Constructing Module I of the Opportunity Atlas: Methodology

The objective of the Opportunity Atlas is to measure the average outcomes (e.g., earnings) of children who grow up in each neighborhood in America, by demographic subgroup (race, gender, and parental income). We focus on the neighborhoods where people **grew up** rather than where they live as adults because recent studies have established that the neighborhood in which a child grows up has substantial causal effects on his or her prospects of upward mobility, whereas where one lives as an adult has smaller effects.

This document provides a summary of the methods we use to construct these estimates; for a more detailed and comprehensive description, see the full paper.

Data Construction

We combine three sources of anonymized data linked by and housed at the Census Bureau:

- The 2000 and 2010 Decennial Census short form.
- Federal income tax returns for 1989, 1994, 1995, and 1998-2015.
- The 2000 Decennial Census long form and the 2005-2015 American Community Surveys (ACS). The Census long form asks a longer list of questions to a randomly selected subset of the population (covering approximately one-sixth of households). The American Community Survey asks a similar set of questions in each year 2005-2015 to a randomly selected subset of the population (approximately 1.5% of all households in each year, with a different subset each year).

Starting from these data, we construct an analysis sample of 20.5 million Americans born between 1978-1983 who are in their mid-thirties today. We map these individuals back to the Census tracts (geographic units consisting of about 4,200 people on average) that they lived in through age 23. Then, for each of the 70,000 tracts in America, we estimate children's outcomes across a range of measures.

We measure parent and child income using their percentile ranks in national income distribution. For instance, consider a child born in 1980. That child's income as an adult is ranked compared with the adult incomes of all other children in our sample who were born in 1980, and the child's parents' income is ranked compared with the incomes of all other parents of children in our sample who were born in 1980. We use percentile ranks rather than actual dollar amounts because they yield more precise, stable estimates (Chetty, Hendren, Kline, and Saez 2014).

Statistical Methods

Our goal is to estimate average outcomes (such as earnings or incarceration rates) for each demographic group in each Census tract in the United States. We face two primary challenges in constructing these estimates, which we overcome using standard statistical models.

1. The first is a data limitation: because each subgroup and income level is not represented in each Census tract, it is not possible to simply calculate average outcomes for each tract by children's race, gender, and parental income rank.

   We address this issue with a statistical (regression) model that first estimates the general relationship between parental income and children's outcomes in each tract and then uses this relationship to predict the outcomes of children for all parental income percentiles. We use national data to help estimate the shape of this relationship.

   To illustrate how our model works, suppose that a given Census tract does not contain any parents at the 25th income percentile, but contains many parents at other nearby values (e.g., the 24th and 26th percentiles). We cannot directly measure average outcomes for children whose parental incomes are at the 25th percentile, so instead our model uses the data from the other percentiles to project the hypothetical outcomes of children with parental incomes at the 25th percentile. We rely on data from national statistics – where there are plenty of children at all of these parental income percentiles – to help guide this projection (e.g., should the prediction for the 25th percentile be closer to that of the 24th, closer to the 26th, or halfway in between).

   However, we do not use statistical models to project from the outcomes of one racial or gender group to another. For instance, when projecting the outcomes of black men at the 25th parental income percentile, we use data on black men at other parental income percentiles but never data on children from other races or genders.

2. The second complication is that we must account for movement across tracts, as children often live in more than one Census tract during their childhood. We do so by assigning children to tracts in proportion to the amount of time they spent there during childhood. For instance, if a child spent half their childhood in one tract and half in another, their outcomes would count half as much for each tract as the outcomes of a child who spent their entire life in either tract.

Using these statistical models, we obtain estimates of average outcomes in each Census tract and subgroup (by race, gender, and parental income).

Publicly Available Estimates

We take three final steps before releasing the publicly-available statistics in the Opportunity Atlas.

1. We do not publish estimates based on 20 or fewer children to comply with federal data disclosure standards. In practice, estimates based on so few observations would be highly imprecise. Since a tract rarely contains subgroups of less than 20 children, excluding such estimates omits relatively little data. Our estimates that aggregate racial and gender groups cover 99.9% of individuals, while our race-specific estimates cover 96% of individuals.

2. Second, to protect privacy we add small random numbers (known as “noise”) to all our estimates to protect the privacy of individuals associated with each tract. This “noise” is typically quite small and does not affect the estimates meaningfully, but it can sometimes have a disproportionate effect on tracts with few children. For these reasons, specific estimates should be interpreted with caution, particularly in small subgroups and for selected outcomes where we display warnings that margins of error may be high.

3. We convert our estimates from percentiles back to dollars where applicable in order to facilitate interpretation. For instance, if we estimate that children at the 25th parental income percentile growing up in a given tract end up at the 50th percentile of children’s income on average, then we report in the Atlas that the average income for these children is $50,000, corresponding to the 50th percentile of the children’s national income distribution in 2015.

We follow this procedure to report estimates for a broad range of outcomes, including earnings (household income in 2014-2015), incarceration rates (as measured in the 2010 Decennial Census on April 1, 2010), and educational attainment (high-school degree or 4-year-college degree). We provide estimates for all children in a tract, as well as for the following demographic subgroups:

- **Parental Income** - We provide estimates at five specific household income levels. The Basic version of the Atlas includes three income levels: *Low* (25th percentile or $27,000/year), *Middle* (50th percentile or $55,000/year), and *High* (75th percentile or $94,000/year). The “Advanced” version additionally includes *Lowest* (0th percentile or $2,200/year) and *Highest* (100th percentile or $1,500,000/year).

- **Child Race** - We provide estimates grouped by five race and ethnicity categories. Categories include *Hispanic* (based on reported ethnicity, including all races) as well as four racial categories for non-Hispanic individuals: *White*, *Black*, *Asian*, and *American Indian* (including Alaskan Natives). Non-Hispanic individuals who report two or more races, Native Hawaiian or Pacific Islander, or Some Other Race are combined in a sixth category (“Other”) which is available in data files from our website but not included in the online Atlas.

- **Child Gender** - We also report estimates by the child’s gender (*Male* and *Female*).

In addition to Census tract-level estimates, we also release estimates at the county and commuting zone (CZ) levels, which are constructed using methods analogous to those described above.

Because outcomes in the American Community Survey (such as college graduation rates) are available for only a smaller sample of individuals, we report our estimates for these outcomes at the county and CZ levels only. We also suppress certain estimates which are measured with substantial error at the tract level, such as the fraction of individuals who reach the top 1% of the income distribution and incarceration rates for women.

A list of all the variables available in the Opportunity Atlas – as well as detailed variable definitions are available in Section II of the research paper and on this data page.

For further details on the methods, please refer to Section III of the paper or these slides.

Constructing Module II of the Opportunity Atlas: Methodology

The objective of Module II of the Opportunity Atlas is to measure the changes in average outcomes (e.g., household income) of children who grow up in each neighborhood in America, by demographic subgroup (race, gender, and parental income). We focus on the neighborhoods where people **grew up** rather than where they live as adults because recent studies have established that the neighborhood in which a child grows up has substantial causal effects on his or her prospects of upward mobility, whereas where one lives as an adult has smaller effects.

This document provides a summary of the methods we use to construct these estimates; for a more detailed and comprehensive description, see the full paper.

Data Construction

We combine three sources of anonymized data linked by and housed at the Census Bureau:

- The 2000 and 2010 Decennial Census short forms.
- Federal income tax returns for 1984, 1989, 1994, 1995, and 1998-2019.
- The 2000 Decennial Census long form and the 2005-2019 American Community Surveys (ACS). The Census long form asks a longer list of questions to a randomly selected subset of the population (covering approximately one-sixth of households). The American Community Survey asks a similar set of questions in each year 2005-2019 to a randomly selected subset of the population (approximately 2.5% of all households in each year, with a different subset each year).

Starting from these data, we construct an analysis sample of Americans born between 1978-1992 and measure their outcomes at age 27 (between 2005-2019). We map these individuals back to the counties that they lived in through age 18.[1] Then, for each county in America, we estimate children's outcomes across a range of measures.

We measure parent and child income using their percentile ranks in the national income distribution. For instance, consider a child born in 1980. That child's income as an adult is ranked compared with the adult incomes of all other children in our sample who were born in 1980, and the child's parents' income is ranked compared with the incomes of all other parents of children in our sample who were born in 1980. We use percentile ranks rather than actual

[1] We omit children who cannot be linked to parents (0.4% of children); children whose mean parental income is zero or negative, since this is typically due to large capital losses and is a proxy for significant wealth (3.7% of children); and children with missing childhood location information (3.4% of children).

dollar amounts because they yield more precise, stable estimates (Chetty, Hendren, Kline, and Saez 2014).

Statistical Methods

Our goal is to estimate changes in average outcomes (such as household income) for each demographic group in each county in the United States. We face two primary challenges in constructing these estimates, which we overcome using standard statistical models.

1. The first is a data limitation: because each subgroup and birth cohort is not represented in each county, it is not possible to simply calculate changes in average outcomes for each county by children's race, gender, and parental income rank.

   We address this issue with a statistical (regression) model that first estimates the general relationship between parental income and children's outcomes in each county and birth cohort, and then uses this relationship to predict the outcomes of children for all parental income percentiles in the same county and birth cohort. We use national data to help estimate the shape of this relationship.

   To illustrate how our model works, suppose that a given county does not contain any parents at the 25th income percentile, but contains many parents at other nearby values (e.g., the 24th and 26th percentiles). We cannot directly measure average outcomes for children whose parental incomes are at the 25th percentile, so instead our model uses the data from the other percentiles to project the hypothetical outcomes of children with parental incomes at the 25th percentile. We rely on data from national statistics – where there are plenty of children at all of these parental income percentiles – to help guide this projection (e.g., should the prediction for the 25th percentile be closer to that of the 24th, closer to the 26th, or halfway in between).

   However, we do not use statistical models to project from the outcomes of one racial or gender group to another. For instance, when projecting the outcomes of Black men at the 25th parental income percentile, we use data on Black men at other parental income percentiles but never data on children from other races or genders.

   We then estimate changes in average outcomes across birth cohorts as a linear trend, separately for each county and subgroup. We also report the endpoints of this trend as the average outcomes for children in the 1978 and 1992 birth cohorts.

2. The second complication is that we must account for movement across counties, as children sometimes live in more than one county during their childhood. We do so by assigning children to counties in proportion to the amount of time they spent there during childhood. For instance, if a child spent half their childhood in one county and half in another, their outcomes would count half as much for each county as the outcomes of a child who spent their entire life in either county.

Using these statistical models, we obtain estimates of changes in average outcomes in each county and subgroup (by race, gender, and parental income).

Publicly Available Estimates

We take three final steps before releasing the publicly available statistics in Module II of the Opportunity Atlas.

1. We do not publish estimates based on 20 or fewer children to comply with federal data disclosure standards.[2] In practice, estimates based on so few observations would be highly imprecise. Since a county rarely contains subgroups of fewer than 20 children, excluding such estimates omits relatively little data.

2. To facilitate comparisons within a subgroup across counties, we also report the statistical reliability of outcomes in our data. Statistical reliability is a function of both the number of children in a county and subgroup, and the variation in outcomes across counties within that subgroup. For example, if there was little variation in outcomes across counties, cross-county comparisons will be sensitive to statistical noise and statistical reliability will be low, even if each individual county's estimates are based on a modestly large number of children.

   In the online Atlas data tool, we omit estimates below a statistical reliability threshold of 0.3. We do not impose this restriction in our online data tables. Our count and reliability restrictions collectively omit fewer than 8% of children in our sample for subgroup (race and gender)-specific estimates, and fewer than 0.1% of children in our sample for pooled (race and gender) estimates.

3. We convert our estimates from percentiles back to dollars where applicable to facilitate interpretation. For instance, if we estimate that children at the 25th parental income percentile growing up in a given county end up at the 50th percentile of children's income on average, then we report in the Atlas that the average income for these children is $35,550, corresponding to the 50th percentile of the children's national income distribution in 2023 dollars.

We follow this procedure to report estimates for changes in children's outcomes in adulthood, including household income and individual income (measured at age 27). We provide estimates for all children in a county, as well as for the following demographic subgroups:

- **Parental Income** - We provide estimates at five specific household income levels. The Basic version of the Atlas includes three income levels: *Low* (25th percentile or $33,320/year), *Middle* (50th percentile or $69,520/year), and *High* (75th percentile or

[2] Specifically, we require that for each county and subgroup (by race and gender), there must be at least 20 children across all parental income percentiles and at least 4 birth cohorts.

$122,040/year). The “Advanced” version additionally includes *Lowest* (0th percentile or $1,160/year) and *Highest* (100th percentile or $1,840,000/year).

- **Child Race** - We provide estimates grouped by five race and ethnicity categories. Categories include *Hispanic* (based on reported ethnicity, including all races) as well as four racial categories for non-Hispanic individuals: *White*, *Black*, *Asian*, and *American Indian* (including Alaskan Natives).

- **Child Gender** - We also report estimates by the child’s gender (*Male* and *Female*).

In addition to county-level estimates, we also release estimates at commuting zone (CZ) level, which are constructed using methods analogous to those described above.

We also report covariates (e.g., change in poverty rates) derived from publicly available data like the Decennial Census and the ACS. In the online Atlas data tool, we omit the bottom 1% of counties and CZs by population, as well as estimates below the 1st percentile or above the 99th percentile of the national distribution. In the online data tables, we do not impose this restriction.

A list of all the variables available in Module II of the Opportunity Atlas, as well as detailed variable definitions, are available in Section II of the research paper and on this data page.

For further details on the methods, please refer to Appendix A of the paper.