Working Papers
www.mmg.mpg.de/workingpapers

Max Planck Institute for the Study of Religious and Ethnic Diversity

Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften

MMG Working Paper 21-01 • ISSN 2192-2357

Kasey Zapatka[1]
(The Graduate Center, City University of New York)

Superdiversity in Metropolitan New York - Technical Report

Kasey Zapatka (The Graduate Center, City University of New York)
Superdiversity in Metropolitan New York – Technical Report

MMG Working Paper 21-01

Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften,
Max Planck Institute for the Study of Religious and Ethnic Diversity
Göttingen

ISSN 2192-2357 (MMG Working Papers Print)

Download: **www.mmg.mpg.de/workingpapers**

MPI zur Erforschung multireligiöser und multiethnischer Gesellschaften
MPI for the Study of Religious and Ethnic Diversity, Göttingen
Hermann-Föge-Weg 11, 37073 Göttingen, Germany
Tel.: +49 (551) 4956 - 0
Fax: +49 (551) 4956 - 170

www.mmg.mpg.de

info@mmg.mpg.de

Abstract

This report describes the data and materials used to produce the *Superdiversity in Metropolitan New York* visualization website (**www.superdiv-newyork.mmg.mpg.de**), launched in 2020. This includes the data sources, design principles and methods, as well as the properties and categories of the variables used in the visualizations presented on the website.

Keywords: - still missing -

Authors

- still missing -

Contents

Background

The *Superdiversity in Metropolitan New York* website expands the scope of the original *Superdiversity visualization* (**www.superdiv.mmg.mpg.de**) to include New York City and its surrounding metropolitan area. The New York City-based team consisted of Professor Philip Kasinitz, Professor Van Tran, and Doctoral Candidate Kasey Zapatka, from The Graduate Center, City University of New York. Together, they worked with the site's original co-creators to compare the diversification of the New York metropolitan area with the cities included in the original *Superdiversity Visualization*: Vancouver, Sydney, and Auckland.

This technical report does not discuss the conceptual ideas behind the website—such as the definition of superdiversity or its practical implications—but instead describes the data sources and methods for the interactive visualization on the website. For more on the background of the original *Superdiversity visualization*, we direct readers to the original **Superdiversity and Cities - Technical Report.**[2]

Data Sources

We used two main types of data to build these visualizations: administrative and survey data. The administrative data are collected and maintained by the US Department of Homeland Security, whereas the survey data are from the US Census Bureau.

Administrative Data

Many countries maintain administrative data on visas granted to individuals entering their territories. Statistical data on immigration have been published annually by the US government since the 1860s. While the federal agency responsible for recording these data has changed over the years—along with the content, format, and title of the annual publications—immigration data are currently published in the annual ***Yearbook of Immigration Statistics*** by the Office of Immigration Statistics in the Office of Strategy, Policy, and Plans of the Department of Homeland Security.

---

1 Direct correspondence to Kasey Zapatka via email at **kzapatka@gradcenter.cuny.edu**.

2 Hiebert, Dan. 2019. "Superdiversity and Cities - Technical Report." MMG Working Paper 19-05. Gottingen, Germany. **https://www.mmg.mpg.de/499294/wp-19-05**.

The Department of Homeland Security (DHS) reports the annual number of people from each country around the world who were admitted to the United States. For simplicity, we have aggregated these data into three general categories based on individual status:[3]

- **Legal Permanent Resident**: This category includes immigrants who received their "green card" and have been granted lawful permanent residence in the United States. They could have obtained their lawful permanent residence status due to (1) a family member sponsoring their vis (family-based) a; (2) an employer sponsoring their visa (employment-based); (3) they were a part of the Diversity Visa Program (diversity-based); or (4) for some other reason.[4]
- **Refugee and Asylee Status:** This category includes immigrants entering the United States either as a refugee or those granted asylum while in the United States. Both groups are unable or unwilling to return to their country of nationality because of fear of persecution based on their race, religion, nationality, social group status, or their political opinion.[5] Refugee applicants are screened outside of the US, whereas asylum applicants apply at a US port of entry upon arrival. While the DHS data differentiates between affirmative and defensive asylum applicants,[6] we combined both categories in our analyses.
- **Temporary Non-Immigrant Admission:** This category includes individuals who entered the US under a temporary non-immigrant status such as a student visa, a temporary worker visa, for diplomatic reasons, and a residual category for all other classes of temporary admission. We decided not to include business and tourism visas because the sheer number—especially from specific countries such as Mexico—would overwhelm the visualizations. We created a residual category that included the children and spouses of foreign government officials, the minor chil-

3 We excluded naturalization data because legal permanent residents are eligible to apply for naturalization after the required five-year waiting period but not all of them will do so.

4 For specific year(s), this category includes persons entering under the Amerasian, former H-1 registered nurse, Cuban/Haitian entrant, Soviet and Indochinese parolee, and 1972 Registry provisions. It also includes individuals whose removal or deportation was canceled and those who were covered under **IRCA legalization**, which granted legal permanent resident status to persons unlawfully residing in the US before January 1982 or seasonal agricultural workers residing in the US for at least 90 days before May 1982.

5 See the Department of Homeland Security's 2018 **Annual Flow Report** for more on refugees and asylees.

6 Affirmative asylum applicants apply within a year of their arrival in the US, regardless of immigration status or entry method. By contrast, defensive asylum applicants apply during removal proceedings in immigration court to defend against deportation.

dren of fiancees, and individuals whose admission status was unknown. Unfortunately, data suppressed to protect confidentiality creates some minor discrepancies in country-level totals. Also note that the majority of short-term admissions from Canada and Mexico are excluded by the DHS and therefore from our calculations.

We compiled data by sending region and sending country. Following DHS classifications, we adopted the following categories for sending region: Africa, Asia, Europe, North America, Oceania, and South America. Sending country indicates "country of birth" in data for Legal Permanent Residents admissions, "country of nationality" for Refugees and Asylee admissions, and "country of citizenship" for other Non-Immigrant admissions. Since data at the metropolitan area are only available for Legal Permanent Residents, we limited our analysis to the national scale to maintain data consistency across the three immigrant categories.

Our analyses draw on data tables from the 1998-2019 *Yearbook of Immigration Statistics*. *Yearbook* tables are publicly available and are released as they become available. They are compiled together as a PDF and made available in September following the end of the previous fiscal year.

Census Data

All of the remaining website visualizations are based on data originating with the US Census Bureau. As required by Article 1, Sections 2 and 9 of the US Constitution, the Census Bureau conducts a national census designed to count every person in the country. Beginning in 1790 and conducted every ten years since, these constitutionally mandated enumerations are used to determine representation in US House of Representatives.

In the early years of the census, respondents were only asked basic questions about sex, race, and age, for example. However, as the list of research questions grew and became more detailed, it became known as the census "long-form" survey. Although the decennial census provided a trove of useful information for researchers and data users, demand quickly grew for more timely data. In response to this growing demand, the US Census Bureau developed the American Community Survey (ACS), which uses a rolling sample design to give a snapshot of US population every year. First introduced after the 2000 Decennial Census, it was designed to replace the decennial long form survey.[7] Today, the decennial census continues to count every

7 More information on the design and implementation of the American Community Survey **can be found on their website**.

living person in the United States every ten years and **asks a few basic questions** (household size, sex, age, and race-ethnicity), whereas the ACS is conducted every year using a smaller sample of approximately 3.5 million individuals and **asks more detailed questions** (e.g., educational attainment, internet access, transportation, and marriage status, among others).

Unit of Analysis. We use US Census data for two different levels of analysis: individual-level microdata and census tract-level aggregations. While US census microdata are available **directly from the US Census using their new API**, they can be quite difficult to work with because of the changes to variable measurement and coding as well as different sample sizes across years. To circumvent arduous tasks like harmonizing variables, assigning uniform variable codes, and developing complicated weighting schemes, we use microdata obtained from **IPUMS** that have already addressed these issues. [8]

As a part of the University of Minnesota's **Institute for Social Research and Data Innovation**, IPUMS has received many federal grants over the last 25 years that have enabled them to integrate, harmonize, and make publicly available the country's most comprehensive census database. IPUMS has also received grants from other government agencies like the National Institute of Health, the National Science Foundation, and the Food and Drug Administration to conduct comparative analyses, create targeted reports, and support their maintenance of their comprehensive database. These individual-level microdata from IPUMs are the underlying data for the Sankey Diagram, the Bubble Chart, and the Intersectionality Dashboard.

We use data aggregated to the census tract-level to create the *Choropleth Map*. These are the smallest geographical unit for which reliable estimates from the 2014-2018 ACS 5-year estimates can be obtained.[9] Census tracts are small statistical subdivisions of a county that do not cross county boundaries. While they can vary in size from 1,200 to 8,000 people, the average population is between 4,000 and 5,000 people. Census tract boundaries can vary in size depending on population density, but the intention is that boundaries are generally maintained to facilitate comparisons over time. When they do change, tracts are either split due to population growth

8 Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2021. **https://doi.org/10.18128/D010.V11.0**

9 While some researchers use smaller, geographic units like census blocks or census block groups, census tracts in the NY-NJ-PA metropolitan area are the smallest but most reliable estimates appropriate for our analyses. The **US Census Bureau's website** has more on how margins of error and confidence intervals are calculated.

or merged as a result of population loss. Sophisticated techniques to address tract boundary changes were not necessary in our analysis because we did not use these data in any of our analyses that compare change over time.

New York Metropolitan Area. We define the spatial scope for this project as the New York-Newark-Jersey City, NY-NJ-PA Metropolitan Statistical Area (NY-NJ-PA metropolitan area).[10] Under these delineations, a metropolitan area is defined as a "region consisting of a large urban core together with surrounding communities that have a high degree of economic and social integration with the urban core."[11]

One difficulty in simultaneously working with individual-level microdata and census tract-level aggregated estimates is ensuring equivalent metropolitan area definitions for each analysis. This was complicated because microdata and census tract aggregations do not have a shared geographic variable to allow users to approximate metropolitan area delineations between estimates. To ensure consistent metropolitan boundaries for both analyses, we relied on another census-defined geography—**Public Use Microdata Areas (PUMAs)**, which generally follow county groups or census-defined "places" and contain at least 100,000 people. Using the MET2013 variable to delineate consistent boundaries for NY-NJ-PA metropolitan area for the 2014-2018 sample, we listed all the PUMAs within the metropolitan area. Then, we located each census tract in their respective 2014-2018 PUMAs using a crosswalk from the US Census Bureau.[12] This process ensured equivalent NY-NJ-PA metropolitan areas for the choropleth maps, which only use 2014-2018 ACS estimates.[13]

---

10 The **MET2013 variable** follows the **2013 Office of Management and Budget's metropolitan area delineations**. These were the most recent delineations incorporated into census data at time of analysis and were harmonized using crosswalks to create consistent samples from 2000 or later.

11 Since metropolitan area boundaries change over time, IPUMS relies on a census-defined geography called **Public Use Microdata Areas (PUMAs)** to harmonize metropolitan area boundaries. Data users and researchers can use the IPUMS variable, **MET2013**, to delineate harmonized metropolitan boundaries from 2000 onward. However, since PUMAs are the only sub-state-level variable available in census microdata, IPUMS cannot exactly identify each respondent's residence in a given metropolitan area. So, if the majority of a PUMA's population in which an individual resides is located within a metro area, then that individual is designated as residing within the given metro area. Since this protocol can yield two different types of errors, IPUMS provides a variable, **MET2013ERR**, to report the mismatch in the share of residents living in the metropolitan area and the share not using Census summary data. For the NY-NJ-PA metropolitan area, there is less than a 1% coverage error.

12 We used the "**2010 Census Tract to 2010 PUMA Relationship File**" as a crosswalk.

13 Although PUMAs can change over time, no longitudinal comparisons were made using PUMAs.

Survey Variables

The population for our sample not only encompasses private households, but also **group quarters**, which include institutions like military bases, hospitals, and prisons in which a number of unrelated individuals are living. While only 4.13 percent of respondents in our sample lived in group quarters, we included them to ensure equal representation across diverse populations. However, since respondents living in group quarters are sampled as individuals, many household-level variables like household income or variables created as a result of household level variables (e.g., housing affordability) are not available. All survey data from the US Census data was self-reported, unless otherwise noted.

Another decision that affected the scope of our sample was to include respondents **who were still in school** when they were surveyed. While this can complicate a clear interpretation of some educational outcomes (e.g., a student enrolled in college might declare they are in college but drop out after the survey), we felt this population should be included since students are an important sub-population that constitute about 25 percent of our sample. This is especially important given that approximately 400,000 city residents are enrolled in New York City's public university system, the City University of New York. Since we use census data to estimate the overall educational distribution of the sample—not educational attainment among a specific population, including respondents who were still in school accounts for the level of education that these respondents had achieved.

Since the ACS is conducted on a rolling basis every year, it provides us a timelier sampling frame than the last decennial census, which was conducted in 2010. Moreover, its design allows users to combine estimates together over multiple years to increase sample size and "statistical reliability of the data for less populated areas and small population subgroups."[14] We therefore use the 2014-2018 5-year ACS estimates—the most recent data available at the time of analysis—to build the data visualizations used for this. [15] In each sampled household, the Census Reference Person is expected to provide information about each individual in the household as well as information about the household as a whole (e.g., ownership status).

---

14 More information on the use of American Community Survey estimates **can be found on their website**.

15 2015-2019 ACS 5-Year Estimates were made available after initial analyses were completed. However, since the ACS is essentially a rolling-five-year summary, we didn't think including these data instead would have made any difference since we would essentially be swapping 2014-data out with 2019 data with the other four years remaining the same.

The remainder of this section reproduces the specific census questions and explains the variables used.

Sex: "What is this person's sex?" (coded in our data as Male, Female).

Age: "What is this person's age and what is this person's date of birth?" Babies are reported as age 0 when less than 1 year old. In several visualizations, populations were limited to working age individuals by limiting the sample to individuals between the ages of 18 and 64. (Micro-data: age is reported as a continuous variable; census tract aggregations: reports the number of individuals belonging to specific age brackets: 18-24, 25-34, 35-44, 45-54, 55-64, all else are consider not working age).

Language: "Does this person speak a language other than English at home?" (Yes, No). If the answer is "Yes", then the respondents is asked: "What is this language?

Year of immigration: "When did this person come to live in the United States?" (Before 1980; 1980-1990; 1991-2000; 2001-2010; 2011-2018).

Ancestry and ethnic origin: "What is this person's ancestry or ethnic origin?" Respondents could list as many responses as they wished to self-report; however, only the first two responses were reported. Respondents were instructed not to report religion because the census is not allowed to collect information on religion. Compound responses like "Pennsylvanian Dutch" or "French Canadian" were treated as one response. Generally, responses listed "Uncodable" or "Other" were usually religion.

Geographic Mobility: "Did this person live in this house or apartment one year ago?" In combination with other census questions, census officials recode this variable as (1) same house, (2) moved within state, (3) moved between states (4) moved from abroad one year ago (Same house vs. not the same house one year ago).

Race, ethnicity, and Hispanic origin: Several variables in our analysis rely on the major categories of race, ethnicity, and Hispanic origin in US society. To create these variables, we combine two variables: **race** and **Hispanic origin**. In response to questions about race ("What is this person's race?"), respondents can enter in their own response and mark more than one box. In response to questions about Hispanic origin ("Is this person of Hispanic, Latino, or Spanish origin?"), respondents can enter their own response or chose from a few common answers.

Allowing respondents to self-report and check more than one box can make analyzing race-ethnicity in the US is very complicated. To simplify, we combine both race, ethnicity, and Hispanic origin into a single variable that is mutually exclusive and exhaustive. This approach records all Hispanic responses as one category; reports those who identify *only* as White, Black, and Asian (i.e. non-Hispanic) in three sep-

arate categories; and marks all other or multiple responses as a residual category: non-Hispanic Other.[16] This results in five mutually exclusive and exhaustive categories (non-Hispanic White, non-Hispanic Black, nonHispanic Asian, non-Hispanic Other, and Hispanic). Those that self-report as Hispanic are analyzed separately using the Hispanic variable. This framework was used throughout this report.

In two visualizations (the Bubble Chart and the Intersectionality Dashboard) we also explore the outcomes for the top ten national origin groups in metropolitan New York in 2014-2018. We determine these top ten groups by creating a weighted frequency table listing total population estimates for each group in the NY-NJ-PA metropolitan area. Responses are based on individual self-reporting of **ancestry or ethnic origin** as either a first or second option. We use both responses to increase sample size and reliability of our estimates. In descending order by group size, these ten ethnic groups are Dominican, Puerto Rican, Chinese,[17] Mexican, Indian,[18] Ecuadorian, Jamaican, Colombian, Haitian and Filipino.

Birthplace: "Where was this person born?" Respondents were given the option of (1) inside the US or (2) to name a country outside of the US (native vs foreign-born)

Education: "What is the highest degree or level of school this person has COMPLETED?" Note that the universe of respondents only includes respondents 25 years or older for census tract-level data (Less than high school degree; High school degree or GED; Postsecondary education without a university degree; University degree or higher).

Employment status: Respondents were asked, "LAST WEEK, did this person work for pay at a job (or business)?" and "LAST WEEK, did this person do ANY work for pay, even for as little as one hour?" Answers to these questions are used in simplified variables to indicate whether a person is employed (including active self-employment), unemployed, or not in the labor force. Note that while universe of respondents only includes respondents 16 years or older for census tract-level data, we use total population as the denominator when calculating employment rates to

16 American Indian/Native American, other major race, two or more races, and three or more major races are all included in the "non-Hispanic Other" category.

17 To create a variable that counted Chinese ancestry, we collapsed first and second responses for Chinese, Cantonese, Manchurian, Mandarin, Mongolian, Tibetan, Hong Kong, and Macao into one "Chinese" category. Taiwanese was not counted as Chinese.

18 To create variable that counted Indian ancestry, we collapsed first and second responses for Asian Indian, Andaman Islander, Andhra Pradesh, Assamese, Goanese, Gujarati, Karnatakan, Keralan, Maharashtran, Madrasi, Mysore, Naga, Pondicherry, Pondicherry, and Tamil into one "Indian" category.

capture employment as a function of all people, including those not in the labor force (i.e., homemakers or those who have permanently moved out of the labor force). This presents a clearer picture of employment inequality among groups than traditional US methods (Employed vs. not employed).

Home ownership: Respondents were asked, "Is this a house, apartment, or mobile home" and to check which box is applicable: (1) Owned by you or someone in this household with a mortgage or loan (including home equity loans)?, (2) Owned by you or someone in this household free and clear (without a mortgage or loan), (3) Rented, or (4) Occupied without payment of rent? (Homeowner vs. not a homeowner).

Household income: Household income data are based on a respondent's total pretax, self-reported personal income (or losses) from all sources for the previous year. Household income reports all incomes from all household members that are over the age of 15 from the previous year and differs from family income, which reports the household incomes of family members living in the household. For individuallevel microdata, exact income values are reported. For tract-level aggregations, the US Census reports distribution of household income divided into 16 different income buckets, where the number of households falling into each bucket are reported within each census tract. Thus, poor areas will have a preponderance of their population in lower-income buckets, while more affluent areas will see more of their tract's households falling into higher-income buckets. Negative and zero incomes were included in this measure and all monetary values were inflation-adjusted to 2018 dollars (Microdata: household income is reported as a continuous variable; census tract aggregations: number of households falling into each of the 16 buckets).

Housing affordability: This variable was calculated as the ratio of annual household housing costs to annual household income (housing costs/household income). Since a household includes all individuals in a household, household income is the sum of all self-reported personal incomes of each individual in the household. Housing costs were calculated separately for renters and homeowners, using **monthly gross rent** (which includes utilities) and **selected monthly owner costs** (which is the total of all mortgage payments, property taxes, various insurance payments, utilities, and fuel costs). Each measure of monthly housing cost was converted to an annual measure before calculating housing affordability. So that our measure is comparable with prior work, households where the ratio of annual housing costs to annual household income exceeded 30 percent were considered to be unaffordable. In creating this measure, negative and zero incomes were included, since households could have debt that could impact the affordability of their living situations. All monetary values

were inflation-adjusted to 2018 dollars (Living in affordable housing vs. not living in affordable housing).

Low income: We measure low-income populations in relation to the poverty line for that given year. So, each family's total income for the previous year is calculated as a percentage of the poverty thresholds established by the Social Security Administration in 1964, subsequently revised in 1980, and adjusted for inflation to current values. Each family member is assigned the same poverty code. Poverty status is adjusted for family size, the number of children in the family, and the age of the householder. Individuals who live in a household where the family income is 5 times greater than the poverty line are top-coded, receiving the same score of 501, or 501% of the poverty line. We take a conservative estimate of what it means to be low-income and classify individuals living at or below 150% of the poverty line as low-income (Living at or below 150% of poverty line vs. above 150% of the poverty line).

Data Visualizations

Tree Diagram

The Tree Diagram is the only visualization that uses administrative data from the DHS. Data for the three categories described above (legal permanent residents, refugee and asylees, and non-immigrants admissions) were collected, cleaned, and organized in two spreadsheets for each year, with origin countries and their respective regions arranged as rows and the categories of entry as columns. The first spreadsheet contains information on the number of refugee/asylees and legal permanent residents (broken down by the number family-visas, employment visas, diversity visas, and other visas). The second spreadsheet records data on non-immigrant admissions for students, temporary workers, diplomats, and an additional other/unknown status.

Stamen used several tools to build the code that converts these data into interactive visualizations and decided on the colour palette (see the section on website code and design, below). The overall design, with the graph situated above the shifting rectangle of source countries, was defined through a series of conversations between Stamen and the co-authors of the project.

Sankey Diagram

Sankey visualizations show flows between two variables where line widths are proportional to volume. We the Sanky Diagram to visualize patterns of interaction between ethnicity/ancestry, and language spoken at home for residents of the NY-NJ-PA metropolitan area population. This visualization uses microdata from the 2000 Decennial Census and the 2014-2018 5-year ACS estimates.

Unlike with the Bubble Chart and the Intersectionality Dashboard, we only use the first response to survey question about ancestry/ethnicity to avoid unnecessary complication that would results with multiple responses. Responses to questions about the language spoken at home were divided into a larger subgroup based on language family.[19] About 6% of our sample didn't report speaking a language, likely because they were too young to speak a language. Respondents for whom the language reported was not classifiable were included in an „Other and Unknown" category, which amounted to less than 0.1 percent.

Stamen Design created the code to convert the Excel matrices into Sankeys, using a combination of Javascript libraries and their own work (see below). The code for visually highlighting particular groups on the Sankey, and the magnifying tool, is an innovation designed by Stamen.

Bubble Chart

The Bubble Chart is designed to provide an overview of the relationship between ethnic diversity and socio-economic outcomes in the NY-NJ-PA metropolitan area. Since employment is one the key outcomes, the scope of the visualization was limited to working age population (individuals between the ages of 18 and 64). Similar cross tabulations were calculated for working-age immigrants that arrived, on a rolling basis, to the United States within the last five years.[20] The construction of the Bubble Chart followed the same process as with the original **Superdiversity and Cities - Technical Report** and was designed entirely in-house by Stamen and built using Python code.

---

19 Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2020. **Ethnologue: Languages of the World**. Twenty-third edition. Dallas, Texas: SIL International.

20 Since the ACS data we use follow a rolling-sample design and combine data from each of the years between 2014 and 2018, "within the last five years" means within the last five years of each respective year of data. So, for 2014, immigrants who arrived within the last five years could have arrived in any year between 2010 and 2014.

Data for the Bubble Chart were created using cross tabulations between one of the race-ethnicity variables (e.g., non-Hispanic White, non-Hispanic Black, non-Hispanic Other, etc.) or top ten national origin groups (e.g., Dominican, Chinese, Mexican, etc.) and each of the four socio-economic indicators: Working, University degree; Low income; and Home ownership. Each of these indicators was treated as a binary variable (e.g., University degree was a binary variable that respondents had a college degree or higher). Note that we break with traditional US practices of calculating employment rate using labor force population as the denominator. Instead, we divide by total population to capture employment as a function of all people including those not in the labor force (i.e., homemakers or those who have permanently moved out of the labor force). This presents a clearer picture of employment inequality within groups than traditional US methods. Data for this chart were cleaned and formatted in Stata and exported to Excel spreadsheets for use on the website by Stamen.

Bivariate Choropleth Maps

Two sets of maps were created for NY-NJ-PA metropolitan area using census-tract-level data (4,683 census tracts). We use census tracts because they are the finest geographical scale available with reliable estimates. The first set of "traditional" maps used similar variables as in the rest of the project, but with estimates collected at the census tract level. The "traditional" maps included data on the number of immigrants, the number of recent immigrants (arrived in the US in 2010 or later), the number of high- income households (households with incomes above approximately $150,000 or the 80th percentile),[21] and the number individuals belonging to each of the five race-ethnicity groups (non-Hispanic White, nonHispanic Black, non-Hispanic Asian, non-Hispanic Other, and Hispanic). The population universe for immigrants, recent immigrants, and race-ethnicity variables was the total population in each census tract, but for high-income, it was the total number of households for which there was data.

---

21 We used individual-level microdata to approximate the lower threshold of the 80th percentile for the entire NY-NJ-PA metropolitan area. This approximated $160,000, but since census tract-level data are only available aggregated to the geographic level and distributed into one of 16 pre-determined income buckets, the closest we could come to the $160,000 threshold was $150,000. However, we calculated that separating the data in this way captured approximately 21 percent of high-income households, which we feel is very close to the incomes at or above the 80th percentile.

Calculations were standardized using a Location Quotient (LQ) statistic, where a number was derived for each cell by dividing the percentage of the estimate in a census tract by the percentage of estimated total population.[22] For example, if 10 percent of the people in the metropolitan area identify as Chinese in origin, the LQ for a census tract with 5 percent Chinese would be 0.5 (0.05/0.10 = 0.50). Similarly, in a tract where 25 percent of its population were Chinese, the LQ would be 2.5 (0.25/0.10 = 2.5). LQ is a measure of relative concentration in a census tract in relation to the overall metropolitan area population.

LQ values were calculated for each census tract for each variable and Stamen created a map visualization of all six variables using Mapbox software and Python programming. This involved merging the census tract-based data with polygon shapefiles provided by the US Census Bureau for the NY-NJ-PA metropolitan area .[23] A blue colour ramp was chosen to illustrate the degree of concentration of each variable across census tracts. Each tract was classified based on ranking LQ values into 15 categories (quantiles), where the lowest LQ values were lighter blues and higher LQs were darker blues. In practice, maps are dominated by low-intensity coloured areas because all zero values (i.e., where a particular group is absent from a geographical area) are in the lowest quantile, and zero values are very common given the extraordinarily detailed geographical areas.

A second set of "superdiversity" maps were also created to highlight areas of high social complexity. These maps included data on the number of distinct ancestries present in each census tract; the share of each census tract's population that moved into the tract within the last year; the number of individuals in each income bucket; the decade of arrival of each tract's immigrants (i.e., arrived before 1990, 1990-1999, 2000-2010, or after 2010); and the number of individuals with varying levels of educational attainment (less than school degree, high school degree, postsecondary education without a degree, college degree or higher).

Ancestry was standardized by dividing the number of different ancestral/ethnic groups present in a census tract. This ensured that a tract's diversity was not dependent on its population but that the diversity statistic was adjusted for a tract's pop-

---

22 See the page 24 of the original **Superdiversity and Cities - Technical Report** for more on location quotients.

23 US Census data were accessed using the **tidycensus package in R**. Kyle Walker and Matt Herman (2021). tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames. R package version 1.0. **https://CRAN.R- project.org/package=tidycensus**

ulation (otherwise, census tracts with larger populations would by definition have greater diversity since they are more likely to have more distinct ethnicity/ancestral groups). This measure was intended to illustrate the ethnic diversity in each census tract. Mobility was calculated by dividing the number of a tract's new residents in the last year by its total population to represent the degree of "churn" in a census tract.

Using Simpson's Generalized Index of Entropy (SI), we calculated a measure of diversity for each census tract for each of the three remaining variables (recent immigration, educational attainment, and household income). The formula for SI is expressed as:

$$\mathrm{SI}=1-\sum_{k=1}^{k}(P_k)^2$$

where P equals the proportion of each k^{th}subgroup within a census tract for a given variable. Our calculations for this census tract yield an SI value, where higher values indicate a higher degree of diversity across the different groups of a variable. Together these calculations indicate the degree of educational, income, or immigration diversity in a given census tract.

The bivariate choropleth maps illustrate the degree of diversity of two out of five variables at the same time (i.e. diversity in education, income, immigration, and ethnicity, as well as geographic mobility). The Mapbox software used by Stamen requires these diversity/mobility values to be ranked as low/medium/high values, since the software can only depict a total of nine colours in any given map. Accordingly, each variable was ranked into three equal-sized tercile and each census tract was assigned a ranked order value of 1, 2, or 3 based on its location on the distribution for the NY-NJ-PA metropolitan area. Blank values were assigned to observations with missing data and were excluded from our calculations.

Intersectionality Dashboard

The Intersectionality Dashboard helps users explore the relationship between diversity and socioeconomic outcomes in greater detail. The logic for the dashboard visualization follows that of a multiple regression model in that users can select values for a set of independent (control) variables and then visualize the results for a number of dependent (outcome) variables.[24] Control variables include: age group (18 to 24,

24 Note that this visualization only follows the logic of multiple regression, but no regression models are in fact being estimated.

25 to 34, 35 to 44, 45 to 54, and 55 to 64; sex (male or female); race-ethnicity (non-Hispanic White, non-Hispanic Black, non-Hispanic Asian, non-Hispanic Other, Hispanic); the top ten ethnic groups (Dominican, Puerto Rican, Chinese, Mexican, Indian, Ecuadorian, Jamaican, Colombian, Haitian and Filipino); residence in one of the five boroughs of New York City proper (Manhattan, Brooklyn, Queens, the Bronx, Staten Island, or outside NYC proper); and immigrant arrival cohort (pre-1980, 1980-1990, 1991-2000, 2001-2010, and 2011-2018).

The response variables are all based on statistical likelihoods. That is, once a person with particular characteristics is selected in the website dropdown (e.g., a 35-44 year-old Chinese immigrant male who arrived to the US between 2001-2010 and lives outside of New York City proper), the dials provide a visual image of the probability that such an individual has a college degree; is employed; is not experiencing low income (i.e., does not live in a household whose family income is less than 150 percent of the poverty line); speaks English at home; has secured affordable housing; and has realized home ownership. In each case, the dial is set with the average for the entire working-age population as the vertical (middle) value, and the probability for each indicator is shown as a deviation from this this point. Therefore, for example, if the selected type of person is more likely to own a home than the average working-age person, then the image on the dial would be to the right of the center point.

Assembling the data required to produce this visualization required several steps. First, data for each combination of control variables had to be cross tabulated separately for each response variable, resulting in 40 unique tables. Second, each table was then formatted so that each set of rows represented a unique combination of all the control variables (in total there were 16,758 rows across all 40 tables). Each row had a column for the number of observations in that cross tabulation as well as a value for each of the control variables. Third, the raw values for each response column were transformed into z-scores, separately for the male, female, and the total for the NY-NJ-PA metropolitan area. A z-score was calculated given the following formula:

$$Z = (x - \overline{x})/sd$$

where x is a given observation, x is the population mean, and sd is the weighted standard deviation. A z- score was calculated for each observation by subtracting the overall total population mean from each observation and then dividing by the weighted standard deviation. Population means and the weighted standard deviations were calculated using estimates separately for the male, female, and total NY-NJ-PA metropolitan area population.

In practice, this means that when a user selects the variable "female" from the drop-down menu, the response variable on each dial is scaled such that the midpoint indicates the average expected value for all working-age females. Each dial will adjust as the user selects more "control" variables. So, for example, if the user selected the profile for a 45-54-year-old non-Hispanic native-born female, one of the dials would show the likelihood that person has a college degree compared with all working-age females in total.[25]

Website Design and Coding

For more on website design and coding see the original **Superdiversity and Cities - Technical Report.**

Acknowledgement

We wish to thank the following people and institutions for their help in assembling the reams of information behind this website, and its design and implementation. Steven Vertovec and the Max Planck Institute for the Study of Religious and Ethnic Diversity provided crucial funding support. Daniel Hiebert provided invaluable technical guidance on data and variables. Stamen Design provided crucial expertise on website design and visualization. CUNY Graduate Center and the Center for Urban Research also provided crucial funding and institutional support.

25 Note that intersections that have less than 100 observations will appear greyed out because that is too few observations to make any inferences.