Introduction to the NYC Geodatabase (nyc_gdb)
ArcGIS Version

Frank Donnelly, Geospatial Data Librarian, Baruch College CUNY

Aug 10, 2015

Abstract

This tutorial provides an introduction to the NYC Geodatabase (nyc_gdb), a resource for mapping and analyzing city-level features and data in GIS. The database comes in two formats: a Spatialite geodatabase built on SQLite that can be used in open source software like QGIS and the Spatialite GUI, and a personal geodatabase built on MS Access that can be used in ArcGIS. This document explains their content and structure (which are identical for both formats) and demonstrates how you can use them to explore and map data. Experience with using GIS software is presumed. For detailed metadata for all objects in the database and information about updates, see the document "NYC Geodatabase (nyc_gdb) Data Dictionary". The databases and associated documentation are available at `https://www.baruch.cuny.edu/confluence/display/geoportal/NYC+Geodatabase`.

This document contains a general overview of the database format and structure, and specific instructions for using the MS Access Personal Geodatabase with ArcGIS.

Contents

1 Rights **1**

2 Purpose **2**

3 The Databases **2**

3.1 Formats 2

3.2 Structure 3

4 MS Access Personal Geodatabase and ArcGIS **6**

4.1 Adding Personal Geodatabase Data 7

4.2 Example: Mapping Geodatabase Data 7

1 Rights

Disclaimer: Every effort was made to insure that the data, which was compiled from public sources, was processed accurately for inclusion in the NYC Geodatabase. The creator, Baruch College, and CUNY disclaim any liability for errors, inaccuracies, or omissions that may be contained therein or for any damages that may arise from the foregoing. Users should independently verify the accuracy of the data for their purposes.

The database and associated documentation are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike license CC BY-NC-SA `http://creativecommons.org/licenses/by-nc-sa/4.0/` (cc) BY-NC-SA. You are free to share and to adapt the work as long as you cite the source, do not use it for commercial purposes, and release adaptations under the same license.

2 Purpose

The goals of the NYC Geodatabase project are:

- To provide new users with a resource for learning GIS and experimenting with data
- To provide intermediate users with a resource for expanding their GIS skills, into the realms of spatial SQL and database management
- To provide NYC users with a foundational dataset that they can build on for their specific research
- To provide all users with an example of an open source Spatialite database, to demonstrate its capabilities
- To build a foundation for creating additional GIS instructional programs at Baruch College, CUNY
- To create a basic dataset for in-house work at Baruch College, CUNY

The NYC Geodatabase (nyc_gdb) is a resource designed for basic geographic analysis and thematic mapping within the five boroughs of New York City. It contains geographic features and data compiled from several public sources. All of the features were transformed to share a common coordinate reference system (CRS) that is appropriate for the area: NAD 83 NY State Plane Long Island (feet); EPSG code 2263. Subsets of large features like water, greenspace, and public facilities were created and Census geographies like tracts, ZCTAs, and PUMAs were geoprocessed to create land-based boundaries . Census data from the 2010 Census, American Community Survey (ACS), and ZIP Code Business Patterns are stored in tables that can be easily related to geographic features. Transit and public facility point data were gathered from several city agencies and transformed into spatial data that can be used for reference or analysis for measuring distance, drawing buffers, or counting features within areas.

The database contains many foundational map layers and data that can be readily used, but was also constructed so users could build on that foundation and extend it for their own projects. All of the boundaries are based on the 2010 Census, which allows users to easily add additional layers from the Census TIGER files or to extend the study area beyond the five boroughs. The database also serves as an educational tool for introducing spatial databases and SQL.

The dataset is appropriate for thematic and reference mapping at a city and borough level and for thematic mapping at a sub-borough level. While it can be used for creating detailed reference maps at a sub-borough level, it is not the ideal choice for this purpose, given the degree of generalization in the TIGER Line files (in terms of the detail of the line work for the coast line and the number of water and landmark features selected for inclusion). Users will have to judge for themselves based on their intended purpose.

The database will be updated bi-annually: Census American Community Survey (ACS) data in the winter, Census ZIP Business Patterns and NYC transit data in the summer, and NYC public facilities in both winter and summer. Features created from the Census Bureau TIGER shapefiles (statistical areas and landmarks) and 2010 Census data are stable and won't be updated until after the 2020 Census.

The databases and documentation are available at `https://www.baruch.cuny.edu/confluence/display/geoportal/NYC+Geodatabase`.

3 The Databases

3.1 Formats

A geodatabase (or spatial database) is a relational database that has been enhanced to hold spatial objects or geographic features. The NYC Geodatabase (nyc_gdb) comes in two formats for use with different GIS software. The Spatialite version (.sqlite) is an open source format built on the SQLite database, and can be used in open source GIS software like QGIS. Spatialite also has its own command line and windows-based tools (the Spatialite GUI) which allow users to manipulate the data in a relational database environment. The personal geodatabase (.mdb) is a proprietary format created by ESRI; it is built on the Microsoft Access database and can be used in an ArcGIS environment. The walls have been falling in recent times; ESRI began supporting the Spatialite format beginning with ArcGIS version 10.2, and personal geodatabases can at least be accessed with later 2.x versions of QGIS.

Both formats give users a better way to organize and structure their data relative to shapefiles and individual data tables, as multiple features and attributes can be stored in a single database file and can be easily related to each other. Both formats are simple, file-based databases that can be created, copied, and distributed easily. Unlike enterprise-level databases (i.e. ArcGIS enterprise geodatabases, PostGIS), file sizes are functionally limited to a few gigabytes, and individual user permissions cannot be specified. Desktop databases are not an ideal choice for data that must be accessed simultaneously from many computers over a network filesystem.

The open source Spatialite database has the added advantage of allowing users to perform spatial queries in addition to regular SQL queries. Examples include calculating areas and distance and evaluating geographic relationships like adjacency and overlap. Thus, Spatialite is able to extend the geographic selection and analysis capabilities of open source GIS, which is still developing these capabilities. In contrast, the capabilities of the MS Access personal geodatabase are limited; in the proprietary world the ArcGIS software does most of the heavy lifting via the ArcToolbox, and the database serves as a simple container for organizing and storing data.

Both geodatabases can also be accessed and manipulated using regular relational database tools, like MS Access or the SQLite Manager (available as a Firefox plugin), but these tools can only be used for traditional SQL queries and nothing spatial. When a geodatabase is created each respective program (Spatialite GUI or ArcGIS) populates the new database with tables and relations that manage and support any geographic features that are inserted. When working with the database care must be taken to not alter or remove these tables.

3.2 Structure

Objects in the database are categorized and named with a prefix to differentiate different types of features. A brief list of the database objects is provided below; for full details consult the nyc_gdb Data Dictionary. In most instances normal form for relational databases is violated in order to provide a resource that is readily usable for mapping and analysis; names and codes are repeated in some tables to facilitate identification and selection, and percent values are pre-calculated and provided with totals. This is particularly valuable for the American Community Survey data, where calculating margins of error for derived values like percentages is a difficult task that's cumbersome to perform within a relational database.

3.2.1 A Tables

Objects that begin with the prefix "a" are geographic features that represent points, lines, and areas. The census statistical areas are designed so that they can be joined with "b" tables that contain census data, and all census area features are generalized so that they represent land areas. Public facility point features contain identifying information like names and addresses as well as one variable, like capacity or ridership, that can be measured or mapped; features from the NYC Dept of City Planning's Facilities dataset are taken "as is" and are not verified for accuracy or omissions. Each table contains a unique identifier.

All "a" tables have a column called "bcode" that indicates which borough the feature is in, to facilitate select queries. The bcode is the US Census ANSI / FIPS code for the county.

- 36005 - Bronx County (Bronx)
- 36047 - Kings County (Brooklyn)
- 36061 - New York County (Manhattan)

- 36081 - Queens County (Queens)
- 36085 - Richmond County (Staten Island)

a_boroughs : The five boroughs of NYC, from the census counties file

a_colleges : From the NYC Dept City Planning Facilities database

a_facilities : Selection of airports and other large public facilities from the census landmarks file

a_greenspace : Selection of large parks, wildlife areas, and cemeteries from the census landmarks file

a_hospitals : From the NYC Dept City Planning Facilities database

a_libraries : Public libraries from the NYC Dept City Planning Facilities database

a_metro_counties : Counties in the NYC Metropolitan CSA, from the census counties file

a_path_stations : NYC PATH Stations from NJ Transit with ridership data from PANYNJ

a_pumas2010 : Public Use Microdata Areas; census statistical areas designed to have approx 100k residents. Boundaries from the 2010 Census were used for the first time in the 2012 ACS

a_roads : All roads in NYC from the census roads file

a_schools_private : From the NYC Dept City Planning Facilities database

a_schools_public : From the NYC Dept City Planning Facilities database

a_subway_complexes : Single or multiple stations with shared entrances and passages where riders can freely transfer, and for which the MTA publishes ridership statistics

a_subway_complexes_srvnotes : Notes on service disruptions that impact the ridership statistics in a_subway_complexes

a_subway_stations : Individual stations represented by distinct platforms for specific trains

a_tract_popcenters : Population centers / centroids for census tracts based on the 2010 Census. Represents the center of the population's distribution within each tract

a_tracts : 2010 census tracts; census statistical areas designed to have an ideal size of 4,000 residents, with a range of 1,200 to 8,000. Tracts can be aggregated to Neighborhood Tabulation Areas (NTAs) created by the City

a_train_stations : LIRR and Metro North stations in NYC, from the MTA

a_water_coastal : Selection of major coastal water from the census water file, used to create land boundaries for the census layers

a_water_lakes : Selection of major lakes from the census water file

a_zctas : 2010 ZIP Code Tabulation Areas; census statistical areas created by aggregating census blocks based on postal addresses, to create geographic approximations of USPS ZIP Codes

3.2.2 B Tables

Objects that begin with the prefix "b" are non-spatial data tables from the US Census, with data reported as values and percentages. These tables can be joined to geographic "a" features so that quantities can be mapped and evaluated spatially. The unique identifier field for the 2010 Census and ACS "b" tables is "GEOID2", which is the census FIPS code for that area. The unique ID for the Business Patterns tables is "ZCTA5", the five-digit Census ZCTA number. Tables are named based on their geography, dataset, and year. Column names are codes that uniquely identify each variable. For each dataset there is an index table named with the suffix "lookup", that relates column codes to variable names.

The American Community Survey (ACS) is an ongoing sample survey of the population that's tabulated annually for 1, 3, and 5-year periods. The values are published as estimates with a 90% confidence interval and margins of error (+/-). In this database, data from the ACS are from the 5-year series; the year indicates the year of release and final year of the estimate range (i.e. 2012 represents 2008-2012 5-year data). There are two data tables for each geography that represent a subset of the four demographic profiles (tables DP02 through DP05). Each individual variable is named based on its subject and consists of four adjacent columns in this order: the estimate itself (identified by the suffix "E"), a margin of error for the estimate (suffix "M"), a percent total (suffix "PC"), and a margin of error for the percent total (suffix "PM"). The lookup table correlates the column heading with the variable name. This data is updated annually.

- b_YEARacs_lookup
- b_pumas_YEARacs1
- b_tracts_YEARacs1
- b_zctas_YEARacs1
- b_pumas_YEARacs2
- b_tracts_YEARacs2
- b_zctas_YEARacs2

The decennial census is a 100% count of the population taken on April 1st. Data from the 2010 Census represents all the data in the demographic profile (table DP01). Each variable has two values: the actual count (named with the prefix HD01) and a percent total (named with the prefix HD02). The percent totals are stored in a separate table with the suffix "pct". The lookup table correlates the column headings (created by the Census) with the variable names, while the footnotes table contains footnotes referenced for certain variables in the index. This data will not be updated until the 2020 Census. Decennial census data is not tabulated at the PUMA level.

- b_2010census_lookup
- b_tracts_2010census
- b_zctas_2010census
- b_2010census_footnotes
- b_tracts_2010census_pct
- b_zctas_2010census_pct

The Census Bureau compiles the ZIP Code Business Patterns (ZBP) data from the Business Register, which contains a record for each business establishment with paid employees in the US; an establishment is a single physical location at which business is conducted or services or industrial operations are performed. ZBP data is stored in two tables: the "emp" table provides the total number of establishments, employees, and payroll (for the first quarter and annually in $1,000s of dollars) and the "ind" table provides a count of establishments based on type of business, as classified by the North American Industrial Classification System (NAICS). The names that are correlated with the two-digit NAICS sector codes are in the "indcodes" table. The records in these tables represent US Census ZCTAs and *not* USPS ZIP Codes. The data was aggregated from ZIP Codes (as published in the ZBP) to ZCTAs. A table that cross-walks ZIP Codes to ZCTAs is included for user reference. This data is updated annually.

- b_zctas_YEARbiz_emp
- b_zctas_YEARbiz_ind
- b_zctas_YEARbiz_indcodes
- b_zips_to_zcta

3.2.3 C Tables

Objects that begin with the prefix "c" are geographic features that represent the actual boundaries for census statistical areas. Other than transforming the projection to match the other database features, these features and their attributes have not been altered in any way from the original TIGER shapefiles from the Census. They are included in case the user wishes to depict the actual boundaries (that encompass land and water) for reference. They should not be used for mapping census data.

- c_bndy_boroughs
- c_bndy_metro_counties
- c_bndy_pumas2010
- c_bndy_tracts

3.2.4 X Tables

Objects that begin with the prefix "x" are "extra" geographic features that represent the original source data for some of the "a" features. The "x" features are included in case the user wants to add additional features that are not part of the generalized "a" layer.

x_landmarks : this layer was used to create the facilities and greenspace layer and includes all 2010 Census landmarks in NYC

x_nad83_boroughs : this borough layer does not share the same coordinate reference system as the other layers in the database; it is in simple NAD 83. It is included to provide a frame of reference for users who need to plot latitude and longitude data

x_water : this layer was used to create the coastal and lakes water layers and includes all the 2010 Census water layers in the greater metro area

3.2.5 Other Tables

"d_ntas_2010census" is not a table, but a view that is included for the sake of example. It joins the a_tracts layer to the b_tracts_2010census table and groups basic population and housing data by Neighborhood Tabulation Areas defined by the City.

"z_metadata" is a table that describes the name and source of all of the tables in the database, along with the year that the feature or table was last updated.

All other objects are core parts of the geodatabase designed to manage and support geographic features. The table names will differ between the MS Access and SQLite versions. These tables should not be removed or altered, otherwise the database could be rendered unusable.

4 MS Access Personal Geodatabase and ArcGIS

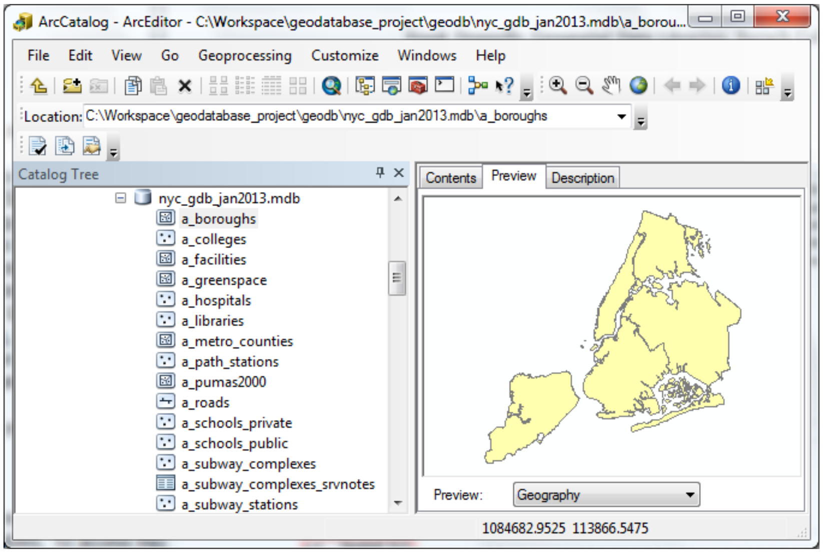
ArcGIS is an established, proprietary suite of GIS software that runs on MS Windows. You can find information about ArcGIS on ESRI's website at `http://resources.arcgis.com/en/help/getting-started/`. MS Access personal geodatabases are one of the two single-user geodatabases that ArcGIS supports (the other is the file database, created by ESRI) `http://www.esri.com/news/arcnews/winter0809articles/the-geodatabase.html`. Unlike the open source alternatives, the Access personal geodatabase is only used for organizing, structuring, and storing data and not for performing geospatial operations; the ArcGIS software handles the latter directly via tools in the ArcToolbox and menus in ArcMap, and can read and write directly to the database.

4.1 Adding Personal Geodatabase Data

There are two ArcGIS applications that you can use for working with the geodatabase:

ArcMap : For manipulating data, geoprocessing, conducting analyses, and making maps.

ArcCatalog : For creating, previewing, and editing geodatabases, and also for importing and exporting features and tables. To access the geodatabase you connect to the folder where it is located (under File > Connect Folder) and use the file menu tree to expand and view the contents. Unique icons identify the type of feature, and selecting the feature allows you to preview the geography and attributes.



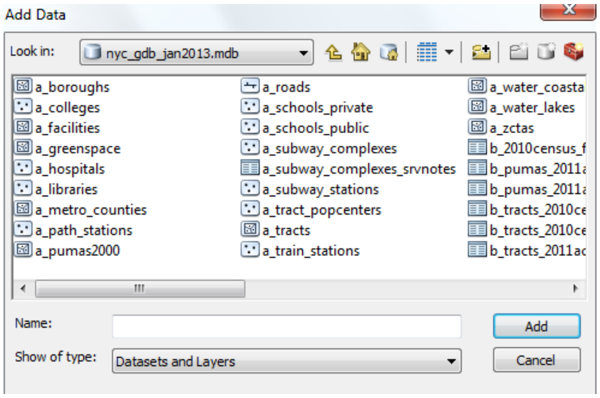
Location: C:\Workspace\geodatabase_project\geodb\nyc_gdb_jan2013.mdb\a_boroughs

Preview: Geography

4.2 Example: Mapping Geodatabase Data

The following example illustrates how to add personal geodatabase features, join data tables to features, and map data in ArcGIS. These instruction were written using ArcGIS 10.1 in an MS Windows environment.

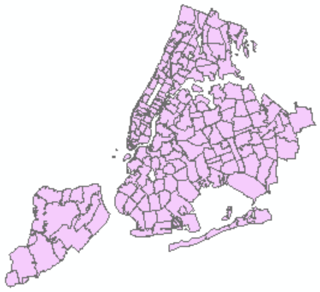
1. Launch ArcMap. Hit the Add Data ✚ button. Browse to the folder where you've placed the database. In ArcGIS geodatabases appear as grey cylinders: nyc_gdb_jan2013.mdb. Doubleclick on the database to see its contents.



Look in: nyc_gdb_jan2013.mdb

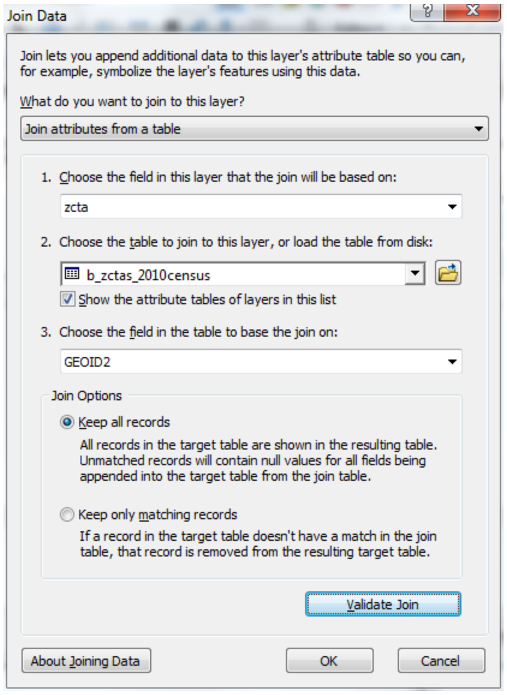
Show of type: Datasets and Layers

2. Select the a_zctas layer in the list and hit the Add button. This adds the ZCTAs to the project.



3. Hit the Add Data button again. Select the b_zctas_2010census table. Hit the Add button to add it to the project.

4. Doubleclick on the a_zctas layer in the layers list to open the Layer Properties menu. Select the Joins & Relates tab. In the Joins section on the left hit the Add button. In the Join Data menu choose zcta in box 1 as the field in this layer that the join will be based on, choose b_zctas_2010census in box 2 as the table to join to this layer, and choose GEOID2 in box 3 as the field in the table to base the join on. Under Join Options keep the radio button that says Keep all records selected. Hit OK. The Joins & Relates tab now shows the details of the join.



Join Data

Join lets you append additional data to this layer's attribute table so you can, for example, symbolize the layer's features using this data.

What do you want to join to this layer?

Join attributes from a table

1. Choose the field in this layer that the join will be based on:

zcta

2. Choose the table to join to this layer, or load the table from disk:

b_zctas_2010census

☑ Show the attribute tables of layers in this list

3. Choose the field in the table to base the join on:

GEOID2

Join Options

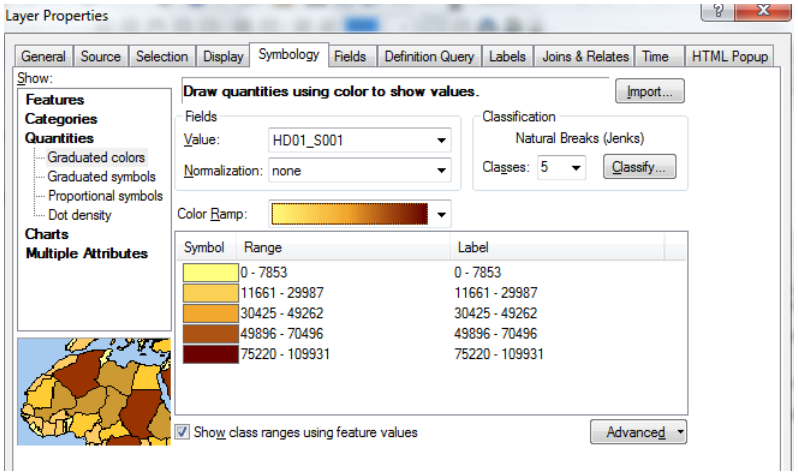
● Keep all records

All records in the target table are shown in the resulting table. Unmatched records will contain null values for all fields being appended into the target table from the join table.

○ Keep only matching records

If a record in the target table doesn't have a match in the join table, that record is removed from the resulting target table.

5. In the Layers Properties menu switch to the Symbology tab. In the Show menu on the left change the option from Single symbol to Quantities, and choose the option for Graduated colors. Under Fields hit the drop down and select the HD01_S001 column, which represents total population in 2010. In the classification menu on the right keep the default to Natural Breaks (Jenks). Check the box that says Show class ranges using feature values. Hit OK.



Draw quantities using color to show values.

Fields

Value: HD01_S001

Normalization: none

Classification

Natural Breaks (Jenks)

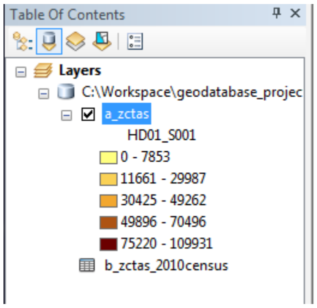
Classes: 5

Color Ramp:

Symbol	Range	Label
🟨	0 - 7853	0 - 7853
🟨	11661 - 29987	11661 - 29987
🟧	30425 - 49262	30425 - 49262
🟫	49896 - 70496	49896 - 70496
🟥	75220 - 109931	75220 - 109931

☑ Show class ranges using feature values

6. If you save the project ArcGIS will remember to add these objects and join them each time you open the project. We know that HD01_S001 is total population; we can verify this and see what the other columns are by viewing the b_2010census_lookup table. Instead of mapping totals we could map percentages stored in the b_zctas_2010census_pct table. To fill in the white space and cover non-residential areas we could add the a_facilities and a_greenspace features and layer them over top of the ZCTAs (selecting and dragging the items in the layer list changes their drawing order).



Layers

C:\Workspace\geodatabase_projec

☑ a_zctas

HD01_S001

🟨 0 - 7853

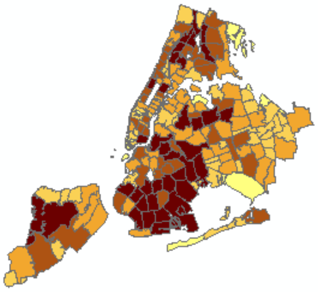
🟨 11661 - 29987

🟧 30425 - 49262

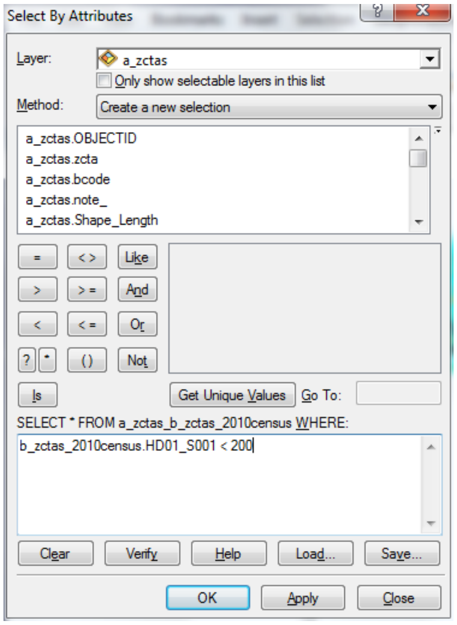
🟫 49896 - 70496

🟥 75220 - 109931

b_zctas_2010census



7. Under the Selection menu, the option to Select By Attributes allows you to perform SQL queries, while Select By Location allows you to do spatial selections. For example, to select non-residential ZCTAs that have less than 200 people, go to Selection > Select By Attributes. The basic SQL statement is already defined at the bottom of the menu as `SELECT * FROM a_zctas_b_zctas_2010census WHERE`. To add a where clause, you can use the fields and operators boxes to add elements, or you can type them directly in the SQL box: `b_zctas_2010census.HD01_S001 < 200`. Hit OK to see the result.



Select By Attributes

Layer: a_zctas

☐ Only show selectable layers in this list

Method: Create a new selection

a_zctas.OBJECTID
a_zctas.zcta
a_zctas.bcode
a_zctas.note_
a_zctas.Shape_Length

SELECT * FROM a_zctas_b_zctas_2010census WHERE:

```
b_zctas_2010census.HD01_S001 < 200
```

8. By default the selected areas are outlined in blue. You can view the records for the selection by selecting a_zctas in the layers menu, right clicking, and opening the attribute table. You can clear the selection by hitting the Clear Selection button.