

# Introduction to the NYC Geodatabase (nyc\_gdb) Open Source Version

Frank Donnelly, Geospatial Data Librarian, Baruch College CUNY

July 8, 2013

## 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 [http://www.baruch.cuny.edu/geoportal/nyc\\_gdb/](http://www.baruch.cuny.edu/geoportal/nyc_gdb/).

This document contains a general overview of the database format and structure, specific instructions for using the open source SQLite / Spatialite database with QGIS and the Spatialite GUI, and a basic introduction to Spatialite.


## Contents

<b>1</b>	<b>Rights</b>	<b>1</b>
<b>2</b>	<b>Purpose</b>	<b>2</b>
<b>3</b>	<b>The Databases</b>	<b>3</b>
3.1	Formats . . . . .	3
3.2	Structure . . . . .	3
<b>4</b>	<b>Spatialite and QGIS</b>	<b>6</b>
4.1	Adding Spatialite Data . . . . .	7
4.2	Example: Mapping Spatialite Data . . . . .	8
4.3	Spatial Queries in QGIS . . . . .	10
<b>5</b>	<b>Working With the Spatialite GUI</b>	<b>10</b>
5.1	Basic Tasks . . . . .	10
5.2	Spatial SQL . . . . .	15
5.3	Extending the DB . . . . .	21
5.4	Reference & Links . . . . .	28

## 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/3.0/> . 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 Plan 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 geoprocesed 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 [http://www.baruch.cuny.edu/geoportal/nyc\\_gdb/](http://www.baruch.cuny.edu/geoportal/nyc_gdb/).

## 3 The Databases

### 3.1 Formats

A geodatabase 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.

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 NYC DOITT

**a\_pumas2000** : Public Use Microdata Areas; census statistical areas designed to have approx 100k residents. Boundaries from the 2000 Census are still in use in the 2010 and 2011 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 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 "b" tables is "GEOID2", which is the census FIPS code for that area. 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 "index", 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. 2011 represents 2007-2011 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 index table correlates the column heading with the variable name. This data is updated annually.

- b\_YEARacs\_index
- 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 index 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\_index
- b\_2010census\_footnotes
- b\_tracts\_2010census
- b\_tracts\_2010census\_pct
- b\_zctas\_2010census
- 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. This data is updated annually.

- b\_zctas\_YEARbiz\_emp
- b\_zctas\_YEARbiz\_ind
- b\_zctas\_YEARbiz\_indcodes

### 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\_pumas2000
- 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 Spatialite and QGIS

QGIS is free, open source, cross-platform GIS software, available for download from <http://www.qgis.org/>. For an introduction to QGIS try the GIS Practicum at <http://www.baruch.cuny.edu/geoportal/practicum/>.

With QGIS you can use a Spatialite database to:

- Add, view, and symbolize geographic features and tables
- Layer Spatialite features with other spatial files (like shapefiles)
- Use QGIS selection, analysis, and geoprocessing tools
- Create new shapefiles or tables from Spatialite layers
- Create new Spatialite layers and tables from scratch
- Edit Spatialite layers by modifying or adding features and attributes

- Create basic SQL and spatial SQL queries
- Create basic spatial views stored within the project (selections with criteria, table joins)

You can't do the following in QGIS with Spatialite:

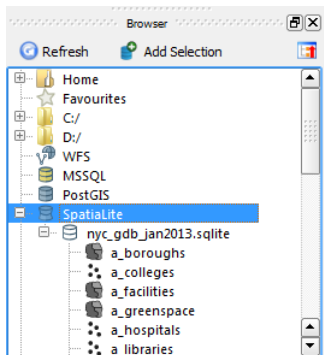
- Create new Spatialite databases
- Create new Spatialite layers from shapefiles or queries
- Create complex SQL and Spatial SQL queries (aggregation, calculated fields)


NOTE: It is possible to perform some of these operations with 3rd party plugins, but it's not recommended. QGIS natively supports exporting an individual shapefile out as a new Spatialite database, creating a database that consists of a single layer. It currently does not support exporting multiple files to a new database, or exporting single or multiple files to existing databases.

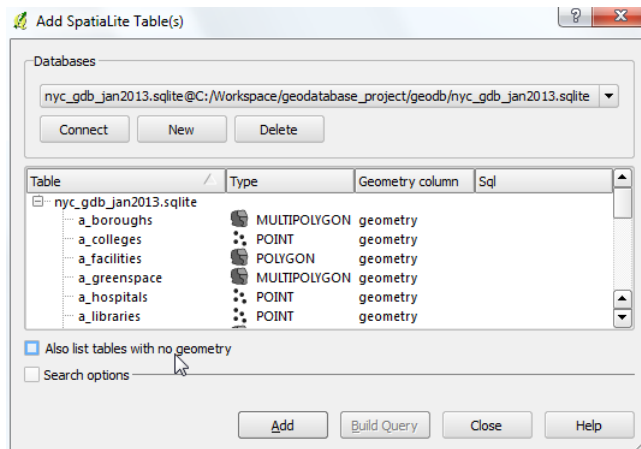
## 4.1 Adding Spatialite Data



There are a few different methods for adding Spatialite data to QGIS:

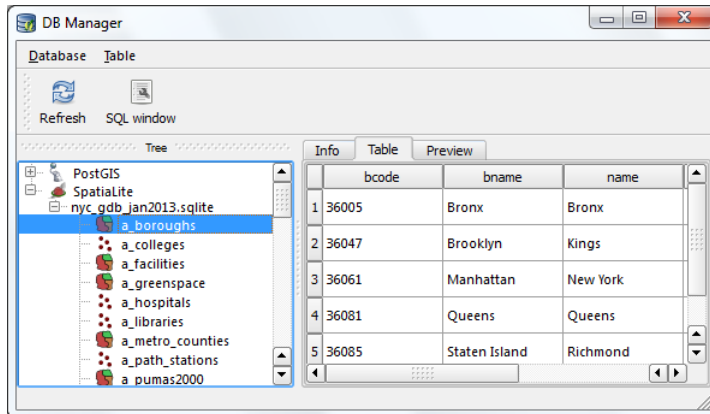
**Browser** : The data browser was introduced in QGIS 1.8. It allows you to view your entire file system within QGIS so you can easily drag and drop features into a project. The browser allows you to connect to databases by selecting the database type and browsing to its location to establish a connection. You can see and add Spatialite geographic features, but not tabular ones. If the browser is not visible, you can activate it by right-clicking on an empty area of the toolbars and checking Browser.



**Add Spatialite Layer Button** : This button  is on the Layers toolbar. Press it and browse through your file system to connect to the database. Once connected you can add geographic and tabular features to your project.




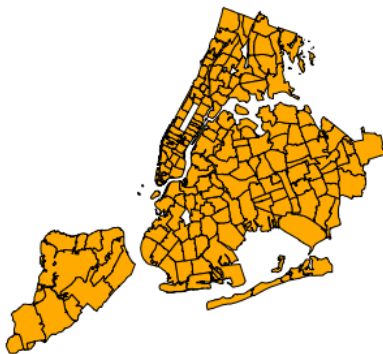
**DB Manager** : This plugin allows you to access a geodatabase and interact with it in a separate window - you can preview tables and geometry and drag and drop both into your project. You can also run SQL and spatial SQL queries in the query window, and create temporary spatial queries or views that you can view in the project window (such as selecting a subset of records that meet a certain criteria and viewing the features as a new layer). To activate the plugin go to Plugins > Manage Plugins and select the DB Manager. This will add a button  to the toolbar and a Database entry on the file menu. You must connect to your database using the browser or the  toolbar button before accessing it via the manager.




## 4.2 Example: Mapping Spatialite Data

The following example illustrates how to add Spatialite features, join data tables to features, and map data in QGIS. These instructions were written using QGIS 1.8 in an MS Windows environment, and will vary when using other operating systems (i.e. Mac laptop users should substitute "right-click" with a two-fingered single click) or older versions of QGIS.

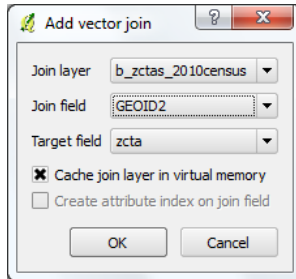
1. Launch QGIS. Hit the Add Spatialite Layer  button on the toolbar. Hit the New button and browse to the folder where you've placed the database. Select it and then hit the Connect button. You should see a list of all the geographic features in the database.
2. Select the `a_zctas` layer in the list and hit the Add button. This adds the ZCTAs to the project.



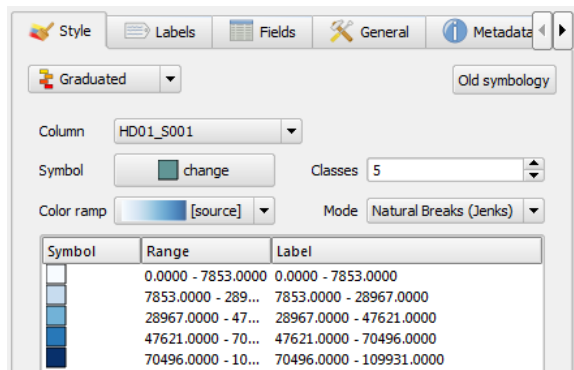
3. Hit the Database Manager  button. Expand the Spatialite DB tree by hitting the plus sign, then expand the `nyc_gdb` database to see its contents. Scroll through the list to the tables and select `b_zctas_2010census`. Add it to the project by holding down the left mouse button and dragging it into the view, or right click and choose the option that says Add to QGIS canvas. Close the DB Manager.



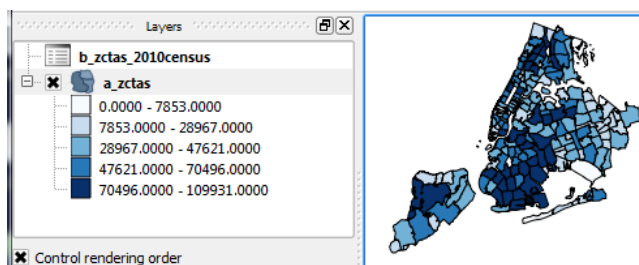
4. Doubleclick on the a\_zctas layer in the layers list to open the Layer Properties menu. Select the Joins tab. Hit the green plus symbol. In the Add vector join menu choose b\_zctas\_2010census as the join layer, GEOID2 as the join field, and zcta as the Target field. Hit OK to save the join. GEOID2 and zcta are the unique identifiers in the table and zcta features respectively; they both contain the 5 digit ZCTA number (you could verify this by viewing the table and attribute table of the features).



5. In the Layer Properties menu switch to the Style tab. Change the symbol dropdown from Single Symbol to Graduated. The default Column is HD01\_S001, which is the total 2010 population. In the dropdown menu on the right change the Mode to Natural Breaks. Choose a color ramp. Hit Classify, then hit OK.




6. Expand the a\_zctas layer in the layers menu to see the classification categories. Select the layer name, right click, and check Show Feature Count to display the number of features in each class.



7. If you save the project QGIS 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\_index 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\_greenpace features and layer them over top of the ZCTAs (selecting and dragging the items in the layer list changes their drawing order).

NOTE: In QGIS there are limitations to what you can do with tabular data that has been joined to geographic features; some limits are options that are not available, such as the ability to run certain QGIS tools (like Basic Statistics under the Vector > Analysis tools) and others are due to bugs (as of QGIS 1.8, some classification methods in the Style menu will not classify joined data properly). To overcome these limitations you must permanently fuse the data to the geography by creating a new feature. You can do this by creating a new table / feature in Spatialite and populating it with a query, or you can use QGIS to save the joined geography and data as a new shapefile.

### 4.3 Spatial Queries in QGIS


You can use the SQL Window in the Database Manager  to perform many of the SQL and Spatial SQL queries that are demonstrated in the next chapter on the Spatialite GUI - with two caveats. First, there are usually two different ways for referencing spatial commands: you can use the prefix ST\_ (so the function for calculating area would be ST\_Area) or you can simply use the alias for the function and drop the prefix (Area). The Database Manager in QGIS 1.8 does NOT recognize the ST\_ prefix; you MUST drop the prefix from each spatial command to run queries.

Second, if you are creating new calculated or aggregated fields that you want to map, you'll need to use the Spatialite GUI to create these and make them permanent columns in a new table. More on that in the next section.

## 5 Working With the Spatialite GUI

The Spatialite GUI is free, open source, cross-platform software that allows you to work with a geodatabase in a relational database environment and to conduct spatial SQL queries. The simplest way to get it up and running is to download a pre-compiled binary for your OS at <http://www.gaia-gis.it/spatialite-2.4.0-4/binaries.html>. Alternatively you can download the source files from [https://www.gaia-gis.it/fossil/spatialite\\_gui/index](https://www.gaia-gis.it/fossil/spatialite_gui/index). The nyc\_gdb was created using Spatialite version 2.4.0 RC-4, and this tutorial covers GUI version 1.4.0. Both versions are a couple years old, but are stable and were selected to make the software approachable for new users; the extensive Spatialite Cookbook was written using these versions, and they were the last to be released as easy-to-install binaries for all operating systems (binaries for current versions are only released for MS Windows; Mac and Linux users must download and compile from the source).

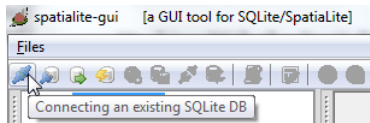
This section demonstrates basic SQL and spatial SQL that you can use when working with the database. For a fuller treatment of SQL with Spatialite visit the Spatialite tutorials, cookbook, and reference guide at <https://www.gaia-gis.it/fossil/libspatialite/index>. There are a number of good tutorials on non-Spatial SQL with SQLite, such as <http://zetcode.com/databases/sqlitetutorial/>, as well as a few good books - *Using SQLite* by Jay Kreibich provides a clear intro for new database users as well as a thorough reference. The complete reference for SQLite syntax is available at: <https://www.sqlite.org/lang.html>.

Most of the spatial SQL functions are prefaced with the prefix ST\_. This chapter follows that convention as it makes these commands easier to identify. However, you can drop the prefix and Spatialite will be able to interpret it. If you run any spatial SQL functions in QGIS in the Database Manager  you MUST drop the prefix; otherwise the functions won't work.

### 5.1 Basic Tasks

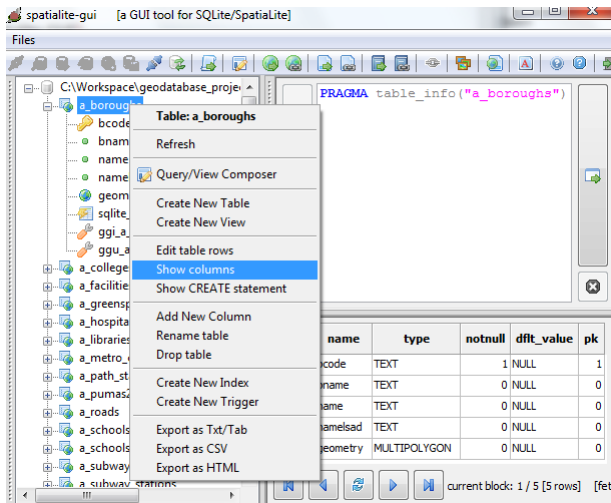
#### 5.1.1 Connect to database

Launch the Spatialite GUI. Hit the Connect to database button, browse through your file system to where the database is stored, and select it. Whenever you subsequently launch Spatialite it will automatically connect to this database, unless you explicitly hit disconnect on the toolbar.



### 5.1.2 View table metadata

In the menu tree on the left hit the plus symbol beside `a_boroughs` to expand and display its attribute list. The primary key that uniquely identifies each record looks like a key, and the column where geometry is stored looks like a globe. Click on `a_boroughs` and right click to open a menu of options. Choose the Show Columns option. This lists each column, its data type, and any constraints (i.e. whether null values are permitted, what the primary key is, etc).



You can also use this menu to edit individual rows in a table, add new columns, change the name of a table, and delete tables. You cannot delete columns or change the name or data type of columns; this is something that SQLite does not support (more on this later). These menu options are shortcuts that save you the step of having to enter SQL code; you could type the code yourself in the SQL window. For example, the code for viewing the table metadata for the boroughs feature:

```
1 PRAGMA table_info (a_boroughs)
```

### 5.1.3 View table records

To view the actual records in a table you have to execute a SELECT query. To view the records in the boroughs table type the following into the SQL box, then execute the code by hitting the large Execute SQL Statement button to the right of the SQL window.

```
2 SELECT *
   FROM a_boroughs
```

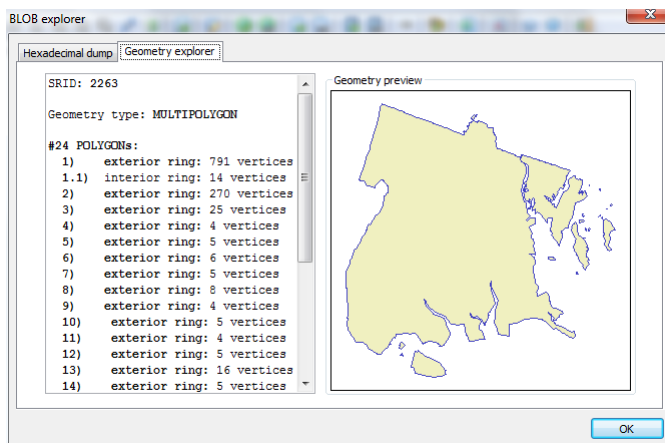
	bcode	bname	name	namelsad	geometry
1	36005	Bronx	Bronx	Bronx County	BLOB sz=21452 GEOMETRY
2	36047	Brooklyn	Kings	Kings County	BLOB sz=32270 GEOMETRY
3	36061	Manhattan	New York	New York County	BLOB sz=17402 GEOMETRY
4	36081	Queens	Queens	Queens County	BLOB sz=42097 GEOMETRY
5	36085	Staten Island	Richmond	Richmond County	BLOB sz=17544 GEOMETRY

current block: 1 / 5 [5 rows] [fetched in 00:00:00.035]

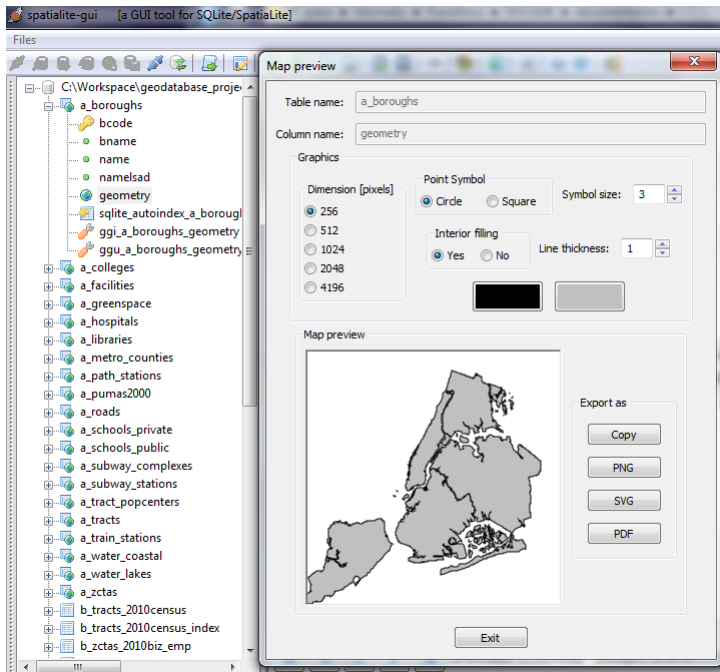
NOTE - executing the SELECT \* statement on tables with a large number of columns or records (for example, census tract data tables) may take time to process.

#### 5.1.4 Preview table geometry

In the table results window (underneath the SQL window), in the geometry column for the Bronx, click on the actual data value that says BLOB sz=21452 GEOMETRY to activate it, then right click on it and choose BLOB Explore. In the BLOB Explorer menu that appears the default is the Hexadecimal tab - click on the Geometry Explorer tab to see a preview of the feature and details about the geometry. Hit OK to close the window.



Recent versions of the software allow you to preview an entire layer. In the object tree on the left, expand the columns for the boroughs. Select the geometry column, right click, and chose the Map Preview option The preview gives you the option to export a picture of the layer out as an image. Hit Exit to close the window.

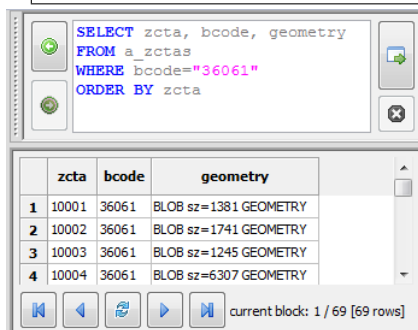


### 5.1.5 Basic select query

Let's perform a basic, non-spatial SQL query. Select all ZIP Code Tabulation Areas in Manhattan and order them by ZCTA number. We'll use the bcode column, which contains the census ANSI / FIPS code for each borough. Type the code into the SQL box and hit execute to see the results.

3

```
SELECT zcta, bcode, geometry
FROM a_zctas
WHERE bcode="36061"
ORDER BY zcta
```



You can also do aggregate queries. This will count the number of ZCTAs by borough:

4

```
SELECT bcode, COUNT(zcta) AS zctas
FROM a_zctas
GROUP BY bcode
```

	<b>bcode</b>	<b>zctas</b>
<b>1</b>	36005	25
<b>2</b>	36047	38
<b>3</b>	36061	69
<b>4</b>	36081	69
<b>5</b>	36085	13

### 5.1.6 Identify census variables

In order to work with the census data in the "b" tables, we need to identify the variable that we want and find the code that identifies that column. Let's say we are interested in total population by ZCTA for the 2010 census. First, we need to look in the index table for the 2010 Census:

5 

```
SELECT *
FROM b_2010census_index
```

	<b>count_id</b>	<b>count_name</b>
<b>1</b>	GEOID	Id
<b>2</b>	GEOID2	Id2
<b>3</b>	GEOLABEL	Geography
<b>4</b>	HD01_S001	Number; SEX AND AGE - Total population
<b>5</b>	HD01_S002	Number; SEX AND AGE - Total population - Under 5 years
<b>6</b>	HD01_S003	Number; SEX AND AGE - Total population - 5 to 9 years

The index table tells us that the total population for all sexes and ages is identified by the code HD01\_S001. If we scroll to the right of the table we'll see the column IDs for the percent totals. For the 2010 Census data, codes that begin with the prefix HD01 indicate that the value is a number. Percent totals, which begin with the prefix HD02, are stored in a separate "pct" table. The suffix, S001, indicates the variable number. HD01\_S001 and HD02\_S001 are the total population and percent total respectively, HD01\_S002 and HD02\_S002 are the total population under 5 years of age and the percent total under five years of age, etc.

Now that we know that HD01\_S001 holds total population, we can query the ZCTA 2010 Census table for that value and sort the results from most to least populated. Using AS allows us to create an alias for cryptic column names. The first three columns in the table are standard geographic identifiers; in the "b" tables GEOID2 is always the unique identifier.

6 

```
SELECT geoid, geoid2, geolabel, HD01_S001 AS pop2010
FROM b_zctas_2010census
ORDER BY pop2010 DESC
```

	<b>GEOID</b>	<b>GEOID2</b>	<b>GEOLABEL</b>	<b>pop2010</b>
<b>1</b>	8600000US11368	11368	ZCTA5 11368	109931
<b>2</b>	8600000US11226	11226	ZCTA5 11226	101572
<b>3</b>	8600000US11373	11373	ZCTA5 11373	100820

### 5.1.7 Table join

Let's query the ZCTAs in Manhattan that have a population that's greater than 200 people, to eliminate ZCTAs that are non-residential (encompassing large office buildings and public facilities). To do that we'll need to join our ZCTA census table to our ZCTA geographic features, which have the borough code field. There are two different ways to

write a JOIN statement; you can be implicit:

```
7 SELECT zcta, bcode, HD01_S001 AS pop2010
   FROM a_zctas, b_zctas_2010census
   WHERE zcta=geoid2 AND bcode="36061" AND pop2010 > 200
   ORDER BY pop2010
```

Or explicit:

```
8 SELECT zcta, bcode, HD01_S001 AS pop2010
   FROM a_zctas AS Z
   JOIN b_zctas_2010census AS C
   ON (Z.zcta = C.geoid2)
   WHERE bcode="36061" AND pop2010 > 200
   ORDER BY pop2010
```

The result is the same:

	zcta	bcode	pop2010
1	10162	36061	1685
2	10006	36061	3011
3	10004	36061	3089
4	10282	36061	4783

current block: 1 / 45 [45 rows]

In this example our two tables did not share any column names in common. If they did, we would have to explicitly say which columns the tables were in using the convention `tablename.column`, and we would want to use aliases to keep things brief:

```
9 SELECT a_zctas.zcta AS zcta, a_zctas.bcode AS bcode,
   b_zctas_2010census.HD01_S001 AS pop2010
   FROM a_zctas, b_zctas_2010census
   WHERE zcta=b_zctas_2010census.geoid2 AND bcode="36061" AND pop2010 > 200
   ORDER BY pop2010
```

## 5.2 Spatial SQL

### 5.2.1 Retrieve Coordinates

We can use the geometry column in the subway features to find the northernmost station in the subway system:

```
10 SELECT stop_name, trains, bcode
   FROM a_subway_stations
   WHERE MbrMaxY(geometry) IN (
     SELECT Max(MbrMaxY(geometry))
     FROM a_subway_stations)
```

	stop_name	trains	bcode
1	Wakefield-241 St	2	36005

We can alter this statement to get the nth most point for each cardinal direction:

- MbrMaxY and Max for northernmost
- MbrMinY and Min for southernmost
- MbrMaxX and Max for easternmost
- MbrMinX and Min for westernmost

How did that work? Spatial SQL utilizes the geometry of spatial features to perform a variety of geographic operations. The geometry is stored in the geometry column as a binary object, or BLOB in database-speak. If we convert that object into human-readable form, we would recognize the data as a series of coordinates. We can extract the coordinates of subway stations from the geometry column:

```
11 SELECT stop_name, trains, bcode, ST_AsText(geometry) AS coordinates
    FROM a_subway_stations
```

	stop_name	trains	bcode	coordinates
1	Van Cortlandt Park-242 St	1	36005	POINT(1012291.155985 263271.208046)
2	238 St	1	36005	POINT(1011660.704076 261601.441833)
3	231 St	1	36005	POINT(1010566.908282 259483.047364)
4	Wakefield-241 St	2	36005	POINT(1025543.987933 268346.145574)
5	Nereid Av	2 5	36005	POINT(1024508.570441 266615.235549)

Since our features are projected in NY State Plane, the coordinates that are returned are from that system, and the XY units are represented in feet. If we wanted something more familiar, like longitude and latitude, we can transform the coordinates to a general coordinate system like NAD 83:

```
12 SELECT stop_name, trains, bcode,
    ST_AsText(Transform(geometry,4269)) AS coordinates
    FROM a_subway_stations
```

	stop_name	trains	bcode	coordinates
1	Van Cortlandt Park-242 St	1	36005	POINT(-73.898583 40.889248)
2	238 St	1	36005	POINT(-73.90087 40.884667)
3	231 St	1	36005	POINT(-73.904834 40.878856)
4	Wakefield-241 St	2	36005	POINT(-73.85062 40.903125)
5	Nereid Av	2 5	36005	POINT(-73.854376 40.898379)

4269 is the EPSG code that uniquely identifies the NAD 83 coordinate system. The database's library of coordinate systems is stored in the table spatial\_ref\_sys, which is automatically created when the database is created. The geometry in each table is tied to a specific coordinate system stored in the database. If you wanted to look up the definition there's a button on the toolbar that allows you to search the spatial reference table by name or EPSG code; alternatively you could write the statement yourself:

```
13 SELECT *
    FROM spatial_ref_sys
    WHERE srid = 4269
```

	srid	auth_name	auth_srid	ref_sys_name	proj4text
1	4269	epsg	4269	NAD83	+proj=longlat +ellps=GRS80 +datum=NAD83 +no_defs

We can also determine which coordinate systems each of our layers are in by querying the geometry\_columns table (EPSG 2263 is NY State Plane Long Island feet):



14

```
SELECT *
FROM geometry_columns
```

	f_table_name	f_geometry_column	type	coord_dimension	srid
1	a_boroughs	geometry	MULTIPOLYGON	XY	2263
2	a_metro_counties	geometry	MULTIPOLYGON	XY	2263
3	a_facilities	geometry	POLYGON	XY	2263
4	a Greenspace	geometry	MULTIPOLYGON	XY	2263

Looking up the coordinates for point features is pretty straightforward. We could do the same for lines and polygons, but the output will be much longer as the coordinates for every node will be listed; if the feature is very large or complex the database won't display the coordinates at all. If we wanted a frame of reference for polygon features, we could return bounding box coordinates instead. Bounding boxes, also known as minimum bounding rectangles (MBRs), are drawn around features to enclose them entirely and represent their minimum extent; the four sets of coordinates represent the corners of the box. This will give us the bounding box of each borough in longitude and latitude:

15

```
SELECT bname, bcode, ST_AsText(Transform(ST_Envelope(geometry),4269)) AS bbox
FROM a_boroughs
```

	bname	bcode	bbox
1	Bronx	36005	POLYGON((-73.933453 40.785787, -73.764908 40.785567, -73.76445 40.915076, -73.933324 40.915297, -73.933453 40.785787))
2	Brooklyn	36047	POLYGON((-74.042457 40.570097, -73.833564 40.569985, -73.833142 40.73895, -74.042565 40.739063, -74.042457 40.570097))
3	Manhattan	36061	POLYGON((-74.04728 40.683912, -73.907264 40.683884, -73.906992 40.879275, -74.047419 40.879302, -74.04728 40.683912))
4	Queens	36081	POLYGON((-73.962808 40.549663, -73.701124 40.54928, -73.699996 40.800638, -73.962668 40.801023, -73.962808 40.549663))
5	Staten Island	36085	POLYGON((-74.25585 40.495909, -74.051333 40.496182, -74.051451 40.649483, -74.256438 40.649209, -74.25585 40.495909))

As a shorthand reference for linear features you can retrieve the start and end point of a line using ST\_StartPoint and ST\_EndPoint on the geometry column. Unlike the previous examples, the coordinates would be retrieved in separate fields, one set for the beginning and one for the end.

### 5.2.2 Calculate distances

We can select all the subways that are within 1/2 mile of ZCTA 10010:

16

```
SELECT stop_name, trains,
ST_DISTANCE(a_zctas.geometry, a_subway_stations.geometry) AS dist
FROM a_zctas, a_subway_stations
WHERE zcta = "10010"
AND distance <= 2640
ORDER BY dist
```

	stop_name	trains	distance
1	23 St	6	0.000000
2	23 St	F M	0.000000
3	23 St	N R	0.000000
4	28 St	6	150.952836
5	28 St	N R	329.946622
6	23 St	1	899.626723

Distance represents the minimum distance, and for polygons distances are measured from the edge to the nearest point; stations located inside the polygon are selected and assigned a distance of zero. Since the coordinate reference system for our features is in feet, our input and output units are in feet (.5 miles = 2,640 feet). If we want output in

miles we can do the conversion and rounding:

```
17 SELECT stop_name, trains,
    ROUND((ST_Distance(a_zctas.geometry, a_subway_stations.geometry))/5280,2) AS dist
    FROM a_zctas, a_subway_stations
    WHERE zcta = "10010" AND dist <= 2640
    ORDER BY dist
```

	stop_name	trains	distance
1	23 St	6	0.000000
2	23 St	F M	0.000000
3	23 St	N R	0.000000
4	28 St	6	0.030000
5	28 St	N R	0.060000
6	23 St	1	0.170000

### 5.2.3 Evaluate geographic relationships

Spatialite allows you to evaluate several geographic relationships between features: Equals, Disjoint, Touches, Within, Overlaps, Crosses, Intersects, Contains, and Relate. (for a good explanation of these see [http://workshops.opengeo.org/postgis-intro/spatial\\_relationships.html](http://workshops.opengeo.org/postgis-intro/spatial_relationships.html)). For example, instead of measuring the distances of subway stations from ZCTAs, we may want to know which stations are within a ZCTA:

```
18 SELECT stop_name, trains
    FROM a_zctas, a_subway_stations
    WHERE zcta = "10010" AND ST_Within (a_subway_stations.geometry, a_zctas.geometry)
```

	stop_name	trains
1	23 St	6
2	23 St	F M
3	23 St	N R

Or which ZCTA every station is in:

```
19 SELECT zcta, stop_name, trains
    FROM a_zctas, a_subway_stations
    WHERE ST_Within (a_subway_stations.geometry, a_zctas.geometry)
    ORDER BY zcta, stop_name
```

	zcta	stop_name	trains
1	10001	28 St	1
2	10001	28 St	N R
3	10001	34 St-Herald Sq	B D F M
4	10001	34 St-Herald Sq	N Q R
5	10001	34 St-Penn Station	1 2 3
6	10001	34 St-Penn Station	A C E

Or how many stations are in each ZCTA (have to count the stations using their unique ID, as many stations have duplicate names):

```

20 SELECT zcta, COUNT(stop_id) AS stations
   FROM a_zctas, a_subway_stations
   WHERE ST_Within (a_subway_stations.geometry, a_zctas.geometry)
   GROUP BY zcta
   ORDER BY stations DESC, zcta

```

	<b>zcta</b>	<b>stations</b>
<b>1</b>	11207	13
<b>2</b>	11101	12
<b>3</b>	11201	12
<b>4</b>	11206	9

These examples evaluate point layers within other polygon layers, but you can relate any type of layer to any other, and for some operations you can evaluate features within a single layer. Let's say we want to list every ZCTA and its neighboring ZCTAs - since we're referencing features in the layer more than once we have to give each instance a distinct alias:

```

21 SELECT zcta1.zcta AS zcta, zcta2.zcta AS neighbor
   FROM a_zctas AS zcta1, a_zctas AS zcta2
   WHERE ST_Touches (zcta1.geometry, zcta2.geometry)

```

	<b>zcta</b>	<b>neighbor</b>
<b>1</b>	10001	10010
<b>2</b>	10001	10011
<b>3</b>	10001	10016
<b>4</b>	10001	10018
<b>5</b>	10001	10119
<b>6</b>	10001	10199

#### 5.2.4 Spatial Index

When determining adjacency, overlap, or any other relationship the database evaluates each feature one at a time. Feature number 1 is checked against every single feature to determine their relationship, then feature number 2 is checked against every single feature, etc. This operation doesn't take long with our ZCTA features as there are only about 200 of them, but we would NOT want to run this process on our census tracts; there are over 2,000 of those and we would be waiting a few minutes.

The census tract layer in nyc\_gdb includes a spatial index that makes this process more efficient, but utilizing the index requires a different set of SQL statements. The spatial index is an R Tree, or bounding box index; rectangles have been drawn around each tract to encompass it, and these rectangles become the search frame. Whenever a feature is evaluated, it is compared to its neighbors in the R Tree index based on bounding boxes, rather than to every single feature in the layer.

So, if we want to generate a neighbor list for all the census tracts in NYC, we can use the spatial index to speed up the process:

```

22 SELECT tract1.tractid AS tract, tract2.tractid AS neighbor
    FROM a_tracts AS tract1, a_tracts AS tract2
    WHERE ST_Touches(tract1.geometry, tract2.geometry)
    AND tract2.ROWID IN (
        SELECT pkid FROM idx_a_tracts_geometry
        WHERE pkid MATCH RTreeIntersects (MbrMinX(tract1.geometry), MbrMinY(tract1.geometry),
        MbrMaxX(tract1.geometry), MbrMaxY(tract1.geometry)))

```

Some things to note: `idx_a_tracts_geometry` is a table in the database where the index is stored, and its unique identifier is `pkid`. The `ROWID` is a unique integer ID field that SQLite automatically assigns to every record in a table. `ROWID` is "invisible" in the sense that you don't see it in the table and it's not listed in the metadata as a column, but it exists and you can call it in a SQL statement. The spatial index relies on the `ROWID`, so we have to call it for this operation. As we've seen before, `Mbr` is the minimum bounding rectangle, and we look for the minimum and maximum values for both X and Y coordinates.

The census tracts are the only layer in this database with a spatial index; since the other layers have a small number of features an index was not created. If you want to use the tracts in a spatial comparison with another layer and want an index for that 2nd layer, you can create one:

```

23 SELECT CreateSpatialIndex("layer_name", "geometry")

```

### 5.2.5 Calculate area and density

We can calculate areas for all the ZCTAs; the default will be in square feet since our CRS is in feet, but we can do the extra work to get square miles:

```

24 SELECT zcta, ST_Area(geometry) AS sqft,
    (ST_Area(geometry))*0.0000000358700643 AS sqmi
    FROM a_zctas
    ORDER BY sqmi DESC

```

	zcta	sqft	sqmi
1	10314	386100312.702513	13.849443
2	10312	216029373.971924	7.748988
3	11234	210358174.577227	7.545561
4	10306	208988895.334777	7.496445

If we join the ZCTAs to our census table we can calculate population density; we'll modify our query to select just residential ZCTAs based on population:

```

25 SELECT zcta, HD01_S001 AS pop2010,
    HD01_S001 / ((ST_Area(geometry))*0.0000000358700643) AS density
    FROM a_zctas, b_zctas_2010census
    WHERE zcta=geoid2 AND pop2010 > 200
    ORDER BY density DESC

```

	zcta	bcode	pop2010	density
1	10162	36061	1685	148232.400385
2	10028	36061	45141	143685.635309
3	10075	36061	26121	141548.203054
4	10128	36061	60453	128442.218728

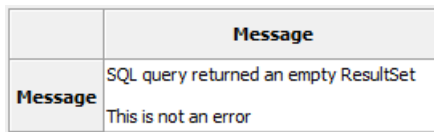
In addition to area you can also calculate perimeter (ST\_Perimeter) for polygons and length (ST\_Length) for linear features.

## 5.3 Extending the DB

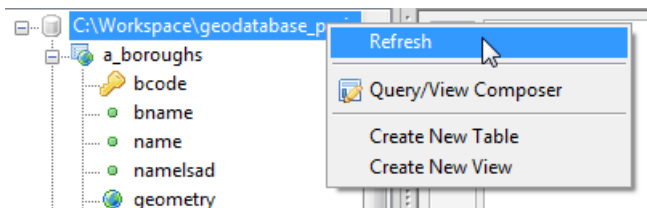
### 5.3.1 Create new view

As these queries can get quite involved you may want to save some of them to re-use in the future. When you're in Spatialite you can use the arrow keys to the left of the SQL statement box to cycle through queries you've created, but once you close the software your statements will not be saved. To save them, you have to create a view:

```
26 CREATE VIEW d_zctas_2010pop AS
    SELECT zcta, bcode, HD01_S001 AS pop2010
    FROM a_zctas, b_zctas_2010census
    WHERE zcta=geoid2 AND pop2010 > 200
    ORDER BY pop2010
```



Whenever you run a CREATE, INSERT, or ALTER statement it's normal to get the Empty Results Set message (as you haven't done any selection). After you've run the statement, you have to refresh the database in order to see the view. Select the name of the database at the top of your tree in the menu on the left, right click, and choose refresh.



Scroll down and you can see your view listed in the menu; we've named it with the prefix "d" so that it doesn't get mixed up with the other objects in our database. A view saves your SQL statement - not the data that is shown as a result of executing it. So whenever you open a view you're executing a saved SQL statement.

### 5.3.2 Create spatial view

If you save a query as a spatial view you'll be able to view the results and map the data in QGIS. This requires two steps. First you write the CREATE VIEW statement. In this statement you must include the geometry column and the invisible ROWID unique integer column (automatically created for every SQLite table; it's hidden but call-able) for the geographic feature.

```
27 CREATE VIEW d_zctas_spatial AS
    SELECT a_zctas.ROWID, zcta, bcode, HD01_S001 AS pop2010, geometry
    FROM a_zctas, b_zctas_2010census
    WHERE zcta=geoid2 AND pop2010 > 200
    ORDER BY pop2010
```

Second, you need to register the view in a special table in the database that contains geometry for the views:

```
28 INSERT INTO views_geometry_columns  
  (view_name, view_geometry, view_rowid, f_table_name, f_geometry_column)  
  VALUES ("d_zctas_spatial", "geometry", "ROWID", "a_zctas", "geometry")
```

Refresh the database and you should see the view in the tree. Now if you launch QGIS and connect to the database, `d_zctas_spatial` will be listed as a geographic feature. You can add it to a project and symbolize the data just like you would any other layer.

If you'd rather not go through all the steps of creating views, spatial or non-spatial, Spatialite has a menu driven view builder that you can use instead. If you select the database at the top of the tree, then right click, you can launch the Query/View composer.

**LIMITATIONS WITH SPATIAL VIEWS:** At this stage you cannot create map-able spatial views that have calculated fields, like area or density from our earlier example. In the view, calculated fields do not have a data type associated with them; QGIS is able to read the columns but can't evaluate what they are, and thus they can't be classified or symbolized. Using the SQL CAST operator to assign them a type has no effect. Also, you cannot create spatial views using aggregated queries that use the GROUP BY clause. If you create calculated fields or aggregates that you wish to map, you will have to create a brand new table with that data.

### 5.3.3 Create table from query

In many implementations of SQL it's possible to write a CREATE TABLE AS statement, where a SELECT statement with a query follows the CREATE statement and the result is written to a new table. While you can do this in SQLite it is NOT always recommended, because you will not be able to specify a primary key or data types for columns (thus, any calculated fields would not have a specified type).

If you have a query or view that you would like to turn into a table, or you have a table where you want to drop or rename columns or change data types (as noted earlier, SQLite does not allow this as part of the ALTER TABLE statement), the following are the standard SQLite steps that you need to take:

1. Create a new, blank table where you specify your column names, their data types, a primary key, and any constraints that you wish to add.
2. Add a geometry column for spatial tables.
3. Insert the data from a query, table, or view into the new table.
4. Drop the old table or view (only if it's redundant or no longer needed).

Let's say that we want population density for ZCTAs. We need to turn the query into a permanent table so we can map the data in QGIS. First, we'll create the blank table that will hold our data; for data types the `zcta` is TEXT (since it's an identifier that may begin with a zero, and not a quantity), `pop2010` is an INTEGER (whole numbers) and `density` is a REAL (number with decimals):

```
29 CREATE TABLE d_zctas_density (  
  zcta TEXT NOT NULL PRIMARY KEY,  
  bcode TEXT,  
  pop2010 INTEGER,  
  density REAL)
```

Then we add a geometry column. We need to specify the type of geographic features as well as the coordinate system. POINT, MULTIPOINT, LINestring, MULTILINESTRING, POLYGON, and MULTIPOLYGON are the six primary options. Since one ZCTA feature may consist of several separate polygons, it's declared as a MULTIPOLYGON. The code for our CRS is 2263 (NAD 83 New York Long Island State Plane feet).

30 `SELECT AddGeometryColumn ( "d_zctas_density", "geometry", 2263, "MULTIPOLYGON", "XY")`

Refresh the database to see the table in the tree. Now we can insert the data into our table; we'll nest a subquery into our INSERT query to select the data we need. This makes our statement look a little busy; if we were inserting pre-existing data from a table or view it would look more straightforward.

31 `INSERT INTO d_zctas_density (zcta, bcode, pop2010, density, geometry)  
SELECT zcta, bcode, HD01_S001, HD01_S001/((ST_Area(geometry))*0.000000358700643),  
geometry  
FROM a_zctas, b_zctas_2010census  
WHERE zcta=geoid2 AND HD01_S001 > 200`

Refresh the database to see the new spatially-enabled table, d\_zctas\_density in the list. View all the data:

32 `SELECT *  
FROM d_zctas_density`

	zcta	bcode	pop2010	density	geometry
1	10001	36061	21102	33959.490640	BLOB sz=1381 GEOMETRY
2	10002	36061	81410	92573.511829	BLOB sz=1741 GEOMETRY
3	10003	36061	56024	97188.806436	BLOB sz=1245 GEOMETRY
4	10004	36061	3089	5518.808120	BLOB sz=6307 GEOMETRY

And view the metadata:

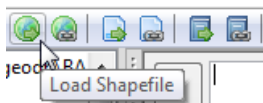
33 `PRAGMA table_info (d_zctas_density)`

	cid	name	type	notnull	dflt_value	pk
1	0	zcta	TEXT	1	NULL	1
2	1	bcode	TEXT	0	NULL	0
3	2	pop2010	INTEGER	0	NULL	0
4	3	density	REAL	0	NULL	0
5	4	geometry	MULTIPOLYGON	0	NULL	0

This feature can be added to a QGIS project and symbolized like any other layer.

### 5.3.4 Import Data

You can load shapefiles, delimited text files, and dbf files into the database via the Spatialite GUI. For each format you have a choice of loading the file or creating a virtual table; the latter creates a link to the external file without actually loading it into the database.



When you load a shapefile you must specify what coordinate system it uses in the SRID input box; change the default from -1 to 2263, the EPSG code for NAD83 New York Long Island (ftUS). Before you load a shapefile you must prep it by transforming its CRS to that projection, so it will match the database layers (you can use the Transform function covered in the final section, or you can use any GIS software to reproject files). The geometry of the shapefile

will be loaded into a default column named "Geometry".

When you load delimited text files you'll be asked to specify the delimiter (tab, comma, etc.) and whether or not text fields are surrounded by quotes (a common convention for preserving text values). If you have numeric values that need to be saved as text (like ZIP Codes or FIPS codes) you will need to prep your text data so that these columns are enclosed in quotes. Microsoft Excel is poor for working with text or csv files; it automatically saves anything that looks like a number as a number and doesn't allow you to quote text fields when saving a file as delimited text. Consider using either a text editor or Calc, the spreadsheet that's part of the LibreOffice suite <http://www.libreoffice.org/>. Calc does a better job with text files and also supports the dbf format (which Excel does not).

When you import data Spatialite does not give you the option to include or exclude columns, rename them, specify data types, designate primary or foreign keys, or enforce constraints. It imports all columns as is and makes assumptions about how to cast the data. If you want to insure that you have a well formed database you'll have to follow the steps in the previous section: import the data, create an empty table to hold the values you want, load the values into that table, and drop the imported data table. After you import your data, these are the generic steps:

```
34 CREATE TABLE newtable (  
    newid TEXT NOT NULL PRIMARY KEY,  
    otherid TEXT,  
    value1 INTEGER,  
    value2 REAL)  
  
SELECT AddGeometryColumn ( "newtable", "geometry", 2263, "TYPE OF GEOMETRY", "XY")  
  
INSERT INTO newtable (newid, otherid, value1, value2, geometry)  
SELECT shapeid, label, popvar, housevar, geometry  
FROM shapefile  
  
DROP shapefile
```

### 5.3.5 Create new features from existing ones

You can use Spatialite to do some geoprocessing. For example, we can merge features that share a common attribute. If you look at the census tract features, you'll see that each census tract is affiliated with an area called a neighborhood tabulation area or NTA; a statistical area created by the City of New York for presenting census data for neighborhoods.

```
35 SELECT * FROM a_tracts
```

	tractid	tractnum	name	namelsad	nta	nta_name
1	36005000100	000100	1	Census Tract 1	BX98	Rikers Island
2	36005000200	000200	2	Census Tract 2	BX09	Soundview-Castle Hill-Clason Point-Harding Park
3	36005000400	000400	4	Census Tract 4	BX09	Soundview-Castle Hill-Clason Point-Harding Park
4	36005001600	001600	16	Census Tract 16	BX09	Soundview-Castle Hill-Clason Point-Harding Park
5	36005001900	001900	19	Census Tract 19	BX39	Mott Haven-Port Morris

Let's say we want to create a brand new layer for the neighborhood areas; we can do that by merging the census tracts that share a common nta code. To create the new layer, we follow the same steps as before - first we create a new table that will hold the features:



```
36 CREATE TABLE a_ntas (
    nta TEXT NOT NULL PRIMARY KEY,
    nta_name TEXT,
    bcode TEXT)
```

Second - we add a column to the new table to hold the geometry:

```
37 SELECT AddGeometryColumn ( "a_ntas", "geometry", 2263, "MULTIPOLYGON", "XY")
```

Third - copy the data from the tracts table into the new one, and while doing so merge the features:

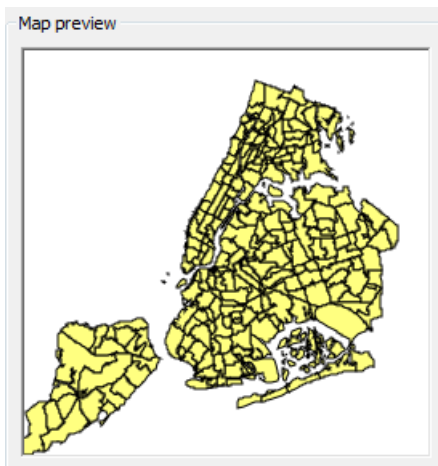
```
38 INSERT INTO a_ntas (nta, nta_name, bcode, geometry)
    SELECT nta, nta_name, bcode, ST_Union(geometry)
    FROM a_tracts
    GROUP BY nta
```

Union is the command that does the merging, and Group By specifies which column should be used for dissolving the features. Hit the Refresh button. Query the table - you should have a table with 195 neighborhood records:

```
39 SELECT * FROM a_ntas
```

	nta	nta_name	bcode	geometry
1	BK09	Brooklyn Heights-Cobble Hill	36047	BLOB sz=1229 GEOMETRY
2	BK17	Sheepshead Bay-Gerritsen Beach-Manhattan Beach	36047	BLOB sz=8698 GEOMETRY
3	BK19	Brighton Beach	36047	BLOB sz=1805 GEOMETRY

You can now add this layer to QGIS just like any other feature. You can also preview it in Spatialite - in the table of contents on the left select a\_ntas, expand the columns list, select the geometry column, right click, choose Map Preview.



### 5.3.6 Importing Coordinate Data and Transforming Projections

In this section we'll cover several operations: importing a data table, converting data stored as XY coordinates into geometry, transforming the coordinate system of a layer, adding columns, and updating data. We'll import the population centers of New York's counties into the database. Open your computer's text editor (i.e. Notepad on MS Windows), copy the data below and paste it into the editor. \*\*Make sure to hit the enter key at the end of the last

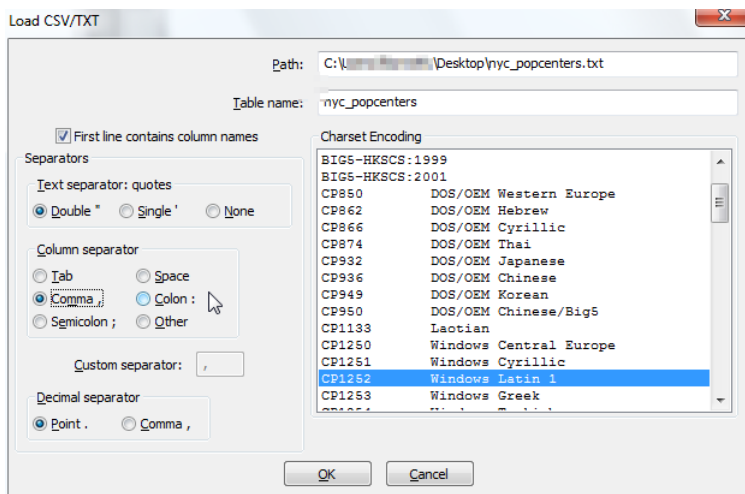
record (Richmond), so it will be recognized as a record. Save the file as nyc\_popcenters.txt.

STATEFP,COUNTYFP,COUNAME,STNAME,POPULATION,LATITUDE,LONGITUDE

```
"36","005",Bronx,New York,1385108,+40.849354,-073.882363
"36","047",Kings,New York,2504700,+40.650523,-073.954811
"36","061",New York,New York,1585873,+40.777958,-073.966303
"36","081",Queens,New York,2230722,+40.721294,-073.833451
"36","085",Richmond,New York,468730,+40.588523,-074.137304
```

(Given the small amount of data, we could have created a blank table and used a SQL INSERT statement to populate it. But since it's more likely that you'll be importing data rather than creating it manually, we'll use this example).

First we'll load the text file into the database - hit the load CSV / TXT file button on the toolbar. Browse to the location where you saved the file and select it. In the menu check the box that says First line contains column names, select the Comma option under Column separator, and verify that the Text separator: quotes option is set to Double. Hit OK.



Query the data to make sure it loaded properly.

```
40 SELECT * FROM nyc_popcenters
```

	PK_UID	STATEFP	COUNTYFP	COUNAME	STNAME	POPULATION	LATITUDE	LONGITUDE
1	0	36	005	Bronx	New York	1385108	40.849354	-73.882363
2	1	36	047	Kings	New York	2504700	40.650523	-73.954811
3	2	36	061	New York	New York	1585873	40.777958	-73.966303
4	3	36	081	Queens	New York	2230722	40.721294	-73.833451
5	4	36	085	Richmond	New York	468730	40.588523	-74.137304

Next we'll create the table that will hold our data. We'll have just one column to hold our id code and we'll drop the state name as it's unnecessary.

```

41 CREATE TABLE a_borough_popcenters (
    bcode TEXT NOT NULL PRIMARY KEY,
    cname TEXT,
    pop2010 INTEGER,
    latitude REAL,
    longitude REAL)

```

Next we create a geometry column for the table.

```

42 SELECT AddGeometryColumn ("a_borough_popcenters", "geometry", 2263, "POINT", "XY")

```

Then we'll insert the data from the original table into our new one.

```

43 INSERT INTO a_borough_popcenters (bcode, cname, pop2010, latitude, longitude,
    geometry)
    SELECT (statefp || countyfp), couname, population, latitude, longitude,
    Transform(MakePoint(longitude, latitude, 4269),2263)
    FROM nyc_popcenters

```

There are a few things going on here. First, we're using || to concatenate the state and county codes into a single code. Second, we're taking the longitude and latitude coordinates (longitude is always X, latitude is Y) and using MakePoint to create geometry out of them; we have to define the geometry as NAD 83 (EPSG code 4269) because that's the format the coordinates are in (most lat / long data from the US government is in NAD 83). Lastly, we're taking that new geometry and transforming it to NY Long Island feet (EPSG 2263) because that's the coordinate system that the rest of our data is in, and we want to be able to layer this data over our other layers.

In this case we nested the MakePoint command inside the Transform command. Ordinarily, if we were reprojecting layers that have geometry, we'd create a new table, add a geometry column with the system we want, then copy the geometry from the existing table into the new one while transforming it: Transform (geometry, EPSG code).

Refresh the database, then query the data to make sure it looks OK.

```

44 SELECT * FROM a_borough_popcenters

```

	bcode	cname	pop2010	latitude	longitude	geometry
1	36005	Bronx	1385108	40.849354	-73.882363	BLOB sz=60 GEOMETRY
2	36047	Kings	2504700	40.650523	-73.954811	BLOB sz=60 GEOMETRY
3	36061	New York	1585873	40.777958	-73.966303	BLOB sz=60 GEOMETRY
4	36081	Queens	2230722	40.721294	-73.833451	BLOB sz=60 GEOMETRY
5	36085	Richmond	468730	40.588523	-74.137304	BLOB sz=60 GEOMETRY

Delete the original table since we don't need it any longer, then refresh the database.

```

45 DROP TABLE nyc_popcenters

```

Our new layer has county names but no borough names. We could have created a new column in the MAKE TABLE statement to hold the names, but we can add additional columns after the fact:

```

46 ALTER TABLE a_borough_popcenters ADD COLUMN bname TEXT

```

Now we can add the borough names. We can simply copy the names for the Bronx and Queens from the county

column, since the names are identical; for the other three boroughs we have to change them manually. You could use the GUI interface to modify the tables, or use SQL:

```

47 UPDATE a_borough_popcenters
   SET bname=cname WHERE bcode IN ("36005","36081")
48 UPDATE a_borough_popcenters
   SET bname="Brooklyn" WHERE bcode="36047"
49 UPDATE a_borough_popcenters
   SET bname="Manhattan" WHERE bcode="36061"
50 UPDATE a_borough_popcenters
   SET bname="Staten Island" WHERE bcode="36085"
51 SELECT * FROM a_borough_popcenters

```

	bcode	cname	pop2010	latitude	longitude	geometry	bname
1	36005	Bronx	1385108	40.849354	-73.882363	BLOB sz=60 GEOMETRY	Bronx
2	36047	Kings	2504700	40.650523	-73.954811	BLOB sz=60 GEOMETRY	Brooklyn
3	36061	New York	1585873	40.777958	-73.966303	BLOB sz=60 GEOMETRY	Manhattan
4	36081	Queens	2230722	40.721294	-73.833451	BLOB sz=60 GEOMETRY	Queens
5	36085	Richmond	468730	40.588523	-74.137304	BLOB sz=60 GEOMETRY	Staten Island

You can add this layer to QGIS and overlay it with the boroughs to see where the population centers are; they happen to be close to where the geographic centers are, indicating that population is rather evenly distributed in each borough. Staten Island is an exception; the population is clustered a little more to the north of the island. You can use the ST\_Centroid function to calculate the geographic center of geometries; this is useful when you want to measure the distance from a point to the center of a polygon rather than to the edge (the latter is the default for ST\_Distance). Of course, when measuring distances from services to populations in a given area a population centroid is preferable to a geographic one.

## 5.4 Reference & Links

### 5.4.1 Software

Binaries are pre-compiled versions of the software that are ready to go - just download and execute them. Binaries for both the GUI and the command line are available. This tutorial used binaries from version 2.4.0 RC-4 available at <http://www.gaia-gis.it/spatialite-2.4.0-4/binaries.html> because it was the last time binaries were compiled for Mac, Linux, and MS Windows, and because the Spatialite Cookbook (the best reference for Spatialite) was written using this version. The nyc\_gdb was created using 2.4.0, but can be viewed and manipulated in subsequent versions.

Newer binaries are only available for MS Windows. For Mac and Linux the only way to get newer versions is to download and compile the software from the source (this can also be done for MS Windows if you don't want to use a binary). Links for MS Windows binaries and for the source files for the command line (libspatialite) and the spatialite gui are available directly from the homepage at <http://www.gaia-gis.it/gaia-sins/>. You click the sources link and download the zip or tar file with the most recent version number, then follow the instructions for compiling and installing.

The latest stand-alone version of QGIS (1.8 Lisboa) utilizes Spatialite 3.0.1. The differences between Spatialite versions are often not earth-shattering, but you'll want to be aware of them as some functions and commands (particularly for spatial indexes) have changed. See the Spatialite homepage and documentation for details.

#### 5.4.2 SQLite

- Official reference for all SQLite functions: <http://www.sqlite.org/lang.html>
- Good SQLite tutorial: <http://zetcode.com/db/sqlite/>
- Good SQLite & Python tutorial: <http://zetcode.com/db/sqlitepythontutorial/>
- Good book: *Using SQLite* by Jay Kreibich, O'Reilly Media 2010.

#### 5.4.3 Reference for Spatialite Functions

- Spatialite 2.4 (used in this manual): <http://www.gaia-gis.it/spatialite-2.4.0/spatialite-sql-2.4.html>
- Spatialite 3.0 (used in QGIS 1.8): <http://www.gaia-gis.it/gaia-sins/spatialite-sql-3.0.0.html>
- Spatialite 4.0 (most recent): <http://www.gaia-gis.it/gaia-sins/spatialite-sql-4.0.0.html>

#### 5.4.4 Spatialite Concepts and Reference

- Spatialite Cookbook: <http://www.gaia-gis.it/gaia-sins/spatialite-cookbook/index.html>
- Other Spatialite tutorials and docs: <https://www.gaia-gis.it/fossil/libspatialite/wiki?name=misc-docs>
- Scratching Surfaces (blog with good examples): <http://www.surfaces.co.il/?cat=12>
- Spatial Relationships (for PostGIS but largely applies): [http://workshops.opengeo.org/postgis-intro/spatial\\_relationships.html](http://workshops.opengeo.org/postgis-intro/spatial_relationships.html)
- Spatial Geometries (for PostGIS but largely applies): <http://workshops.opengeo.org/postgis-intro/geometries.html>

#### 5.4.5 Forums for Help

- GIS Stack Exchange: <http://gis.stackexchange.com/>
- stackoverflow (for SQLite, SQL, or general database questions): <http://stackoverflow.com/>

#### 5.4.6 Common CRS for NYC Area

- EPSG 2263 (NAD83 / New York Long Island (ft US)): used in the `nyc_gdb` and by most city agencies
- EPSG 26918 (NAD83 / UTM Zone 18 N): another projected coordinate system that's suitable for the region (units are in meters)
- EPSG 4269 (NAD83): basic latitude / longitude system used by US federal government agencies like the Census Bureau
- EPSG 4267 (NAD27): older latitude / longitude system commonly used by US federal government agencies up to the mid 1990s
- EPSG 4326 (WGS84): basic latitude / longitude system common throughout the world; almost identical to NAD83