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. 2022)

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 via 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 "2010 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.

In addition, the Networked Geographical Infrastructure (NetGI) describes the physical linkages between the objects contained in a single level of the GI. These support analysis of connectivity across the streets and neighborhoods of Boston.

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, 2022)
- Land parcels (aggregated from the City of Boston Property Assessment Database, 2022)
- Intersections (from the City of Boston and from TIGER line data, 2013)
- Street segments (from census TIGER line data, 2013)

2010 Census Geographies

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

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).

The NetGI consists of matrices and edge lists, stored in .csv files:

- Travel adjacency matrices (including seven different travel measures between census block groups and between census tracts (based on population weighed centroids)
- Street intersection edge list (list of streets in Boston that intersect with each other. Using the shape files from BARI's 2013 roads)

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.

[1] Land parcels and properties without X and Y coordinated do not include localized and census geographical information. The parcels without geographical information in 2020 and 2021 continue without it in 2022.

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 7
      1.2.3. Geographical Information 8
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
      4.2.1. Identifying Characteristics 17
      4.2.2. 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
      5.2.3. Population weighted centroids 24

5.2.3.1. Description of contents 24
5.2.3.2. Description of variables 24
5.2.3.3. Identification Variables 25
5.2.3.4. Centroid Variables 25
6. Other Geographies 26
6.1. Description of Contents 26
6.2. Description of Variables 26
6.2.1. BPDA Neighborhood Statistical Areas 26
6.2.2. BPDA Planning Districts 26
6.2.3. City Council Districts 26
6.2.4. Election Precincts 27
6.2.5. Election Wards 27
6.2.6. Fire Districts 27
6.2.7. ISD Neighborhoods 27
6.2.8. Police Districts 27
6.2.9. PWD Districts 27
6.2.10. ZIP Codes 28
7. Travel adjacency matrices 29
7.1. Description of contents 29
7.2. Description of files 29
7.3. Additional considerations 30
8. Street intersection edgelist 31
8.1. Description of contents 31

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 177,091 properties in Boston in 2022. 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

[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

containing the property. Geographical information provides further detail on the location of the property and the other geographies that contain it.

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 thru 7 are the parcel, and digits 8 thru 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.
- *ST_NAME_SUF* is the suffix of the street name. This variable contains two-character short-forms of each type of suffix (St, Av, BL, PL, etc…)
- UNIT_NUM is the specific unit number within a multi-unit building.
- *ZIPCODE* is the zip code of the property. .
- *Unit_N* is an estimation of the number of units within the property.
    - *Note*: Tabulated from the City of Boston's Street and Address Management (SAM) system. For cases without unit data in SAM values are imputed (see

[3] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/UWTQ4E

Appendix D for methodology).

- *Unit_N_orig* is a more conservative but less complete estimate of the number of units within the property, based solely on the number of units appearing in the SAM data (left blank for properties without unit-level information in the SAM data).

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 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.
- *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
- *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 the unique land parcels present in the City of Boston's Property Assessment Database (see Section 1). It was constructed by merging all properties that match on *GIS_ID* or on street address. Note that the latter condition may combine certain GIS_IDs into a single Land Parcel. This is important because the City of Boston considers every GIS_ID as its own "land parcel," but for the purposes of this database we have combined a small number that have the same address because they would otherwise be impossible to differentiate. This processing reduced the number of land parcels from 98,873 in the City's original database to a final set of 98,150 unique land parcels. In addition to a CSV, we include a shapefile, which is the shapefile maintained by the City of Boston for its Property Assessment Database.

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 split into four parts: lower number, upper number, street, and zip code. It contains an upper and lower number to allow for a range of addresses along a street.
- *property_N* is the total number of properties in each land parcel.
- *unit_N* is the number of units within the property, summed from the properties therein (see Section 1.2.1).

- *unit_N_orig* is the raw number of units within the property, summed from the properties therein that have units information in the SAM database (see Section 1.2.1).

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.
- *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".
- *R_BLDG_STYL* is the building style for residential properties. The styles are: BL for

Bi-Level, BW for Bungalow, CL for Colonial, CN for Contemporary, CP for Cape, CV for Conventional, DK for Decker, DX for Duplex, L for Tri-Level, Oth for Other, RE for Row End, RM for Row Middle, RN for Ranch, RR for Raised Ranch, SL for Split Level, TF for Two-Family Stack, TD for Tudor, SD for Semidetached, and VT for Victorian It is calculated by taking the mode of the building styles of the properties in the land parcel.

- *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 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 (updated May 9, 2018).
- *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 2018 Property Assessment database's Property Occupancy Code (PTYPE) designations of 304, 305, 309, 342, 953, 954, 958, 959, and 979. This information was cross checked with the City of Boston's Department of Innovation and Technology's "Hospital Locations" dataset on Analyze Boston (updated October 19, 2017).
- *supermkt* indicates whether the land parcel contains a supermarket (1) or not (0).
    - *Note:* Sourced from the City of Boston's 2018 Property Assessment database's Property Occupancy Code (PTYPE) 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 2018 Property Assessment database's Property Occupancy Code (PTYPE) 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 2018 Property Assessment database's Property Occupancy Code (PTYPE) 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 2018 Property Assessment database's Property Occupancy Code (PTYPE) designations of 129, 906, 970, 971, and 379.
- *lib* indicates whether the land parcel contains a library (1) or not (0).
    - *Note*: Sourced from the City of Boston's 2018 Property Assessment database's Property Occupancy Code (PTYPE) designation of 978. This information was cross checked with the City of Boston's Department of Innovation and Technology's "Public Libraries" dataset on Analyze Boston (last updated July 18, 2018).
- *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 2018 Property Assessment database's Property Occupancy Code (PTYPE) designation of 976. This information was cross checked with the City of Boston's Department of Innovation and Technology's "Public Schools" dataset on Analyze Boston (last updated July 10, 2018).
- *police* indicates whether the land parcel contains a police station (1) or not (0).
    - *Note:* Sourced from the City of Boston's 2018 Property Assessment database's Property Occupancy Code (PTYPE) designation of 975. This information was cross checked with the City of Boston's Department of Innovation and Technology's "Police Districts" dataset on Analyze Boston (last updated July 10, 2018).
- *fire* indicates whether the land parcel contains a fire station (1) or not (0).

    - *Note:* Sourced from the City of Boston's 2018 Property Assessment database's Property Occupancy Code (PTYPE) designation of 974. This information was cross checked with the City of Boston's Department of Innovation and Technology's "Fire Departments" dataset on Analyze Boston (last updated July 10, 2018).
- *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 2018 Property Assessment database's Property Occupancy Code (PTYPE) 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.
- *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 (last updated July 10, 2018).
- *hlth_cntr* indicates whether the land parcel contains a community health center location (1) or not (0).
    - *Note:* Sourced from the Massachusetts League of Community Health Centers' "Find a Health Center" web service.
- *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 2018 Property Assessment database's Property Occupancy Code (PTYPE) 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 2013 TIGER lines street segments to only those that match the street name of the property, and then finding the one that is geographically closest to the property. At the end, if there are no name matching, we apply spatial join and nearest neighbor functions from each parcel to street segments, for a 100m buffer around parcel centroid.
- *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.

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 2013 update. It contains 24,891 segments. This dataset is updated every ten years.

Unique Identifier: *TLID*

4.2. Description of Variables

Roads 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.

4.2.1. 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.

- *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-lne 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 2022 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* (13.3% of streets with parcels); 2 = *Mix of Two-Family and Single-Family Residential* (18.2 %); 3 = *Commercial* (17.3%); 4 = *Single-Family Residential Only* (24.1%); 5= *Exempt*

(11.0%); 6 = *Condominiums and Other High-Density Housing* (11.8%); 7 = *Mixed-Use Commercial* (4.3%). See Appendix C for more detail on the average street in each.

- *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.

4.2.2. 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 = 13,813 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 11,075 street segments). 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-present) in which the segment is located (see Section 5).
- *CT_ID_10* is the unique identifier for the census tract (2010-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 "2010 Census Geographies" dataset within BARI's Boston Data Portal. As of the 2010 census, Boston, MA contains 7,288 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 558 block groups, which are themselves nested in 178 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.

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 includes 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 land area. 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]

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

Unique Identifiers: Block: *Blk_ID_10*; Block groups: *BG_ID_10*; Tracts: *CT_ID_10*

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).
- *TRACTCE10* is the code identifying the census tract.
- *BLOCKCE10* is the code identifying the block
- *GEOID10* is the unique identifier (FIPS code) of the region.
- *NAME10* the block name, within the tract.
- *MTFCC10* a class code indicating the type of feature.
- *ALAND10* is the land area of the region (in sq. meters).
- *AWATER10* is the water area of the region (in sq. meters).
- *INTPTLAT10* is the latitudinal coordinate of the region.
- *INTPTLON10* is the longitudinal coordinate of the region.
- *POP100_RE* is the population of the region as of 2010.
- *HU100_RE* is the number of housing units in the region as of 2010.
- *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_10* is the unique identifier for the block (identical to *GEOID10* in the block file, but compatible with the other levels).
- *BG_ID_10* is the unique identifier of the census block group (identical to *GEOID10* in the block file, but compatible with the other levels).
- *CT_ID_10* is the unique identifier of the census tract (identical to *GEOID10* in the block file, but compatible with the other levels).
- *BOSNA_R_ID* is the numerical unique identifier for the Boston Redevelopment Authority Neighborhood Statistical Area in which the region is located (see Section 7).
- *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_NAME* 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).

5.2.3. Population weighted centroids

5.2.3.1. Description of contents

To find more accurate estimates of where individuals live within Boston's block groups and census tracts, population weighted centroids were calculated using the population of the census blocks therein. Populations were obtained using the *demographics* command of the *UScensus2010blk* package[5]. Shape files for all relevant blocks were obtained using the *blocks* command of the *tigris* R package[6]. First, centroids were found for each block. Then, population weighted centroids were calculated for each census block group and tract based on the location of each block's centroid and its proportion of the total population. If a block group had zero population, then an unweighted average of the blocks' centroids was calculated. FIPS code is maintained as the unique identifier for each census block group and tract. Block groups are embedded within tracts, and the data reflect this nesting. The related file is "pop.centroids.csv".

5.2.3.2. Description of variables

The variables are split into the areal units' unique FIPS codes and the coordinates of their population weighted centroids. All 558 block groups appear only once, but the 178 census tracts appear multiple times based on the number of block groups.

---

[5] Almquist, Z. W. (2010). US Census spatial and demographic data in R: the UScensus2000 suite of packages. *Journal of Statistical Software*, *37*(6), 1-31.
[6] Walker, K. tigris: Load Census TIGER/Line Shapefiles into R, 2017. *R package version 0.5*, *1*.

5.2.3.3. Identification Variables

- *BG_ID_10* is the 2010 census block group ID number.
- *BG_ID.char* is the character version of block groups' ID number with "bg" added to the front.
- *CT_ID_10* is the 2010 census tract ID number.
- *CT_ID.char* is the character version of census tracts' ID number if "ct" added to the front.

5.2.3.4. Centroid Variables

- *cen.bg_lat* is the latitude of block groups' centroids.
- *cen.bg_long* is the longitude of block groups' centroids.
- *cen.ct_lat* is the latitude of tracts'' centroids.
- *cen.ct_long* is the longitude of tracts' centroids.

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" dataset. 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.

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-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. Travel adjacency matrices

7.1. Description of contents

Seven different travel measures between census block groups and between census tracts (based on population weighted centroids, see Section 5.2.3; two separate sets of files) were found. Below is a description of each file with additional information for how the measures were collected. In all of the files, the character versions of block group and census FIPS codes from the centroid files are the row and column names; this way the files can be used together easily. See additional considerations (section 9.3 below) for information on when values needed to be re-queried. Note that we are using the term "adjacency matrix" in the network sense, which is distinct from the use of adjacency (i.e.; contiguity) matrix in spatial analyses. If one wishes to use these matrices for spatial analyses, then one can use the R package *spdep* to create the contiguity matrix[7]

The first measure found is the geometric distance (or "as the crow flies") between areal units. Three different travel measures for driving were found using the Google Maps Distance_Matrix API. The API provides travel distances and times between areal units' centroids using the optimized route. The query was set to use the "best guess" model (rather than pessimistic or optimistic). The matrices are asymmetric indicating that the relationships are not always equal. The rows indicate the origin and the columns indicate the destination. Three different measures of public transit were found using the Google Maps Directions API between areal units' centroids. The Directions API was used because the complete directions were needed to calculate the number of boardings of public transit from origin to destination. In these queries, mode was set to 'transit' and departure time was still Noon on August 1st. All of the matrices are asymmetric.

7.2. Description of files

Each of the seven types of travel relations exists for both census block groups and tracts. Below, we describe the contents of each file in terms of census block groups (with the prefix "bgs_"), and the same relations with the prefix "cts_" identifies the files for census tracts.

[7] Bivand, R., Altman, M., Anselin, L., Assunção, R., Berke, O., Bernat, A., & Blanchet, G. (2015). Package 'spdep'. *See ftp://garr.tucows.com/mirrors/CRAN/web/packages/spdep/spdep.pdf.*

- *bgs_crow.flies.csv* contains the geometric distance between block groups (in meters), or "as the crow flies". This was done with the *distGeo* command from the *geosphere* R package[8]. The matrices are symmetric and the diagonal is 0.
- *bgs_drive.dist.csv* contains the driving distance in meters between census block groups.
- *bgs_drive.dur_avg.csv* contains the average driving duration in seconds between census block groups.
- *bgs_drive.dur_wed.csv* contains the driving duration in seconds between census block groups queried for Wednesday at noon two weeks from query data. This was done so that local daily traffic patterns would not impact the route or duration.
- *bgs_transit.distance.csv* contains the transit distance in meters between two census block groups.
- *bgs_transit.duration.csv* contains the transit duration in seconds between two census block groups.
- *bgs_transit.transfers.csv* contains the number of distinct boardings (train or bus) that an individual would take between two census block groups. A value of 0 indicates that walking is fastest, and thus, no boarding is needed.

7.3. Additional considerations

One block group represented Long Island, located in Boston Harbor and managed as part of the Boston Harbor Islands National Recreation Area. Travel via driving and transit to and from this block group was listed as "Unnavigable" in the datasets. Occasionally, Google Maps API returns outputs that state that two locations do not have transit directions between them, or that the API doesn't support all mode='transit' between these two locations, even though the Google Maps website does. This occurred for two pairs of census tracts and over forty pairs of block groups. These points were manually re-queried with the exact latitude and longitude coordinates, producing results for all but twelve pairs.

[8] Hijmans, R. J., Williams, E., & Vennes, C. (2017). geosphere: Spherical Trigonometry. R package version 1.5-7. *Available at hoops://CRAN. R-project.org/package=geosphere*.

For these twelve, the latitude and longitudes of origin and destination were both increased such that they represented points 80 meters to the North and 80 meters to the East of the actual centroid, which was enough for Google Maps to provide transit directions for eight. For the remaining four they offset was increased to 400 meters.

8. Street intersection edgelist

8.1. Description of contents

The file contains the edge list of streets in Boston that intersect with each other. Using the shape files from BARI's 2013 roads, we used the *st_intersect* command from the *sf* package in R to find which streets intersect[9]. From the 24,891 street segments, there are 47,493 intersection pairs between streets. Each pair represents a dyad in a network, where streets are nodes. The edge list contains all of these dyads. The relationship is undirected. The file "streets.intersections.el.csv" contains the unique tiger line IDs for each street. The names have "tl" added to the numeric ID, so that they can be used as the node IDs in the network. Note that the way intersections were found includes any intersections between the streets' shape files; thus, this does not account for over/under passes.

[9] Pebesma, E. (2017). sf: Simple features for R. R package version 0.5-0. *Available at https://cran.r-project.org/web/packages/sf/index.html.*

APPENDIX A. Codes for Land Use[10]

USE CODE	DESCRIPTION
R1	Residential 1 Family
R2	Residential 2 Family
R3	Residential 3 Family
R4	Residential 4 - 6 Units
A	Residential 7 or more Units
RL	Residential Lot
CD	Condominium
CC	Commercial Condominium
CM	Condo Main (Bldg. broken into condo units)
C	Commercial
RC	Mixed Residential Commercial
CL	Commercial Land
CP	Condo Parking
I	Industrial
E	Exempt
EA	Exempt (Chapter 121A)

[10] Use code AH is not officially defined in the City of Boston's documentation. It represents Affordable Housing.

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

APPENDIX C. Characteristics of Clusters of Road Segments Generated by a Cluster Analysis on Parcel Zoning Characteristics.

Streets with one or more parcels (*n* = 13,825) were categorized 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.

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 (note that only prominent land usage types are noted and proportions do not add up to 100%;other land uses may appear on a given street in a cluster in small proportions):

0 = No Parcels (11,066 streets)

1 = *Three-Family Residential with Assorted Other Residential* (1,835 streets with parcels, or 13.3%): 50.4% Three-Family Residential, 11.7% Two-Family Residential, 8.4% Single-Family Residential, 7.5% Residential Lot, 7.3% Condominiums.

2 = *Mix of Two-Family and Single-Family Residential* (2,519 streets with parcels, or 18.2%): 46.2% Two-Family Residential, 22.7% Single-Family Residential.

3 = *Commercial* (2,396 streets with parcels, or 17.3%): 40.8% Commercial, 11.6% Commercial Lot, 10.2% Apartment Buildings (residential with 7 or more units).

4 = *Single-Family Residential Only* (3,335 streets with parcels, or 24.1%): 74.9% Single-Family Residential.

5= *Exempt* (1,519 streets with parcels, 11.0%): 76.7% Exempt.

6 = *Condominiums* (1,625 streets with parcels, 11.9%): 44.3% Condominiums, 12.1% Single-Family Residential, 10.10% Three-Family Residential, 7.0% Two-Family Residential.

7 = *Mixed-Use Commercial* (596 streets with parcels, or 4.3%): 57.6% Residential-Commercial, 9.0% Commercial.

APPENDIX D. Unit Imputation

The City of Boston's Street and Address Management (SAM) system provides information about all properties in Boston, dividing properties into discrete units where possible. 97,855 properties (55%) in the assessment data set were separated into units. We tabulated the number of units for these, and then imputed the number of units for the remaining parcels using two methods.

For certain land uses, we assumed the number of units based on the definition of the land use itself. These included: Single-Family Residential, Two-Family Residential, and Three-Family Residential, which were set equal to the legal number of residential units associated with the designation (e.g., Two-Family Residential = 2 units); Commercial Lots, Residential Lots, and Condo Parking, which were set equal to zero units as they have no buildings on them; Condo Main, set equal to one unit as it is the lobby of a condo building. This accounted for 1,780 parcels that previously did not have unit data.

The remaining 5,560 properties (4%) were distributed across nine land uses: Residential 4, Apartments (residential with 7 or more units), Commercial Condominium, Commercial, Condominium, Exempt, Exempt (121A), Industrial and Residential-Commercial. We used regression-based imputation, leveraging data from assessments including the total assessed value for the property, land, and building (*AV_TOTAL*, *AV_LAND*, and *AV_BLDG*, respectively) and the total gross floor area and living area (*GROSS_AREA* and *LIVING_AREA*, respectively). The last two also had missing values, so we first imputed values for them based on the other variables. We then ran 9 separate generalized linear models, one for each land use, that used assessed value and area to predict the number of units for all cases for which this information was known. The parameters from these models were then used to estimate the number of units (rounded to the nearest whole number) for those properties for which this information on the number of units was missing.

The imputed values were assessed for outliers using Mahalanobis distance. There were no multivariate outliers. However, for 5 records, the prediction was above 700 units which is considered unrealistic after case-by-case assessment. The imputed values for these cases were removed (i.e., an NA value).