Brookings Metro

METHODOLOGY | SEPTEMBER 2022

REDUCING POVERTY WITHOUT COMMUNITY DISPLACEMENT:

INDICATORS OF INCLUSIVE PROSPERITY IN U.S. NEIGHBORHOODS

ROHIT ACHARYA AND RHETT MORRIS

Methodology appendix

The analyses produced for this paper focus on three research questions that look at different aspects of neighborhoods with concentrated poverty:

- **Current state:** How do neighborhoods with concentrated poverty compare to other census tracts using the most recently released data? (Discussed in Methodology Part B below.)
- **Changes over time:** How have neighborhoods with concentrated poverty changed when comparing data from two points in time: 2000 and 2015? (Discussed in Methodology Part C below.)
- **Indicators of large decreases in poverty rates without community displacement:** How were neighborhoods that experienced large decrease in poverty rates with no community displacement different from other poor communities when comparing data from two points in time: 2000 and 2015? (Discussed in Part Methodology D below.)

The following pages document the full methodology used to formulate the analytical results shared in this report.

Methodology Part A: Defining key concepts, data aggregation, and geographic variable/feature creation

IDENTIFYING CENSUS TRACTS AS 'NEIGHBORHOODS WITH CONCENTRATED POVERTY'

Before answering our three research questions, we must first define "neighborhoods with concentrated poverty" and collect the data needed for our analyses.

This paper studies neighborhoods with concentrated poverty, which we define as local communities where at least 30% of residents live in households with incomes below the federal poverty threshold. There are multiple definitions for neighborhoods with "concentrated poverty" or "high poverty."[1] Some researchers define these communities as those with at least 40% of residents living in households with incomes that are less than the federal poverty threshold. Others, such as the recent study on household finances from the Pew Charitable Trusts, use lower definitions of 25% or even 20%.[2] This report follows the example of studies using the threshold of 30%, since analyses show that significant differences in economic and health outcomes can already be seen when comparing neighborhoods that are above and below this threshold, as illustrated on pages 9 to 13.

It should also be noted that the use of federal poverty thresholds based on pre-tax income as a measure for poverty has a number of limitations. For example, it does not account for differences in access to non-cash benefits, such as health insurance, or post-tax income subsidies. However, this is the only measure available at the neighborhood or census tract level for the time periods we chose to study. We believe it to be useful in large part because of the significant differences in outcomes that can be seen across communities when using this measure, as evidenced by the analyses on pages 9 to 13.

In order to study these communities, we use census tracts as the primary geographic unit of analysis in this paper, based on the Census Bureau's 2010 delineation. The average census tract in a residential area contains around 4,000 people, making it roughly equivalent to a large neighborhood.[3] For this reason, this report uses the terms "tract" and "neighborhood" interchangeably.

Census tracts are useful for performing longitudinal analysis because they are "relatively permanent statistical subdivisions of a county" whose boundaries are often consistent over time, and because many federal agencies and other data providers compile and release information at the census tract level. Smaller local geographic subdivisions used by the Census Bureau, such as blocks and block groups, are not as consistent in terms of boundaries over successive years and have much less available data on their attributes.

All the neighborhoods with concentrated poverty considered in the analyses of this report are urban census tracts that meet two criteria:

- **Metropolitan:** All are in counties that are a part of a metropolitan statistical area (MSA). Metropolitan statistical areas are the federal government's designation for urban regions.
- **Residential areas:** All have at least 1,000 people living in them per square mile of land.[4] This removes census tracts in metropolitan areas that are primarily industrial or commercial; made up of park space, water, or drainage areas; or have small residential populations.

When conducting comparative analyses, we used two additional criteria:

- **Minimum size:** Census tracts or neighborhoods with fewer than 500 total residents are removed from the

sample to prevent analyses from being skewed by communities with especially small populations.

- **Unskewed by college student populations:** Areas with high proportions of college students living in them are also removed, which is consistent with the Census Bureau's policy of not counting students living on college and university campuses in calculations of local poverty rates.[5]

AGGREGATING DATA ON THE ATTRIBUTES OF NEIGHBORHOODS WITH CONCENTRATED POVERTY AND OTHER CENSUS TRACTS

In order to conduct analyses on the neighborhoods with concentrated poverty that we identified, we collected data on the attributes of census tracts at different points in time, starting with the most current data available and in one case, going back as far as 1990.

All the data used to answer the project's three research questions are inventoried in the rest of this section. The information we collected on these attributes can be divided into two categories: pre-aggregated data and fixed-location data. An explanation of each of these types of data as well as examples used in our analyses are found below.

COLLECTING AND PROCESSING PRE-AGGREGATED DATA

The first type of data collected for this project is information already aggregated to the specific boundaries of census tracts, school districts, counties, and states. The values for each listed variable in this category were taken directly from the sources below.

The decennial census is a survey conducted by the U.S. Census Bureau once every 10 years to provide an official count of all persons living in the country as well as a limited amount of information about the nation and its residents, such as a person's age, sex, race, and whether they own or rent their home. In previous decades, it also provided more detailed information about additional topics by sampling a smaller percentage of households for a long-form survey, which was ultimately replaced by the American Community Survey. We use the data from this long-form survey to provide information on households and residents in the year 2000.

The American Community Survey is an ongoing survey conducted by the U.S. Census Bureau that provides a wide variety of information about the nation and its residents by sampling approximately 3.5 million households each year. It asks about topics not on the decennial census, such as education, employment, internet access, and transportation. Our analysis uses the five-year estimates from the survey, which uses samples over a 60-month period, and as such, has a higher degree of accuracy about underlying demographics than the one-year estimates, which only aggregate the last year. To approximate the demographics of a neighborhood in 2015, we used the five-year survey spanning from 2013 to 2017, where 2015 is the midpoint. To approximate the current state of neighborhoods with concentrated poverty and make comparisons with other census tracts, we used the most recently released data: the five-year survey spanning from 2015 to 2019.

The data we used from the above two sources were aggregated at the census tract and county levels. More information on these surveys can be found at https://www.census.gov/programs-surveys/acs/about/acs-and-census.html

- Land area
- Population
- Number of households
- Number of total occupied housing units
- Number of occupied housing units by tenure of year householder moved into unit
- Proportion of residents enrolled in colleges or universities
- Number of residents in categories based on age group, race and ethnicity, education, and household income
- Average household income
- Number of residents living in households with incomes below the poverty level (i.e., the poverty rate), as well as those with incomes near the poverty threshold and substantially above it

- Number of children living in households with incomes below the poverty level
- Average household rent
- Number of residents by employment level
- Number of children enrolled in public school
- Number of children in households without internet access
- Number of adults working in occupational group categories or serving in the military
- Number of residents who are disabled
- Number of housing units constructed by decade since 1950
- Number of three- and four-year-olds enrolled in preschool
- Value of owner-occupied homes
- Number of total housing units
- Number of owner-occupied housing units
- Number of vacant housing units
- Number of residents who are self-employed
- Number of residents with varying levels of commute times
- Number of children without health insurance
- Number of adults without health insurance
- Number of residents born outside the United States
- Number of households receiving SNAP benefits
- Number of single-parent households
- Number of 16- to 19-year-olds enrolled in school
- Number of 16- to 19-year-olds employed at a job
- Number of adults who are working full time or part time and have household incomes below the poverty line

Note: Data on high-income occupations shared on page 12 came from the 2019 five-year American Community Survey release. We considered the following occupations to be high-income:

- Computer and mathematical occupations
- Management occupations
- Life, physical, and social science occupations
- Architecture and engineering occupations
- Business and financial operations occupations
- Health diagnosing and treating practitioners and other technical occupations

THE U.S. SMALL-AREA LIFE EXPECTANCY ESTIMATES PROJECT is a partnership of the National Center for Health Statistics, the Robert Wood Johnson Foundation, and the National Association for Public Health Statistics and Information Systems. Our calculations on page 9 were created using an unweighted average from this data for each urban tract in the United States.

The data we used from this source was aggregated at the census tract level. More information on the project can be found at https://www.naphsis.org/usaleep

- Projected life expectancy of children born between 2010 and 2015

OPPORTUNITY INSIGHTS is a non-partisan, not-for-profit organization located at Harvard University. The organization's Opportunity Atlas uses anonymized longitudinal data for people born between 1978 and 1983 to estimate outcomes for this population based on household income percentile rank during their childhood and census tract of residency during childhood using 2010 census tract delineations.

For our analysis on page 9 and 10, we examined two outcomes for these populations: mean percentile rank for household income compared to the national distribution in 2014-2015 and the fraction incarcerated as of April 1, 2010. It should be noted that in order to protect privacy, the estimates provided are not exact. According to the Opportunity Atlas "a small amount of noise is added to each of the estimates; this noise is typically less than one-tenth the standard error of the estimate itself."

For the household income ranks, we then converted the average mean percentile ranks to 2015 dollar values using a crosswalk provided by Opportunity Insights.

More information on the estimates can be found at https://opportunityinsights.org/paper/the-opportunity-atlas/

- Young adult earnings based on childhood household income
- Proportion of young adults who are incarcerated based on childhood household income

U.S. DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT'S AGGREGATED USPS ADMINISTRATIVE DATA ON ADDRESS VACANCIES provides aggregate vacancy and no-stat counts of residential and business addresses that are collected by postal workers and submitted to HUD on a quarterly basis. The U.S. Postal Service collects this data to facilitate efficient mail delivery. While occupancy status is recorded, USPS does not capture any information about the nature of the vacancy or the address itself, other than whether it is a residential or business address.

Our business vacancy analysis uses HUD-aggregated USPS administrative data on address vacancies from the third quarter of 2020. We calculated what percentage of businesses were categorized as "vacant" or "no-stat."

The data we used from this source was aggregated at the census tract level. More information on the project can be found at https://www.huduser.gov/portal/datasets/usps.html

- Number of businesses addresses
- Number of vacant businesses addresses

HEALTH PROFESSIONAL SHORTAGE AREAS are designated by the U.S. Health Resources & Services Administration as having shortages of primary medical care, dental, or mental health providers and may be geographic (a county or service area), populations (e.g., low-income, Medicaid-eligible), or facilities (e.g., federally qualified health center, state or federal prisons).

The data we used from this source was aggregated at the county, census tract, and census place levels. Counties are supersets of tracts, so if a county is designated as having a health professional shortage, then we designated every tract within the county as also having a shortage. Census places correspond to administrative jurisdictions such as cities, which may not cleanly subset into tracts. In these cases, we performed a spatial join to identify which tracts' centers of population intersected with a designated place, and those that did were tagged as also having a shortage. More information on this data can be found at https://data.hrsa.gov/tools/shortage-area/mua-find

- Designation of tracts as medically underserved areas

THE COMMUNITY REINVESTMENT ACT (CRA) requires certain lending institutions to make annual public disclosures of their small business, small farm, and community development lending activity. **The CRA Aggregate and Disclosure** system provides access to each lending institution's individual disclosure statement, aggregate tables covering the lending activity of all institutions subject to CRA for each MSA and non-MSA portion of each state, and national aggregate tables covering the lending activity of all institutions for the entire nation.

Our methodology looked at the number of loans for small businesses with revenue under $1 million. The data we used from this source was aggregated at the census tract level. More information on this data can be found at https://www.ffiec.gov/craadweb/naaginfs.htm

- Number of small business loans

PICTURE OF SUBSIDIZED HOUSEHOLDS DATA, provided by the U.S. Department of Housing and Urban Development's Office of Policy Development and Research, describes households living in HUD-subsidized housing in the United States.

The data we used from this source was aggregated at the census tract level. More information on this data can be found at https://www.huduser.gov/portal/datasets/assthsg.html

- Number of housing units within public housing developments

THE BUREAU OF ECONOMIC ANALYSIS produces economic statistics that enable government and business decisionmakers, researchers, and the American public to follow and understand the performance of the nation's economy. To do this, it collects source data, conducts research and analysis, develops and implements estimation methodologies, and disseminates statistics to the public. This includes national, regional, industry, and international accounts that present essential information on key issues such as economic growth, regional economic development, interindustry relationships, and the nation's position in the world economy.

The data we used from this source was aggregated at the MSA level, which was easy to attribute to individual census tracts since each tract is in only one single MSA. More information on this data can be found at https://www.bea.gov/

- Metropolitan statistical area gross domestic product

THE UNIFORM CRIME REPORTING (UCR) PROGRAM generates statistics on crimes and law enforcement. It includes data from more than 18,000 city, university/college, county, state, tribal, and federal law enforcement agencies. Agencies participate voluntarily and submit their crime data either through a state UCR program or directly to the FBI's UCR Program.

The data we used from this source was aggregated at the county level, which was easy to attribute to individual census tracts since each tract is in only one single county. More information on this data can be found at https://www.fbi.gov/services/cjis/ucr

- County homicide rate

NOTE: In a small number of counties, such as Miami-Dade, Fla., data that was missing from the UCR database was supplemented with information from local sources.

WASHINGTON CENTER FOR EQUITABLE GROWTH researchers released datasets in 2016 of historical state and sub-state minimum wage levels for the United States.

The data we used from this source was aggregated at the state and county level, which was easy to attribute to individual census tracts since each tract is in only one single county and state. More information on this data can be found at https://equitablegrowth.org/working-papers/historical-state-and-sub-state-minimum-wage-data/

- Minimum wage rates by state and county

DEPARTMENT OF EDUCATION'S NATIONAL CENTER FOR EDUCATION STATISTICS provides data at the local education agency level, which can then be crosswalked to tracts using a school district geographic relationship file.

We defined low-performing schools as having high school graduation rates far below the national average. To do our analysis, we used adjusted cohort graduation rates provided for the 2017-18 school year. We then calculated the weighted average high school graduation rate for every tract. Any tract that had a graduation rate below the national average was considered to be in a low-performing school district.

At this point, we then used 2019 five-year American Community Survey data to determine the number of children residing in a tract and the percentage of children attending public school, since our statistics are for public school only. This provided an approximate figure of the percentage of children in low-performing school districts.

The data we used from this source was aggregated at the school district level, which we attributed to individual census tracts using National Center for Education Statistics' crosswalk files. The crosswalk files provide information of the land and water area overlap between a school district and census tracts. Using this, we allocated the school district to the tracts based on how much of the district overlaps with a census tract. Most tracts are covered by only one school district, so they take on the attributes of that school district. In cases in which there were multiple school districts overlapping a tract, we take weighted averages of the school districts' attributes based on the number of students from each district that would

fall into that tract based on the overlap calculated earlier. More information on this data can be found at https://nces.ed.gov/

- School district high school graduation rates

DEPARTMENT OF EDUCATION'S CIVIL RIGHTS DATA COLLECTION PROGRAM collects data on key education and civil rights issues in our nation's public schools, including student enrollment and educational programs and services—most of which is disaggregated by race/ethnicity, sex, limited English proficiency, and disability.

The data we used from this source was aggregated at the school district level, which we attributed to individual census tracts using National Center for Education Statistics' crosswalk files. The crosswalk files provide information of the land and water area overlap between a school district and census tracts. Using this, we allocated the school district to the tracts based on how much of the district overlaps with a census tract. Most tracts are covered by only one school district, so they take on the attributes of that school district. In cases in which there were multiple school districts overlapping a tract, we take weighted averages of the school districts' attributes based on the number of students from each district that would fall into that tract based on the overlap calculated earlier. More information on this data can be found at https://ocrdata.ed.gov/

- Number of students
- Number of police and security officers
- Number of health and social workers per student
- Number of teachers

LONGITUDINAL ANALYSES OF PRE-AGGREGATED DATA

Conducting longitudinal analysis of census tracts was one of the primary challenges we had to overcome in our analysis using pre-aggregated data. Although tracts are meant to be a relatively permanent geographic subdivisions of a county, tracts are split, consolidated, or changed in other ways from the previous boundaries to reflect population growth or decline, geographic changes, or updates in road layouts. These changes can make it difficult to follow the progression of a single tract across successive censuses.

One solution to this problem would be to use areal attribution to compare tracts with changing boundaries over time. For example, if 60% of the land of Tract X from 2000 falls within the boundaries of Tract Y from 2010, then you could assume a 60% attribution of Tract X to Tract Y. The issue with this attribution method is that our research is about people, not land, and population is not uniformly distributed across land.

Therefore, population clusters in varying parts of the tract would not be allocated properly over time. We used the Longitudinal Tract Data Base (LTDB) as our main crosswalk when comparing data from one census survey to the next. Created by Logan, Shultz, and Xu, the LTDB uses both "areal and population interpolation as well as ancillary data about water-covered areas" to provide a crosswalk to link census tracts from 1970 onwards to the most recent boundaries. The LTDB is accessible online here: https://s4.ad.brown.edu/Projects/Diversity/researcher/LTDB.htm

COLLECTING AND PROCESSING FIXED-LOCATION DATA

The second type of data collected for this project is data that provided specific locations using fixed points. This could be for individual businesses, organizations, or facilities in the form of addresses or points of longitude and latitude; or for the boundaries of a park, road, or area designated as specifically underserved in the form of shapefiles.

In some cases, the fixed point data was aggregated by counting the number of instances of the variable occurring within a census tract. For example, data on the presence of interstate highways was calculated by examining whether an interstate ran through a tract or made up one of its boundaries. Similarly, data on gun violence is aggregated by counting the instances that occur within a tract.

In other cases, we looked both within and nearby tracts to see if residents had exposure or access to the measured entity. For example, research has demonstrated that living within 1 mile of a toxic release site is associated with negative health outcomes. In order to assess residents' exposure to toxic releases, it is more appropriate to measure whether a tract is near one of these sites, rather than only measuring if it contains one within its borders.

Similarly, branches of public libraries are not expected to serve only a single tract of just 4,000 or so individuals. To assess residents' access to libraries and their services, it is more appropriate to measure if a tract is near a public library branch (e.g., within a half-mile, 1 mile, etc.) rather than only measuring if it contains one within its borders.

To overcome this issue for single address locations, we superimposed "buffer zones" of varying radii to each address using specific distances related to exposure risk or residential access. We then calculated whether that resulting circle overlapped with the internal population centroid of a tract, and if so, that location was attributed to that tract. This can generally be translated into useful measures in the real world. For example, using a 1-mile radius from library branches, we would be able to calculate which tracts fall within a 20-minute walk. A desired result of this approach is that multiple tracts can be attributed exposure or access to a single entity located at one address.

For determining accessibility to places such as parks, which are not at a particular location (i.e., a point) but span an area, we use a similar process to determine access. We added a buffer to the outline of the shape based on the distance we wanted to measure. For example, a half-mile buffer would be added to the borders of a park to determine locations that are within a 10-minute walk of the park. Then we calculate if that buffered shape overlaps with the population centroid of a tract to determine which tracts it would be attributed to.

To determine the location of the city center of the principal city in each metropolitan area, we used the **GOOGLE MAPS API** to geolocate a principal city's city hall and downtown area using textual search. This data is current as of November 2021. Given that the city hall is traditionally in a city's central business district, we can use that as a proxy for the center of a city. We also corroborate this location by searching for "downtown", which Google Maps lists as an approximate region of a city for each principal city. If the distance between the city hall and downtown results is less than 1 mile, we can assume that we have identified the city center and we use the city hall result as the center point for the city, since it most likely results in a rooftop geolocation instead of an approximation. If the distance is more than 1 mile, we perform a manual investigation to approximate where the downtown area truly should be. In almost all of these latter cases, the downtown result provided a more accurate estimate for the city center than the city hall location.

For these cities we overrode the city hall return with the downtown return:

Anniston, Ala.

Cape Coral, Fla.

Charleston, S.C.

Dayton, Ohio

Eau Claire, Wis.

Elmira, N.Y.

Hanford, Calif.

Kalamazoo, Mich.

Las Vegas

Miami

Muskegon, Mich.

Ogden, Utah

Rapid City, S.D.

Rocky Mount, N.C.

Springfield, Ill.

Virginia Beach, Va.

Youngstown, Ohio

Yuba City, Calif.

For Palm Bay, Fla., we used the downtown area for Melbourne, Fla.

- Location of downtown or central business district of the primary city in each metropolitan area

THE ENVIRONMENTAL PROTECTION AGENCY'S TOXICS RELEASE INVENTORY tracks the industrial management of toxic chemicals that may cause harm to human health and the environment.

We used TRI data for the reporting year 2019 to determine all sites with toxic releases. More information on this data can be found at https://www.epa.gov/toxics-release-inventory-tri-program/tri-researchers#Obtaining%20TRI%20Data%20for%20Research%20Purposes

- Locations of toxic release facilities (assessed using buffer zones)

TIGER/LINE SHAPEFILES are extracts of selected geographic and cartographic information from the Census Bureau's Master Address File (MAF)/ Topologically Integrated Geographic Encoding and Referencing (TIGER) Database (MTDB). The shapefiles include information for the 50 states, Washington, D.C., Puerto Rico, and the Island areas (American Samoa, the Commonwealth of the Northern Mariana Islands, Guam, and the United States Virgin Islands). The shapefiles include polygon boundaries of geographic areas and features, linear features including roads and hydrography, and point features.

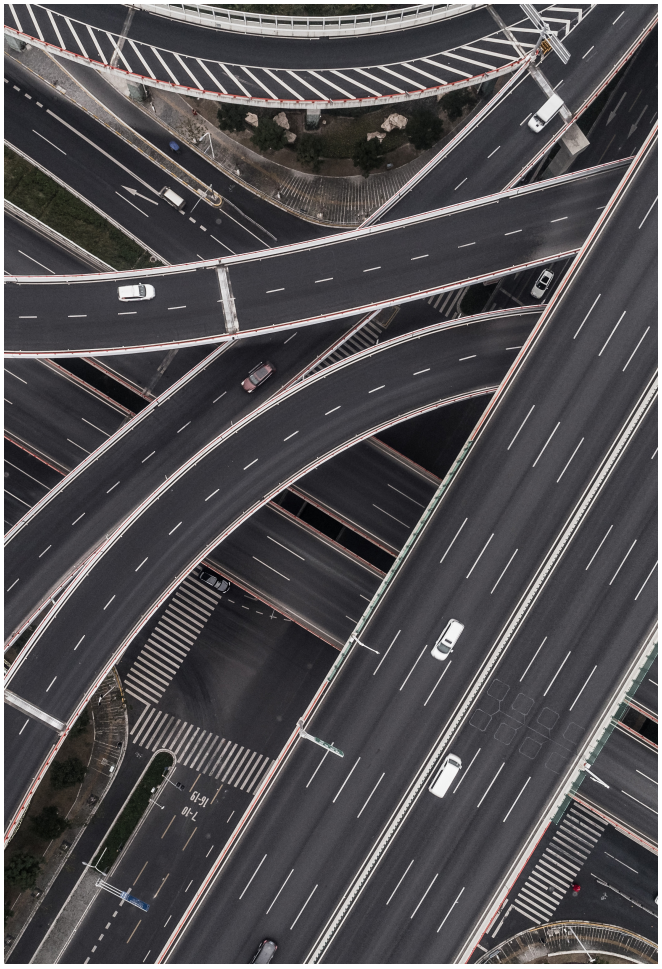
We used this more recent version of this data in summer 2021 to assess whether a tract had likely experienced the construction of interstate highways through existing housing, as noted on page 13. To calculate this, we first identified tracts with interstate highways within them or making up one of their boundaries. We then used data from the American Community Survey five-year release 2015-19 to identify tracts in which the majority of housing units were built before 1960. Tracts with interstate highways that had a majority of housing built before 1960 were assessed to have likely experienced the construction of interstate highways through existing housing.

More information on this data can be found at https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html

- Locations of interstate highways (assessed by counting the presence of interstates within or along the boundaries of a tract)

THE FEDERAL HIGHWAY ADMINISTRATION'S NATIONAL BRIDGE INVENTORY collects annual data from states, federal agencies, and tribal governments in accordance with the National Bridge Inspection Standards and the Recording and Coding Guide for the Structure Inventory and Appraisal of the Nation's Bridges. More information on this data can be found at https://www.fhwa.dot.gov/bridge/nbi.cfm

- Location of all local bridges
- Location of local bridges assessed to be in poor condition or in need of replacement



THE NATIONAL CENTER FOR CHARITABLE STATISTICS provides data primarily from information that tax-exempt nonprofit organizations file with the Internal Revenue Service. More information on this data can be found at https://nccs-data.urban.org/

The data drawn from this source includes the locations of the following types of organizations, which were assessed using buffer zones:

- Community-building organizations
- Visual and performing arts organizations
- Cultural promotion organizations
- Educational support organizations
- Preschools
- Alumni associations
- Parent and teacher groups
- Community health clinics
- Mental health organizations
- Crime and legal services organizations
- Employment-related organizations
- Food assistance programs
- Homeless shelters
- Homeowner associations
- Public safety organizations
- Sports and youth development organizations
- Human services organizations
- Civil rights organizations
- Community improvement organizations
- Neighborhood and homeowner associations
- Economic development organizations
- Community service organizations
- Private foundations
- Corporate foundations
- Membership and mutual benefit organizations

THE GUN VIOLENCE ARCHIVE is an online archive of gun violence incidents collected from over 7,500 law enforcement, media, government, and commercial sources daily in an effort to provide near-real-time data about the results of gun violence. It is an independent data collection and research group with no affiliation with any advocacy organization. More information on this data can be found at https://www.gunviolencearchive.org/about

- Incidents of gun violence (assessed by counting the incidents within the boundaries of a tract)

THE PUBLIC LIBRARIES SURVEY examines when, where, and how library services are changing to meet the needs of the public. This data, supplied annually by public libraries across the country, provides information that policymakers and practitioners can use to make informed decisions about the support and strategic management of libraries. More information on this data can be found at https://www.imls.gov/research-evaluation/data-collection/public-libraries-survey

- Number of public libraries (assessed using buffer zones)

PARKSERVE, from the Trust for Public Land, provides free, easy-to-navigate access to the most comprehensive database on parks ever assembled, including information for 14,000 cities with a combined population of more than 260 million. More information on this data can be found at https://www.tpl.org/parkserve

- Presence of park space (assessed using buffer zones)

THE FEDERAL DEPOSIT INSURANCE CORPORATION (FDIC) provides a list of all FDIC-insured institutions and their branches. More information on this data can be found at https://www7.fdic.gov/idasp/warp_download_all.asp

- Number of retail bank branches (assessed using buffer zones)

THE NATIONAL CREDIT UNION ASSOCIATION compiles data on the credit union system's financial performance, merger activity, changes in chartering

and fields of membership, as well as broader economic trends affecting credit unions. More information on this data can be found at https://www.ncua.gov/analysis/credit-union-corporate-call-report-data/quarterly-data

- Number of credit union branches (assessed using buffer zones)

DIGITAL SCHOLARSHIP LAB OF THE UNIVERSITY OF RICHMOND: In order to evaluate areas that were redlined, we developed a new process for handling data near the end of our aggregation efforts. In the early 20th century, the Home Owners' Loan Corporation (HOLC) "graded" neighborhoods into four zones based upon the perceived riskiness of mortgage investments. These zones were heavily based on racial discrimination and blocked residents from obtaining a path to homeownership. The lowest-graded areas are commonly called "redlined" districts.

Redlined districts do not line up well with tract boundaries, and as a result, we needed to determine what percentage of today's population lived in redlined areas from the early 20th century. First, we created an approximate location for a tract's residents. We determined baseline populations using very granular block-level population data from the 2010 decennial census. Then we extrapolated the population using block group estimates from the most recent American Community Survey (2019). Finally, we used road network maps from the Census Bureau and building footprint maps open-sourced by Microsoft to approximate coordinates for every resident. We then overlaid these coordinates over a shape file of HOLC's "graded" neighborhoods provided by the University of Richmond to understand what percentage of a tract's population fell into each zone.[6]71 We also were able to determine the demographic characteristics of residents living in different zones based on census data.

If a resident fell into the highest-grade zone, we scored that resident with a 4; the next highest grade received a 3, and so on until we reached residents living in redlined zones, which received a score of 1. Residents that did not fall into any kind of graded zone based on the shape file maps received no score. This scoring system allowed us to create a "grade point average" for each tract by averaging all of the residents that received a score. Tracts that had a score less than 2 were considered to be "redlined" tracts. Tracts in which less than 50% of the residents had a score were excluded from the analysis.

- Population-weighted average HOLC "redlining" score

ADDITIONAL DATA TRANSFORMATIONS

We used data science tools and processes to generate multiple variables for many of the topics examined in our analyses. These data transformations were conducted in order to increase the interpretability of our analyses among our intended audience.

Our experience suggests that local leaders and practitioners typically have only a general sense of neighborhood-level data, which is almost always based on relative comparisons with the rest of their city or county, or on binary classifications—e.g., "This neighborhood has one of the highest crime rates in the city," or "That neighborhood has a public park, but this one does not."

To align our data with the real-world frameworks local leaders use, we converted the data we collected into variables based on relative comparisons and also classified data into categories or bins, which resulted in the following five variable types:

- **ACTUAL VALUE:** The continuous or count values of the specified tract provided by the pre-aggregated sources and our own calculations of fixed-location data, or simple arithmetic calculations using that data. For example, the homeownership rate is the total number of owner-occupied housing units in a tract divided by the total number of housing units in a tract.
- **BOOLEAN CLASSIFICATIONS BASED ON ACTUAL VALUES:** Classifications of actual values of the specified tract into binary categories using thresholds specific to each variable type. For example, the presence of nonprofit arts organizations within 1 mile of a tract's internal population centroid was divided into "actual value greater than 0" and "actual value equal to 0."

- **LOCALLY NORMALIZED PERCENTILE RANK VERSUS COUNTY OR METROPOLITAN AREA:** The percentile rank of the specified tract compared to all other urban census tracts in the same county or MSA. This was important because our data exploration indicated that many attributes of tracts showed significant local area effects. For example, homeownership rates in Manhattan and Los Angeles County have a very different distribution than those in smaller cities or counties.
- **LOCALLY NORMALIZED QUARTILE OR TERTILE ASSIGNMENT VERSUS COUNTY OR METROPOLITAN AREA:** Classification of the specified tract's percentile rank into quartile or tertile categories, which approximated local leaders' comparative evaluations of high, medium, and low rates of the presence of different attributes in a community.
- **LOCALLY NORMALIZED BOOLEAN CLASSIFICATIONS BASED ON QUARTILE OR TERTILE ASSIGNMENT:** Classification of the specified tract's quartile or tertile assignment into binary categories, such as "bottom quartile compared to urban tracts in the same county" versus "top three quartiles compared to urban tracts in the same county."

Our literature review and experience working with local leaders also indicated that the attributes of adjacent urban tracts could very likely influence the changes in individual tracts. We first defined tracts as adjacent if they shared a land or water boundary, regardless if the boundary crossing was a line or single point like a corner boundary (also known as queen's case contiguity). Then, we identified all adjacent neighborhoods for every U.S. tract.

In order to assess the attributes of adjacent tracts, we generated two additional variable types:

Adjacent tract values: The continuous or count values of adjacent urban tracts, such as "greatest value," "least value," "differences," and "population-weighted average value" of adjacent urban tracts using queen's case contiguity. For example, this included "greatest proportion of households earning $100,000 or more in annual income among adjacent urban tracts."

Adjacent tract boolean classifications: Classifications of actual values of the adjacent tracts into binary categories using thresholds specific to each variable type. For example, this included "plurality racial or ethnic group in all adjacent tracts matches specified tract."

As a result of these transformations, data collected on a single attribute of census tracts, such as homeownership rates or the proportion of residents from 25 to 34 years old, could often generate more than 10 distinct variables, and in some cases, well over 20 distinct variables. The following table provides examples of this using data on three attributes.

TABLE 1

Examples of derived variables

		Examples of the derived variables created for a single census tract, "Tract A," for three variable types in the project		
Derived varible	**Definition**	**Homeownership rates in 2000**	**Presence of nonprofit arts organizations within 1 mile of the mean population center in 2000**	**Residents who are 25 to 34 years old in 2000**
Actual value	The continuous or count value for a tract collected from the original source	Percentage of housing units that are owner-occupied within the boundaries of Tract A	Number of nonprofit arts organizations within 1 mile of the mean population center of Tract A	Percentage of residents between the ages of 25 and 34 years old in Tract A
Boolean classifications based on actual values	Classifications of actual values into binary categories using thresholds specific to each variable type	N/A	Actual value is greater than 0 vs. equal to 0 Actual value is greater than 1 vs. less than or equal to 1 Actual value is greater than 2 vs. less than or equal to 2	Actual value is greater than 17.2% vs. less than or equal to 17.2% (This is the national proportion of residents who are 25 to 34 years old in 2000.)
Percentile rank versus county or metropolitan area	The percentile rank of the actual value compared to all urban tracts in the same county	Percentile rank compared to urban tracts in the same county Percentile rank compared to urban tracts in the same MSA	Percentile rank compared to urban tracts in the same county Percentile rank compared to urban tracts in the same MSA	Percentile rank compared to urban tracts in the same county Percentile rank compared to urban tracts in the same MSA
Quartile or tertile assignment versus county or metropolitan area	Classification of percentile rank into quartile or tertile categories	*County comparisons:* Quartile compared to urban tracts in the same county Tertile compared to urban tracts in the same county *MSA comparisons:* Quartile compared to urban tracts in the same MSA Tertile compared to urban tracts in the same MSA	Tertile compared to urban tracts in the same county Tertile compared to urban tracts in the same MSA	*County comparisons:* Quartile compared to urban tracts in the same county Tertile compared to urban tracts in the same county *MSA comparisons:* Quartile compared to urban tracts in the same MSA Tertile compared to urban tracts in the same MSA

Table continued on next page.

TABLE 1 CONTINUED

Examples of derived variables

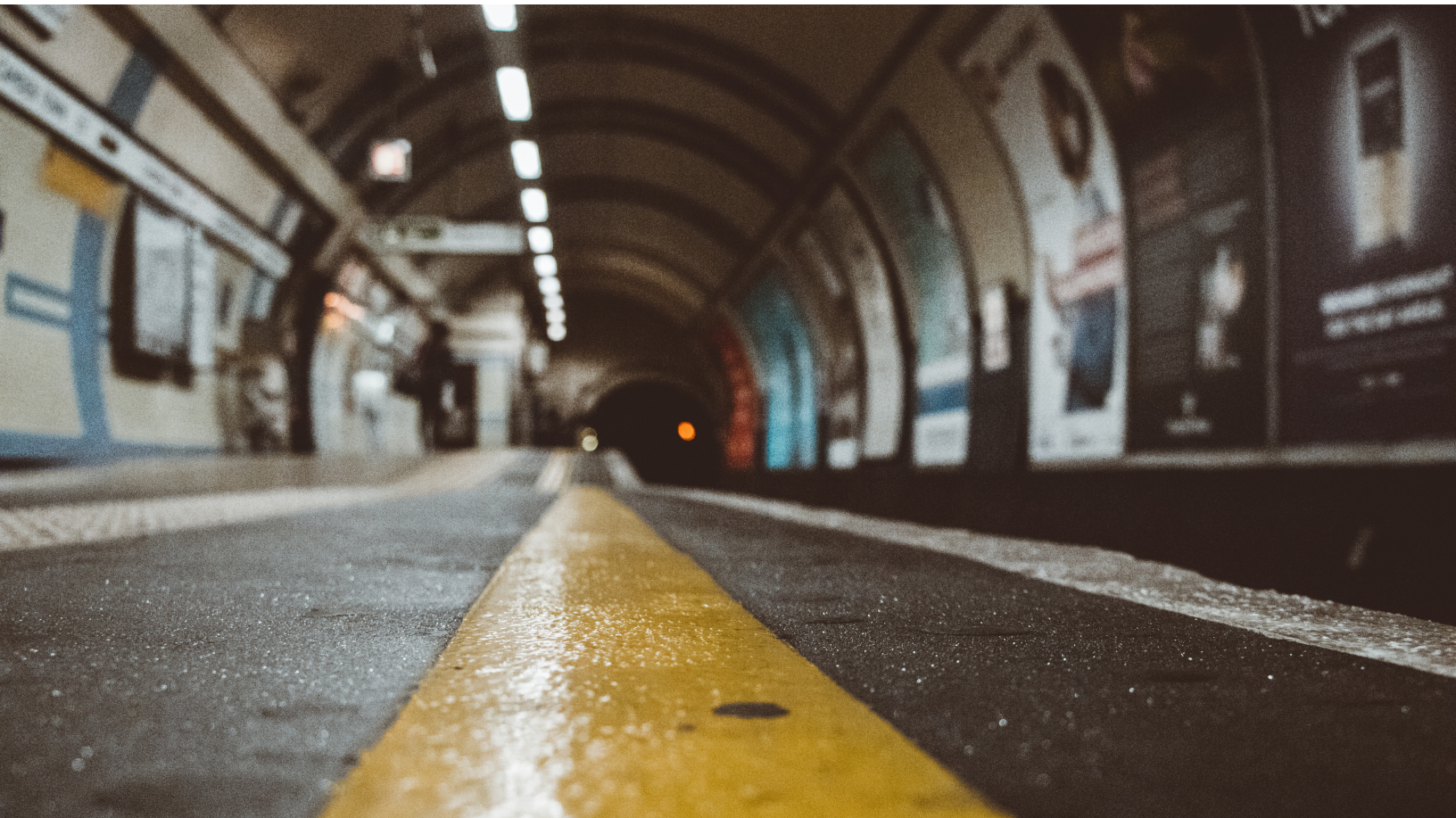
Boolean classifications based on quartile or tertile assignment	Classifications of percentile values into binary categories using quartiles or tertiles	*County comparisons:* Top quartile compared to urban tracts in the same county vs. bottom three quartiles Middle two quartiles compared to urban tracts in the same county vs. top or bottom quartiles Bottom quartile compared to urban tracts in the same county vs. top three quartiles Top two quartiles compared to urban tracts in the same county vs. bottom two quartiles (a.k.a. Above the median percentile vs. median percentile and below) Top tertile compared to urban tracts in the same county vs. bottom two tertiles Middle tertile quartiles compared to urban tracts in the same county vs. top or bottom tertile Bottom tertile compared to urban tracts in the same county vs. top two tertiles *MSA comparisons:* Top quartile compared to urban tracts in the same MSA vs. bottom three quartiles Middle two quartiles compared to urban tracts in the same MSA vs. top or bottom quartiles Bottom quartile compared to urban tracts in the same MSA vs. top three quartiles Top two quartiles compared to urban tracts in the same MSA vs. bottom two quartiles (a.k.a. Above the median percentile vs. median percentile and below) Top tertile compared to urban tracts in the same MSA vs. bottom two tertiles Middle tertile quartiles compared to urban tracts in the same MSA vs. top or bottom tertile Bottom tertile compared to urban tracts in the same MSA vs. top two tertile	*County comparisons:* Top tertile compared to urban tracts in the same county vs. bottom two tertiles Middle tertile quartiles compared to urban tracts in the same county vs. top or bottom tertile Bottom tertile compared to urban tracts in the same county vs. top two tertiles *MSA comparisons:* Top tertile compared to urban tracts in the same MSA vs. bottom two tertiles Middle tertile quartiles compared to urban tracts in the same MSA vs. top or bottom tertile Bottom tertile compared to urban tracts in the same MSA vs. top two tertiles	*County comparisons:* Top quartile compared to urban tracts in the same county vs. bottom three quartiles Middle two quartiles compared to urban tracts in the same county vs. top or bottom quartiles Bottom quartile compared to urban tracts in the same county vs. top three quartiles Top two quartiles compared to urban tracts in the same county vs. bottom two quartiles (a.k.a. above the median percentile vs. median percentile and below) Top tertile compared to urban tracts in the same county vs. bottom two tertiles Middle tertile quartiles compared to urban tracts in the same county vs. top or bottom tertile Bottom tertile compared to urban tracts in the same county vs. top two tertiles *MSA comparisons:* Top quartile compared to urban tracts in the same MSA vs. bottom three quartiles Middle two quartiles compared to urban tracts in the same MSA vs. top or bottom quartiles Bottom quartile compared to urban tracts in the same MSA vs. top three quartiles Top two quartiles compared to urban tracts in the same MSA vs. bottom two quartiles (a.k.a. Above the median percentile vs. median percentile and below) Top tertile compared to urban tracts in the same MSA vs. bottom two tertiles Middle tertile quartiles compared to urban tracts in the same MSA vs. top or bottom tertile Bottom tertile compared to urban tracts in the same MSA vs. top two tertiles

Table continued on next page.

TABLE 1 CONTINUED

Examples of derived variables

Adjacent tract values	Greatest, least, and average values of adjacent urban tracts using queen methodology	N/A	N/A	Greatest value of tract that is adjacent to Tract A Smallest value of tract that is adjacent to Tract A Weighted average value of all tracts that are adjacent to Tract A Difference between the greatest value of tract that is adjacent to Tract A and Tract A itself Difference between the smallest value of tract that is adjacent to Tract A and Tract A itself
Adjacent tract boolean classifications	Binary values of adjacent urban tracts using queen methodology	N/A	N/A	Greatest value of tract that is adjacent to Tract A is in the top county quartile vs. bottom three county quartiles Greatest value of tract that is adjacent to Tract A is in the top county tertile vs. bottom two county tertiles
Total variables derived		**21**	**14**	**29**



Methodology Part B: Analyses of the current state of neighborhoods with concentrated poverty

To answer our first research question (How do neighborhoods with concentrated poverty compare to other census tracts using the most recently released data?), we analyzed the most recently collected, relevant data described in Methodology Part A to create the analyses shared in the text on pages 9 to 14. Additional information of the methods used to create these analyses can be found below.

ESTIMATED REDUCTION IN LIFE (PAGE 9)

"Neighborhoods with concentrated poverty" are defined according to the criteria explained in the first section of this report: census tracts with 30% or more of residents living in households with income below the federal poverty threshold, that are located in metropolitan areas, and have at least 1,000 residents per square land mile.

Tracts with fewer than 500 total residents are removed from the sample, as are those with high proportions of college and graduate school students—i.e., 30% or more of residents over three years old.

"Other U.S. neighborhoods" are defined as census tracts with less than 30% of residents living in households with income below the federal poverty threshold, that are located in metropolitan areas, and have at least 1,000 residents per square land mile. Tracts with fewer than 500 total residents are removed from the sample.

SOURCE: Common Good Labs analysis based on estimates from the U.S. Small-Area Life Expectancy Estimates Project and tract-level data from the U.S. Census Bureau.

INCOME AND INCARCERATION IN EARLY ADULTHOOD BASED ON CHILDHOOD ENVIRONMENT (PAGE 10)

"Neighborhoods with concentrated poverty" are defined according to the criteria explained in the first section of this report.

"Low-income households" are those at the 25th percentile in earnings compared to all U.S. households. "Middle-income households" are those at the 50th percentile in earnings compared to all U.S. households. "High-income households" are those at the 75th percentile in earnings compared to all U.S. households.

SOURCE: Common Good Labs analysis of data from Opportunity Insights and the U.S. Census Bureau.

CHALLENGES FOUND IN NEIGHBORHOODS WITH CONCENTRATED POVERTY (PAGE 11 AND 12)

"Neighborhoods with concentrated poverty" are defined according to the criteria explained in the first section of this report.

"Vacant businesses and storefronts" are defined as business addresses assessed by the U.S. Postal Service to be vacant or "no-stat" during the previous 90 days. A neighborhood has a high proportion of these vacancies if it is at or above the 75th percentile among all urban metropolitan residential areas, excluding those with large numbers of college students.

"Toxic release sites" are defined by the Environmental Protection Agency.

School districts are defined as "low-performing" if their high school graduation rates fall below the national average.

A neighborhood has a high proportion of students in low-performing schools if it is at or above the 75th percentile among all urban metropolitan residential areas, excluding those with large numbers of college students.

"Medically Underserved Areas" are areas designated by the Health Resources and Services Administration for having too few primary care providers, or similar challenges related to health care access.

A neighborhood has a high proportion of households with children that lack internet access if it is at or above the 75th percentile among all urban metropolitan residential areas, excluding those with large numbers of college students.

"Higher-income occupations" are high-wage management and professional service occupations listed on page 4 of the methodology. A neighborhood has a low proportion of these occupations if it is below the 25th percentile among all urban metropolitan residential areas, excluding those with large numbers of college students.

Tracts are considered to be redlined if the average HOLC score was less than "C-grade," where "A-grade" was the highest grade given and redlined areas were assessed as "D-grade."

SOURCE: Common Good Labs analysis of relevant U.S. government data listed above.

COMPARISON OF CHALLENGES FOUND IN THREE NEIGHBORHOODS WITH CONCENTRATED POVERTY (PAGE 14)

"Neighborhoods with concentrated poverty" are defined according to the criteria explained in the first section of this report.

A neighborhood defined to have a high percentage of children who lack health insurance is one in which the proportion of children estimated to be without health insurance is at or above the 75th percentile among all urban metropolitan residential areas.

"Medically Underserved Areas" are areas designated by the Health Resources and Services Administration for having too few primary care providers, or similar challenges related to health care access.

A neighborhood defined to have a high number of three- and four-year-olds not enrolled in preschool is one in which the proportion of children not enrolled in school is at or above the 75th percentile among all urban metropolitan residential areas.

A neighborhood is defined to have a high proportion of children without internet access at home if it is at or above the 75th percentile among all urban metropolitan residential areas in the proportion of children without internet access at home, excluding those with large numbers of college students.

"Vacant businesses and storefronts" are defined as business addresses assessed by the U.S. Postal Service to be vacant or "no-stat" during the previous 90 days. A neighborhood has a high proportion of these vacancies if it is at or above the 75th percentile among all urban metropolitan residential areas, excluding those with large numbers of college students.

"Toxic release sites" are defined by the Environmental Protection Agency.

Methodology Part C: Defining changes over time in neighborhoods with concentrated poverty

To investigate our second research question (How have neighborhoods with concentrated poverty changed when comparing data from two points in time: 2000 and 2015?), we analyzed relevant data collected to describe the year 2000 or 2015, or the period in between those years listed in Methodology Part A and created the analyses shared in the text on pages 16 to 39. Additional information on the methods used to create these analyses can be found below.

DEFINING NEIGHBORHOOD CHANGES OVER TIME (2000 TO 2015)

For the analysis on neighborhood changes over time presented on pages 16 to 25, we examined how poverty levels and residential displacement changed over an approximately 15-year period across census tracts, or neighborhoods, with concentrated poverty. We chose to use a 15-year period because we believed it was long enough for substantial differences to emerge if they did exist, but short enough that policymakers and practitioners would consider the resulting findings relevant to the time horizons in which they think of their own work.

Data for the first year of observation are drawn from the 2000 decennial census. Changes are calculated using information from the same variables reported in the 2017 American Community Survey, which provides an average of the annual data collected from 2013 to 2017 for each census tract. Since the midpoint of this period is 2015, we refer to this as "2015 data" and consider the total time period covered in this analysis to be approximately 15 years.

To calculate the changes in each census tract, we had to address the fact that the Census Bureau sometimes adjusts the boundaries of individual census tracts over time. We used the Longitudinal Tract Database, a common tool in economic and social research, to adjust for the changes in individual tracts when making comparisons across time periods with different tract boundaries.[7]

IDENTIFYING NEIGHBORHOODS WITH CONCENTRATED POVERTY FOR ANALYSIS

The second phase of our analysis used the same categories of analytical criteria as our first phase, described on pages 16 to 25, to identify urban neighborhoods with concentrated poverty in the year 2000 that would be observed again in 2015. A total of 4,334 census tracts met all five of these initial criteria. In addition to these five criteria, we added several additional requirements for the second phase of analysis in order to ensure that the sample used to analyze changes over time was not biased by factors unique to only a small subset of census tracts.

- **NO LARGE-SCALE PUBLIC HOUSING DEVELOPMENTS:** Tracts with large quantities of place-based public housing in 2000 (i.e., 20% or more of all occupied units) were removed from the sample since they would be expected to have high proportions of poverty and low-income residents due to the government's requirements for tenants.
- **NO SIGNIFICANT DEMOLITIONS OF PUBLIC HOUSING:** Similarly, neighborhoods with large decreases in the available public housing units (i.e., a decline of 200 or more) between 2000 and 2015 were also removed, since this likely indicated that a large amount of public housing had been demolished or otherwise taken out of service by a governmental agency. Tracts were also removed for this reason if they reported that the total number of public housing units decreased by 100 while the total number of HUD households decreased by 100 or more from 2000 to 2015.

- **NO MILITARY HOUSING TRACTS:** Tracts with large numbers of military personnel in 2000 (i.e., 20% or more of all residents over 16) were also removed.
- **NO NEIGHBORHOODS IN ALASKA, HAWAII, OR THE NEW ORLEANS AREA:** Finally, tracts in geographic areas believed to be especially unique (namely, those in Alaska, Hawaii, and the New Orleans metropolitan area due to Hurricane Katrina) were also excluded.

A total of 3,673 census tracts met all of these criteria, which was the final sample used in this phase of analysis. We created a categorization scheme for these 3,673 tracts based on our two primary outcomes of interest: changes in poverty and the displacement of existing residents.

- **POVERTY:** We calculated our change in poverty measure as the change in the percentage of people living in poverty (e.g., a shift from 40% to 30% of the population in poverty would be a decrease of 10 percentage points). We chose this method because it measures the level of change as it would be experienced by people in a community. For example, a drop from 60% poverty to 50% in a neighborhood of 4,000 people and a drop from 35% poverty to 25% in a neighborhood of 4,000 people would both represent 400 fewer residents in poverty and would both register as a 10% absolute change.
- **DISPLACEMENT:** Our classification of displacement among existing residents is modeled off the definition used in recent research by the National Community Reinvestment Coalition.[8] This defines displacement as occurring if: 1) the decline of a racial or ethnic group's percentage of the population is more than two standard deviations from their mean percentage of population change among all census tracts nationwide; and 2) the racial or ethnic group's population in a neighborhood declines by at least 5%.

We looked displacement of three specific racial and ethnic groups tracked by the Census Bureau: Asian American, Black, and Latino or Hispanic people. This displacement definition allows us to identify racial groups within neighborhoods that are losing population at rates beyond what can be explained by national demographic changes. Additionally, it prevents incorrectly labeling displacement in scenarios where the total number of people in a racial or ethnic group grows slightly or remains in the same community, but because new emigrants to the community increase the number of residents from other groups, the relative percentage of the racial or ethnic group declines in the area.

We conducted additional analyses to understand how our findings might be affected by issues with high levels of margin of error for data drawn from estimates of Census Bureau data collected from samples of the population. This is particularly important for variables drawn from the American Community Survey, as a number of academics have noted in recent years.

We found that removing tracts due to margin of error was primarily resolved by dropping tracts with populations of fewer than 500 people. For example, if we added an additional standard to remove tracts from our sample when the value for any variable used in our primary definition or in the categorization schema described on the previous page had a coefficient of variation of 30% or more, it would reduce our sample of tracts by only 3%. Removing this small number of additional tracts did not substantial change the findings in our analyses.

Since the primary audience of this report are people working in non-technical roles as local government officials, business leaders, and philanthropic staff, we report our results using the larger sample, since this is easier for our audience to understand and follows the typical conventions of other practitioner-focused reports on this topic.

CATEGORIZING CHANGES IN NEIGHBORHOODS WITH CONCENTRATED POVERTY

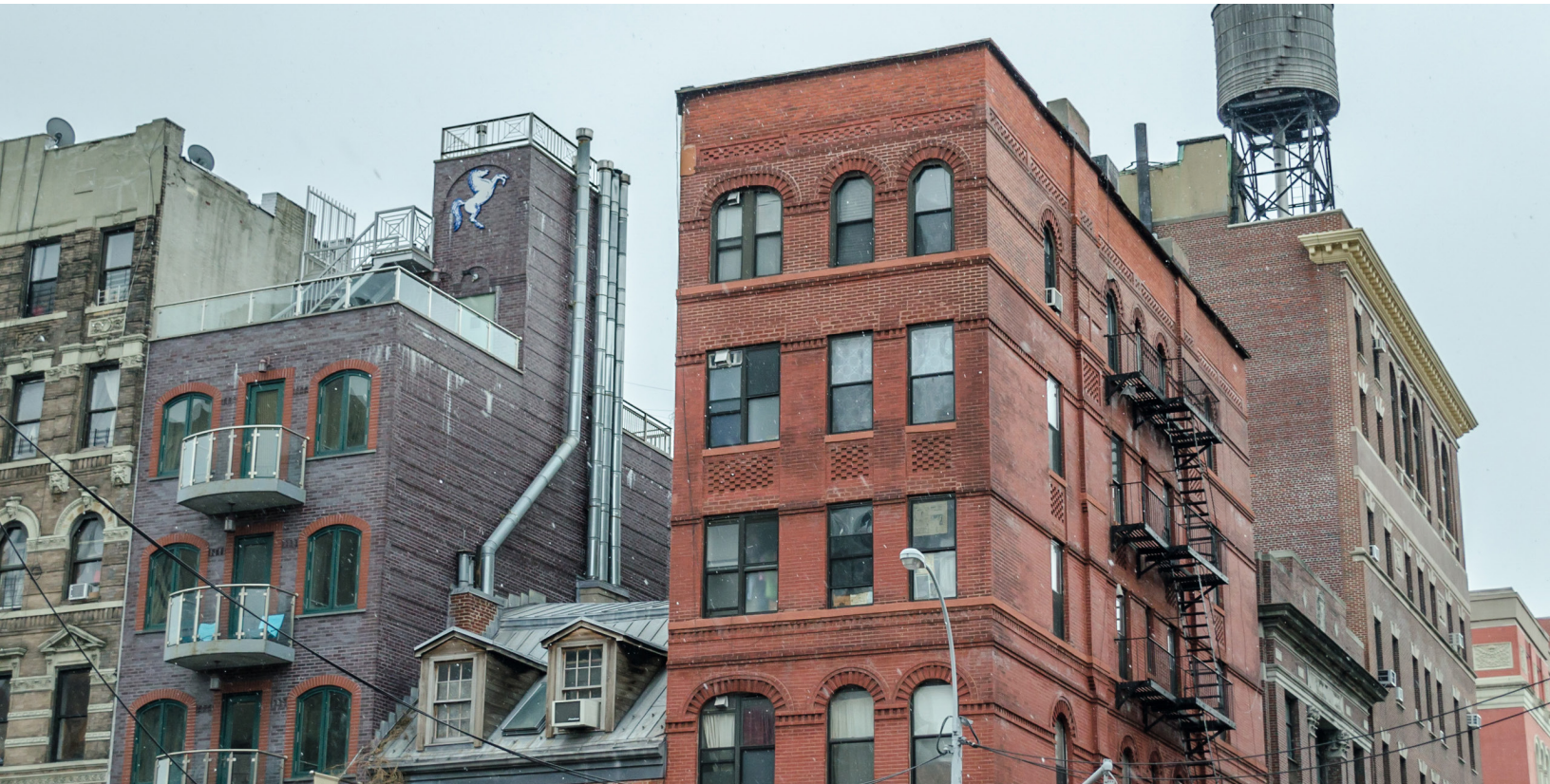
We categorized the changes in the 3,673 neighborhoods with concentrated poverty into six mutually exclusive categories outlined in table 3 of the report. These six types of neighborhood change are the primary focus of our analyses in the second phase of this study:

- Large decrease in poverty rate and no community displacement
- Moderate decrease in poverty rate and no community displacement
- Large increase in displacement
- Moderate increase in displacement
- Large increase in poverty rate
- Moderate increase in poverty rate

In addition to the outcomes on poverty levels and displacement that were used for categorizing the six outcome types, we added two final criteria for the neighborhoods that decreased poverty without displacement:

- **POPULATION CHANGE:** We used population data to measure whether neighborhoods were stable or growing. Neighborhoods with stable or growing populations are those in which the total number of residents in the community has either increased, remained the same, or not decreased by more than one standard deviation compared to all urban census tracts during this period.
- **RESIDENT RETENTION:** We used resident retention data from the American Community Survey to test whether the total levels of people moving out of each community was not above the normal levels for U.S. neighborhoods. Neighborhoods with normal to high retention are those in which the proportion of households from 2000 that are still living in the neighborhood in 2015 is within or above one standard deviation of the average for all urban census tracts during this period.

These two criteria ensure that the neighborhoods where poverty is decreasing were doing so by increasing the incomes of existing residents rather than through the process of abandonment or attrition, or by replacing existing poor residents through much greater than average rates of new people moving in.



Methodology Part D: Identifying indicators associated with significantly reducing poverty without displacing local communities over time (2000 to 2015)

In order to identify specific indicators that answered our final research question (How were neighborhoods that experienced a large decrease in poverty rate with no community displacement different from other poor communities when comparing data from two points in time: 2000 and 2015?), we analyzed relevant data collected to describe the year 2000, the year 2015, or the period in between those years listed in Methodology Part A to create the analyses shared in the text on pages 26 to 39. Additional information of the methods used to create these analyses can be found below.

CRITERIA FOR INDICATORS IDENTIFIED IN PROJECT ANALYSES

In order for neighborhood indicators identified in our analysis to be practically useful to leaders working in government, philanthropy, and local organizing, we designed a methodology that would meet these four criteria:

- **INFLUENCEABLE:** The indicators identified in our analysis should be factors that leaders can actually control or influence, particularly at the local level.
- **GENERALIZABLE:** The indicators identified in our analysis should be adjusted or controlled for local variation to account for the diversity of U.S. metropolitan areas.
- **CONCISE:** The indicators identified in our analysis should be narrowed down from the long list of potential factors collected in our data aggregation to be easily understood by local leaders.
- **INTERPRETABLE:** The indicators identified in our analysis should be translated to match the ways that local leaders and practitioners typically understand their neighborhoods and be modeled transparently to enable us to gather insights on the ways that they interact.

HYPOTHESES GENERATION AND DATA AGGREGATION FOR POTENTIAL INDICATORS ASSOCIATED WITH SIGNIFICANTLY REDUCING POVERTY WITHOUT DISPLACING LOCAL RESIDENTS OVER TIME (2000 TO 2015)

We began by focusing on our first criterion: **influenceable.** The indicators identified in our analysis should be factors that leaders can actually control or influence, particularly at the local level.

To do this, we collected potential hypotheses to answer our third research question through conversations with local government leaders, neighborhood organizers, and researchers over the course of our work in communities across the country. We focused hypothesis development on factors that leaders believed could be both effective and influenceable by targeted policies and programs. We have highlighted a number of the most useful books, articles, papers, and reports that we encountered in our literature review listed in “Recommended additional reading” section.

We drew data from the collection efforts described in Methodology Part A on pages 2 to 15 to test the hypotheses we identified. We focused on data for the year 2000 before observed changes in poverty and displacement took place or at the same time as the changes were occurring from 2000 to 2015.

DATA EXPLORATION OF POTENTIAL INDICATORS ASSOCIATED WITH SIGNIFICANTLY REDUCING POVERTY WITHOUT DISPLACING LOCAL RESIDENTS OVER TIME (2000 TO 2015)

Next, we focused on our second criterion: **generalizable.** The indicators identified in our analysis should be adjusted or controlled for local variation to account for the diversity of U.S. metropolitan areas.

To do this, we conducted exploratory analyses of the data to identify linear relationships, discover factors that should be controlled for, and determine how to best utilize geospatial characteristics. The analyses we conducted in this step included running univariate and bivariate analyses, testing correlations between variables, exploring distributions of data across different sets of metropolitan areas and tract categories, identifying outliers, and visualizing data in charts and maps to discover patterns within and among variables.

The two most important factors are:

- County-level effects, often driven by differences in population density (e.g., homeownership), which we accounted for using the locally normalized variable transformations described earlier on pages 11 to 15
- Displacement, which was associated mostly with external attributes surrounding a neighborhood that are hard for leaders to change—e.g., the physical distance between a neighborhood and the downtown area or central business district of a metropolitan area.

The displacement of communities by wealthier members of other racial or ethnic groups is influenced in a number of ways, primarily by the environment surrounding a neighborhood. We constructed a model to help us understand the characteristics related to increased risk of community displacement for neighborhoods with concentrated poverty studied in the third phase of our analysis. This model created a displacement index and assigned a risk of displacement score to each tract based on the probability of its existing residents being pushed out as defined in phase two. This model was developed using a logistic regression with L1 regularization to reduce model complexity.

We found that the likelihood of displacement in neighborhoods with concentrated poverty from 2000 to 2015 was higher in communities observed to have the following characteristics:

- **The counties in which the observed neighborhood was located had:**
  - Larger numbers of people from racial or ethnic groups in 2000 that were different than the largest racial or ethnic group in the observed neighborhood
  - Higher levels of population growth from 1990 to 2000 among racial or ethnic groups that were different than the largest racial or ethnic group in the observed neighborhood in 2000
  - Greater increases in local GDP from 2001 to 2015
- **At least one neighborhood immediately adjacent to the observed community had:**
  - A majority of residents from a racial or ethnic group in 2000 that was different than the largest racial or ethnic group in the observed neighborhood
  - A relatively high proportion of 25- to 34-year-old residents in 2000 and/or a relatively high proportion of households earning $100,000 or more in 2000
  - Significantly lower commute times in 2000 than the observed community, which indicate that its residents lived comparatively closer to their work
- **The observed neighborhood itself had:**
  - Close proximity to the downtown or central business district of the primary city in its metropolitan area in 2000
  - Lower levels of homeownership in 2000 and/or higher levels of vacant homes in 2000

In addition to the characteristics listed above, the model also included binary variables that noted when neighborhoods were located in the New York City and Los Angeles metropolitan areas. This was added since

the displacement patterns in these large metropolitan areas differed from that of other cities, and it subsequently improved model performance.

ADVANCED ANALYSES FOR INDICATORS ASSOCIATED WITH LARGE REDUCTIONS IN POVERTY WITHOUT DISPLACING RESIDENTS OVER TIME (2000 TO 2015)

Next, we turned our attention to our third criterion: **concise.** The indicators identified in our analysis should be narrowed down from the long list of potential factors collected in our data aggregation to be easily targeted by local leaders.

In order to do this, we utilized newer modeling techniques that are now possible due to increases in computational power and larger sets of data, frequently referred to as "machine learning." We used a specific modeling technique for data exploration called a "random forest classifier" that was first developed in the 1990s by a researcher at Bell Labs and has been refined over the last few decades.[9]

OVERVIEW OF RANDOM FOREST CLASSIFICATION

A random forest is a machine learning prediction model that fits many decision trees on randomly extracted subsets of a dataset in order to identify different classes or categories of outcomes by creating separation criteria from the data. The model then merges these decision trees together to get a more accurate and stable prediction than any decision tree would obtain on its own.

For example, let's say we are training a decision tree classifier to determine whether an individual attended college given their attributes. Given a lot of data about people who attended and did not attend college, the classifier may determine that living in an urban area may be a great prediction of having attended college. Therefore, the decision tree may bifurcate a population into urban and rural residents as its first step of identifying college attendees. It may then determine that having a household income of $54,000 is also another way to bifurcate the urban subgroup, such that the higher-income subgroup is more likely to have attended college.

This bifurcation process can keep going to break the population into smaller and smaller subgroups until, in theory, each subgroup contains only college attendees or non-college attendees, thereby perfectly separating the two classes. To determine the breakpoints for its separation criteria, the decision tree classifier uses splitting criteria, such as a Gini impurity or entropy metric, both of which are beyond the scope of this explanation.

The term "random forest" derives from the fact that the classifier fits and averages across many decision trees in order to prevent overfitting, since technically a decision tree could perfectly separate classes with an arbitrary number of bifurcations that do not generalize to the whole population (e.g., all individuals with initials RAA attended college). To reduce overfitting, the training procedure randomly chooses extracts of the dataset (i.e., a subset of variables/features and sample of observations) and runs a decision tree classifier on that extract alone.

While this has the effect of reducing the predictive capability of each individual decision tree since it does not have access to the full dataset, the advantage is that each tree provides a distinct decision process. Therefore, when the random forest averages across all of these uncorrelated trees, it is able to reduce the variance due to model overfitting while also identifying the most important predictive features, since they will be the most influential on average across the decision trees. This ranking of influential variables is called feature importance and it helps us understand which variables consistently seem to have some explanatory power in classification.

ADVANTAGES OF RANDOM FOREST CLASSIFIERS FOR NARROWING DOWN VARIABLES/FEATURES

Random forest classifiers offer a number of advantages for identifying the most important factors in sets of data similar to what we aggregated in this project.

- Compared to more traditional, regression-based models, such as logistic regression, they are better suited for narrowing down very large numbers of variables to a smaller list of interacting factors that can be targeted in analyses or programs, often referred to as "feature selection." [10]
- They are also better able to identify linear and nonlinear relationships, both of which we would expect neighborhoods to exhibit.[11]
- Random forests can handle continuous and categorical variables without the need for significant rescaling, transformation, or outlier handling, thereby reducing the need for extensive preprocessing.
- Many of these classifiers are also well suited to work with unbalanced datasets, which are datasets which have a significantly unequal number of instances for each class or outcome. This is very important in our analysis since the outcome we are concerned with (a large decrease in poverty without displacing residents) is much less common than the other types of change found in neighborhoods with concentrated poverty over time.

The use of random forest classifiers also has disadvantages compared to other commonly used analytical techniques. For example, the computational resources needed to store the model increase as you have more training examples.

Another drawback is that random forests are non-parametric, meaning that you cannot describe and interpret the model in a set number of parameters in the same way that you might with a logistic regression. For this reason, random forest classifiers are often referred to as a "black box." It is possible to see the inputs that go into the model and the results it produces, but it is very difficult to understand or interpret the specific ensemble, or combination, of decision trees that generate the model's results.[12] Because of this limitation in interpretability, we used random forest classifiers only as a tool for narrowing down the long list of potential factors collected in our data aggregation, also known as "feature selection."[13]

IMPLEMENTATION OF THE RANDOM FOREST CLASSIFIER TO SELECT FEATURES/VARIABLES

To apply the random forest classifier, we used the scikit-learn package in Python. We used a dataset made up of information on the targeted outcome in each neighborhood (i.e., the presence or absence of a large decrease in the poverty rate with no community displacement) as well as several dozen features/variables we aggregated that corresponded to specific hypotheses identified as influenceable. These features/variables were all factors measured in 2000 before the changes in poverty and displacement took place or at the same time as the changes were occurring. They were also adjusted to maximize their explanatory power by normalizing them to all neighborhoods in the same metropolitan area, if exploratory analyses explained on pages 11 to 15 indicated this increased their explanatory power. Finally, we also included the output score from the likelihood of displacement as of 2000 for each neighborhood to capture the effects of the factors described on page 22.

We randomly grouped our observations into training and testing datasets using an 80/20% split in order to simultaneously determine the out-of-sample accuracy of the prediction model and identify any potential overfitting. We also tuned the model's hyperparameters for the maximum depth of the trees and the minimum samples per leaf to further reduce the potential for overfitting. To identify optimal hyperparameter setups, we looked at Receiver Operating Characteristic (ROC) AUC and Precision-Recall Curve AUC scores. We then ran the random forest across 5,000 trees.

We ran multiple versions of the random forest model using different seeds for the 80/20% split and found that a consistent group of variables appeared atop the feature importance list when predicting for

neighborhoods with large decreases in poverty rates and no community displacement. The top 10 features were:

- GDP growth in the MSA of a neighborhood from 2001 to 2015 (Note: GDP data at the metropolitan area level was not available for the year 2000.)
- Likelihood of displacement score for a neighborhood in 2000 drawn from the logistic regression model explained on pages 22 and 23
- Homicide rate per 100,000 residents in the neighborhood's county in 2000
- Number of nonprofit health care facilities within a 30-minute walking distance of the center of a neighborhood in 2000
- Number of community-building organizations within a 30-minute walking distance of the center of a neighborhood in 2000
- Percent change in the total housing units per square mile in a neighborhood from 1990 to 2000
- Percentage of abandoned residential units in a neighborhood 2000*
- Percentage of a neighborhood within a 30-minute walking distance of park space in 2000
- Percentage of residential units in a neighborhood that were owner-occupied in 2000*
- Percentage of residents who were self-employed in a neighborhood in 2000*

Two factors should be noted: 1) The 10 features/variables above are listed in alphabetical order, since the purpose of this random forest classifier was only to select the features that would be used in the final analyses that would provide an interpretable model; and 2) All features denoted with an "*" are locally normalized. The value used in the model is the percentile rank of the specified tract compared to all other urban census tracts in the MSA.

CONSTRUCTION OF AN INTERPRETABLE MODEL

Finally, we focused on our fourth criterion: **interpretable**. The indicators identified in our analysis should be modeled transparently and using variables/features that match the ways that local leaders and practitioners typically understand their neighborhoods. We did this using three final steps: data transformation, combinatorial search, and external confirmation.

DATA TRANSFORMATION

We began by transforming the data on the 10 most important features identified in the random forest classifier to match the ways that local leaders and practitioners typically understand their neighborhoods. Our experience suggests that most decisionmakers have only a general sense of neighborhood-level data, which is almost always based on relative comparisons with the rest of their city or region, or on binary classifications—e.g., "This neighborhood has one of the highest vacancy rates in the city," or "That neighborhood has a public park, but this one does not."

To align our data with the real-world frameworks local leaders use, we converted the variables the random forest classifier identified from continuous to categorical variables. For example, each neighborhood's percentile rank of owner occupancy is expressed as a continuous percentage in the data, but we created categories to bin the variable into values:

- "High" (top local quartile)
- "Moderate" (middle two local quartiles)
- "Low" (bottom local quartile)
- "Not high" (bottom three quartiles)
- "Not low" (top three quartiles)
- "Median or below" (bottom two quartiles)
- "Above the median" (top two quartiles)

The presence of community-building organizations in a neighborhood was also converted into two values: "present" or "absent." (An explanation of the variable transformations used in our analyses is described in Methodology Part A on page 2.)

COMBINATORIAL SEARCH

These data transformations enabled us to use a combinatorial search for our final analysis using the 10 features the random forest classifier identified as most important. Combinatorial problems are well known in operations research (e.g., the traveling salesman problem) and in industry, particularly in fields such as transportation and logistics.[14]

Since each of the 10 variables had been categorized to have between two and seven potential values, a combinatorial search modeling approach enabled us to compare the purity or prevalence of our desired outcome (i.e., a large decrease in poverty with no displacement as measured from 2000 to 2015) in neighborhoods that had every possible combination of the categorized values for the 10 features.

A combinatorial search traverses the entire solution space, which allows us to examine smaller subsets of the features. For example, we could look at the future prevalence of our desired outcome in the sample of neighborhoods only where owner occupancy was "high" and community-building organizations were "present," with no other specified values for any of the other features included.

This requires a great deal of computational power, which would have made it very difficult and time consuming to conduct in the past. However, recent advances in software and the use of multiple cores in cloud computing made it possible for us to calculate the prevalence of our desired outcome for every potential combination of these 10 features—more than 1 million combinations, in total—in a very short period.

Though calculating the prevalence of our desired outcome across more than 1 million different combinations of the selected features required a great deal of computational power, it yielded results that are completely transparent and easy to interpret. The resulting output provides clear information on which combinations of features are best for finding combinations that were particularly common among neighborhoods with our desired outcome optimizing the future prevalence of our desired outcome and which combinations are also more and less common among the neighborhoods with concentrated poverty in our sample. This was the reason we selected it to meet our final criterion.

We focus our interpretation of the combinatorial search model's output on pattern recognition across the combinations with the highest levels of prevalence for our desired outcome and the largest number of neighborhoods that met their specified combinatorial criteria. To understand the interaction between these two factors, consider the following simplified comparison:

- "Combination A" included only one neighborhood from the sample and had 100% prevalence of the desired outcome (i.e., one neighborhood met the criteria specified in this combination and that neighborhood also had a large decrease in the poverty rate with no community displacement as measured in 2015).
- "Combination B" included 100 neighborhoods from the sample and had 20% prevalence of the desired outcome (i.e., 100 neighborhoods met the criteria specified in this combination and 20 of these neighborhoods had large decreases in poverty rate with no community displacement as measured in 2015).

In this case, we would be much more interested in examining "combination B" than "combination A" in our pattern recognition analysis.

Analysis of the combinations with the best performance and the largest number of neighborhoods enabled us to better understand how different features interacted together, and yielded several important insights:

- Given the size of our sample (3,673 neighborhoods), it was difficult to combine more than seven or eight features in any specific combination. Combinations that included more than eight features typically had between 0 and 10 total neighborhoods that met their specified criteria and were therefore not useful.

- There was no single feature or combination of features that had significantly higher future prevalence of our desired outcome than what could be found in other top-performing combinations, and a very large number of neighborhoods that met its criteria. This supports the idea promoted by other researchers and practitioners that there is not a single "silver bullet" associated with large decreases in poverty without displacement in neighborhoods with concentrated poverty.
- Three features primarily related to the external environment around each neighborhood seemed to function as factors that were "necessary, but not sufficient" for our desired outcome. If any of them were missing, the prevalence of the target outcome in the future dropped significantly, but when these features were present, it was not associated with very large increases in future prevalence. These specific features and their values were:
  - Positive GDP growth in the MSA of a neighborhood from 2001 to 2015
  - Fewer than 25 homicides per 100,000 residents in the neighborhood's county in 2000
  - Low likelihood of displacement score for a neighborhood in 2000
- Five other features related to the internal attributes of each neighborhood tended to be associated with larger increases in the future prevalence of our desired outcome, as long as the three external "necessary, but not sufficient" features were also present and when they are combined in greater numbers with each other. They seem to function as factors that are "activated" by the presence of the right external characteristics with potential "magnifying effects" when combined with one another. These specific features and their values were:
  - At least one community-building organization within 1 mile of the center of a neighborhood in 2000
  - High or moderate levels of residents who were self-employed in a neighborhood in 2000, defined as at or above the 25th percentile of all urban neighborhoods in the same metropolitan area
  - High or moderate levels of residential units in a neighborhood that were owner-occupied in 2000, defined as at or above the 25th percentile of all urban neighborhoods in the same metropolitan area
  - Low or moderate levels of abandoned residential units in a neighborhood in 2000, defined as below the 75th percentile of all urban neighborhoods in the same metropolitan area
  - Positive growth in the total housing units per square mile in a neighborhood from 1990 to 2000

Additional information on each of these features and their combined results can be found in the text on pages 26 to 35.

EXTERNAL CONFIRMATION

We wanted to properly contextualize and interpret the findings from the combinatorial analyses for our readers, so we took three final steps to help confirm and explain our findings:

- Literature reviews: We conducted a review of existing literature and research on each of the individual eight indicators highlighted in the combinatorial analysis. As expected, we found that there was very little evidence of how these indicators interact. However, we identified a number of papers that demonstrated others had previously found individual relationships between each factor and poverty reduction, as noted in the text on pages 32 and 33.
- This provided confidence that the individual mechanisms found to interact together in our analyses are well connected to the outcome of poverty reduction without displacement of existing residents, since these researchers used different analytical approaches, sets of data, and usually focused on different periods of time.
- Site visits: To further refine our thinking, we also conducted site visits of neighborhoods in more than 20 U.S. cities: Baton Rouge, La.; Boston; Charlotte, N.C.; Chicago; Columbus, Ohio; Erie, Penn.; Houston; Jackson, Miss.; Knoxville, Tenn.; Louisville, Ky.; Nashville, Tenn.; New Orleans;

New York; Philadelphia; Portland, Ore.; Pittsburgh; San Francisco, Calif.; Seattle; Virginia Beach, Va.; Washington, D.C.; and Youngstown, Ohio. We walked, biked, and drove through neighborhoods in these cities that had experienced different types of changes that are outlined earlier in our methodology, to better understand how different features could interact and to improve our understanding.

- Roundtable discussions: We also held roundtable discussions with expert panels made up of researchers at think tanks that engage with place-based poverty to ensure that the findings were novel, but also connected to the experience and understanding of others working in the sector. These conversations included staff from the Brookings Institution, American Enterprise Institute, Center for American Progress, Center for Economic and Policy Research, Niskanen Center, Washington Center for Equitable Growth, and Urban Institute. A partial list of the roundtable discussion participants can be found in the Acknowledgements section on page 44 of the report.

Appendix

TABLE 2

Frequency of challenges in U.S. neighborhoods

	Neighborhoods with concentrated poverty	Other neighborhoods
Percentage of local bridges used for vehicle traffic that are rated as being in poor condition by the Federal Highway Administration	7.3%	5.6%
Proportion of neighborhoods that existed in 1960 (i.e., 50% or more of current housing units existed in 1960) and had interstates built through them	15.1%	6.8%
Average number of loans for businesses with less than $1 million in revenue	25.4	58.0
Average number of police and security officers stationed at schools per 10,000 students	3.2	0.8
Proportion of non-institutionalized residents who are disabled	17.7%	11.7%

SOURCE: Common Good Labs analysis of data from the 2019 American Community Survey, 2019 Community Reinvestment Act National Aggregates, National Bridge Inventory from the U.S. Department of Transportation Federal Highway Administration, and Civil Rights Data Collection.

TABLE 3

Federal government programs offered to support low-income communities in metropolitan areas

Recent examples of single-issue, one-size-fits-all programs

Program	Time period	Funding	Goal	Limitations
Opportunity zones	2017 to present	$11.2 billion in tax incentives, as of 2019	Encourage private investment in real estate and local businesses located in poor communities via tax incentives	Only 16 percent of eligible census tracts have received any funding through the program.
New Markets Tax Credit	2000 to present	$57.5 billion in tax credits, as of 2019	Attract private investment into businesses and nonprofit entities located in low-income communities using tax credits	Less than 10 percent of eligible communities have received investments through the program.
Enterprise Communities & Empowerment Zones Federal Grant Program	1994 to 2001	$1.8 billion in grants	Use one-time funding on locally selected programs to increase job opportunites in structurally disadvantaged communities	Program data is quite limited, but 12 metropolitan areas received over half the grant funding.*
HOPE VI	1993 to 2011	$6.7 billion in grants	Replace government-owned housing projects for low-income residents with privately-owned housing for mixed-income residents using federal grants	The number of units removed or converted was equal to less than 10 percent of all government-owned housing and less than one-third of all government-owned housing built before 1950.
EB-5 Immigrant Investor Visa	1990 to present	No direct grants or tax incentives	Increase private investment in underserved geographical areas by offering foreign investors prederential access to U.S. immigration visas	Over 95 percent of investments made by immigrants in the program go through Regional Centers, which had $0 in investment in 25 states according to the last review of the program.

SOURCE: Opportunity Zones: Campbell, Sophia and Wessel, David. "Little Evidence of Increased Demand for Property in Opportunity Zones so Far." Brookings, 15 Mar. 2021, https://www.brookings.edu/blog/up-front/2021/03/15/little-evidence-of-increased-demand-for-property-in-opportunity-zones-so-far/. Kennedy, Patrick, and Wheeler, Harrison. "Neighborhood-Level Investment from the U.S. Opportunity Zone Program: Early Evidence." Forthcoming. 15 April 2021. https://www.dropbox.com/s/zt1ws7e2py4hxsn/oz_kennedy_wheeler.pdf?dl=0. Freedman, Matthew, Khanna, Shantanu, and Neumark, David. "The Impacts of Opportunity Zones on Zone Residents." NBER, November 2021. https://www.nber.org/papers/w28573. New Markets Tax Credit: Congressional Research Service. "New Markets Tax Credit: An Introduction." 27 March 2019, https://fas.org/sgp/crs/misc/RL34402.pdf. Empowerment Zones and Enterprise Communities: Empowerment Zones, Enterprise Communities, and Renewal Communities: Comparative Overview and Analysis. https://www.everycrsreport.com/reports/R41639.html#_Toc286402522. Accessed 20 Dec. 2021. U.S. General Accounting Office. "Community Development Federal Revitalization Programs Are Being Implemented, but Data on the Use of Tax Benefits Are Limited." March 2004, https://www.gao.gov/assets/gao-04-306.pdf. Hope IV: HOPE VI Data Compilation and Analysis | HUD USER. https://www.huduser.gov/portal/pdredge/pdr-edge-research-032017.html. Accessed 20 Dec. 2021. USHMC 95: Public Housing: Image Versus Facts. https://www.huduser.gov/periodicals/ushmc/spring95/spring95.html. Accessed 20 Dec. 2021. Haltiwanger, John, et al. The Children of HOPE VI Demolitions: National Evidence on Labor Market Outcomes. National Bureau of Economic Research, Nov. 2020. Crossref, doi:10.3386/w28157. About HOPE VI - Public and Indian Housing - HUD | HUD.Gov / U.S. Department of Housing and Urban Development (HUD). https://www.hud.gov/program_offices/public_indian_housing/programs/ph/hope6/about#4. Accessed 20 Dec. 2021. EB-5: Congressional Research Service. "EB-5 Immigrant Investor Visa." 16 December 2021, https://fas.org/sgp/crs/homesec/R44475.pdf. U.S. Department of Commerce. "Estimating the Investment and Job Creation Impact of the EB-5 Program" 10 January 2017, https://www.commerce.gov/sites/default/files/migrated/reports/estimating-the-investment-and-job-creation-impact-of-the-eb-5-program_0.pdf.

TABLE 4

Prevalence of each indicator in neighborhoods with concentrated poverty

Indicator	Neighborhoods with a large decrease in poverty rate and no community displacement from 2000 to 2015	All others neighborhoods with concentrated poverty
Positive economic growth in the local metropolitan area from 2001 to 2015	5.4%	0.0%
Lower homicide rates in the local county in 2000	5.7%	1.9%
Low risk of displacement from nearby neighborhoods in 2000	5.5%	2.4%
Higher rates of homeownership in 2000	8.1%	4.0%
Lower levels of residential vacancy in 2000	6.9%	4.2%
Increased housing density in 2000	7.8%	3.3%
Higher rates of self-employment in 2000	6.5%	4.3%
Presence of community building organizations	8.1%	4.7%

SOURCE: Common Good Labs analysis of data from U.S. Census Bureau's 1990 and 2000 Decennial Censuses, Bureau of Economic Analysis Metropolitan Area Gross Domestic Product, Federal Bureau of Investigation's Uniform Crime Reporting Database, U.S. Department of Housing and Urban Development's Aggregated USPS Administrative Data on Address Vacancies, U.S. Department of Housing and Urban Development's Public Housing Data, and National Center for Urban Statistics Data.

TABLE 5

Frequency of challenges found in neighborhoods with concentrated poverty

	Neighborhoods with a large decrease in poverty rate and no community displacement from 2000 to 2015	All others neighborhoods with concentrated poverty
Proportion of neighborhoods that were in historically redlined zones	34.2%	36.0%
Proportion of neighborhoods that existed in 1960 (i.e., 50% or more of current housing units existed in 1960) and had interstates built through them	15.5%	15.8%
Proportion of neighborhoods that are designated as medically underserved	74.1%	75.3%
Proportion of tracts that are within one mile of facilities that release toxic emissions	30.1%	29.7%

SOURCE: Common Good Labs analysis.

Endnotes

1 Beznow, August, and Fikri, Kenan. "The Expanded Geography of High-Poverty Neighborhoods." Economic Innovation Group, April 2020, https://eig.org/wp-content/uploads/2020/04/Expanded-Geography-High-Poverty-Neighborhoods.pdf. Shapiro, Isaac, et al. *Basic Facts on Concentrated Poverty*. Center on Budget and Policy Priorities, 3 Nov. 2015. https://www.cbpp.org/sites/default/files/atoms/files/11-3-15hous2.pdf

2 Sharkey, Patrick, and Felix Elwert. "THE LEGACY of DISADVANTAGE: MULTIGENERATIONAL NEIGHBORHOOD EFFECTS on COGNITIVE ABILITY." Ajs; *American Journal of Sociology*, vol. 116, no. 6, 1 May 2011, pp. 1934–1981, www.ncbi.nlm.nih.gov/pmc/articles/PMC3286027/.

3 Bureau, US Census. "Census Tracts" *Census.Gov*, https://www2.census.gov/geo/pdfs/education/CensusTracts.pdf. Accessed 20 Dec. 2021.

4 Bureau, US Census. "The Urban and Rural Classifications." *Census.Gov*, https://www2.census.gov/geo/pdfs/reference/GARM/Ch12GARM.pdf. Accessed 20 Dec. 2021. https://www.census.gov/content/dam/Census/library/publications/2021/demo/p70br-172.pdf

5 Other policy research organizations, such as the Initiative for a Competitive Inner City also routinely take steps to remove undergraduate and graduate students at colleges and universities from local poverty calculations: https://icic.org/wp-content/uploads/2020/04/ICIC_CARESAct_Brief_f.pdf

6 Robert K. Nelson, LaDale Winling, Richard Marciano, Nathan Connolly, et al., "Mapping Inequality," American Panorama, ed. Robert K. Nelson and Edward L. Ayers, accessed June 11, 2020, https://dsl.richmond.edu/panorama/redlining/#loc=5/39.1/-94.58&text=downloads.

7 All tract analysis adjusted the 2000 tracts to be comparable to those in the second time period, referred to as 2015, which used the boundaries set out in the 2010 decennial census. The Longitudinal Tract Database (LTDB) is maintained by a team of sociologists and geographers and based at Brown University. More information on the LTDB and its methodology can be found here: https://www.tandfonline.com/doi/full/10.1080/24694452.2016.1187060

8 *Shifting Neighborhoods: Gentrification and Cultural Displacement in American Cities » NCRC.* https://ncrc.org/gentrification/. Accessed 20 Dec. 202

9 Tin Kam Ho. "Random decision forests." Proceedings of 3rd International Conference on Document Analysis and Recognition (1995).; Leo Breiman. "Random forests." Machine Learning (2001).

10 Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot. "Variable selection using random forests." Pattern Recognition Letters (2010); Leo Breiman. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)." Statistical Science (2001).; Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. "Random forest versus logistic regression: a large-scale benchmark experiment." BMC Bioinformatics (2018).

11 Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. "Random forest versus logistic regression: a large-scale benchmark experiment." BMC Bioinformatics (2018). Theodos, Brett. "Examining the Assumptions behind Place-Based Programs." *Urban Institute*, 8 June 2021, https://www.urban.org/research/publication/examining-assumptions-behind-place-based-programs

12 7 Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal. "Explaining Explanations: An Overview of Interpretability of Machine Learning." 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (2018).

13 We explored other tools to aid with random forest interpretability, such as permutation importance and SHAP values, but settled on feature importance for the sake of simplicity. Since our primary focus was on feature selection associated with a single outcome (i.e., large decrease in poverty with no displacement) we chose to run single-class prediction random forest classifiers, rather than those for multi-class prediction. However, we also ran classifiers on each of the other two "significant outcomes" (i.e., large increases in poverty and large increases in displacement) to better understand which features were most important for neighborhood change in these instances as well.

14 Yoshua Bengio, Andrea Lodi, Antoine Prouvost. "Machine Learning for Combinatorial Optimization: a Methodological Tour d'Horizon." European Journal of Operations Research (2020).

B | Brookings Metro

1775 Massachusetts Ave NW,
Washington, DC 20036
(202) 797-6000
www.brookings.edu