Documentation for the Urban Institute's Home Mortgage Disclosure Act Neighborhood Summary Files: Census Tract Level

Revised 12/20/2023

The Home Mortgage Disclosure Act (HMDA) is a federal act that requires mortgage lenders to keep records of loan applications and lending practices and submit data to regulatory authorities. HMDA reporting allows government regulators to analyze information on mortgage loans and mortgage lending trends. HMDA data are used to understand credit accessibility, fair lending, and the mortgage market. For example, Urban researchers have used HMDA data to analyze the state of the mortgage market, evaluate racial trends in homeownership, and understand how well GSEs serve minority borrowers. HMDA data are also an important research tool understanding housing market dynamics and neighborhood change. The data are of special interest to the National Neighborhood Indicators Partnership (NNIP) because the data include information about the location (census tract) of the property for each individual loan application. Individual loan application data can be aggregated to the census-tract level to provide insights on the demographic and economic characteristics of people purchasing homes with mortgage loans.

These data files are very large and cumbersome. The 2021 file has over 26 million observations and is 10 gigabytes. To make these data accessible to more analysts, the Urban Institute has published a public use census-tract level file with selected indicators that focus on monitoring neighborhood change. This is a brief description of the source data and datasets provided by Urban. If we see demand for this data, we will continue to update as additional years of HMDA are released.

Source Data

HMDA's Snapshot National Loan Level Dataset contains the applications for a calendar year as of a fixed date for all HMDA reporters, as modified by the Consumer Financial Protection Bureau to protect applicant and borrower privacy. For example, the 2021 Snapshot National Loan Level Dataset reflects all the loan applications reported for the calendar year 2021 as of April 30, 2022. For more information on the source data, see the full documentation from the Consumer Financial Protection Bureau.

Definition and Creation of Variables

The indicators included in this dataset are primarily focused on loans that meet the following criteria, though a handful of indicators that are proxies for neighborhood change have been included that reflect other types of loans (e.g., those for investment purchases).

1. Loan completed the origination process (action_taken = 1 - Loan originated)
2. Loan is for a home purchase (loan_purpose = 1 – Home Purchase)
3. Loan is first lien (lien_status = 1 – Secured by a first lien)
4. Home is single family or 1-4 unit building (derived_dwelling_category = Single Family (1-4 Units):Site-Built or Manufactured)
5. Home is owner occupied (occupancy_type = 1 – Principal Residence)

Within this universe, the Urban Institute uses a three-step process to create HMDA neighborhood variables. First, we use information from individual loan applications to flag characteristics of that application based on several key measures described below. Second, we combine those characteristic flags to indicate whether a loan application meets all the criteria for that indicator. And third, loan applications that meet the criteria are counted for each census tract.

Counts are then created for the number of loans:

1. By race/ethnicity
2. By income level (relative to surrounding geographic area)
3. By age group
4. By race/ethnicity AND relative income

The median loan amount and median borrower income are also calculated using the same universe as described above.

An additional set of three variables provides counts for investment loans: those loans where the loan completed the origination process (action_taken = 1 - Loan originated) and the loan is for an investment property (occupancy_type = 3 - Investment Property). These variables include total investment loans and investment loans broken down by buildings with one to four units and buildings with five or more units.

Geography

This data set is published at the census tract level. Census tract codes follow the format: 2-digit state FIPS code (s), 3-digit county FIPS code (c), and 6-digit census tract FIPS code (t), with the form: sscccttttttt.

However, approximately one percent of records on the Snapshot National Loan Level Dataset are missing census tract information. For that reason, attempting to aggregate all of the records with valid census tracts in a county or state will not add up to the total number of records in that geography.

Rather than drop these records, we created special census tract codes for them to facilitate accurate summaries of larger geographies. For partial geographic matches we use the following format:

- If a record has a known county it is assigned: sscccXXXXXX.
- If a record only has a known state, it is assigned: ssXXXXXXXXX.
- If a record has no geography details at all, it is assigned XXXXXXXXXXX.

The US Census Bureau changes census tract boundaries and identifiers over time to reflect population changes; systematic changes occur as part of each decennial census. HMDA data for the years 2018 to 2021 are published using census tract codes that correspond to the 2010 decennial census, whereas

HMDA data published for the years 2022 and thereafter are published using census tract codes that correspond to the 2020 decennial census. Due to this shift, HMDA data at the census tract level from 2021 and before are not directly comparable to HMDA data from 2022 and after. To emphasize this difference, the census tract variable in the Urban files is named census_tract_2010 or census_tract_2020 depending on the vintage of the census tract geographies at which HMDA data is reported for a given calendar year.

We will consider in the future adjusting 2018 to 2021 years of the census tract-level HMDA indicators to 2020-based census tracts so they can be compared with HMDA indicators for the years 2022 and the rest of the decade that will be reported using 2020 census tract definitions.

Race/Ethnicity

To create the indicators about race and ethnicity, we evaluate the race and ethnicity variables for both the applicant and co-applicant. A record may have up to five race/ethnicities each for the applicant and co-applicant. In the source data, the field series we use are applicant_race_1 - applicant_race_5; applicant_ethnicity_1 - applicant_ethnicity_5; co_applicant_race_1 - co_applicant_race_5; and co_applicant_ethnicity_1 – co_applicant_ethnicity_5. Note that the source data file includes two other variable series about race on the data file that we do not use: derived race (Derived_race) and race based on observation (applicant_race_observed).

The following steps are used to create the "Household race/ethnicity" indicator:

First, we determine a race/ethnicity category for the applicant:

- IF applicant_ethnicity is Hispanic then they are coded: Hispanic
- ELSE IF their race is missing or is marked as "Not applicable" or "Not provided" then they are coded as: Race not available
- ELSE IF the applicant indicated identifying with multiple races (using fields applicant_race_2 - applicant_race_5) then they are coded as: "Multiple" races
- ELSE they are coded as the race indicated in applicant_race_1

Second, the same logic is followed to determine a race/ethnicity category for the co-applicant (if there is a co-applicant).

Third, the values determined in steps one and two are used to determine the household race/ethnicity category for each record:

- IF there is a co-applicant AND both the applicant and co-applicant race are available AND they are not the same, then the record is coded as: "Mixed" races
- ELSE IF the applicant and co-applicant's race/ethnicity are the same OR there is no co-applicant, then the record is coded as the applicant's race/ethnicity

Age

To create the age indicators, we evaluate the age variables for both the applicant (ageapplicant) and co-applicant (co-applicant_age). Both the raw ageapplicant and co-applicant_age variables have seven levels: less than 25; 25-34; 35-44; 45-54; 55-64; 65-74, and greater than 74. We collapse these categories into: less than 25; 25-44; 45-64; and greater than 65. By collapsing age categories, we are

able to classify more records as belonging to a single age category as opposed to a "mixed age" category as described in the next paragraph.

When both the applicant and co-applicant ages fall into the same category, the record is assigned that age category. When either the applicant or co-applicant age is missing, but one age variable is available, the record is assigned the non-missing age category. When the applicant and co-applicant variables fall into different age categories, the record is classified as "mixed age". In all other cases, the record's age indicators are classified as missing.

Relative Income

We provide indicators based on relative income categories based on the area median family income for the larger geographic area in which a census tract is located. Using the relative income level is helpful in comparing across areas of the country with different housing costs and other costs of living. Income limit data are published by the Department of Housing and Urban Development (HUD) and can be acquired from the HUD USER website.

For most of the country, we use the area median income levels for metropolitan areas (for counties in those areas) or counties (for non-metropolitan areas). However, in parts of the Northeast, income limit data are unique to subcounty divisions. We take a two-part approach to merging income limits with the Snapshot National Loan Level Dataset:

- Where income limits are published at the county level, we merge income limits using the county identifier because every census tract falls within a single county and no others.
- Where income limits are published at the subcounty level, we first assign each tract to the subcounty division that encompasses the greatest share of that tract's land area (because some tracts fall in multiple subcounty divisions) and then we merge income limits at the subcounty division level. We identify tract-subcounty division land area intersections based on a geographic crosswalk from the Missouri Census Data Center's Geocorr 2022: Geographic Correspondence Engine.

We compare the borrower's income to the Area Median Income to determine whether the applicant falls into the category of:

- Very low income (0-50 percent of Area Median Income)
- Low income (50.1-80 percent of Area Median Income)
- Moderate income (80.1-120 percent of Area Median Income)
- High income (120.1 percent or greater of Area Median Income)

Owner-occupied Housing Units

We provide an indicator of the percent of owner-occupied housing units based on American Community Survey (ACS) 2015-2019 5-year data, to provide context for how much of the housing market our HMDA indicator set covers. For example, if the census tract is only 10 percent owner-occupied, the numbers around characteristics of owner-occupied borrowers are not very meaningful for neighborhood change. The indicator is calculated by dividing the number of owner-occupied housing units by the number of total occupied units (these variables are also included in the neighborhood HMDA dataset). Data from the 2015-2019 ACS are included on the 2018 to 2021 data files because this is the last year of ACS data

that is available based on 2010 census tract geographies. For the 2022 data file, we use the 2017-2021 ACS data with census tract geographies based on 2020 definitions of census tracts.

Revision History

Year of Data	Access Date for Raw Data	Date Published
2018	02/08/2023	12/19/2023
2019	02/08/2023	12/19/2023
2020	02/08/2023	12/19/2023
2021	02/08/2023	12/19/2023
2022	05/10/2023	12/19/2023

Citation and License

Urban Institute. 2023. Home Mortgage Disclosure Act Neighborhood Summary Files: Census Tract Level. Accessible from https://datacatalog.urban.org/dataset/home-mortgage-disclosure-act-neighborhood-summary-files-census-tract-level. Data originally sourced from HMDA Snapshot National Loan-Level Dataset, developed at the Urban Institute, and made available under the ODC-BY 1.0 Attribution License. Consumer Financial Protection Bureau (2018) [computer file]. Washington, DC: Snapshot National Loan Level Dataset, accessed February 8, 2023 at https://ffiec.cfpb.gov/data-publication/