2022 COMMUNITY RESILIENCE ESTIMATES

Detailed Technical Documentation

Updated January 2024

Small Area Estimates Program

Social, Economic, and Housing Statistics Division

U.S. Census Bureau, Department of Commerce

Contents

1. Introduction

1.1 Overview

The general framework of the Community Resilience Estimates (CRE) follows a standard area-level approach from small-area estimation: a weighted average is taken of a direct estimator and an indirect estimator to produce a composite estimator, see Rao and Molina (2015) for more information on small-area estimation methods. The composite estimator is less volatile than either of the two original estimators. In the case of CRE, the direct estimates are 2022 American Community Survey (ACS) estimates for the number of individuals within a given tract who possess zero, one or two, or three or more components of social vulnerability (SV). The indirect, or synthetic, estimates are developed from applying modeled proportions to auxiliary population data. Ensuing sections will look at each step in more detail.

1.2 Data Sources

There are three sources of data for CRE: the 2020 Decennial Census, the Population Estimates Program (PEP) of the United States Census Bureau, and ACS One-Year Estimates.

1.2.1 Decennial Census

Every ten years, the United States Census Bureau conducts the Decennial Census. The intent of the Decennial Census is to enumerate the United States resident population for the purposes of apportionment in the United States House of Representatives. Information collected during the Census includes demographic data such as age, race, and sex. The 2022 CRE estimates use the Decennial Census Public Law 94-171 (PL94) Redistricting Data summary tables and Demographic Housing Characteristics File (DHC) tables from the 2020 Decennial Census. We treat the DHC and PL94 data as being free from error even though data from both releases was injected with noise for disclosure avoidance purposes (United States Census Bureau, 2023).

1.2.2 PEP

The Population Estimates Program of the United States Census Bureau produces postcensal resident population estimates for geographies within the United States and Puerto Rico that update the size and demographic composition of the population since the last decennial census. Population estimates are published annually for households and Group Quarters (GQ) and are broken down by age, race, sex, and Hispanic origin. The 2022 CRE estimates use internal data which is also broken down by household/GQ type. CRE methodology assumes county-level PEP estimates are measured without error.

1.2.3 ACS

The ACS samples United States households yearly to gather information on housing, population, demographics, health, and finances. We used internal ACS microdata from the 2022 ACS survey to produce ACS estimates for various geographies. Unlike the Decennial and PEP estimates, ACS estimates contain error which is estimated using replicate weights.

1.3 Notation

CRE creates unpublished tract-level population estimates for combinations of

- four age groups
  - 0-17,18-44,45-64,65+
- five race/ethnicity categories
  - White non-Hispanic
  - Black non-Hispanic
  - American Indian/Alaska Native non-Hispanic
  - Other non-Hispanic
  - Hispanic
- two sex categories

broken down by households and seven Group Quarter (GQ) types:

1. correctional facilities
2. juvenile institutions
3. nursing homes
4. institutional GQs, such as in-patient hospice facilities and military treatment facilities.
5. college dormitories
6. military barracks
7. non-institutional GQs, such as group homes.

Only GQ types 3., 4., 6., and 7. are in-universe for CRE. Let AR be all 20 possible age x race/ethnicity combinations and ARS all 40 possible age x race/ethnicity x sex combinations. We will refer to an element of AR or ARS as a demographic combination or, simply, a demographic.

For a given ACS respondent y, w_y will denote the respondent's ACS weight, and for a demographic j, social vulnerability component k, and geographic area g let

- $ACS_{j,k}$ represent the collection of all ACS respondents with a demographic combination j and social vulnerability component k.
- $ACS_{g,j,k}$ be the subset of respondents from $ACS_{j,k}$ with residence in g.

- $ACSPOP_{g,j}$ denote the ACS population estimate for demographic combination j and geography g.

An area's total in-universe population, i.e. households and all in-scope GQs, will be denoted by $POPUNI_{g,j}$ and $POPUNI_g$. Similar descriptive names are used for other population counts, such as 2010 Decennial counts and PEP estimates.

We use $dg \in \{0,12,3\}$ to differentiate between different SV groups where:

- 0 denotes the zero SV group,
- 12 denotes the one or two SV group, and
- 3 denotes the three or more SV group.

For instance, $ACS_{g,j,dg}$ is the subset of ACS respondents in demographic j and geography g in the dg SV group.

1.4 Assumptions for Components of Social Vulnerability

Residents in nursing homes and other types of institutional GQs are only eligible for a subset of the ten components of social vulnerability:

- crowding
- low income to poverty ratio (IPR)
- communication barrier
- age 65 and over
- presence of a disability.

We assume the residents do not possess any of the remaining five social vulnerability components. The crowding social vulnerability component is assumed to hold for all residents in these GQs. For military barracks we estimate only one social vulnerability component, the crowding component, i.e. we assume all residents in military barracks have this social vulnerability component.

2. Auxiliary Population Estimates

In the 2018 (experimental) and 2019 versions of CRE, the starting points of our population estimates were tract-level 2010 Decennial tabulations by age x race/ethnicity x sex for households and the seven GQ types. The only results from the 2020 Decennial Census available in time for use in CRE 2021 were the PL94 summary tables. The PL94 summary tables contain population totals for different age x race/ethnicity combinations down to the block-level but lack a great deal of information previously used for the CRE auxiliary population. The CRE 2021 methodology imputed this missing information using 2010 Decennial data. In May 2023, more detailed demographic data for a variety of geography levels from the 2020 Census became available from the DHC file. In the next two subsections, we describe how the PL94 and DHC data were used to create the CRE 2022 population estimates.

2.1 Initial 2020 Population Estimates

We want to create 2020 tract-level population estimates broken down by age x race/ethnicity x sex x household/GQ type. Our focus is three groups of DHC tables:

- PCT13A-PCT13I, PCT13
    - Household population broken down by race/ethnicity, age, and sex. Does not have all non-Hispanic racial categories necessary for this project, for instance, does not have a table for Black non-Hispanic.
- PCT18A-PCT18I, PCT18
    - GQ population broken down by race/ethnicity, age, and sex. Does not have all non-Hispanic racial categories and age is restricted to three ranges (0-17, 18-64, and 65 and over).
- P12A-P12O, P12
    - Total population broken down by age, race/ethnicity, and sex. Includes more race/ethnicity and age categories than the first two groups of tables but isn't broken down by household/GQ type.

To convert the data into a usable form

1. for the PCT13A-PCT13I, PCT13 Tables:
    a. impute Hispanic origin for Black, Native American, Some Other Race, and Two or More Races.
    b. impute non-Hispanic multi-racial categories such as non-Hispanic White x Some Other Race, non-Hispanic Black x Some Other Race, and non-Hispanic Native American x Some Other Race.
    c. translate population totals to valid PEP race/ethnicity categories, more information can be found in (PEP, 2012).
2. for the PCT18A-PCT18I, PCT18 Tables:
    a. perform the same processes as in 1. but also impute the 18-44 and 45-64 age ranges.

To deal with 1a, we impute Hispanic origin using the P12 collection of tables and for 1b we impute the multi-racial categories using PL94 data. Then, we make the race/ethnicity categories consistent with those found in PEP, e.g. we distribute categories involving Some Other Race to the categories listed in Section 1, to obtain our final household population estimates. Everything for 2 works along the same lines as in 1, except we also impute the 18-44 and 45-64 age groups using PEP data. In this paper, the final 2020 population results are denoted by $POP20$ along with appropriate subscripts to indicate geography, demographic, and household or GQ type.

2.2 2022 Population Estimates

We calculate growth rates by demographic and household/GQ type using Vintage 2022 PEP data

$$gr_{c,j,22} \quad = \quad \frac{PEPPOP2022_{c,j}}{PEP2020BASE_{c,j}},$$

and apply the growth rates to the $POP20$ estimates.[1] Results are summed separately for households and each GQ type at the tract-level to create population estimates $POP22$. The $POP22$ values are adjusted, or raked, so that their sum is equal to PEP totals at the county-level for households and state-level for each of the GQ types. For instance, in the case of households, we define

$$rk_c \quad = \quad \frac{PEPPOP2022_c}{\sum_{t \in c} POP22_t}$$

for a given county c and multiply $POP22_t$ by rk_c to obtain a raked estimate $POPRK22_t$ for each tract t in c. The raked values become our final 2022 population estimates. In the remainder of the document, when we refer to POP or $POPUNI$, we are referring to these raked values.

3. Estimation Layers

3.1 Tract Universe

The 2022 Census geography vintage contains 84,415 tracts which we split into two groups: residential and non-residential. Of the 84,415 total tracts, 83,804 are classified as residential, i.e. the tract's estimated in-universe population is non-zero, and the remaining 611 tracts are assumed to have no in-scope residents. In Sections 4-6, unless otherwise noted, by tract we will mean a residential tract.

3.2 Layer Definitions

Only residential tracts are assigned to *a* layers, and construction of the *a* layers is based on two criteria:

- tract-level population density
- concentration of urban population (UCI) within a given tract and those near it
    - this value is the weighted (by distance) sum of a tract's population and those tracts near it, scaled so that 100-250 is moderate concentration, and 250+ is high.

[1] A small amount of reconciliation is done, e.g. we account for demographic groups added since the Decennial Census to the PEP data by splitting a county demographic total among constituent tracts.

Using the two criteria, we define four urbanization strata[2]:

- UStrat1 consisting of tracts with UCI < 100
- UStrat2 consisting of tracts with UCI 100-250 and high-density proportion less than 0.5
- UStrat3 consisting of tracts with UCI 100-250 and high-density proportion greater than or equal to 0.5
- UStrat4 consisting of tracts with UCI > 250.

The a layers are the urbanization strata broken down by Census Division. Since there are nine Census Divisions and four urbanization strata there are 36 total a layers. Note that a given a layer can cross state and county boundaries and are not necessarily contiguous.

4. Component of Social Vulnerability Estimation

We estimate SV group membership for households and the seventh GQ type together and separate from the remaining GQ types. Define the ratio $r_{a,j,dg,HH7}$ as the proportion of individuals within an a layer a and demographic $j \in AR$ residing in a household or appropriate GQ type with zero, one or two, or three or more components of social vulnerability:

$$r_{a,j,dg,HH7} \quad = \quad \frac{\sum_{y \in ACS_{a,j,dg,HH7}} w_y}{\sum_{y \in ACS_{a,j,HH7}} w_y}.$$

We use the subscript $HH7$ for the ACS data to indicate it is subset to respondents living in households and appropriate non-institutional GQs. The proportion $r_{a,j,dg,HH7}$ is broken up into a marginal and conditional probability using Bayes' Theorem. The marginal is modelled in a linear regression within each a layer and then the joint values are reconstructed.[3] Denote the final modelled value of $r_{a,j,dg,HH7}$ as $\tilde{r}_{a,j,dg,HH7}$. For a given tract t within an a layer a, and $j \in AR$ we estimate the number of individuals in group dg as the following:

$$\hat{Y}_{t,j,dg,HH7} \quad = \quad \tilde{r}_{a,j,dg,HH7} * POP_{t,j,HH7}.$$

Estimation for the remaining GQs is similar (we estimate GQ types 3. and 4. together and GQ type 6. by itself) and yields values $\hat{Y}_{t,j,dg,34}$ and $\hat{Y}_{t,j,dg,6}$. Summing over demographic groups, we have

$$\hat{Y}_{t,0} \quad = \quad \sum_{j \in AR} \hat{Y}_{t,j,0,HH7} + \hat{Y}_{t,j,0,34} + \hat{Y}_{t,j,0,6} \qquad\qquad \hat{R}_{t,0} \quad = \quad \frac{\hat{Y}_{t,0}}{POPUNI_t}$$

[2] The East South Central Division has slightly lowered thresholds, otherwise the populations for some of the estimation layers would be unacceptably small.

[3] The modelling helps to account for differences between ACS population totals and has a limited impact on the final results.

$$\begin{aligned}\hat{Y}_{t,12} &= \sum_{j\in AR} \hat{Y}_{t,j,12,HH7} + \hat{Y}_{t,j,12,34} + \hat{Y}_{t,j,12,6} & \hat{R}_{t,12} &= \frac{\hat{Y}_{t,12}}{POPUNI_t} \\ \hat{Y}_{t,3} &= \sum_{j\in AR} \hat{Y}_{t,j,3,HH7} + \hat{Y}_{t,j,3,34} + \hat{Y}_{t,j,3,6} & \hat{R}_{t,3} &= \frac{\hat{Y}_{t,3}}{POPUNI_t}.\end{aligned}$$

We let $r_{t,0}, r_{t,12}$, and $r_{t,3}$ denote the ACS estimates for the proportion of individuals within tract t who have zero components of social vulnerability, one or two components of social vulnerability, and three or more components of social vulnerability, respectively.

5. Composite Estimators

5.1 Direct Estimate Variance

Estimates for the variance of $r_{t,dg}$, with $dg = 0{,}12$, or 3, can be calculated directly from the ACS microdata using replicate weights. To improve the reliability of these variance estimates, we smoothed them using a generalized variance function (GVF). The formula for CRE's GVF is

$$r\_gvfvar_{t,dg} \quad = \quad C\hat{R}_{t,dg}\left(1-\hat{R}_{t,dg}\right)n_t^q$$

where n_t is the unweighted ACS sample size of tract t and C and q are parameters to be estimated. Note that n_t varies by tract but not values of dg, i.e. n_t is that same for all values of dg within a given tract. To estimate values for C and q we fit the following regression for all tracts with a sample size greater than 25,

$$\log\left(r\_dirvar_{t,dg}\right) - \log\left(\hat{R}_{t,dg}(1-\hat{R}_{t,dg})\right) \quad = \quad \log C + q\log n_t.$$

We used the estimates for q and C to calculate a GVF variance for all tracts, even those with a sample size of 25 or smaller. Table 1 compares the GVF standard errors with the original standard errors from the ACS for the three different SV groups. The distributions of GVF standard errors have means and medians similar those of the ACS, but the standard error of the distributions tends to be smaller.

Table 1: GVF and ACS Standard Errors by SV Groups

Group	Mean		Median		Standard Error	
	ACS	GVF	ACS	GVF	ACS	GVF
0	0.112	0.102	0.111	0.100	0.038	0.021
12	0.120	0.108	0.115	0.105	0.037	0.024
3	0.089	0.089	0.080	0.086	0.050	0.023

Source: 2022 Community Resilience Estimates.

5.2 MSE Estimation

To estimate the mean squared error (MSE) of our synthetic estimators $\hat{R}_{t,dg}$, we used a method-of-moments (MoM) approach, roughly following the derivation outlined in (Rao & Molina, p. 43-44). If T is an arbitrary collection of tracts T, N_T in total, then our basic MoM equation for MSE is:

$$MSE_{T,dg} \quad = \quad \frac{1}{N_T}\left(\sum_{t\in T}\left(\hat{R}_{t,dg} - r_{t,dg}\right)^2\right) - \frac{1}{N_T}\sum_{t\in T} r\_gvfvar_{t,dg}.$$

A MoM variance estimator can be negative, and the choice of T, i.e. the layer of aggregation, must be made with care. Additionally, for a fixed tract t, the actual number of individuals with zero, one or two, and three or more social vulnerability components have a multinomial relationship and this places additional constraints on $MSE_{T,dg}$. For the time being we'll ignore the selection of T, but note that to address the other concerns

- we assume the MSE for a given SV-group proportion estimate, $\hat{R}_{t,dg}$, is proportional to $\hat{R}_{t,dg}\left(1 - \hat{R}_{t,dg}\right)$, and that the constant of proportionality can be assumed stable over a wide range of SV-groups and tracts.
- we calculate the multiplicative constant of proportionality at an aggregate level, averaging across both SV-groups and tracts.

Under this strategy, we modified the previous equation to average across SV-groups as well as tracts:

$$MSE_T \quad = \quad \frac{1}{3N_T}\sum_{dg=0,12,3}\left(\sum_{t\in T}\left(\hat{R}_{t,dg} - r_{t,dg}\right)^2\right) - \frac{1}{3N_T}\sum_{dg=0,12,3}\sum_{t\in T} r\_gvfvar_{t,dg}.$$

The estimated constant, denoted F_T, was calculated as

$$\begin{aligned} PmP_T &= \frac{1}{3N_T}\sum_{dg=0,12,3}\sum_{t\in T}\hat{R}_{t,dg}\left(1 - \hat{R}_{t,dg}\right) \\ F_T &= \frac{MSE_T}{PmP_T} \end{aligned}$$

so that for an individual tract t in T and each dg, $MSE_{t,dg} = F_T\hat{R}_{t,dg}\left(1 - \hat{R}_{t,dg}\right)$. Returning to the choice of T, we created two national urbanization layers: UStrat1 collapsed with UStrat2 and UStrat3 collapsed with UStrat4. The layers were chosen empirically, and result in estimates which are stable.

5.3 Composite Formula

Our final estimator is a shrinkage, or composite, estimator which is a weighted average of $\hat{R}_{t,dg}$ and $r_{t,dg}$ for each tract t and value of dg:

$$\begin{aligned} wt_{t,dg} &= \frac{r\_gvfvar_{t,dg}}{r\_gvfvar_{t,dg} + MSE_{t,dg}} \\ \tilde{R}_{t,dg} &= w_{t,dg}\hat{R}_{t,dg} + \left(1 - w_{t,dg}\right)r_{t,dg}. \end{aligned}$$

See (Rao & Molina, 2015, p. 57-58) for a derivation of the weight formula. For relatively large values of $r\_gvfvar_{t,dg}$, the synthetic estimator is weighted more heavily while for smaller values of $r\_gvfvar_{t,dg}$ the opposite is true.

We quantify the uncertainty of our estimates $\tilde{R}_{t,dg}$ using

$$\begin{aligned} MSE\left(\tilde{R}_{t,dg}\right) &= w_{t,dg}MSE_{t,dg} & MSE\left(\tilde{Z}_{t,dg}\right) &= MSE\left(\tilde{R}_{t,dg}\right)POPUNI_t^2 \\ RMSE\left(\tilde{R}_{t,dg}\right) &= \sqrt{MSE\left(\tilde{R}_{t,dg}\right)} & RMSE\left(\tilde{Z}_{t,dg}\right) &= MSE\left(\tilde{R}_{t,dg}\right)POPUNI_t \\ MOE\left(\tilde{R}_{t,dg}\right) &= RMSE\left(\tilde{R}_{t,dg}\right) * 1.645 & MOE\left(\tilde{Z}_{t,dg}\right) &= MOE\left(\tilde{R}_{t,dg}\right)POPUNI_t \end{aligned}$$

where $\tilde{Z}_{t,dg} = \tilde{R}_{t,dg}POPUNI_t$. As seen in Table 2 and Table 3, the direct, indirect, and composite estimates are all very similar but the CVs for the composite estimates are better than those for the ACS and the indirect estimates. The statistics are calculated over all tracts with ACS sample.[4]

Table 2: Comparison of Estimates by SV Group

Group	Mean			Median		
	ACS	Indirect	Comp.	ACS	Indirect	Comp.
0	0.338	0.342	0.343	0.332	0.344	0.339
12	0.440	0.443	0.442	0.434	0.438	0.436
3	0.222	0.215	0.216	0.184	0.209	0.198

Source: 2022 Community Resilience Estimates.

[4] To better gauge the variance reduction, we use the indirect estimate in the denominator of all the CVs.

Table 3: Comparison of CVs by SV Group

Group	P25			Median			P75		
	ACS	Indirect	Comp.	ACS	Indirect	Comp.	ACS	Indirect	Comp.
0	0.249	0.283	0.188	0.296	0.332	0.220	0.357	0.381	0.257
12	0.215	0.247	0.164	0.240	0.267	0.178	0.269	0.287	0.192
3	0.365	0.405	0.273	0.416	0.466	0.308	0.473	0.518	0.343

Source: 2022 Community Resilience Estimates.

5.4 Zero Values

Non-residential tracts are tracts that have an estimated population of zero. In earlier versions of CRE, e.g. CRE 2021, county-level values were used as rate estimates for these tracts, while all count estimates were set to zero. For CRE 2022, in addition to setting count estimates to zero, non-residential tracts have their rate estimates set to zero as well.

5.5 Tract-County Relationship

To obtain a point estimate $\tilde{Z}_{c,dg}$ for a given county c, we can simply sum over the county's tracts,

$$\tilde{Z}_{c,dg} = \sum_{t \in c} \tilde{Z}_{t,dg}$$

Geographically close tracts will tend to be correlated, and, in particular, tracts within a given county are likely correlated so we can't assume

$$MSE(\tilde{Z}_{c,dg}) = \sum_{t \in c} MSE(\tilde{Z}_{t,dg})$$

For the purposes of determining margins of error at the county-level, CRE assumes a fixed correlation ρ for all tracts within a given county and that the mean squared error of all tracts within a given county are roughly equal. The value of ρ, and more generally the spatial correlation between tracts, is a topic of ongoing research. For CRE versions 2021 and earlier, ρ was set to .4 and for 2022 it has been set at .2.

With all these assumptions in mind, we define

$$\begin{aligned} MSE(\tilde{Z}_{c,dg}) &= \sum_{t \in c} MSE(\tilde{Z}_{t,dg})\,(1 + \rho(N_c - 1)) \\ RMSE(\tilde{Z}_{c,dg}) &= \sqrt{MSE(\tilde{Z}_{c,dg})} \\ MOE(\tilde{Z}_{c,dg}) &= RMSE(\tilde{Z}_{c,dg}) * 1.645. \end{aligned}$$

with N_c the number of tracts in c. Similar to the case for tracts, we divide $MSE(\tilde{Z}_{c,dg})$, $RMSE(\tilde{Z}_{c,dg})$, and $MOE(\tilde{Z}_{c,dg})$ by the appropriate power of $POPUNI_c$ to obtain $MSE(\tilde{R}_{c,dg})$, $RMSE(\tilde{R}_{c,dg})$, and $MOE(\tilde{R}_{c,dg})$. For higher geographies, we assume mean squared errors are summable, i.e. county estimates are independent and state estimates are independent.

6. Ranking

We rank counties and tracts based on the proportion of individuals within the county or tract who have three or more components of social vulnerability. A rank of one indicates the county/tract has the lowest proportion of individuals with three or more components of social vulnerability and larger rankings indicate higher proportions. Dividing the rankings by the total number of tracts/counties being ranked and multiplying the results by 100 translates the rankings into percentiles. Using bootstrapping principles based on those outlined in Wright, T. Klein, M. and Wieczorek (2014), we obtain lower/upper bounds for the rankings and percentiles.

6.1 County

For a county c and $1 \leq i \ \leq 100{,}000$ we draw a random variable $\tilde{R}_{c,3,i} \sim N(\tilde{R}_{c,3}, MSE(\tilde{R}_{c,3}))$. For each i, we rank the counties according to $\tilde{R}_{c,3,i}$. This gives us 100,000 different rankings for each county, from which we obtain upper and lower bounds for the county's ranking. These rankings and bounds are converted to percentiles, and a subset of counties with the largest percentiles are published.

6.2 Tract

Tracts are similar to counties, except we also model the correlation between tracts of a given county. For each county c and $1 \leq i \ \leq 100{,}000$ we draw a vector of rates $\vec{\boldsymbol{r}}_{c,i} \sim MVN(\vec{\boldsymbol{r}}_c, \Sigma)$ where $\vec{\boldsymbol{r}}_c$ is a vector with each component a value $\tilde{R}_{t,3}$ for a tract t within c and Σ is a covariance matrix constructed from the values for $MSE(\tilde{R}_{t,3})$ for all tracts within c and an assumed between-tract correlation. Just as with the counties, we develop rankings (percentiles) for each i, from which we obtain upper and lower bounds for each tract's rank (percentile).

7. References

Population Estimates Program, U.S. Census Bureau. (2012). Modified Race Summary File Methodology Statement. Retrieved 4/26/2023 from: https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/modified-race-summary-file-method/mrsf2010.pdf.

Rao, J.N.K. & Molina, I. (2015). *Small Area Estimation* (Second Edition). John Wiley & Sons Incorporated.

U.S. Census Bureau. (2023). Disclosure Avoidance and the 2020 Census: How the TopDown Algorithm Works. Retrieved 11/28/2023 from: https://www2.census.gov/library/publications/decennial/2020/census-briefs/c2020br-04.pdf

Wright, T. Klein, M. and Wieczorek, J. (2014). Ranking Populations Based on Sample Survey Data. *Research Report Statistics # 2014-07*. Center for Statistical Research and Methodology, U.S. Census Bureau.