Boston Area Research Initiative ● Northeastern & Harvard Universities
E-Mail: bari@northeastern.edu ● Web: www.bostonarearesearchinitiative.net

A Guide to Using the Geographical Infrastructure for the City of Boston, MA (v. 2024)

Overview

The Boston Area Research Initiative's Geographical Infrastructure for Boston (GI) is a database that organizes and links information for places and areas within Boston, MA at 17 different levels of geographic organization—including land parcels, streets, census geographies, and other administrative regions. The levels are organized in a hierarchy, with smaller geographic variables nested within larger ones (e.g., land parcels nested in census tracts). This is coordinated by using variables that act as unique identifiers at each level. As a composite, the database is intended to facilitate aggregate calculations across levels of the hierarchy and the linkage and analysis of data from different sources that reference the same geographical units. In particular, the database makes it possible to connect data sets generated by the City of Boston with census geographies and data.

It should be noted that the Geographical Infrastructure comprises three separate datasets within the Boston Area Research Initiative's (BARI) Boston Data Portal. Localized geographies (from properties to streets) are contained in the "Geographical Infrastructure for the City of Boston" dataset, census geographies (i.e. blocks, block groups, and tracts) are contained in the "2020 Census Geographies" dataset and administrative geographies (i.e. zip codes, planning districts, etc.) are contained in the "Administrative Geographies for the City of Boston" dataset. These latter two datasets remain largely unchanged from year to year while the first is annually updated.

The levels are each documented in a .csv and in most cases a *.shp (shapefile for GIS), and include:

Geographical Infrastructure for the City of Boston[1]

- Properties (from the City of Boston Property Assessment Database, Fiscal Year 2024)
- Land parcels (aggregated from the City of Boston Property Assessment Database, Fiscal Year 2024)

[1] Land parcels and properties without X and Y coordinates do not include localized and census geographical information.

- Intersections (from the City of Boston and from TIGER line data, 2013)
- Street segments (from census TIGER line data, 2023)

2010 Census Geographies

- Census blocks, block groups, and tracts (from Census 2010)

2020 Census Geographies

- Census blocks, block groups, and tracts (from Census 2020)

Administrative Geographies for the City of Boston

- Ten ways that the City of Boston divides the city into administrative districts for planning, elections, and operations (maintained in a separate dataset from the GI).

This documentation contains a section for each of these groupings, describing the contents and variables. For each, the unique identifier variables used to link the files are noted.

Table of Contents

1. Properties 5
   1.1 Description of Contents 5
   1.2 Description of Variables 5
      1.2.1 Identifying Characteristics 6
      1.2.2 Property and Building Characteristics 6
      1.2.3 Geographical Information 7
2. Land Parcels 9
   2.1 Description of Contents 9
   2.2 Description of Variables 9
      2.2.1 Identifying Characteristics 9
      2.2.2 Property and Building Characteristics 10
      2.2.3 Geographical information 13
3. Intersections 15
   3.1 Description of Contents 15
   3.2 Description of Variables 15
      3.2.1 Identifying Characteristics 15
      3.2.2 Geographical Information 16
4. Street Segments 17
   4.1 Description of Contents 17
   4.2 Description of Variables 17
      1.1.1. Identifying Characteristics 17
      4.2.1 Geographical Information 19
5. Census Geographies 21
   5.1 Description of Contents 21
   5.2 Description of Variables 21
      5.2.1 Identifying Characteristics 22
      5.2.2 Geographical Information 23

6. Other Geographies 25
6.1 Description of Contents 25
6.2 Description of Variables 25
6.2.1 BPDA Neighborhood Statistical Areas 25
6.2.2 BPDA Planning Districts 25
6.2.3 City Council Districts 25
6.2.4 Election Precincts 26
6.2.5 Election Wards 26
6.2.6 Fire Districts 26
6.2.7 ISD Neighborhoods 26
6.2.8 Police Districts 26
6.2.9 PWD Districts 26
6.2.10 ZIP Codes 27
7. Appendix 28
7.1 Appendix A: Codes for Land Use (LUC) 28
7.2 Appendix B: Codes for Road Types 29
7.3 Appendix C. Characteristics of Clusters of Road Segments 31
7.4 Appendix D: Unit Imputation 32

1. Properties

1.1 Description of Contents

The City of Boston's Assessing Department is responsible for determining the value of each property in the city. To this end, it maintains parcel ownership and value information to ensure fair assessment of both taxable and non-taxable property in Boston. Assessing records are compiled and reviewed annually to reflect changes to parcels because of new construction, remodeling, and changes in ownership. This constitutes the City of Boston Property Assessment Database, which acts as the fundamental dataset for BARI's geographical infrastructure.

The data contained herein are a modified version of the original Property Assessment Data Set, and describe the parcel-specific address, ownership, and land use for the 182,235 properties in Boston in 2024. They include variables and measures from the original data, some of which are cleaned or modified by BARI, as well as derived measures based on original variables.[2]

No shapefile is included for Properties because the polygons associated with the City of Boston's Property Assessment Database are Land Parcels. Each property exists on only one land parcel from the Land Parcels database, but there may be multiple properties on a Land Parcel. The information necessary to split the Land Parcels appropriately into properties is not available, and in many cases, properties may exist on top of one another on different floors. Note that the definition of Land Parcel in this database is slightly different from that of the City of Boston's original data files in that we combine a small number of land parcels that are differentiated by the City (see Section 2 for more detail).

Unique identifier: *PID (Property ID)*

1.2 Description of Variables

Property variables are split into three categories: identifying characteristics, property and building characteristics, and geographical information. Identifying characteristics include variables regarding the basic identity and attributes of the address. Building characteristics include information on the physical attributes of the building containing the property. Geographical information provides further detail on the location of the property and the other geographies that contain it.

[2] Tax rate calculation information published by the City of Boston Assessing Department through the department's website: http://www.cityofboston.gov/assessing/taxrates.asp

The Tax Assessor's department maintains many properties and building characteristic variables, but here we only include those relevant to its geographic utility. For access to the full breadth of variables and their documentation see BARI's Property Assessment Database[3].

1.2.1 Identifying Characteristics

- *PID* is the 10-digit property identification number, unique to each property. The first two digits indicate the Ward, digits 3 through 7 are the parcel, and digits 8 through 10 are the sub-parcel.
- *CM_ID* is the 10-digit property number of the main condo building property. All condo units in each building are related to this number.
- *GIS_ID* is another 10-digit property identification number. It is the unique identifier for the plot the property is in.
    - *Note*: This differs slightly from *Land_Parcel_ID*, which combines multiple *GIS_ID* values based on certain criteria (see Section 2).
- *ST_NUM* is the street number of the property.
- *ST_NAME* is the street name of the property.
- UNIT_NUM is the specific unit number within a multi-unit building.
- *ZIPCODE* is the zip code of the property.

1.2.2 Property and Building Characteristics

- *LU* is the Land Use type for the property. Codes for land use can be found in Appendix A.
- *OWN_OCC* is a one-character code that indicates if an owner receives a residential exemption for the property. A "Y" indicates that the owner claims to live within the property (a.k.a. the property is "owner-occupied") and a "N" indicates the opposite.
- *YR_BUILT* is the year in which the property was built. The original dataset held many

[3] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XQE1RX

properties whose year of construction was listed as zero. It was fixed by updating the YR_BUILT variable, which now contains a "NA" value where it previously showed a "0".

- *YR_REMOD* is the year in which the property was last remodeled. For some properties the year of its most recent remodel was listed as zero. It was fixed by updating the YR_REMOD variable, which now contains a "NA" value where it previously showed a "0".
- *LAND_SF* is the total size of the property in square feet. This is also known as the lot size.
- *GROSS_AREA* is the gross floor area for commercial properties.
- *NUM_FLOORS* is the number of levels in the structure that is located on the property.

1.2.3 Geographical Information

Geographical information for the properties comes from multiple sources to coordinate them with census geographies. First, latitude and longitude of the property was derived as the centroid of the containing land parcel, based on the *.shp provided by the City of Boston (see Section 2). Properties were then linked to the appropriate census TIGER line road segment, defined as the nearest road segment to the XY coordinate that matched on street name. This was done instead of geocoding because City of Boston land parcels will be more accurate about the location of numbers along the segments of a street than census TIGER line data. Last, all properties were linked to the containing census block and higher geographies by spatial overlay.

- *X* is the geo-coded longitude of the property.
- *Y* is the geo-coded latitude of the property.
- *Land_Parcel_ID* is the unique identifier for the land parcel the property is in.
- *TLID* is the identifier for the census TIGER line road segment containing the property (See Section 2.2.3).
- *Blk_ID_10* is the 2010 census Block ID number.
- *BG_ID_10* is the 2010 census Group ID number.
- *CT_ID_10* is the 2010 census Tract ID number

- *Blk_ID_20* is the 2020 census Block ID number.
- *BG_ID_20* is the 2020 census Group ID number.
- *CT_ID_20* is the 2020 census Tract ID number
- *NSA_NAME* is the name of the Boston Redevelopment Authority Neighborhood Statistical Area in which the region is located.
- *BRA_PD* is the name for the Boston Redevelopment Authority planning district in which the region is located.

2. Land Parcels

2.1 Description of Contents

This dataset contains parcel-level information for each unique land parcels present in the City of Boston's Property Assessment Database (see Section 1). Each parcel is uniquely identified by the *Land_Parcel_ID*, which is derived by merging properties that share the same street address (street number, street name, and ZIP code). *Land_Parcel_ID* is assigned as the most frequently occurring *GIS_ID*, which the City of Boston provides as a parcel identifier. In cases of a tie, the lowest *GIS_ID* is selected. For addresses without a street number, the original *GIS_ID* is retained as the *Land_Parcel_ID*. This approach is necessary as the City of Boston treats each *GIS_ID* as a separate parcel. However, for this database, we have grouped parcels with the same address that would otherwise be impossible differentiate. This processing reduced the number of unique land parcels from 98,531 in the City's original database to a final set of 97,785 unique land parcels. In addition to a CSV, we include a shapefile, based on the land parcel shapefile maintained by the City of Boston.

Unique Identifier: *Land_Parcel_ID*

2.2 Description of Variables

Land parcel variables are split into three categories: identifying characteristics, property characteristics, and geographical information. Identifying characteristics include variables regarding the basic identity and attributes of the address. Building characteristics include information on the physical attributes of the buildings contained by the land parcel. Geographical information provides further detail on the location of the parcel and the other geographies that contain it.

2.2.1 Identifying Characteristics

- *Land_Parcel_ID* is the unique identifier for land parcels.
- *full_address* is most common address within the parcel, comprises the street number, street name, and zip code. For parcels with properties classified as "Condo Main" (CM), the full address is taken from the property designated as Condo Main.
- *property_N* is the total number of properties in each land parcel.

- *unit_N* is an estimation of number of units within the land parcel.
    - Note: Estimated by aggregating units that shares the same Land_Parcel_ID from the City of Boston's Street and Address Management (SAM) system. This process proved to be prone to error, however, so we pre-determined the number of units for certain Land Use values (e.g., single-family residential properties all were attributed 1 unit; see Appendix D for more detail).
- *unit_N_orig* is a more conservative but less complete estimate of the number of units within the land parcel based solely on the number of units sharing the same Land_Parcel_ID in the SAM data (left blank for parcels without unit-level information in the SAM data).

2.2.2 Property and Building Characteristics

- *AV_LAND* is the assessed value of the land.
- *AV_BLDG* is the total assessed value for the building on the properties on the parcel.
- *AV_TOTAL* is the total assessed value for the properties on the parcel. It is a summation of the assessed values of the land and building.
- *LAND_SF* is the total size of the properties in square feet on the parcel. This is also known as the lot size.
- *GROSS_AREA* is the gross floor area for commercial properties on the parcel.
- *LIVING_AREA* is the total living area for residential properties. on the parcel.
- *LU* is the designated land use of the address. The parcel level LU is determined by the most common LU of properties in this parcel (see Section 1.2.2 and Appendix D).
- *OWN_OCC* is the proportion of properties in the land parcel that are occupied by their owner.
- *NUM_FLOORS* is the number of levels in the structure that is located on the property.
- *YR_BUILT* is the year in which the property was built. The original dataset held many properties whose year of construction was listed as zero. It was fixed by updating the

YR_BUILT variable, which now contains a "NA" value where it previously showed a "0".

- *YR_REMOD* is the year in which the property was last remodeled. For some properties the year of its most recent remodel was listed as zero. It was fixed by updating the YR_REMOD variable, which now contains a "NA" value where it previously showed a "0".
- *owner_address* is the street address of the owner of the land parcel. It is calculated by taking the mode of the owner's street address for all properties in the land parcel.
- *comcenter* indicates whether a land parcel contains a Boston Centers for Youth and Family (BYCF) community center (1) or not (0).
    - *Note*: Sourced from the City of Boston's Department of Innovation and Technology's "Community Centers" dataset on Analyze Boston (from the May 2018 update).
- *medhos* indicates whether the land parcel contains a hospital, medical facility, medical office, nursing facility, rehabilitation facility, and/or a convalescent facility (1) or not (0).
    - *Note:* Sourced from the City of Boston's 2024 Property Assessment database's Property Occupancy Code (LUC) designations of 304, 305, 309, 342, 953, 954, 958, 959, and 979.
- *supermkt* indicates whether the land parcel contains a supermarket (1) or not (0).
    - *Note:* Sourced from the City of Boston's 2024 Property Assessment database's Property Occupancy Code (LUC) designation of 324.
- *parking* indicates whether the land parcel contains a public or private parking lot or garage (1) or not (0).
    - *Note:* Sourced from the City of Boston's 2024 Property Assessment database's Property Occupancy Code (LUC) designations of 116, 336, 337, 338, 339, 387, 961, and 962.
- *vacant* indicates whether the land parcel contains a publicly-owned or privately-owned vacant property (1) or not (0).
    - *Note:* Sourced from the City of Boston's 2024 Property Assessment database's Property Occupancy Code (LUC) designations of 012, 130, 131, 132, 202, 211, 390, 391, 392, 393, 440, 441, 442, 986, and 987.

- *rel* indicates whether the land parcel contains a property owned by a religious institution/organization and/or housing for a religious institution/organization (1) or not (0).
    - *Note:* Sourced from the City of Boston's 2024 Property Assessment database's Property Occupancy Code (LUC) designations of 129, 906, 970, 971, and 379.
- *lib* indicates whether the land parcel contains a Boston Public Library (1) or not (0).
    - *Note*: Sourced from the City of Boston's 2024 Property Assessment database's Property Occupancy Code (LUC) designation of 978.
- *bps* indicates whether the land parcel contains a Boston Public School or other variant of a public school (1) or not (0).
    - *Note:* Sourced from the City of Boston's 2024 Property Assessment database's Property Occupancy Code (LUC) designation of 976.
- *police* indicates whether the land parcel contains a police station (1) or not (0).
    - *Note:* Sourced from the City of Boston's 2024 Property Assessment database's Property Occupancy Code (LUC) designation of 975.
- *fire* indicates whether the land parcel contains a fire station (1) or not (0).
    - *Note:* Sourced from the City of Boston's 2024 Property Assessment database's Property Occupancy Code (LUC) designation of 974.
- *private* indicates whether the land parcel contains a private or taxable school, college, or university location (1) or not (0).
    - *Note:* Sourced from the City of Boston's 2024 Property Assessment database's Property Occupancy Code (LUC) designations of 123, 124, 378, 904, 937, 942, 951, and 977.
- *mbta_stop* indicates whether the land parcel contains an MBTA stop (1) or not (0).
    - *Note:* Sourced from shapefiles on MBTA Rapid Transit and Bus Routes provided by MassGIS (from the December 2022 update).
- *college* indicates whether the land parcel contains a college or university location (1) or not (0).

- *Note:* Sourced from the City of Boston's Department of Innovation and Technology's "Colleges and Universities" dataset on Analyze Boston (from the December 2023 update).

- *hlth_cntr* indicates whether the land parcel contains a community health center location (1) or not (0).
  - *Note:* Sourced from shapefiles on Community Health center locations provided by MassGIS (from the December 2024 update).
- *sub_house* indicates whether the land parcel contains a property owned by the Boston Planning and Development Agency (1) or not (0).
  - *Note:* Sourced from the City of Boston's 2024 Property Assessment database's Property Occupancy Code (LUC) designation of 908.

2.2.3 Geographical information

- *X* is the x-coordinate of the land parcel, in longitude, drawn from the centroid of the parcel's polygon in the shapefile.
- *Y* is the y-coordinate of the land parcel, in longitude, drawn from the centroid of the parcel's polygon in the shapefile.
- *TLID* is the identifier for the segment of road containing the properties on the parcel.
  - This is found by matching the 2023 TIGER lines street segments to only those that match the street name of the property and finding the segment that is geographically closest to the property. If there is no matching name, we find all street segments within 200m of the parcel centroid and select the closest street. Parcels centroids that have no street segments within 200m have TLID listed as NA.
- *Blk_ID_10* is the 2010 Census Block ID number.
  - This is found by spatially overlaying the longitude and latitude of the property onto the Census Blocks shapefile.
- *BG_ID_10* is the 2010 Census Group ID number.
- *CT_ID_10* is the 2010 Census Tract ID number.
- *NSA_NAME* is the name of the Boston Redevelopment Authority Neighborhood Statistical Area in which the region is located (see Section 7).

- *BRA_PD* is the name for the Boston Redevelopment Authority planning district in which the region is located (see Section 7).

3. Intersections

3.1 Description of Contents

Intersections comprise two datasets. One is a catalog of all intersections (*Intersections*; 11,116) generated by the census TIGER line roads data from 2013. The other is a master list of intersections maintained by the City of Boston (*Intersections (City)*; 8,421) that has then been merged with intersections within the census TIGER line road map (to the same intersection name when possible, otherwise to the closest point). These datasets are updated every 10 years.

Unique identifiers: *ObjectID* (census intersections), *propid* (City intersections).

3.2 Description of Variables

Variables are largely equivalent between the two files. The two datasets can be connected via their unique identifiers. They are split into two categories: identifying characteristics and geographical information. Identifying characteristics includes variables regarding the basic identity and attributes of the intersection. Geographical information provides further detail on the location of the intersection and the other geographies that contain it.

3.2.1 Identifying Characteristics

- *ObjectID* is the unique identifier for census-generated intersections.
- *Propid* is the unique identifier for City-defined intersections.
- *Address*, indicating the two roads that form the intersection.
- *TLID* is the unique identifier for the road segment to which the intersection is attributed.
    - *Note:* Intersections are attributed to the road that forms the intersection that is most predominant in terms of main/non-main street and zoning.
- *X* gives the x-coordinate of the intersection
- *Y* gives the y-coordinate of the intersection
- *Main* indicates whether the primary road segment is a main street or not.

- *RoadType* gives the zoning status of the road. See *Zoning* in *Roads* (Section 4).
    - *Note:* Possible values for this are Commercial, Residential, Exempted, Independent, or None.

3.2.2 Geographical Information

- *BG_ID_10* is the unique identifier for the census block group in which the intersection is located.
- *CT_ID_10* is the unique identifier for the census tract in which the intersection is located.
- *BG_ID_20* is the unique identifier for the census block group in which the intersection is located.
- *CT_ID_20* is the unique identifier for the census tract in which the intersection is located.

4. Street Segments

4.1 Description of Contents

This dataset contains a complete list of all street segments in Boston, MA, as defined by census TIGER Line data as of the 2023 update. It contains 24,891 segments. This dataset is updated every ten years.

Unique Identifier: *TLID*

4.2 Description of Variables

Street variables are split into two categories: identifying characteristics and geographical information. Identifying characteristics includes variables regarding the basic identity and attributes of the address. Geographical information provides further detail on the location of the road and the other geographies that contain it. Because roads often form the borders between regions, nesting in higher levels was done using the road centroids.

1.1.1. Identifying Characteristics

- *TLID* is the unique identifier for the road segment.
- *STATEFP* is the unique identifier (FIPS code) for the state containing the road segment (25 = Massachusetts).
- *COUNTYFP* is the unique identifier (FIPS code) for the county containing the road segment (025 = Suffolk).
- *TFIDL* and *TFIDR* are the unique identifier for the left and right sides of the road segment, respectively.
- *MFTCC* is a code provided by the census for the type of road. These types are listed in Appendix B.
- *FULLNAME* is the name of the road.

- *SMID*. Is a geographic ID code provided by the Census that is equal to 2758 for all streets in Boston.
- *LFROMADD, LTOADD, RFROMADD,* and *RTOADD* give the address ranges for the left and right sides of the road.
- *ZIPL* and *ZIPR* are the ZIP codes containing the left and right sides of the street segment, respectively.
- *Length* is the length of the segment in meters.
- *CLASS* describes the nature of the road, as provided by Massachusetts Dept. of Transportation.
    - *Note:* Values taken: 1 – Limited access highway; 2 – Multi-lane highway, not limited access; 3 – Other numbered route; 4 – Major road – arterials and collectors; 5 – Minor street or road with road inventory information; 6 – Minor street or road with no road inventory information.
- *RDTYPE* is an extended version of the *CLASS* variable, provided by Massachusetts Dept. of Transportation.
    - *Note:* 1-6 remain the same. 7 – Ramp; 8 – Tunnel; 9 – Tunnel for limited access highway; 10 – Tunnel for a multi-lane highway, not limited access; 11 – Tunnel for other numbered route.
- *Main* indicates if a road segment is considered part of a main street ("1" = Main)
    - *Note:* Based on MassGIS' *Class*, with all classes less than 5 denoted as main streets.
- *Cluster* indicates the cluster number to which the segment was attributed, from a cluster analysis based on the zoning characteristics of the parcels on the street. More details on the clusters are available in Appendix C.
    - *Note*: Clusters were redetermined before the 2018 release based on the greater specificity of parcels and land uses. For 2024 we kept all categories from previous years, except when streets had changes in parcels. In such cases we placed the street segment in the cluster whose centroid was the most similar in terms of proportions of land use. Values taken: 1 = *Three-Family Residential with Assorted*

Other Residential (13.15% of streets with parcels); 2 = *Mix of Two-Family and Single-Family Residential* (17.77%); 3 = *Commercial* (17.02%); 4 = *Single-Family Residential Only* (23.77%); 5= *Exempt* (11.65%); 6 = *Condominiums and Other High-Density Housing* (12.22%); 7 = *Mixed-Use Commercial* (4.42%). See Appendix C for more details.

- *DeadEnd* indicates whether a street segment is a dead end, defined as any street segment that is included in only one intersection.
- *parcel_N* is the number of parcels on the street segment.
- *property_N* is the number of properties on the street segment, as a sum of all properties contained in the parcels on the street.
- *unit_N* is the number of units on the street segment, as a sum of all units contained in all properties contained in the parcels on the street.
    - *Note*: Based on the imputed *unit_N* variable at these lower levels of aggregation (see Section 1).
- *unit_N_orig* is the number of units on the street segment, as a sum of all units contained in all properties contained in the parcels on the street.
    - *Note*: Based on the non-imputed *unit_N_orig* variable at these lower levels of aggregation (see Section 1), meaning there could be undercounting on streets with multiple properties with no unit-level information in the SAM.

4.2.1 Geographical Information

The following variables indicate which census regions (block groups and tracts) the street segment falls in. These were calculated in two ways. First, street segments with at least one parcel attributed to census geographies took the modal value for census block group of the parcels sitting on it (*n* = 13,850 street segments). Tract was then determined as the one containing that census block group. Streets with no parcels on them were attributed to the census geography that contained its centroid (possible for all but 24 of the remaining 11,041street segments for 2010 census geographies; all were linked for 2020 census geographies). This process ensured that a street was not attributed to a region that contained few or none of its parcels, as those are the

locations that would presumably drive most of the activity on the street that would be of interest for analysis. In a substantial minority of cases, the street segment had more parcels in a census block group other than the one containing its centroid (2,483 street segments – 17% of segments with parcels – and 10% of all street segments).

This methodology makes it theoretically possible, though not likely, for a street to change block group and census tract between years due to parcel changes. For example, suppose a street is located at the border of two block groups (BG001 and BG002) and it has three parcels associated with it in one year, two of which are in BG001. If two new parcels are constructed in BG002 in the following year, then it now has three parcels there versus two in BG001, shifting its attribution from the one block group to the other.

Note: This methodology was new in the 2018 release. Previous releases used the centroid to link all streets to census geographies.

- *BG_ID_10* is the unique identifier for the census block group (2010-2019) in which the segment is located (see Section 5).
- *CT_ID_10* is the unique identifier for the census tract (2010-2019) in which the segment is located (see Section 5).
- *BG_ID_20* is the unique identifier for the census block group (2020-present) in which the segment is located (see Section 5).
- *CT_ID_20* is the unique identifier for the census tract (2020-present) in which the segment is located (see Section 5).

5. Census Geographies

5.1 Description of Contents

As previously noted, all *.csv and *.shp (shapefiles) pertaining to Census Geographies are maintained in a separate "2020 Census Geographies" dataset within BARI's Boston Data Portal. As of the 2020 census, Boston, MA contains 5,975 city blocks, defined as any piece of land bounded by streets or water on all sides, and not divided by any streets or water. These are nested in 581 block groups, which are themselves nested in 207 census tracts. Census geographies are a natural hierarchy, with the unique identifier at each level being an extension of the level above it, indicating the specific region and all of the higher-order regions that contain it. This dataset is updated every ten years. For similar details on 2010 census geographies, please see previous releases of the Geographical Infrastructure.

5.2 Description of Variables

Census variables are split into two categories: identifying characteristics and geographical information. Variables are consistent across all three census levels, though some variables appear in one file and not in others. Identifying characteristics include variables regarding the basic identity and attributes of the block. Geographical information provides further detail on the location of the block and the other geographies that contain it. Nesting started with the block level, because all geographies, census or otherwise, conform to block boundaries. Linking to non-census geographies was done by identifying the location of the centroid of each block.

Because block groups and tracts do not nest perfectly within non-census geographies, this linking was done by identifying the region that contained the plurality of the block group or tract's population, based on census block locations. Some non-census geographies crossed over census the boundaries of a particular level too often, however, to make such linkages reliable, and thus were omitted: for block groups, *Precincts* were not linked; for tracts, *Precincts*, *Neighborhood Statistical Areas*, and *ZIP Codes* were not linked.

Demographic and socioeconomic data for census geographies can be found elsewhere in BARI's Massachusetts's Census Indicators data set on the Boston Data Portal, updated with each U.S. Census release.[4]

Unique Identifiers: Block: *Blk_ID_20*; Block groups: *BG_ID_20*; Tracts: *CT_ID_20*

[4] https://dataverse.harvard.edu/dataverse/Massachusetts_Census_Indicators

5.2.1 Identifying Characteristics

- *STATEFP* is the unique identifier (FIPS code) for the state containing the region (25 = Massachusetts).
- *COUNTYFP* is the unique identifier (FIPS code) for the county containing the region (025 = Suffolk).
- *TRACTCE20* is the code identifying the census tract.
- *BLOCKCE20* is the code identifying the block
- *GEOID20* is the unique identifier (FIPS code) of the region.
- *NAME20* the block name, within the tract.
- *MTFCC20* a class code indicating the type of feature.
- *ALAND20* is the land area of the region (in sq. meters).
- *AWATER20* is the water area of the region (in sq. meters).
- *INTPTLAT20* is the latitudinal coordinate of the region.
- *INTPTLON20* is the longitudinal coordinate of the region.
- *POP20* is the population of the region as of 2020.
- *HOUSING20* is the number of housing units in the region as of 2020.
- *Type* describes the type of neighborhood the region is within.
  - *Note:* Possible values are Residential, Downtown, Institutional (e.g., industrial, college campuses), and Park. Only for block groups and tracts.
- *Res* indicates whether the block group is generally a residential area (based on *Type*; "1" = yes).

5.2.2 Geographical Information

- *Blk_ID_20* is the unique identifier for the block (identical to *GEOID20* in the block file, but compatible with the other levels).
- *BG_ID_20* is the unique identifier of the census block group (identical to *GEOID20* in the block file, but compatible with the other levels).
- *CT_ID_20* is the unique identifier of the census tract (identical to *GEOID20* in the block file, but compatible with the other levels).
- *NSA_NAME* is the name of the Boston Redevelopment Authority Neighborhood Statistical Area in which the region is located (see Section 7).
- *BRA_PD_ID* is the unique identifier for the Boston Redevelopment Authority planning district in which the region is located (see Section 7).
- *BRA_PD* is the name for the Boston Redevelopment Authority planning district in which the region is located (see Section 7).
- *ZIPCODE* refers to the ZIP code in which the region is located (see Section 7).
- *City_Counc* is the unique identifier for the city council district in which the region is located (see Section 7).
- *WARD* is the unique identifier for the election ward in which the region is located (see Section 7).
- *PRECINCTS* is the unique identifier for the election precinct in which the region is located (see Section 7).
- *ISD_Nbhd* is the name for Boston Inspectional Services Department neighborhood in which the region is located (see Section 7).
- *Police_Dis* is the unique identifier for the police district in which the region is located (see Section 7).

- *Fire_Distr* is the unique identifier for the fire district in which the region is located (see Section 7).
- *PWD* contains a numerical unique identifier for Public Works districts, followed by the name of the Public Works District in which the region is located (see Section 7).

6. Other Geographies

6.1 Description of Contents

Other geographies were provided by the City of Boston and are separately maintained from the GI in the "Administrative Geographies for the City of Boston 2024" dataset.[5] They include: traditional neighborhoods defined by the Boston Planning and Development Agency (*BPDA Neighborhood Statistical Areas, BPDA Planning Districts*); election board regions (*City Council Districts, Election Precincts, Election Wards*); and districts for City operations (*Fire Districts, Inspectional Services Department (ISD) Neighborhoods, Police Districts, Public Works Department (PWD) Districts,* and *ZIP Codes*). For each, only a shape file with the unique identifiers is included. The unique identifiers are listed below. All of them were acquired in 2023.

6.2 Description of Variables

6.2.1 BPDA Neighborhood Statistical Areas

- *OBJECTID* is the unique identifier of the polygon object.
- *Name* is the corresponding neighborhood name.
- *Acres* is the total square acreage of the neighborhood.
- *Neighborhood_ID* is the numerical neighborhood designation that corresponds to *Name*.
- *SqMiles* is the total square mileage of the neighborhood.

6.2.2 BPDA Planning Districts

- *ID* is the numerical unique identifier. Denoted as *BRA_PD_ID* in lower-level files.
- *PD* is the name of the planning district. Denoted as *BRA_PD* in lower-level files.

6.2.3 City Council Districts

- *DISTRICT* is the unique identifier. Denoted as *CITY_COUNC* or *CITY_COUNCIL* in lower-

[5] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GDET2L

level files.

- *Councillor* is the individual holding the seat as of the most recent election predating January 1, 2018.

6.2.4 Election Precincts

- *PRECINCT* is the unique identifier.
- *WARD_PRECINCT* is a concatenation of the ward number and the precinct number within the ward.

6.2.5 Election Wards

- *WARD* is the unique identifier.
- *CNT_WARD* is the number of precincts in each ward.

6.2.6 Fire Districts

- *DISTRICT* is the unique identifier. Denoted as *Fire_Distr* in lower-level files.

6.2.7 ISD Neighborhoods

- *Name* is the unique identifier.

6.2.8 Police Districts

- *ID* is the unique identifier and corresponds to the police district identification number.
- *DISTRICT_* is the unique identifier for the police district of the block. Denoted as *Police_Dis* in lower-level files.

6.2.9 PWD Districts

- *PWD* is the unique numerical identifier for each public works district.
- *NAME* is the unique name of the neighborhood/district

- *COMBO* contains a numerical unique identifier for Public Works districts, followed by the name of the Public Works District. Denoted in lower-level files as *PWD*.
- *DIST* is a shortened variant of the *PWD* identifier

6.2.10 ZIP Codes

- *ZIP5* is the zip code. Denoted as *ZIPCODE* or *ZIP* in lower-level files.

7. Appendix

7.1 Appendix A: Codes for Land Use (LUC)

USE CODE	DESCRIPTION
A	Residential 7 or more units
AH	Agricultural/Horticultural
C	Commercial
CC	Commercial condominium
CD	Residential condominium unit
CL	Commercial land
CM	Condominium main (physical structure housing all related condo units with no assessed value)
CP	Condo parking
E	Tax-exempt
EA	Tax-exempt (121A)
I	Industrial
R1	Residential 1-family
R2	Residential 2-family
R3	Residential 3-family
R4	Residential 4 or more family
RC	Mixed use (res. and comm.)
RL	Residential land

7.2 Appendix B: Codes for Road Types

MTFCC	Feature Class Full Name	Count
H1100	Connector (Hydrography)	198
H3010	Stream/River	12,855
H3020	Canal, Ditch or Aqueduct	520
L4010	Pipeline	7
L4020	Powerline	246
L4110	Fence Line	48
L4130	Point-to-Point Line (Miscellaneous Linear)	4
L4140	Property/Parcel Line (Including PLSS)	296
P0001	Nonvisible Linear Legal/Statistical Boundary	11,393
P0002	Perennial Shoreline	8,441
P0004	Other non-visible bounding Edge (e.g., census water boundary, boundary of an areal feature)	2,584
R1011	Railroad Feature (Main, Spur, or Yard)	2,269
R1051	Carline, Streetcar Track, Monorail, Other Mass Transit Rail)	46
S1100	Primary Road	1,409
S1200	Secondary Road	8,543
S1400	Local Neighborhood Road, Rural Road, City Street	89,971
S1500	Vehicular Trail (4WD)	47
S1630	Ramp	2,070
S1640	Service Drive usually along a limited access highway	59

S1710	Walkway/Pedestrian Trail	162
S1740	Private Road for service vehicles (logging, oil, fields, ranches, etc.)	675
S1750	Private Driveway	5
S1780	Parking Lot Road	16
S1820	Bike Path or Trail	1

7.3 Appendix C. Characteristics of Clusters of Road Segments

Generated by a Cluster Analysis on Parcel Zoning Characteristics.

Streets with one or more parcels were initially categorized in the 2018 Geographical Infrastructure release by the land use of those parcels, based on a cluster analysis conducted with the *kmeans* command in R. The analysis generated seven clusters based on the proportion of each of 17 land use types (see Appendix A). Also tested were a 9-cluster solution and an analysis based on size of properties. The selected solution was the most interpretable and also most informative regarding differences between street segments in crime and disorder. An eighth "cluster" was then created as those streets with no parcels. In all subsequent releases of Geographical Infrastructure data after 2018, the 7-cluster categories are retained to that of the previous release, except when streets had changes in parcels. In such cases we placed the street segment in the cluster whose centroid was the most similar in terms of proportions of land use.

The eight clusters, in the order they are referenced by the *Cluster* variable in the streets data set (values 0-7), and their main property types on the average street were as follows:

0 = No Parcels (10,825 streets, or 43.49%)

1 = *Three-Family Residential with Assorted Other Residential* (1,849 or 13.15% streets with parcels)

2 = *Mix of Two-Family and Single-Family Residential* (2,500 or 17.77% streets with parcels)

3 = *Commercial* (2,394 or 17.02% streets with parcels)

4 = *Single-Family Residential Only* (3,343 or 23.77% streets with parcels)

5= *Exempt* (1,639 or 11.65% streets with parcels)

6 = *Condominiums* (1,719 or 12.22% streets with parcels)

7 = *Mixed-Use Commercial* (622 or 4.42% streets with parcels)

7.4 Appendix D: Unit Imputation

The City of Boston's Street and Address Management (SAM) system is a master repository for all addresses in Boston, dividing properties into discrete units where possible. This allows us to assign a specific number of units (*unit_N*) per street address by tabulating the number of rows (each representing a unit) sharing the same *Land_Parcel_ID* in the SAM dataset.

For certain land uses, we assumed the number of units based on the definition of the land use itself. For parcels with Single-Family Residential (R1), Two-Family Residential (R2), Three-Family Residential (R3), and Four-Family Residential (R4), *unit_N* is set equal to the legal number of residential units associated with the designation (e.g., Two-Family Residential = 2 units). Commercial Lots (CL), Residential Lots (RL-RL), and Condo Parking (CP), have *unit_N* set equal to zero units as they have no buildings on them. Condo Main is set equal to one unit as it is the lobby of a condo building. Residential Condominium (CD), Commercial Condominium and Agricultural/Horticultural (AH) are set to number of properties within the parcel.

The land uses of parcels are determined by the most common land use on the property level except for a few cases. If the most common land use is either Commercial Lot (CL), Residential Lot (RL-RL), or Condo Parking (CP), we assigned the second most common land use and unit_N is assigned based on this land use. In this way, we avoid *unit_N* values being inappropriately set to zero. If multiple land uses (with the exception of CL, RL-RL and CP) are tied as the most frequent, ties are broken with the following scheme: RC > CC > C > CD > A > R4 > R3 > R2 > R1 > EA > E > I > AH > CM, where RC takes precedence over CC and so on.