated result 4](#_Toc126856663)

[Potential over-representation after aggregation 6](#_Toc126856664)

[PA ZCTA-level SVI 2020 6](#_Toc126856665)

Currently, our SVI calculation logic is stored in R functions , and we’re in the process of package development. In responding to the request for PA SVI at the ZCTA level from 2017-2021, we’d like to include this report to provide the rationale and approaches used in validating the results. Any suggestions and feedback are greatly appreciated.

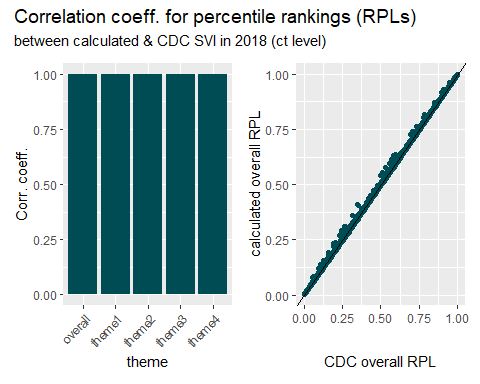
This report will cover the following sections:

* **SVI calculation and validation:** Brief introduction to SVI calculation and variables preparation which are validated by comparing our county-/tract-level results to CDC’s SVI (example: PA, tract level, 2018);
* **PA ZCTA-level SVI 2018:** Summary of the approaches for aggregating CDC tract-level data to ZCTA level and comparing it with our PA ZCTA-level result for 2018;
* **PA ZCTA-level SVI 2020:** Comparing aggregated CDC ZCTA-level SVI with our PA ZCTA-level result for 2020.

## SVI calculation and validation

As included in a separate R script (“function collection.R”), get\_census\_data()(using {tidycensus} under the hood) and get\_svi() were used to calculate SVI from census data. The variables required for SVI calculation were either extracted from the dictionary published by CDC when SVI was released (for 2014, 2016, 2018, 2020), or modified from the adjacent years to account for minor changes (for 2015, 2017, 2019, 2021). Details about the variables preparation are included in a separate script (“svi variable prep.R”).

As part of the validation process for our R functions, county and census tract level SVI for certain states are calculated and compared with [CDC-released SVI](https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html) for the same year (if available) or adjacent year, making sure the two versions of SVIs by GEOID are well correlated. For example, our calculated SVI for PA in 2018 at census tract level is highly consistent with the CDC data for 2018, with a correlation coefficient above 0.99 for overall and theme-specific SVI.



## PA ZCTA-level SVI 2018

With SVI calculated at the ZCTA level in PA from 2017-2021, here we’d like to check the results against the CDC-published SVI data for PA in the years that they are available (2018 and 2020). Since SVI data released by CDC are either at county or census tract level for each state, we need to add ZCTA information to the SVI tables and aggregate the data from census tract level to ZCTA level. ZCTA to census tract (ct) crosswalk were modified from the relationship files (decennial) from [census.gov](https://www.census.gov/geographies/reference-files/time-series/geo/relationship-files.2020.html#list-tab-Z1E5E6KE884J6MIBIQ).

### Add ZCTA information to CDC-released SVI

Using ZCTA to census tract crosswalk for 2010, we’re matching each census tract in the 2018 CDC SVI table with a corresponding ZCTA (first few tracts shown below):

| ST | STATE | ST\_ABBR | STCNTY | COUNTY | GEOID | ZCTA |
| --- | --- | --- | --- | --- | --- | --- |
| 42 | PENNSYLVANIA | PA | 42001 | Adams | 42001030101 | 17019 |
| 42 | PENNSYLVANIA | PA | 42001 | Adams | 42001030101 | 17316 |
| 42 | PENNSYLVANIA | PA | 42001 | Adams | 42001030101 | 17324 |
| 42 | PENNSYLVANIA | PA | 42001 | Adams | 42001030101 | 17372 |
| 42 | PENNSYLVANIA | PA | 42001 | Adams | 42001030102 | 17316 |
| 42 | PENNSYLVANIA | PA | 42001 | Adams | 42001030102 | 17350 |

We could briefly check if there’s any GEOID (ct) that didn’t get matched with a ZCTA code (none):

# A tibble: 0 × 125  
# … with 125 variables: ST <dbl>, STATE <chr>, ST\_ABBR <chr>, STCNTY <dbl>,  
# COUNTY <chr>, GEOID <dbl>, ZCTA <chr>, LOCATION <chr>, AREA\_SQMI <dbl>,  
# E\_TOTPOP <dbl>, M\_TOTPOP <dbl>, E\_HU <dbl>, M\_HU <dbl>, E\_HH <dbl>,  
# M\_HH <dbl>, E\_POV <dbl>, M\_POV <dbl>, E\_UNEMP <dbl>, M\_UNEMP <dbl>,  
# E\_PCI <dbl>, M\_PCI <dbl>, E\_NOHSDP <dbl>, M\_NOHSDP <dbl>, E\_AGE65 <dbl>,  
# M\_AGE65 <dbl>, E\_AGE17 <dbl>, M\_AGE17 <dbl>, E\_DISABL <dbl>,  
# M\_DISABL <dbl>, E\_SNGPNT <dbl>, M\_SNGPNT <dbl>, E\_MINRTY <dbl>, …

### Aggregating ct data to ZCTA level

Generally speaking, ZCTAs represent larger areas than census tracts, but census tracts do not necessarily nest within ZCTAs – sometimes one census tract could correspond to multiple ZCTAs (while one or more of those ZCTAs also show up in other census tracts). In such cases, we might be aggregating values of a larger area to represent a smaller area. Considering the complicated relationship between ZCTAs and census tracts, here we’re subsetting the ct-specific GEOIDs with only one matching ZCTA (different ct can be matched to the same ZCTA).

The potential caveat is that we might be over-representing some census tracts for a ZCTA when it’s supposed to include multiple census tracts, but some of the census tracts are excluded because they’re also matched to other ZCTAs. An example will be provided later in this document.

#### Variables (“E\_xx” and “EP\_xx”)

For individual variables, we’re taking sum for the “E\_xx” values by the ZCTA as they mostly represent counts, whereas for “EP\_xx” values, we’re taking the mean as they represent a percentage of the total. Before 2018, the exception is “E\_PCI”, which is the same value as “EP\_PCI”, representing per capita income, and we’re taking the mean of that. First 15 rows of the aggregated table is shown below:

| ZCTA | var\_name | var\_zcta |
| --- | --- | --- |
| 15001 | E\_AGE17 | 3179.00 |
| 15001 | E\_AGE65 | 3583.00 |
| 15001 | E\_CROWD | 65.00 |
| 15001 | E\_DISABL | 3269.00 |
| 15001 | E\_GROUPQ | 507.00 |
| 15001 | E\_HH | 8067.00 |
| 15001 | E\_HU | 9586.00 |
| 15001 | E\_LIMENG | 104.00 |
| 15001 | E\_MINRTY | 4974.00 |
| 15001 | E\_MOBILE | 222.00 |
| 15001 | E\_MUNIT | 613.00 |
| 15001 | E\_NOHSDP | 1025.00 |
| 15001 | E\_NOVEH | 993.00 |
| 15001 | E\_PCI | 24178.67 |
| 15001 | E\_POV | 3483.00 |

#### Percentile ranking (“RPL\_xx”) by theme

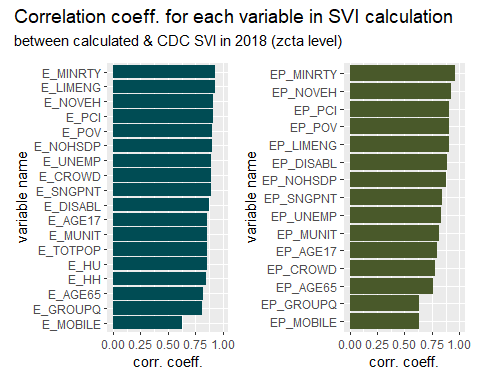
For percentile ranking, for now we’re taking the mean by the ZCTA and we’ll get a table with first few rows looking like this: (I tried weighted by population and it didn’t seem to help too much with correlation.)

| ZCTA | cdc\_RPL\_themes | cdc\_RPL\_theme1 | cdc\_RPL\_theme2 | cdc\_RPL\_theme3 | cdc\_RPL\_theme4 |
| --- | --- | --- | --- | --- | --- |
| 15001 | 0.6241000 | 0.6000500 | 0.6283000 | 0.5522167 | 0.5931333 |
| 15003 | 0.7153667 | 0.6803333 | 0.6176667 | 0.5735667 | 0.6555000 |
| 15005 | 0.3658000 | 0.3362000 | 0.5917000 | 0.1410000 | 0.5667000 |
| 15009 | 0.2006250 | 0.2957250 | 0.5000000 | 0.1411000 | 0.2125000 |
| 15010 | 0.7475750 | 0.7720000 | 0.6890250 | 0.5234500 | 0.6250750 |
| 15012 | 0.4354000 | 0.5984000 | 0.2847000 | 0.4928000 | 0.2996000 |
| 15017 | 0.4439667 | 0.4159333 | 0.5477667 | 0.2490000 | 0.5918000 |
| 15021 | 0.5301000 | 0.7184000 | 0.7387000 | 0.2730000 | 0.2379000 |
| 15022 | 0.7778667 | 0.6937000 | 0.9548667 | 0.2902000 | 0.7480667 |
| 15024 | 0.1514000 | 0.2757000 | 0.3307000 | 0.0088000 | 0.3312000 |

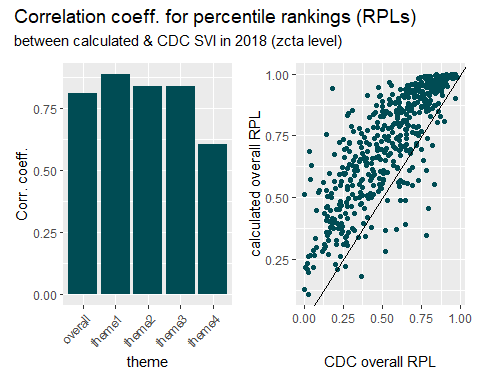
### Correlation: aggregated CDC data vs. calculated result

To compare the aggregated CDC SVI data with our calculated SVI, we’re joining the CDC data with our result by each ZCTA, and we could check the correlation between the two versions of SVI in all ZCTAs.

For individual variables, below shows the correlation coefficient for each variable:



For RPLs, we’ll compare RPL for each theme and overall RPL for all themes. Here is a plot for the overall RPL for all themes of each ZCTA:



### Potential over-representation after aggregation

In the plot above, we notice a data point that is close to 1 in calculated result, but less than 0.25 in CDC data. That point corresponds to ZCTA 15207. In CDC data, this ZCTA has the following census tracts:

| GEOID | ZCTA | RPL\_THEMES | RPL\_THEME1 | RPL\_THEME2 | RPL\_THEME3 | RPL\_THEME4 |
| --- | --- | --- | --- | --- | --- | --- |
| 42003151600 | 15207 | 0.1821 | 0.3716 | 0.0388 | 0.5829 | 0.2389 |
| 42003151700 | 15207 | 0.1931 | 0.3452 | 0.0986 | 0.7330 | 0.0942 |
| 42003310200 | 15207 | 0.3420 | 0.6360 | 0.5629 | 0.1507 | 0.1653 |
| 42003310300 | 15207 | 0.1966 | 0.2538 | 0.2947 | 0.5303 | 0.1544 |
| 42003488500 | 15207 | 0.4132 | 0.4261 | 0.4568 | 0.1379 | 0.6456 |
| 42003562300 | 15207 | 0.9781 | 0.8806 | 0.9496 | 0.7079 | 0.9900 |
| 42003562900 | 15207 | 0.8141 | 0.7159 | 0.9327 | 0.6773 | 0.5611 |
| 42003980500 | 15207 | -999.0000 | -999.0000 | 0.4865 | 0.0000 | 0.0000 |

But among these census tracts, almost all of them match to other ZCTAs, and the ones end up included in the aggregation are the following two (and the -999 gets further excluded):

| GEOID | ZCTA | RPL\_THEMES | RPL\_THEME1 | RPL\_THEME2 | RPL\_THEME3 | RPL\_THEME4 |
| --- | --- | --- | --- | --- | --- | --- |
| 42003151600 | 15207 | 0.1821 | 0.3716 | 0.0388 | 0.5829 | 0.2389 |
| 42003980500 | 15207 | -999.0000 | -999.0000 | 0.4865 | 0.0000 | 0.0000 |

That’s why the “aggregated” SVI for ZCTA15207 in fact only represents one of the census tracts in the area, leading to a discrepancy with our calculated result.

## PA ZCTA-level SVI 2020

Similarly, we could aggregate and compare CDC-released SVI for 2020 at the census tract level with our calculated result at the ZCTA level.

First, we’re joining the CDC SVI table to a new ZCTA to census tract crosswalk updated in 2020. Here is a glance at the table:

| ST | STATE | ST\_ABBR | STCNTY | COUNTY | GEOID | ZCTA |
| --- | --- | --- | --- | --- | --- | --- |
| 42 | Pennsylvania | PA | 42001 | Adams | 42001030101 | 17019 |
| 42 | Pennsylvania | PA | 42001 | Adams | 42001030101 | 17316 |
| 42 | Pennsylvania | PA | 42001 | Adams | 42001030101 | 17324 |
| 42 | Pennsylvania | PA | 42001 | Adams | 42001030101 | 17372 |
| 42 | Pennsylvania | PA | 42001 | Adams | 42001030103 | 17316 |
| 42 | Pennsylvania | PA | 42001 | Adams | 42001030103 | 17372 |

And a brief check for unmatched GEOIDs for reference:

| ST | STATE | ST\_ABBR | STCNTY | COUNTY | GEOID | ZCTA |
| --- | --- | --- | --- | --- | --- | --- |
| 42 | Pennsylvania | PA | 42055 | Franklin | 42055010200 | NA |
| 42 | Pennsylvania | PA | 42055 | Franklin | 42055011302 | NA |

Secondly, we’ll aggregate the CDC data from ct level to ZCTA level, including the variables and percentile rankings by theme (first 15 rows of each table is shown below).

Aggregated data for all variables:

| ZCTA | var\_name | var\_zcta |
| --- | --- | --- |
| 15001 | E\_AGE17 | 2169 |
| 15001 | E\_AGE65 | 2927 |
| 15001 | E\_CROWD | 45 |
| 15001 | E\_DISABL | 2666 |
| 15001 | E\_GROUPQ | 91 |
| 15001 | E\_HBURD | 1719 |
| 15001 | E\_HH | 6669 |
| 15001 | E\_HU | 8026 |
| 15001 | E\_LIMENG | 76 |
| 15001 | E\_MINRTY | 4315 |
| 15001 | E\_MOBILE | 87 |
| 15001 | E\_MUNIT | 657 |
| 15001 | E\_NOHSDP | 951 |
| 15001 | E\_NOVEH | 941 |
| 15001 | E\_POV150 | 3532 |

Aggregated data for percentile rankings:

| ZCTA | cdc\_RPL\_themes | cdc\_RPL\_theme1 | cdc\_RPL\_theme2 | cdc\_RPL\_theme3 | cdc\_RPL\_theme4 |
| --- | --- | --- | --- | --- | --- |
| 15001 | 0.6575200 | 0.6267600 | 0.604360 | 0.7102400 | 0.6102400 |
| 15003 | 0.5986333 | 0.6246333 | 0.491500 | 0.6259333 | 0.4723333 |
| 15005 | 0.3201000 | 0.2105000 | 0.462100 | 0.2617000 | 0.5383000 |
| 15009 | 0.2765000 | 0.3638750 | 0.446600 | 0.2078750 | 0.2478500 |
| 15010 | 0.6468800 | 0.5599400 | 0.600580 | 0.5336600 | 0.6685000 |
| 15012 | 0.2837000 | 0.2684500 | 0.405100 | 0.3665000 | 0.3099000 |
| 15017 | 0.5698000 | 0.4555667 | 0.485000 | 0.4363000 | 0.7247333 |
| 15021 | 0.5318000 | 0.6983000 | 0.549600 | 0.2400000 | 0.3299000 |
| 15022 | 0.7126667 | 0.6582333 | 0.610200 | 0.3929333 | 0.7539667 |
| 15024 | 0.0283000 | 0.1303000 | 0.084900 | 0.1210000 | 0.0753000 |
| 15025 | 0.6588000 | 0.6880500 | 0.538575 | 0.7208250 | 0.5044750 |
| 15027 | 0.3743000 | 0.5193000 | 0.166900 | 0.0909000 | 0.5756000 |
| 15033 | 0.8522000 | 0.7750500 | 0.877350 | 0.7497000 | 0.7119000 |
| 15034 | 0.6284000 | 0.7865000 | 0.502400 | 0.2556000 | 0.4559000 |
| 15035 | 0.5636000 | 0.7136000 | 0.764700 | 0.4209000 | 0.1833000 |

After adding ZCTA information and aggregating the CDC data, we’re ready to compare them to our calculated result for 2020. Most of the variables turn out quite consistent between our calculation and aggregated CDC data, despite a few peculiar outliers (shown in the correlation plot), possibly due to over-representation of some census tracts.

