Introduction to GIS Using Open Source Software

Frank Donnelly, Geospatial Data Librarian, Baruch College CUNY [1]

August 2013

Contents

Introduction

Frank Donnelly, Geospatial Data Librarian, Baruch College CUNY
francis.donnelly@baruch.cuny.edu

Last Updated: August 30th, 2013 (4th ed.)

Introduction

This tutorial was created to accompany the GIS Practicum, a day-long workshop offered by the Newman Library at Baruch College CUNY that introduces participants to geographic information systems (GIS) using the open source software QGIS. The practicum introduces GIS as a concept for envisioning information and as a tool for conducting geographic analyses and creating maps. Participants learn how to navigate a GIS interface, how to prepare layers and conduct a basic geographic analysis, and how to create thematic maps.

This tutorial was written using QGIS version 1.8 "Lisboa", a cross-platform (Windows, Mac, Linux) desktop GIS software package. Salient changes from previous versions (1.7) are noted in the text. You can download the software and user manual from the QGIS website at `http://www.qgis.org/`; given differences between versions it's best to use this tutorial with version 1.8 Lisboa. Quick links for downloading both 1.8 and the data used for the tutorial are available at `http://www.baruch.cuny.edu/geoportal/practicum/`. Once you download and unzip the data file, you'll see that the data is separated into different folders for each part of the tutorial.

Objectives

Participants will be able to bring both the tools and the knowledge they gain from this workshop to enhance their projects and the organizations they work for. Specifically, this workshop will enable participants to:

- Add data to GIS software and navigate a GIS interface
- Perform basic geoprocessing operations for preparing vector GIS data

- Convert text-based data to a GIS data format
- Conduct geographic analyses using standard GIS tools and vector data
- Create thematic maps using the principles of map projections, data classification, symbolization, and cartographic design
- Locate GIS data on the web and consider the merits of different data sources
- Demonstrate competency with a specific GIS package (open source QGIS)
- Identify other GIS topics (tools and techniques for analysis), data formats (raster, vector), and software (open source and ArcGIS) to pursue for future study

Outline

- Chapter 1: General introduction and overview of GIS
- Chapter 2: Introduction to GIS Interface (learn how to navigate the interface: adding data, layering data, symbolization, changing zoom, viewing attributes, viewing attribute table, making basic selections, difference between data formats, organizing projects and data)
- Chapter 3: GIS Analysis (using site selection example in NYC, basic geoprocessing tasks, attribute table joins, plotting coordinate data, buffers, basic statistics, advanced selection)
- Chapter 4: Thematic mapping (using US states as an example, map projections, coordinate systems, data classification, symbolization, calculated fields, labeling, map layouts)
- Chapter 5: Going Further with GIS (exploring and evaluating online sources for free data, exploring open source and ArcGIS software resources for learning more)

Organization of this Tutorial

This document is divided into five chapters and subdivided into sections for specific tasks. Each section begins with steps for learning a specific application or process (the what and when), followed by commentary that explains various facets of the process (the how and why). The process and the commentary were separated in order to keep the steps as concise and easy to follow as possible with few digressions; you follow the steps first, and then go back and understand the details of why you followed the steps you did. This tutorial and associated screenshots were created using QGIS in a Windows operating system. The names of certain tools and menus may vary slightly between operating systems, but functionality should be the same.

The following conventions are used throughout:

- Each section begins with steps for learning a specific application or process, followed by commentary that explains various facets of the process.
- Steps are enumerated and begin with *italicized text*.
- A  stop sign appears at the conclusion of a series of steps, to clearly delineate where the steps end and the commentary for that section begins.
- Menus, tabs, and items are capitalized if they appear capitalized in the interface.
- Images of  toolbar buttons appear in the text whenever they are referenced.

- The names of files and layers appear in SMALL CAPS TEXT.
- Urls for websites appear in `typewriter text`.

Changes From Previous Manual

This manual (4th edition) has been updated from the previous manual (3rd edition) primarily by the replacement of the data and examples in Parts 2 and 3 with new examples. Both editions use QGIS 1.8 Lisboa.

Specific changes:

- Part 1 - No major revisions.
- Part 2 - Replacement of all data files with new data based on the 2010 Census TIGER files, substitution of colleges layer with a subway station layer.
- Part 3 - Retained the example of a site selection problem, but replaced comic book stores with coffee shops, PUMAs with ZCTAs, and ACS data with a mix of 2010 Census and ACS data. Rewrote commentary about geographic units and TIGER files in 3.2.2, census data in 3.3.2, and coordinate data in 3.4.2, and site selection in 3.6.2 Added commentary on using spreadsheets for tabular data in 3.3.2.
- Part 4 - Some revisions to commentary on coordinate reference systems in 4.1.2, dropped singlepart and multipart discussion in 4.2.2.
- Part 5 - Updated some broken links.
- Appendix - No major revisions.

Chapter 1

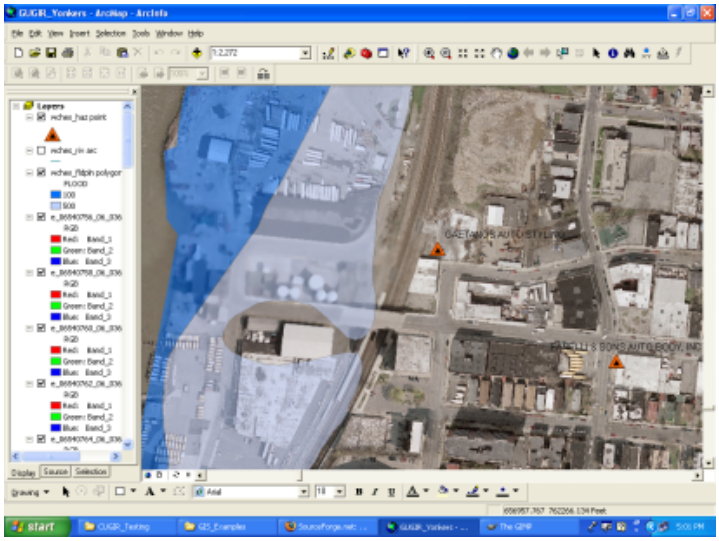
An Overview of GIS

The goal of this chapter is to provide you with a basic foundation in GIS concepts and software in preparation for the rest of the tutorial.

1.1 Basic GIS Concepts

Geographic Information Systems (GIS) are an integrated collection of software and data used to visualize and organize geographic data, conduct spatial analysis, and create maps and other geospatial information. Narrow definitions of GIS focus on the software and data, while broader definitions include hardware (where the data and software is stored), metadata (data that describes the data), and the people who are part of the system and interact with it as creators, curators, and users.

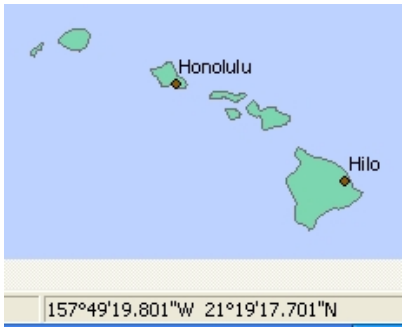
Another definition: GIS is a visual system that organizes information around the concepts of place and location that can be used for geographic analysis, map making, database management, and geospatial statistics. GIS can be applied to virtually any discipline or endeavor.



In a GIS, geographic features are represented as individual files or layers that can be added to a map. These features are not maps in and of themselves, but are the raw materials used for map making and analysis. For much of the 20th century cartographers drew geographic features on individual mylar or acetate sheets and layered those

sheets over a paper base map to create maps. GIS uses the same principles of layering, with individual files consisting of features that can be layered on top of each other in GIS software. GIS software acts as an interface, or window, for viewing and manipulating GIS data. The ability to add different layers is quite powerful, as combining the layers allows for analysis that would be impossible if you were viewing single layers by themselves (see example above of air photo, flood zones, and hazardous sites overlaid).

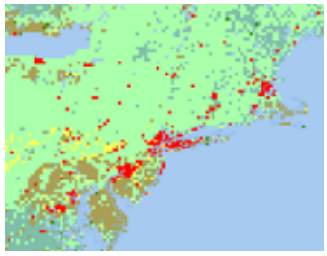
Each GIS file is georeferenced, meaning that the file is actually tied and related to real locations on the earth. If we mouse over Hawaii in GIS software like in the image below, we see its latitude and longitude coordinates. Just as paper maps were drawn based on map projections and coordinate systems, each GIS file has also been created based on a particular projection and coordinate system, which means that files that share the same reference systems can be overlaid. Since projections and coordinate systems are highly standardized, GIS data can easily be shared. If two files do not share the same system, most GIS software can convert files from one system to another so they'll match. This distinguishes map making in GIS versus a graphic design package. Maps created in a graphic design package are just simple lines and shapes with no connection to the earth, and the components of the map can't be easily replicated to make other maps. GIS files used to create maps in a GIS package can readily be shared and used to create any map, because they are tied to the earth using standardized systems.



GIS files are stored in several formats, and each format comes in several different file types. Major formats and files include:

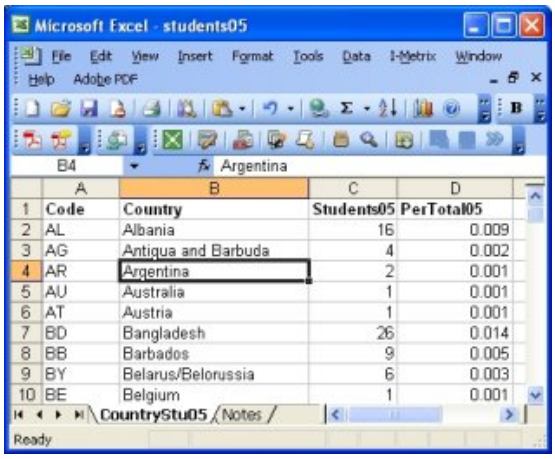
- Raster - a continuous surface that is divided into grid cells of equal size. Each cell appears as a particular color based on some value (i.e. reflected light). Files in the raster format are similar to digital photos. Common raster objects include air photos, satellite imagery, and paper maps that have been scanned. Raster files can also consist of photos or imagery that have been generalized or have had value added to them to create a new layer, like a land use and land cover layer or a grid showing temperature. There are many different file formats, some common ones include Tiffs (.tif), JPEGs (.jpg), and SID (.sid). Unlike regular .tif or .jpg files, GIS raster files are georeferenced.





- Vector - discrete coordinates and surfaces that are represented as individual points, lines, or polygons (areas). Vector files appear to be more "map-like", and are always abstractions rather than actual images (i.e. shapes to represent boundaries, points to represent cities). Common file formats include ESRI shapefiles (.shp), ESRI coverages (.cov), Google KML files (.kml), and Geographic Markup Language files (GML).

- Tables - data tables that contain records for places can be converted to GIS files and mapped in several ways. If the data contains coordinates like latitude and longitude, the data can be plotted and converted to a vector file. If each data record contains unique ID codes for each place, those records can be joined to their corresponding features in a GIS file and mapped. Tables are commonly stored in text files like .txt or .csv, database files like .dbf, or in spreadsheets like Excel.



	A	B	C	D
1	Code	Country	Students05	PerTotal05
2	AL	Albania	16	0.009
3	AG	Antigua and Barbuda	4	0.002
4	AR	Argentina	2	0.001
5	AU	Australia	1	0.001
6	AT	Austria	1	0.001
7	BD	Bangladesh	26	0.014
8	BB	Barbados	9	0.005
9	BY	Belarus/Belorussia	6	0.003
10	BE	Belgium	1	0.001

- Geodatabases - containers that can hold related raster, vector, and tabular data in one place. They are good for consolidating and organizing data, and many can be used for spatial queries and analysis. Geodatabases can be desktop (Microsoft Access .mdb, ESRI file geodatabases .gdb, Spatialite files .sqlite) or server based (PostGIS, ArcSDE).

Raster and vector GIS files exist spatially, in that you can see the grid or shapes and their corresponding location on the earth, but also exist in tabular form. This is particularly valuable in the case of vector files. For example, every feature in a vector file showing state boundaries has an attribute table attached to it that has a record for each state. This attribute table contains columns or fields that store values for each state, such as the state's name, values like population or area that describe it, and ID codes that uniquely identify each one. The names can be used by the GIS to label each state, and the values like population can be thematically mapped.

The ID codes for each state can be used to join the attribute table for the GIS file to a tabular file that contains state-level data. For example, a GIS file of state boundaries with a state code can be joined within GIS using relational database techniques to a text or spreadsheet file that has state-level data and that uses the same codes to identify each state. The data in the table, which was just a regular table with no geospatial geometry, can now be visualized and mapped in GIS. There are number of standard ID codes that can be used for joining data. The two most common families of codes are ANSI / FIPS (created by the US government to identify every single geographic entity in the US; there are also FIPS codes for countries) and ISO (created by the International Standards Organization to identify countries and their subdivisions).

Attribute table - tl_2012_us_state :: 0 / 56 feature(s) selected

	STATEFP	STATENS	GEOID	STUSPS	NAME
0	01	01779775	01	AL	Alabama
1	02	01785533	02	AK	Alaska
2	04	01779777	04	AZ	Arizona
3	05	00068085	05	AR	Arkansas

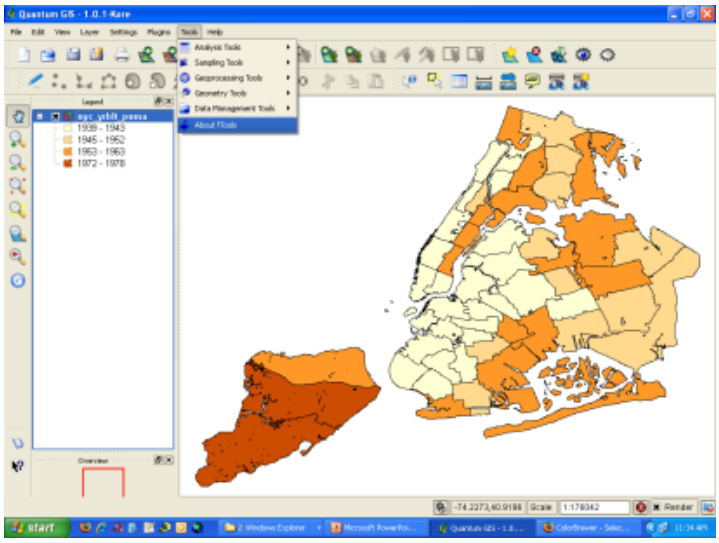
Attribute table - final_demographic :: 0 / 51 feature(s) selected

	GEO_ID2	SUMLEVEL	GEO_NAME	TOTPOP	TOTPOP_ME
0	01	040	Alabama	4633360	NULL
1	02	040	Alaska	683142	NULL
2	04	040	Arizona	6324865	NULL
3	05	040	Arkansas	2838143	NULL

1.2 GIS Software

A standard interface for GIS software has evolved over time. Typically, GIS software has a data view that consists of a table of contents that lists files that have been added to a project, a data window that displays the GIS files, and a set of toolbars and menus for accessing various tools and launching various processes. Dragging the layers in the table of contents changes their drawing order, and right or left clicking on a layer in the table of contents will reveal individual properties for that particular feature. You can also access the attribute table of the feature and a symbol tab for changing how the features are depicted or classified. There are several tools for zooming in and out

to examine different layers and to change the extent of the view.



The way that coordinate systems and projections are handled is different for individual GIS software packages. In general, the options are: define the projection and coordinate system for the project before adding the files, or the project automatically takes the projection of the first file added. If you try to add GIS files that have different projections, some software may try to re-project the data on the fly, while others will simply fail to draw the new layers. Even if the software can correctly draw a layer without the user defining it, or even if it can re-project layers on the fly, users will run into problems later on when trying to manipulate the GIS files. You should always be sure to specify the projection properly and make sure that all files share the same one - most GIS software will give you the ability to re-project data.

GIS software provides users with a variety of ways for querying geographic data, either by selecting records in the attribute table or shapes in the view, or by conducting searches where you build queries to high-light features that contain specific attributes, or that have some relationship with another geographic layer.

GIS software comes with a variety of editing tools that allow you to modify the geometry of GIS files. For example, you can merge features together, break them apart, or clip out or select certain areas to create new files. Collectively these processes are known as geoprocessing. You geoprocess layers in order to prepare raw data for analysis, to create new layers or data, or to simplify layers for cartographic or aesthetic purposes. GIS also provides the ability to edit files on a feature by feature basis.

Most GIS programs have a separate map layout or print layout, where the user can create finished maps with standard map elements like titles, legends, scale bars, north arrows, and accompanying text. Finished maps can be exported out of the GIS as static files, such as pdfs or jpgs.

Users can always save their GIS projects in a GIS project file. The scale and extent of the data view, symbolization and classification assigned to layers, map layouts, and links to GIS files used in the project are stored in the file. It's important to understand that the GIS files themselves are NOT stored inside the project file - the GIS data and the GIS project file exist independently. When adding data to a GIS, you are establishing a link from the GIS project to the GIS data - the GIS data is not stored within the project. Furthermore, changing the colors of the features or classifying them in a certain way has no effect on the actual GIS data files themselves. When you change symbols, you are only changing how the GIS program views the data - you're not changing the data itself.

This is an important concept to grasp. Essentially, the GIS software acts as a window for viewing and working with GIS data, which is stored outside the window. The GIS project file essentially stores the window dressing, of scale and symbolization. You never actually change the GIS data unless you go into an edit mode or conduct an operation that creates a new GIS file. This relationship is of crucial importance when it comes time to move or share files - if you move your project file or your data, the links between them will become broken, and you'll need to re-establish the location between the project and the data in order to repair your project file.

1.3 Open Source

In this tutorial we will be using QGIS, which is free open source software (FOSS). Open source software is an alternative to proprietary software:

- Open source software is free; you don't have to purchase it and you can freely distribute it to anyone else, as opposed to proprietary software which you must purchase and typically can not share with anyone (since it's copyrighted).
- The source code, or actual computer programming, that was used to create the software is transparent, as opposed to proprietary software where the code is hidden and encrypted.
- Under the open source model the programming code is transparent and you are free to change and make improvements to it; this is strictly prohibited with proprietary software.

Open source software can be created in several ways. A programmer or developer creates software from scratch, because they have some need that isn't being met by current software. Over time, as other programmers discover the project they may choose to contribute to building or improving this software, and they rally around the creator and begin to form a group that becomes devoted to the project. The Linux operating system and the Perl programming languages essentially began this way. Alternatively, a group of people who receive support from a business or entrepreneurs take software that was formerly proprietary but is no longer commercially viable, and they build on this product and re-release it as open source. The Mozilla Firefox browser (formerly the proprietary Netscape) and LibreOffice (formerly the proprietary Star Office and a branch of OpenOffice) are examples of the latter.

Why would people want to bother with creating FOSS software?

- It gives programmers a chance to practice their skills
- It gives programmers a way to enhance their prestige for their craft, as they can become known in different programming circles
- Open source is an ethos for some, who believe that software and information should be free
- Some see it as a superior model - since the code is open, there is a better chance that improvements can be made more quickly and that bugs can be discovered more easily than in proprietary software, as open source harnesses the power of the masses
- Businesses may prefer it because it does not tie them to costly, proprietary software that may go out of date or out of business - with open source there is always someone who can take over a project and keep it going, since the code is free and transparent
- If proprietary software for a certain application is inefficient, insufficient, expensive, or non-existent, FOSS software can be created to meet the need

The number of FOSS GIS packages has grown over the course of the last decade, and the Open Source Geospatial Foundation (OSGEO) was created to support the collaborative development of the software and promote its use (`http://www.osgeo.org/`). In this tutorial we will be using Quantum GIS (QGIS), which was initially developed by

a group of volunteers in 2002 as a simple GIS viewer but has evolved into one of the premier FOSS GIS packages.

The advantage of using QGIS for this tutorial: it's free, you can download it, it runs on any operating system, it is mature enough that it supports most essential GIS tasks plus a few intermediate and advanced ones, and it's relatively easy to use.

The disadvantage is that QGIS can't do everything that proprietary software can, is still working out some bugs, and doesn't have the name recognition that software like ArcGIS or MapINFO do. There also isn't as much in the way of documentation or tutorials for QGIS relative to the proprietary options, but all of this is rapidly changing. In the last few years there has been an increase in the number of workshops, online tutorials, and forums as the adoption of QGIS and other FOSS GIS software has grown.

Open software tends to be modular rather than monolithic; you often have several, independent software applications to perform different functions, rather than one, large piece of software that does it all. A typical FOSS GIS workstation may include several applications like QGIS (for viewing data, basic analyses, map making, generally working with vector data), GRASS (a more advanced GIS for doing analyses and modeling and for working with raster data), GDAL / OGR (command line tools for converting files and projections and for basic queries), and a geodatabase application (PostGIS for server-based databases and Spatialite / SQLite for desktop use). Individual FOSS software will often contain a large group of core components as well as a number of plug-ins that were subsequently designed to add new functionality. Plug-ins may be created by the developers or by third parties, and over time can be incorporated as core functions in later versions of the software.

ArcGIS, created by a company called ESRI, has been on the market for several decades and is the dominant, proprietary (non-FOSS) GIS software on the market. It's used by most government agencies and universities. Since it is rather expensive to purchase for individual use, you tend to see it more often in institutional settings. If you are affiliated with a college or university, chances are you'll be able to access it somewhere on your campus. ESRI does distribute trial versions of the software for education and home use. A rival product, MapINFO created by Pitney Bowes, has a smaller but equally dedicated following. If you find that you need to learn one of these products, making the transition from FOSS is relatively straight forward as most GIS software operate under the same properties and principles and share similar user interfaces.

Chapter 2

Exploring the Interface

The goal of this chapter is to familiarize you with the interface and basic features of GIS in general and QGIS in particular. You'll also add and configure some layers that you'll use later in Part 3

2.1 The QGIS Interface

This section will introduce you to the QGIS interface; you will configure the interface in preparation for the rest of this tutorial.

2.1.1 Steps

1. *Launch QGIS Desktop.* (If you're using Microsoft Windows, look under the Start Menu > Program Files > Quantum GIS > Quantum GIS Desktop).
2. *Configure plugins.* We're going to turn off the plugins that we're not going to use, to keep our interface uncluttered. Go to Plugins > Manage Plugins. Click the Clear All Button to unselect all of the plugins. Then scroll through and check these two: Add Delimited Text Layer and ftools. Hit OK.
3. *Configure the toolbars.* Likewise, we're going to turn off toolbars that we won't need. Right click on a blank area of the toolbar to get the toolbar view menu. Make sure the following eight features are checked: Browser, Layers, Attributes, File, Help, Label, Manage Layers, and Map Navigation. If other features are checked, uncheck them. Every time you check or uncheck a feature the toolbar view menu will disappear, so you will need to right click on a blank area of the toolbar to get it back.
4. *Move toolbars.* Move the toolbars around by hovering over the left edge of a toolbar until you see a crosshairs, left click and hold, then drag and drop. Configure the toolbars to your liking (suggestion: try aligning them so you have only two rows of them at the top of the screen and all buttons are visible).

2.1.2 Commentary

Interface Components

1. *Menu Bar*: provides access to various features and functions of the software using a standard hierarchical menu. The location of the menus and menu items is fixed, although if you activate certain plugins they may add an additional menu to the bar.

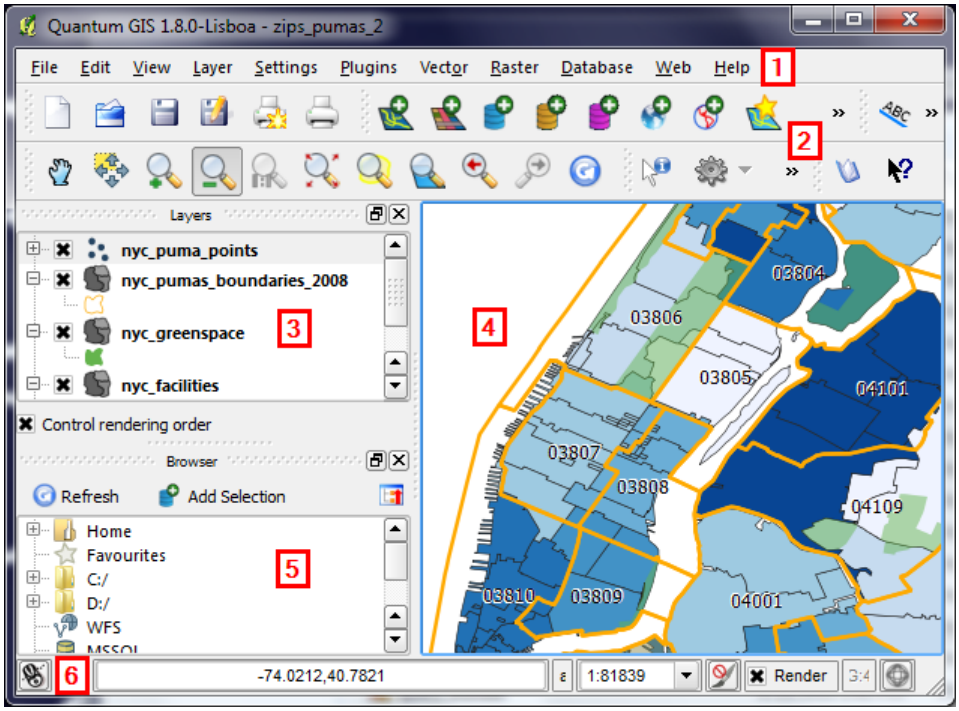
2. *Toolbar*: replicates many of the features and functions in the Menu Bar, providing access to common features in a single click. The location of the toolbars is not fixed; if you hover over the edge of the toolbar and hold down the left mouse button you can drag and dock the toolbar wherever you like (this means that the location of tools on your screen may not match those of other screens, or this tutorial).

3. *Map Legend*: a list of the map layers that are part of your current project. You can check or uncheck layers to turn them on and off, drag them to change the drawing order, select one in order to perform specific tasks on that layer, and right click on a layer to access menus and tools for working with that specific layer. The Map Legend is sometimes referred to as the Table of Contents in other GIS software.

4. *Map View*: geographic display that shows all of your active layers.

5. *Browser*: a new feature in QGIS 1.8, the browser allows you to see your file system and all of your GIS files and databases, and lets you drag files from your file system into your project. In this tutorial the browser occupies this space in the interface but there are other features you can enable, like the Map Overview, that can share or occupy this space.

6. *Status Bar*: shows the current scale of the map view, the coordinates of the current position of the cursor, and the coordinate system used by the project. Progress meters and other messages will appear here as you perform different operations.



- *Want to turn a toolbar off? Wondering where a toolbar went?* If you right click on a blank area of either the Menu Bar or the Toolbar, you'll get a list that shows all of the toolbars, as well as the Map Legend and Map Overview. You can check and uncheck items to turn them on and off.

- *Can't figure out what a button means or does?* If you hover over a button, a small window appears that displays the name of the button. If you select the what's this button and click on any area or item in the interface, you'll get a brief explanation of what it does.

- *Are there hotkeys?* Most menu items and tools can also be accessed by using hotkeys or keyboard shortcuts (for example, CTRL S will save the current project). For a full list of hotkeys, view the QGIS manual. Many of the common Windows shortcuts (like CTRL C for copy and CTRL V for paste) will work in QGIS.

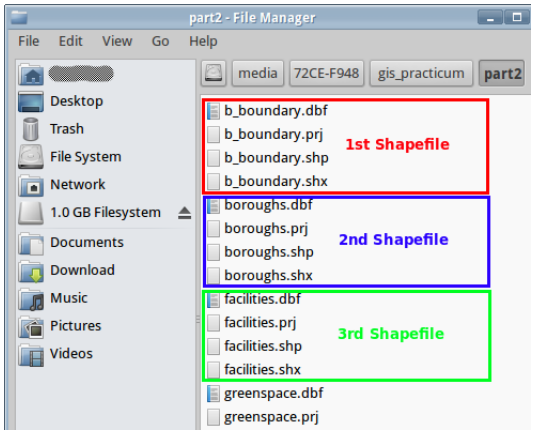
- *Where is the QGIS manual?* These are available on the QGIS website at `http://www.qgis.org/en/documentation/manuals.html`.

2.2 Adding Vector Data

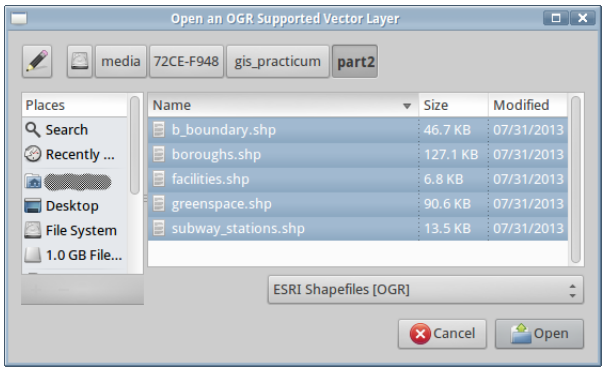
In this section you'll learn how to add vector GIS files (shapefiles) to QGIS and to symbolize them. Shapefiles are a common GIS data format that you'll routinely encounter in your future work.

2.2.1 Steps

1. *Examine your data.* Minimize QGIS for a moment and take a look at the data files under the data folder for part 2 in your operating system's file window. These are shapefiles that we will add to QGIS and work with for this project. There are five shapefiles; each shapefile is composed of multiple files that have the same names but different extensions.



2. *Add the five shapefiles.* Maximize QGIS to return to the program. On the Tool Bar, hit the  add vector layer button. When the Add Vector Layer box appears, hit the Browse button. Browse through the folder list to the data folder for part 2. In the Files of Type dropdown at the bottom of the window make sure the first option, ESRI shapefiles, is selected. Select the first layer in the list, hold down the shift key, then select the last layer. This should select all five shapefiles. Hit Open to add them. Your layers should appear in the Map Legend and Map View.

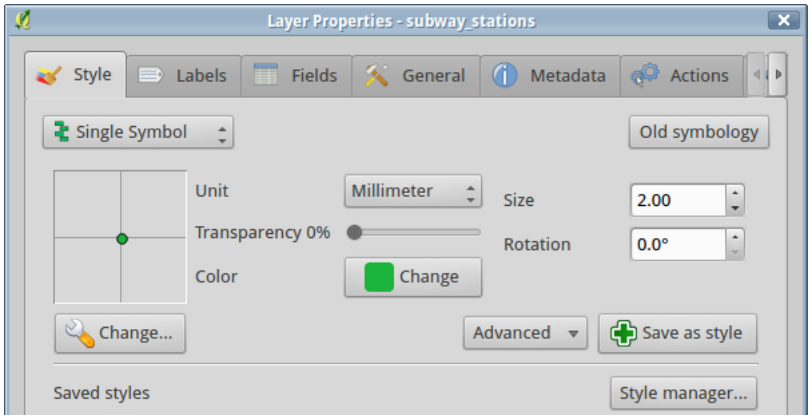


3. *Do your layers look jagged?* If not, skip this step. If so, on the Menu Bar, select Settings > Options > Rendering, and under Rendering Quality check the box that says "Make lines appear less jagged at the expense of some drawing performance", and hit OK.

4. *Experiment with changing the drawing order.* Click on the first layer that's listed in the Map Legend (ML), hold down the left mouse button, and drag it to the bottom of the list. This moves that layer from the top of the drawing order to the bottom; layers in the Map Legend (ML) are stacked on top of each other, and their order in the list determines which are visible relative to others. Move the BOROUGHS layer to the top of the list to see what happens.

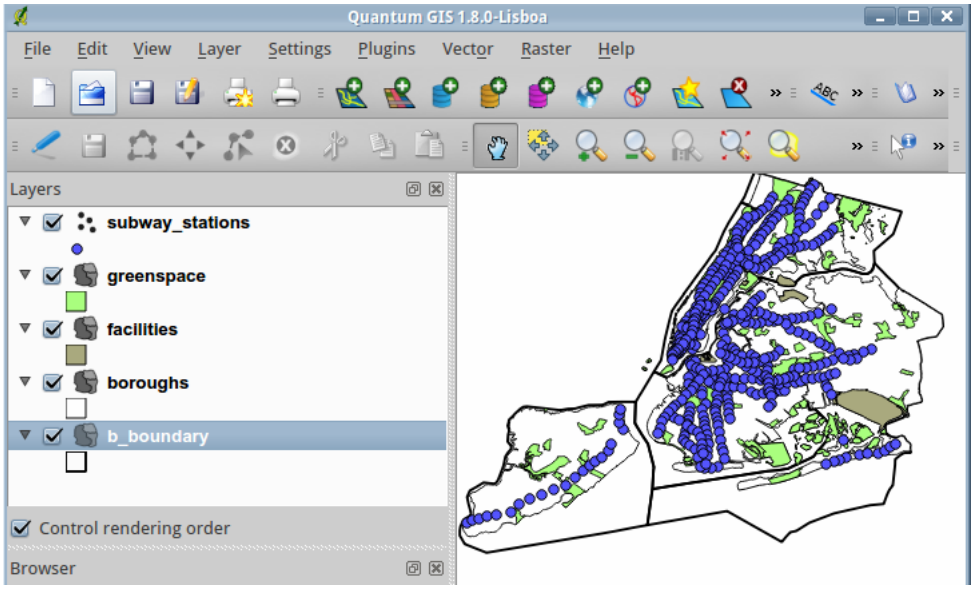
5. *Order the layers.* Drag the layers in the Map Legend (ML) so they appear in this order, from top to bottom: SUBWAY_STATIONS, GREENSPACE (parks and wildlife areas), FACILITIES (airports, ports, prisons), BOROUGHS, B_BOUNDARY (borough legal boundaries).

6. *Change the color for the subway stations.* Double-click on the SUBWAY_STATIONS layer in the ML to open the Layer Properties menu for that layer. Click on the Style tab. The Symbology tab should be set to New symbology by default, which is what we want to work with (there is a toggle button between Old and New on the right-hand side of the window). Click the Change Color button. Change the color to blue by choosing a box in the color palette. Click OK, then OK again on the Style menu.



7. *Change the colors for parks, facilities, and boroughs.* Make the GREENSPACE green, the FACILITIES grey or brown, and the BOROUGHS white. Make sure you're using the New Symbology tab for all your layers.

8. *Give the boundaries no fill.* (i.e. make them hollow with no color). Double-click on the B_BOUNDARY layer in the ML to open the Layer Properties menu for that layer. Click on the Style tab. Click the Change button to open the Symbol Properties window. Change the Fill style dropdown from Solid to No Brush. In the Border Width box, change the value from .26 to .75. Hit OK, and OK again. After completing these steps, your QGIS window should resemble the image below.



2.2.2 Commentary

Shapefiles

A shapefile is a very common file format used for storing vector GIS data. It was created by ESRI, the company that produces ArcGIS (the predominant software in the proprietary GIS market). Shapefiles are an open GIS format that can be used in just about any GIS software package, including QGIS. A shapefile can consist of point, line, or polygon features for a given geographic area, and can never consist of multiple types of geometry (i.e. you can't have a shapefile with points and lines). Polygon features can be single-part (where every individual polygon is an individual feature) or multi-part (where multiple polygons can be grouped together as single features).

Despite it's singular sounding name, a shapefile consists of several individual files. The following three pieces are mandatory:

- .shp file - shape file, contains the geometry
- .shx file - shape index file, an index of the geometry
- .dbf file - attribute file, contains attributes for the features

The following pieces are typically (ideally) included:

- .prj file - a plain text file that contains the projection and coordinate system
- .sbn and .sbx files - spatial index of the features
- .shp.xml file - XML metadata

It is important that all of the pieces of the shapefile are kept together in the same folder, otherwise the file will not work - so be careful when moving files around! Renaming files is often problematic - if you rename one you must rename all of them with the same name, otherwise they won't function together. You can easily rename batches of files with the same name but different extensions if you are familiar with using the command line (i.e. Unix/Linux shell or DOS Command Prompt); it's less tedious than renaming them by hand in a GUI (like Windows Explorer).

Adding Data and Drawing Order

When you add map layers or data to a map view, you are technically not adding data to the window, i.e. copying the file and inserting it into the project. Rather, you are establishing a link between the GIS interface and the files, which exist independently from the software. When you use GIS software to change the symbolization of the layers (colors, outline, labels, etc) you are not modifying the data file itself; you are simply telling the software to display the layers in a certain way. The software is essentially a window for viewing the data files. The only way to change the data files themselves (their geometry or attributes) is within an editing mode which you must specifically launch.

For much of the 20th century maps were created by taking individual layers on translucent mylar sheets and laying them over top of a paper base map. For example, an outline of the United States with boundaries of each state could serve as a paper base map, with individual mylar sheets layered on top that had rivers and cities. The order of the sheets determined which features appeared on top, covering up other features. GIS functions the same way; the order of the layers determines which appear on top. If you move a polygon layer with a solid fill (i.e. boroughs) over top of a point layer (i.e. of subway stations), you will not see the stations as the borough layer is covering it up. In order to show both layers, you would have to move the stations layer on top of the boroughs. Alternatively, you could make the boroughs layer hollow by removing the fill, which would allow the stations layer to be visible if it was on the bottom. You would typically use a hollow fill for a polygon if you wanted to display it's boundaries on top of another polygon layer that has a fill.

Old and New Symbols and Labels

For the past several years QGIS has offered two versions of symbology in the software: the original version (old) that descends from the first versions of QGIS, and an improved, experimental version. At this stage the new version is relatively stable and offers features that are difficult to live without. Similarly, QGIS has maintained two labeling engines: the old one is available under the labels tab for each layer and the new one, which is superior by far, is available from the toolbar. This tutorial uses just the new versions for symbols and labels, as the old versions are slated to disappear when QGIS 2.0 is released.

2.3 Exploring the Map View

In this section you'll learn how to navigate the map view.

2.3.1 Steps

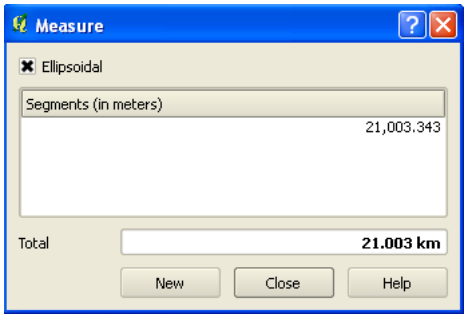
1. *Experiment with the Zoom tools.* Try each of the zoom tools in the Menu Bar.

   -  Pan - move around the map by holding the left mouse button down and drag (does not change the zoom)
   -  Pan Map to Selection - move map to selected features without changing the zoom (skip this one for now)
   -  Zoom In - click to zoom in once, draw a box to zoom in to an area, or use the mouse wheel.
   -  Zoom Out - works the same as the Zoom In tool
   -  Zoom to Native Pixel Resolution - will zoom to the optimal scale for rasters (skip this one for now)
   -  Zoom Full - will zoom the window to the maximum extent of all visible layers
   -  Zoom to Selection - zooms to selected features (skip this one for now)
   -  Zoom to Layer - zooms to the maximum extent of the feature currently selected in the ML

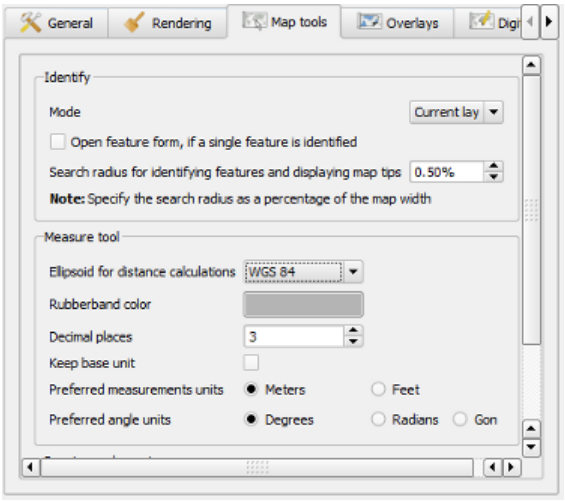
- Zoom last - returns to your previous zoom
- Zoom next - moves you forward to your next zoom (if you've already used zoom last)
- Refresh - redraws the screen (useful if your layers didn't draw completely or properly)

2. *Notice change in coordinates.* Move the cursor around the map. In the Status Bar (below the Map View) notice how the coordinates change; coordinates for the map are provided based on the position of the cursor. If you hover over the box that says EPSG: 4269, a pop-up window tells us the coordinate reference system (CRS) of the project based on our layers is NAD 83. In NAD 83 coordinates are measured in degrees of latitude and longitude. We'll cover this later in the tutorial. The scale box can also be used to change the zoom (a higher number to zoom out and a lower number to zoom in).

3. *Measure some distances.* Use the zoom tools to center Manhattan in your map window. Select the measuring distance tool in the toolbar. You'll notice that crosshairs will appear. Click on the northern tip of Manhattan. This will open the Measure window. Drag the crosshairs to the southern tip of Manhattan. As you do this, you'll see a black line is drawn from the original point you clicked on and the measurement window will update with distances in meters and kilometers. If you click on the southern tip of Manhattan it will lock the line segment and allow you to draw a second segment from the second point. Close the menu when you've finished experimenting.



4. *Change your measurement units.* (Note - in some implementations of 1.8 the following will cause the software to crash - save your work first). Go to Settings > Options > Map Tools tab. In the Measure Tool section under Preferred measurement units select the feet radio button. Under Ellipsoid for distance calculations dropdown change the values from WGS 84 to GRS 80 (What's this? See below). Hit OK. Try the measuring distance tool again and your units will be in feet and miles.



2.3.2 Commentary

Measuring Distances and Area

QGIS lets you measure distance, area, and angles. The measurement tool bases its measurement on the ellipsoid, or approximation of the earth's shape, in the Ellipsoid for distance measurements dropdown. If this ellipsoid doesn't match the one your project or layers are in you can get false measurements. How did we know that NAD 83 uses GRS 80? If you go to Settings > Project Properties and look at the definition for NAD 83, you'll see that GRS 80 is listed within the definition as the ellipsoid.

We'll cover coordinate systems and projections later on in this tutorial. The default CRS used by QGIS (WGS 84) is actually quite similar to NAD 83; in this case any errors in measurement would be quite small. The ellipsoids you are most likely to use are GRS 80 (when layers are in NAD 83) and WGS 84 (when layers are in WGS 84).

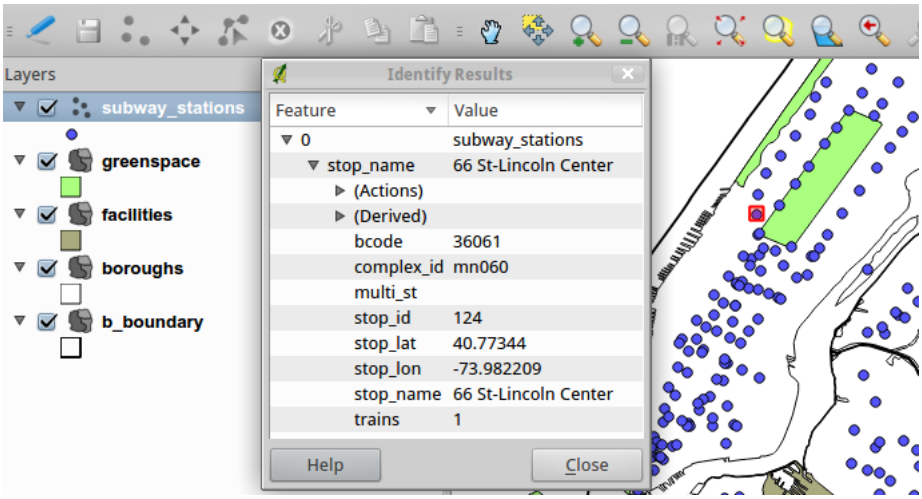
2.4 Exploring Features

In this section you'll learn how to explore and interact with features in the Map View and Attribute table.

2.4.1 Steps

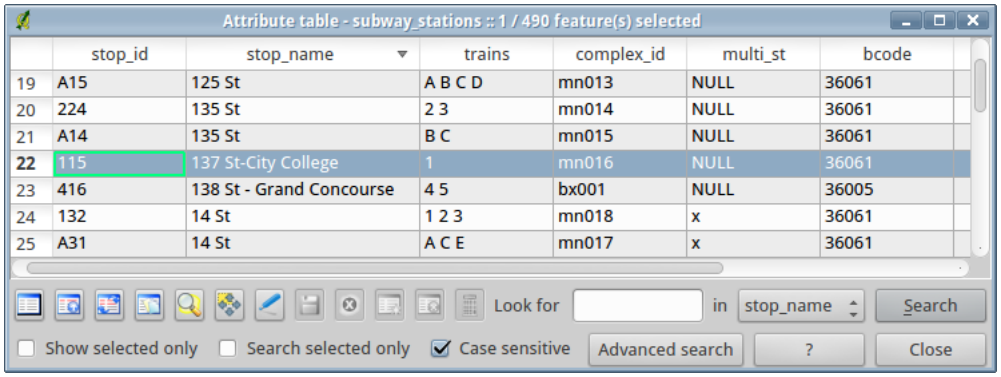
1. *Identify features.* Hit the identify features button in the toolbar. Select the BOROUGHS layer in the ML. Click on Manhattan. Manhattan is hi-lited and information about that feature is displayed. Click on The Bronx to change the selection.

2. *Identify features from a different layer.* Make the SUBWAY_STATIONS layer the active layer by selecting it in the ML. Click on any station in the map view to get information about that station. Where is this information coming from?

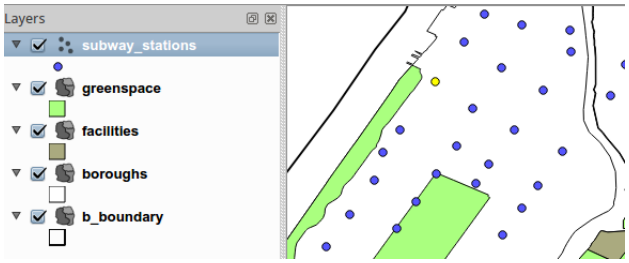


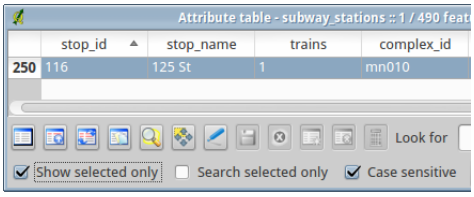
3. *Open the attribute table.* With the subway layer still selected in the ML, right click on the layer and select Open attribute table (alternatively, you could click the open attribute table button on the toolbar). For every station (feature) in the subway layer, there is a record for the station in the attribute table of that layer. Explore the table by scrolling across it and down.

4. *Select a feature from the table.* Sort the table by clicking on the field (column) heading that contains the name of the station (stop_name). Click on the record for 137-St City College in the table. Close the attribute table. Zoom to the area around City College in Harlem, just north of Central Park, and you'll see it is selected. (Note - you can select multiple records from the table by holding down the CTRL key and selecting records one by one, or select a range by selecting a record, hold the SHIFT key, and select the last record).

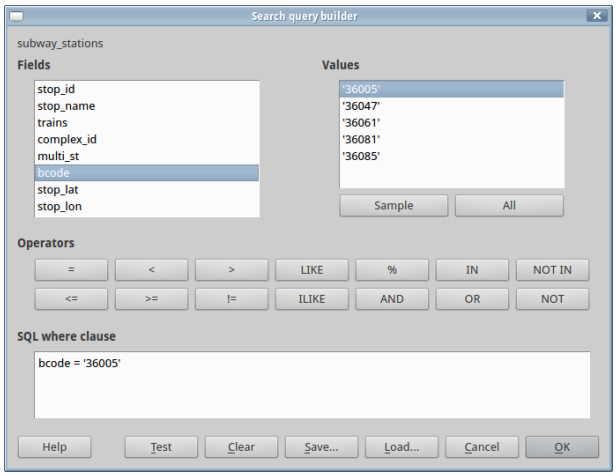


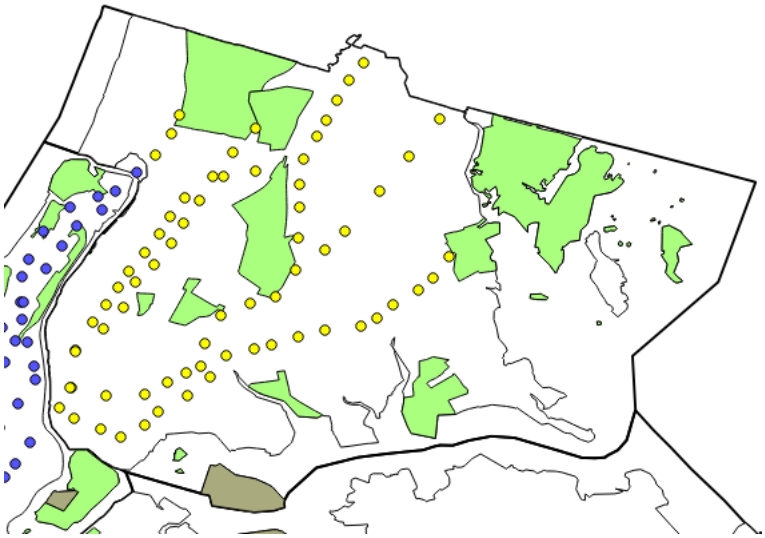
5. *Select a feature from the map.* With the SUBWAY_STATIONS layer still selected in the ML, hit the select feature button in the toolbar. Then select the station that is southwest of 137-St City College and just east of the northern boundary of Riverside Park. Hit the open attribute table button. Click the checkbox that says Show Selected Records Only. This reveals the record for the 125 St station for the 1 Train; this is the station that you've selected in the Map View. These two steps demonstrate that the table and map are linked, and you can select features in one and display them in the other. (Note - you can select multiple features by holding down the CTRL key and clicking on features one by one, or by hitting the dropdown beside the select feature button and choosing one of several options).





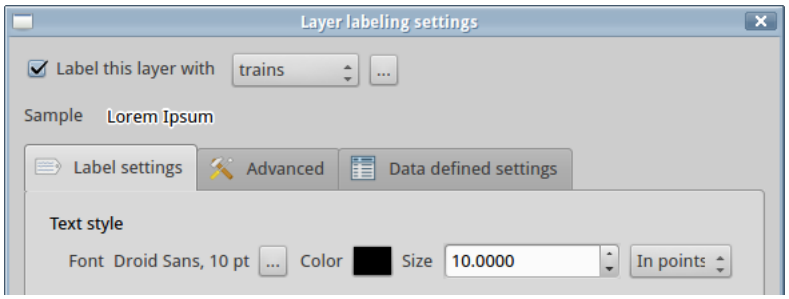
6. *Select Features by Attribute.* With the attribute table for the SUBWAY_STATIONS open, click the Advanced button in the lower right-hand corner. This opens the query builder window, which allows you to select features based on shared attributes. In the Fields box, double-click the bcode field, which adds it to the SQL Clause box at the bottom. Click on the equals sign in the Operators section. Hit the All button under the Values box to display all of the unique values for the bcode field. Double-click on the '36005' value listed in the value field. Your statement in the SQL Clause box should read bcode = '36005'. Click OK. You've just selected all of the subway stations that are located in the Bronx (36 is the census code for NY State, 005 is the code for Bronx County). Close the attribute table and you'll see the stations selected in the map.





7. *Clear selected features.* Click the clear selected features button on the toolbar to remove selected features in the active layer (the active layer is the currently selected layer in the ML hi-lited in blue - in this case, the SUBWAY_STATIONS). Alternatively, you could click on an area of the map that has no stations to clear the features, or you could clear the current selection from the attribute table.

8. *Labeling features.* Attributes stored in the table can also be used to label features. Select the SUBWAY_STATIONS layer in the ML to activate it. Click the labels button on the toolbar. Check the box that says Label this layer with and in the dropdown choose the trains column. Click OK. Explore the map a little, and notice how the labels shift as you zoom in. We'll experiment more with labeling later on. Click the labels button again and uncheck the box to turn the labels off.



2.4.2 Commentary

Attribute Tables

Every vector feature has a record in the attribute table; you can't have a feature without an attribute or vice versa. In a shapefile, the geometry is stored in the .shp file, an index of the geometry is in the .shx file, and the attributes are stored in a .dbf file. As we'll explore throughout this tutorial, attributes can be used for selecting, symbolizing, and

labeling features in layers.

In GIS software attribute tables are managed and handled in the same manner as tables in a relational database. Each column has a data type associated with it which determines the kind of data that can be stored in that column and the types of operations that can be performed on it. Data types include strings (aka text) and various types of numeric fields (integers for whole numbers, reals for numbers with decimal places, etc). When you use the Query Builder to select features, like bcode = '36005', you are actually creating SQL code, which is a standard language for manipulating data in a database. The code '36005' must be surrounded by quotes, as the data is an identifier saved as a text or 'string' field; if we were querying actual numeric values we would not use quotes.

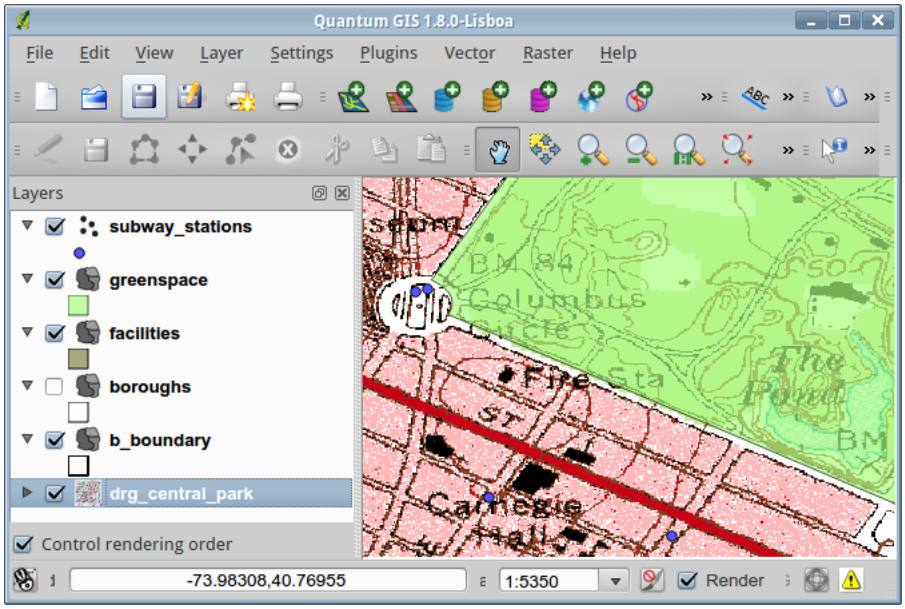
2.5 Adding Raster Data

In this section you'll get a very brief introduction to raster data.

2.5.1 Steps

1. *Add raster data.* Hit the  add raster layer button on the toolbar. Browse to the data folder for part 2, select the DRG_CENTRAL_PARK.TIF file and add hit open. Once the layer is added, drag it to the bottom of the ML.

2. *Explore raster layer.* Select the DRG_CENTRAL_PARK layer in the ML. Right click on the layer and select Zoom to best scale (100%). Uncheck the BOROUGHS layer in the ML to turn it off. Select the GREENSPACE layer in the Map Legend. Double click to open the Layer Properties and go to the Style tab. Drag the transparency slider to 30% and click OK. Explore the area of the map around Central Park and note how the raster layer lines up with the other layers. When you're finished exploring the map, uncheck the raster layer in the ML to turn it off, turn the boroughs layer back on, and return the transparency of the parks layer back to zero.



2.5.2 Commentary

Raster Data

Raster layers differ from vector layers in many ways including composition (continuous surface of pixels versus discrete geometric areas), file formats (many raster formats versus relatively few vector formats), resolution (optimal scale for raster layers matters more than vector layers), size (raster files tend to be much larger), and attribute tables (raster layers do not have attribute tables; the color of individual pixels denotes feature values). Given the differences in format, the tools for working with vector and raster layers are distinct (if you double click on the raster layer to open its properties, you'll see that most of the menu options are different from the vector layers).

Many geographic objects are represented in raster formats including satellite imagery, aerial photography, paper maps that have been scanned and digitized, and imagery that has been interpreted to represent value-added data that does not conform to political boundaries, such as land use and land cover and population density.

There are a number of great plugins for working with rasters, like the gdal plugin for performing raster analysis and the georeferrencing plugin, which allows you to convert non-GIS image files (i.e. a scanned paper map) to a raster GIS file by assigning coordinates to it. Given the time constraints of this tutorial, we're not going to cover rasters beyond this point. It was introduced here to give you a more complete picture of GIS capabilities and data formats.

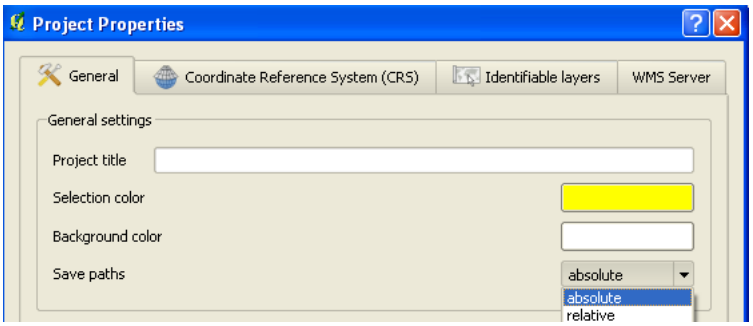
The raster used in this exercise is a DRG (digital raster graphic) which is a digitized, georeferenced version of the USGS' topographic maps. USGS topos are useful for studying elevation and terrain (particularly in non-urban areas) and for providing a frame of reference for overlaying vector layers or creating new ones; however most of the DRGs are several decades old and should be used with that fact in mind. The DRG was stored in a special .tif format called a GeoTIFF; a lossless image file that has georeferencing information (coordinates and map projection) embedded in it.

2.6 Saving Your Project

You'll learn how to save your project.

2.6.1 Steps

1. *Verify paths of files are relative and not absolute.* Under Settings > Project Properties > General Tab, for the last option in the General Settings area labeled as Save Paths, verify that the selected drop down item is relative.



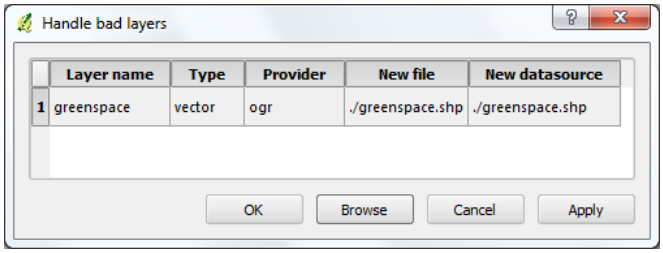
2. *Save your project.* Hit the save project button. Navigate to the data folder for part 2, and save your project there as PART2.QGS. The project file saves the symbolization, labeling, and current zoom for your data, and links to your data files (shapefiles); the shapefiles themselves are NOT stored inside your project file and exist independently. In order to use your project in the future, the project file and the shapefiles you used must be kept together.

2.6.2 Commentary

Project Files

When you add data to a project file you are not saving the data (shapefiles) inside the project; you are saving links to those files. Elements like symbolization, data classification, the extent of your last zoom, and any finished maps you create are stored in the project file. When you click on the project file to open it, the software looks at the paths to your data, re-establishes the links, and then applies the settings (symbols, zoom, etc) that you have saved in your project file. This relationship is of crucial importance when it comes time to move or share files - if you move your project file or your data the links between them can become broken, and you'll need to re-establish the location between the project and the data in order to repair your project file.

If you open a project in QGIS and your project file can't find the data, because the data has been moved or renamed, the software will give you the opportunity to restore the link by asking you to browse through your file folders and select each file that corresponds to a layer you have in the ML of your project. Once you restore the links, you can save the project and it will save the new links.



Paths to files can be stored as absolute links or as relative links. An absolute link contains the complete path of a file, such as F:\My_Stuff\GIS_Practicum\part2\data\boroughs.shp. Use absolute paths when you're working in an established environment where you know that you won't need to move data and projects around, or in situations where your project files won't be stored directly above or in the same folder as your data. Absolute paths are a bad choice if you know you'll be moving data around; they're a particularly bad choice if you're working on a usb drive in a MS Windows environment, as the paths can change as you move from machine to machine (i.e. F:\My_Stuff... on one machine becomes E:\My_Stuff... on another machine; QGIS won't be able to locate the files stored on F:\My_Stuff because it doesn't exist that way on the 2nd machine).

Relative paths save the directory and file information for the folder the project file is in (i.e. path would be .\boroughs.shp) and all folders below it (i.e. path would be .\data\boroughs.shp). Since anything above the project's directory is omitted, relative paths are a good choice if you know that you'll be sharing your project data or moving it around. Relative paths are a bad choice if your data is not going to be stored underneath your project folders (i.e. it's stored above the project directory, in a parallel directory, or another drive or server all together).

Think carefully about where to save project files in relation to your data, and once you've created your project file keep project files and data in a consistent place. Also remember that you must keep all of the individual components of a shapefile together (.shp, .shx, .dbf, .prj, etc); otherwise the shapefile will not function. If you want to share your project file with someone, you will also have to send them your data; the project file cannot exist independently from the data. You can share views or maps you've created in a static format (image file or PDF) that is separate from your project and data files; we'll explore that later in this tutorial.

The QGIS project file (.qgs) is actually just an XML file. If you open the project file in a text editor, you'll be able to see the structure of the file and all of its elements and attributes.

Chapter 3

Geographic Analysis

The goal of this chapter is to introduce some analysis and geoprocessing features and techniques using a site selection problem as an example. Over the course of this exercise you'll learn how to: create a new project from an existing one, create a subset of a layer and process it to create land boundaries, join an attribute table to a shapefile, map the attributes of a shapefile, take a list of coordinates and convert it to a shapefile, draw buffers around a set of features, and select features based on their attributes and their spatial relationship to other features. You'll also get an introduction to the new QGIS Browser.

The object of this particular exercise is to identify potential areas within New York City for locating a neighborhood coffee shop. Market research suggests that the primary demographic group that drinks coffee and visits coffee shops are women aged 18 to 49. Based on this research we will identify neighborhoods where this group represents a high percentage of the total population, and areas where median household income is neither too high (indicating that rent would be prohibitively expensive) or too low (where people would be less likely to have disposable income for coffee). We will also focus on areas that are within close proximity to subway stations as these tend to be high-traffic commercial areas, while avoiding areas where competitors already exist.

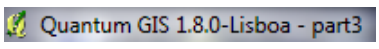
3.1 Creating New Project From Existing One

This section will show you how to create a new project from an existing one and will set the working environment for the rest of part 3.

3.1.1 Steps

1. *Open project. Launch QGIS.* Hit the  open project button (or go to File > Open Project). Browse through your folders to the QGIS project file you created for part 2, and select it to open it.

2. *Save Project As.* Once your project has loaded, hit the  save project as button (or File > Save Project As). Browse to the data folder for part 3. Save the project in that folder as PART3.QGS. Hit save. You've now saved a new copy of your old project, and are currently working in this new copy (you can tell by looking at the title at the top of the window, where the project name is listed). We will work with this new project, PART3.QGS, for this part of the tutorial.



3. *Remove a layer.* We don't need the raster layer for this exercise. Select the DRG_CENTRAL_PARK layer it in the Map Legend (ML). Right click on the layer in the ML and select Remove (or, hit the  remove layer button on the

toolbar).

4. *Zoom out and save.* Hit the  zoom to full extent button to zoom out to the full extent of your layers. Then hit the  save button.

3.1.2 Commentary

Saving Projects and Removing Layers

Use the Save button to save the current project, and the Save As button to save the current project as a new copy with a different project name. Save As saves you the effort of starting from scratch if you have an existing project that you can use to branch off from. When you remove a layer from a project you're just severing the link between a particular project and that data; you're not actually deleting the data itself.

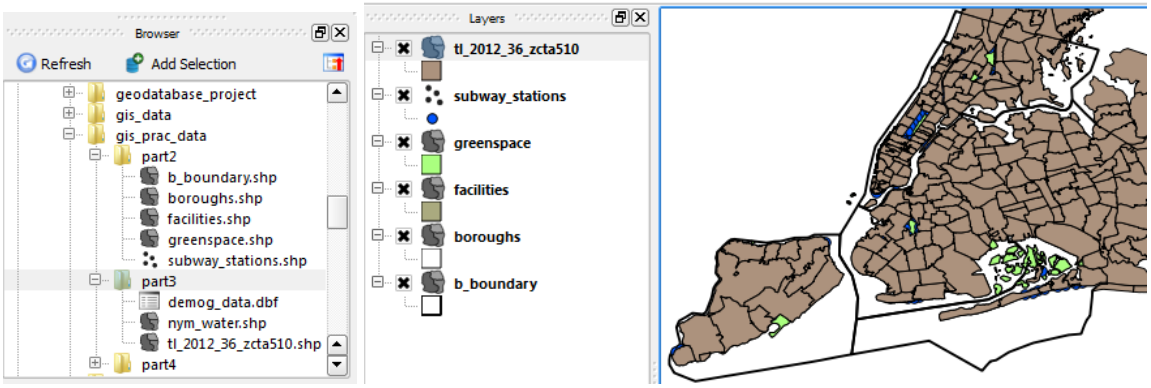
3.2 Geoprocessing Shapefiles

In this section you'll learn how to process a shapefile to prepare it for analysis. This is a common GIS task; normally when you download publicly available shapefiles you'll have to do some processing to make them usable for your projects. You'll also learn how to use the QGIS Browser to work with your files.

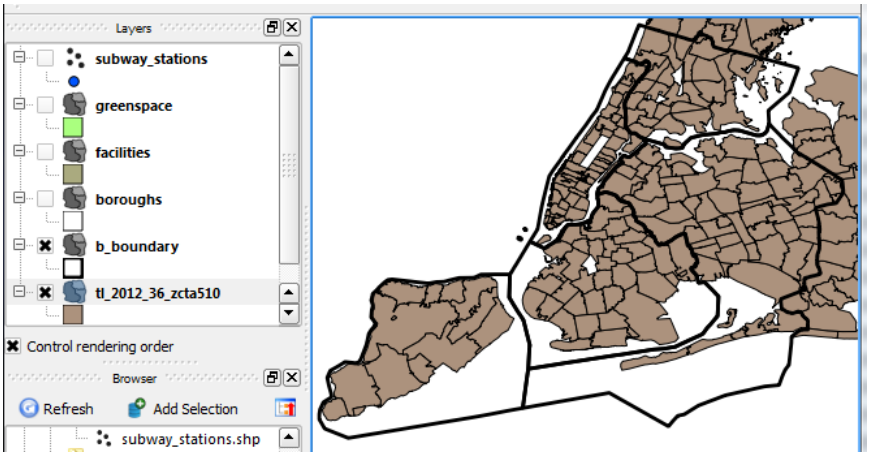
You'll be processing a boundary file for ZIP Code Tabulation Areas (ZCTAs) which we'll use to approximate neighborhoods. ZCTAs are statistical boundaries created by the US Census Bureau to approximate USPS ZIP Codes. The file was downloaded from the US Census TIGER Line Files.

3.2.1 Steps

1. *Add the ZCTA shapefile using the Browser.* The browser is a new feature in version 1.8 that we can use to see our file system within QGIS and add data directly to our project, instead of using the toolbar buttons to add  vector or  raster data. The browser has folders that represent hard and external drives as well as icons for connecting to various geodatabases and web services. Use the plus buttons to expand the folder tree where your project files are stored, and drill down to the part3 folder. Select the ZCTA layer, which is called TL_2012_36_ZCTA510.SHP. You can either right click and choose the option to add it to your project, or you can hold down the left mouse button and drag it into the Map Window.

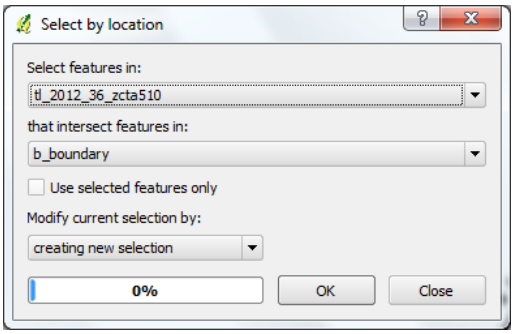


2. *Organize layers.* By default the new layer is drawn over top of the currently selected layer; if no layer is currently selected it is drawn on top of all of them. Select the ZCTA layer in the Map Legend (ML) and drag it to the top of the legend. Hit the  zoom to layer button. You'll see the ZCTA layer covers all of NY state, but we only need ZCTAs for NYC. We'll do some operations and create a new file that just has the NYC ZCTAs. Select the B_BOUNDARY layer in the ML and hit the  zoom to layer button. Select the ZCTA layer in the ML, and drag it to the bottom of the ML. Check the boxes beside all of the other layers GREENSPACE, FACILITIES, SUBWAYS, BOROUGHS to turn them off for now.

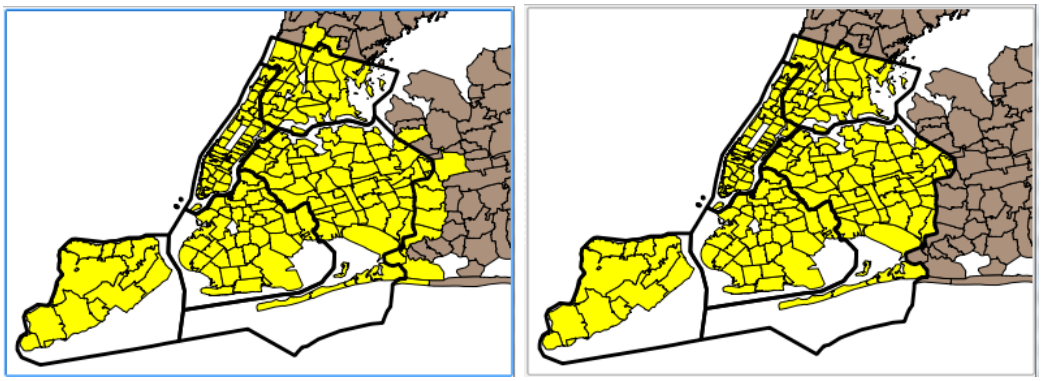


3. *Activate the fTools plugin.* If you haven't done so already, go to Plugins > Manage Plugins, and make sure the fTools plugin is checked. This will make the Vector menu appear on the menu bar.

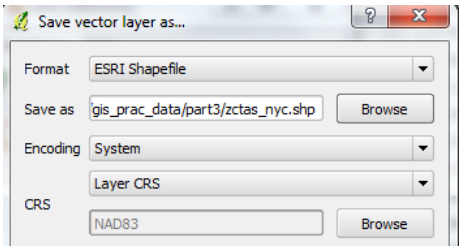
4. *Select ZCTAs within the b_boundary layer.* Go to Vector > Research Tools > Select by Location. Select features in the ZCTA layer (TL_2012_36_ZCTA510) that intersect features in the borough boundary layer (B_BOUNDARY), and keep the default for Modify current selection by creating new selection. Click OK. You'll see that all ZCTAs within and touching the NYC boroughs have been selected. Close the Select by Location menu when finished.



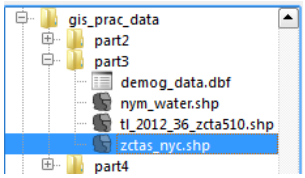
5. *Remove ZCTAs outside NYC from selection.* Select the ZCTA layer in the ML. Hit the  select features button. **While holding down the CTRL key**, click on each of the ZCTAs that are outside of the dark NYC boundary one by one to unselect each one. There are a few ZCTAs that have a small amount of area inside Queens but are primarily outside of NYC; unselect them as well. If you unselect a ZCTA by mistake, just click it again to re-select it. If you inadvertently unselect all of the ZCTAs (by letting go of the CTRL key and selecting a feature), you'll have to redo the previous step with the Select by Location tool to to reselect all of them.



6. *Save selection as new layer.* Select the ZCTA layer in the ML. Right click and choose the Save Selection As option. In the Save Selection as Menu, save the new layer as an ESRI shapefile. Browse and save it in your part 3 folder as ZCTAS_NYC. Leave the Encoding as the default. Notice that the new file will be given the same CRS as the current layer, which is in NAD 83. Hit OK.



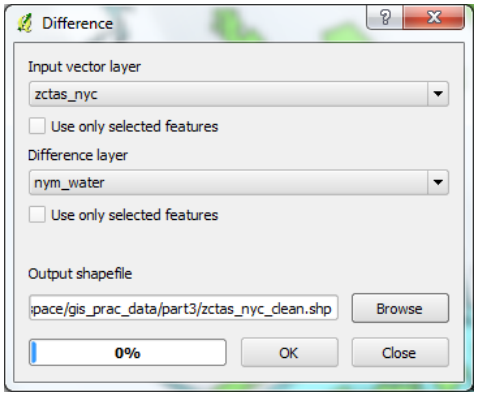
7. *Add new layer to map.* Whenever we save a layer as a new layer, we have to manually add it to our project. Use the browser to drill down to the part 3 folder. You should see the new file ZCTAS_NYC there; if you don't hit the Refresh refresh button just above the browser. Drag the new file into the window to add it, or select it, right click, and Add Layer (Can't find the browser? Make sure it's turned on - right click on a blank section of the toolbar, and check the browser to turn it on).



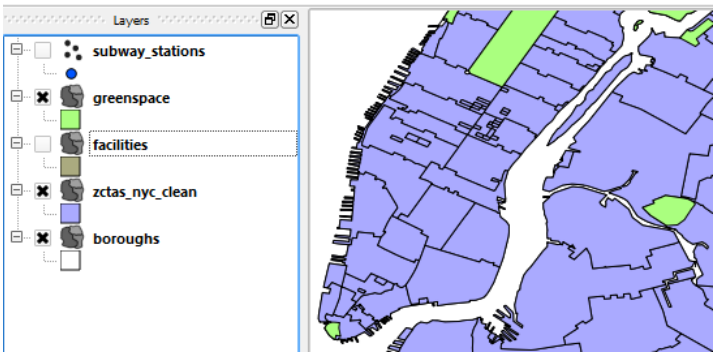
8. *Tidy up your layers.* As we create new layers we can remove the old ones. We're finished with the ORIGINAL ZCTA layer for NY state and the B_BOUNDARY layer. Select each one in the ML, right click and remove it. Drag the ZCTAs FOR NYC to the bottom of the ML. Turn the GREENSPACE and BOROUGHS layer back on by checking the box beside them. Save your project at this point.

9. *Inspect the ZCTA layer relative to the boroughs.* You will usually need to modify statistical boundaries so that they depict just land areas - for example the B_BOUNDARY layer represented the legal boundaries of the boroughs, while the BOROUGHS layer represents just land. By definition, ZCTAs consist of land where street addresses are located (thus parks, uninhabited islands, and water areas are not covered). However the boundaries are still not perfect - for example if you zoom in to the area around Newtown Creek, a small tributary of the East River that separates northern Brooklyn from Queens, you'll see that this is represented as land in the ZCTA layer, but is devoid of land in the borough layer (as the creek is located there).

10. *Clean the ZCTA layer by subtracting water from it.* Use the browser to go to the part 3 data folder and add the layer NYM_WATER. Drag it to the top of the ML. On the menu bar go to Vector > Geoprocessing Tools > Difference. Select ZCTAS_NYC as the Input vector layer, NYM_WATER as the Difference layer, and Browse and save the new file in your part 3 data folder as ZCTAS_NYC_CLEAN. Hit OK. When prompted to add the layer to the project, say Yes. Close the difference menu.



11. *Clean up.* Select the NYM_WATER layer in the ML, right click and remove it. Do the same for the ZCTAS_NYC layer. Then drag the new ZCTAS_NYC_CLEAN layer to the bottom of the ML. If you look at the area around Newtown Creek both the ZCTAs and boroughs will overlap identically. At this point, you have a brand new ZCTA layer just for NYC that represents land boundaries, which match the borough land areas. Uncheck the BOROUGH layer to turn it off. Save your project.

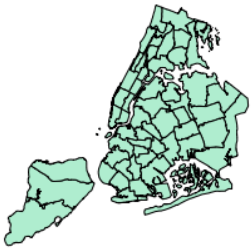


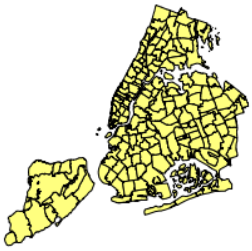
3.2.2 Commentary

Geographic Units

For this exercise we're working with ZIP Code Tabulation Areas (ZCTAs) which are statistical areas created by the US Census Bureau that were designed to approximate USPS ZIP Codes. ZIP Codes are often used for aggregating and publishing data, particularly within business fields like marketing and real estate, because they are common areas that most people are familiar with and because address-based data can easily be summarized by ZIP Code. While they are often used to approximate neighborhoods they are often not the best choice for that purpose; ZIP codes vary tremendously in size, shape, and population and were created for delivering mail, not for studying or delineating neighborhoods. Some ZIP Codes represent clusters of PO boxes or large, individual organizations; since these lack any geographic area they are omitted from ZCTAs.

Here is a summary of some of the most common geographic areas for thematic mapping in the US (most countries will have some corollaries):

	**Counties** - Legal subdivisions of states, counties are commonly used for mapping national or regional distributions given the large amount of data that's available for them. New York City is unique as it's the only US city composed of multiple counties, and for historical reasons the five NYC counties are referred to as boroughs. City agencies classify data using borough names (Bronx, Brooklyn, Manhattan, Queens, Staten Island) while federal agencies like the US Census Bureau use county names (Bronx, Kings, New York, Queens, Richmond). Dataset availability: 2010 Census, 1, 3, and 5-year ACS, pop estimates
	**PUMAs** (Public Use Microdata Areas) - Statistical areas created by the US Census Bureau to have approximately 100,000 residents; they're created by aggregating census tracts. In urban areas they can represent subdivisions of cities, in suburban areas they represent subdivisions of counties, and in rural areas they are often aggregates of several counties. There are 55 PUMAs in NYC; they are occasionally referred to as sub-boroughs. Dataset availability: 1, 3, and 5-year ACS

	**ZCTAs** (ZIP Code Tabulation Areas) - Statistical areas created by the US Census Bureau to approximate areal USPS ZIP Codes. The Census Bureau creates ZCTAs by aggregating small statistical areas called census blocks based on the location of addresses within the blocks. While not always ideal for representing neighborhoods, ZIP Codes are often used for this purpose since most people are familiar with them. ZCTAs do not correspond with other census geographies. Dataset availability: 2010 Census and 5-year ACS
	**Census Tracts** - Statistical areas created by the US Census Bureau to have approximately 4,000 residents (with a range of 1,200 to 8,000). Tracts can be used for analyzing patterns within counties, cities, and neighborhoods and can be aggregated to create neighborhood-like areas; many cities create official neighborhood or sub-municipal areas based on tracts. The NYC Department of City Planning has taken the 2,000 plus census tracts in the city and aggregated them to create 195 Neighborhood Tabulation Areas (NTAs) for presenting and publishing census data. Dataset availability: 2010 Census and 5-year ACS

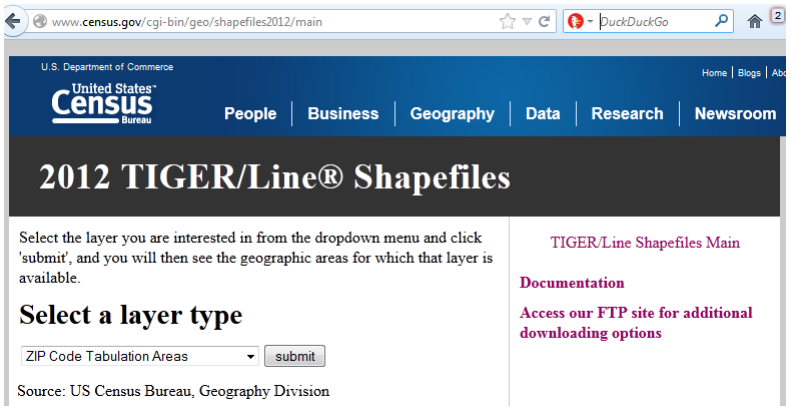
The choice of a geographic unit is an important decision; it's often a balance between the availability of data for an area, the suitability of the unit for the analysis, the amount of work that has to be invested in processing and analyzing the data, and the final outputs that will be created (tables, charts, maps) to explain the data.

For example, we used ZCTAs for this exercise because the demographic data availability is good, there is a work-able number of them in the city (approximately 200), and they are commonly used in real estate applications. Other geographic areas like PUMAs or census tracts would have given us units of equal population size to study, but would have presented other shortcomings. PUMAs are large enough that they can mask a lot of variability within each area, and tracts are small enough that our end result (ideal areas for locating the store) would be more precise, but more geographically fragmented. Combinations of tracts, like Neighborhood Tabulation Areas, would have given us recognizable neighborhood areas that are relatively equal in population and similar in size, but would require additional work to aggregate the demographic data that we will use.

TIGER Line Files

The Census Bureau creates and maintains legal, statistical, and administrative boundaries for all geographic areas that it publishes data for. It also creates and maintains geographic features such as water, roads, and landmarks that are used when creating statistical boundaries. These files were originally in a vector format created by the census called Topologically Integrated Geographic Encoding and Referencing or TIGER. The Census now provides this data in shapefile format. The files are in the public domain and can be downloaded for free at `http://www.census.gov/geo/maps-data/data/tiger.html`

The ZCTAs used in this tutorial were downloaded from the Census TIGER site, as were most of the other files used in this exercise. The borough file is a subset of the TIGER county file for New York State, while the facilities and green space layers are aggregations and selections from the TIGER landmarks file for each of the five counties. All three layers were geoprocessed for this tutorial to convert legal boundaries to land boundaries, using a subset of the TIGER water features.



We were able to add the ZCTA layer directly to our project because it shares the same geographic coordinate system as our other layers - NAD 83. The default coordinate reference system (CRS) for all Census TIGER files is NAD 83. We'll discuss and work with map projections later on in this tutorial.

The Census Bureau makes minor updates to boundaries and issues new TIGER files each year, but major changes occur at the beginning of each decade as the decennial census is released. There are often minor changes to statistical areas (like census tracts and ZCTAs) within a year or two of the decennial release to correct errors, but after that these areas are fixed and do not change until the next ten-year census. In contrast, updates to legal boundaries (like states, counties, or municipalities) are made on an annual basis. The TIGER files used in this exercise are from the 2012 TIGER / Line Shapefiles, which are based on 2010 Census geography.

Geographic Selection

One of the strengths of GIS is the ability to perform spatial queries on features; i.e. select all areas that intersect other areas. This is one area where QGIS is still developing. The Select by Location feature of the fTools plugin only allows you to select features that intersect other features. However, several other spatial query options exist in other GIS packages, such as selecting features that border each other, or that are within or have their center within other features (the latter would have been the preferred option for selecting ZCTAs within NYC boroughs). QGIS does have a Spatial Query plugin that can be activated in the plugins menu, and provides other options such as: crosses, disjoint, intersects, touches, and within. However, the tool isn't perfect and seems to have trouble when making selections between two polygon layers, which is why it wasn't demonstrated in this tutorial (although it works better when selecting points or lines in relation to polygons). It isn't clear if this is a shortcoming with the tool, or with shapefiles that aren't perfectly formed. If you need spatial query options beyond intersect, you can use other open source software such as the GDAL / OGR command-line tools or a geodatabase (PostGIS or SpatiaLite).

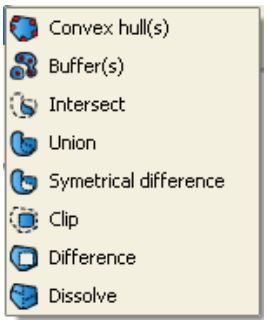
It's pretty common that you'll download geographic data that covers an area that is wider than you need. Since GIS data is malleable, it usually makes sense to grab data for a larger area and select out just the portions you need, if you can't find a layer that consists just of the areas you want; this is something to keep in mind when you search for data on the web.

Geoprocessing

It's also rather common that you'll download shapefiles that represent boundaries, but these boundaries will often incorporate land and water. If your intention is to show the actual boundary lines for reference purposes, then you will want to use the files as is. However, if you want to map the distribution of phenomena by area you'll want to process the boundaries to remove water as that phenomena isn't likely distributed there (i.e. there are no people living in the harbor). You'd also want to alter the boundaries if you're creating maps and want the user to be able to clearly understand the areas you're depicting. The Difference tool accomplishes this by subtracting the areas of

bodies of water from the boundaries, resulting in features that show the outline of land.

This is merely one application and tool in the geoprocessing toolkit. Geoprocessing is essentially a GIS operation to manipulate the spatial aspects of GIS data. In the broad sense it includes layer overlay, feature selection, data conversion, and topology processing. In a more narrow sense that we're using here, it refers specifically to topology processing; modifying the actual geometry (points, lines, and areas) of features and files. Via the ftools plugin, QGIS has the following Geoprocessing tools for vector layers (running each tool creates a new layer; it does not modify existing layers):



- Convex Hulls - creates the smallest possible convex polygon enclosing a group of objects
- Buffers - creates an equal zone around specific features at a specified distance
- Intersect - creates new layer based on the area of overlap of two layers
- Union - melds two layers together into one while preserving features and attributes of both
- Symmetrical Difference - creates new layer based on areas of two layers that do not overlap
- Clip - cuts a layer based on the boundaries of another layer
- Difference - subtracts areas of one layer based on the overlap of another layer
- Dissolve - merges features within a single layer based on common attributes in the attribute table

In addition, there are also some geoprocessing tools under the Geometry Tools menu in ftools that convert or break polygons apart into simpler features (like lines or points) and under the Data Management Tools menu (for aggregating many shapefiles into one file; the opposite of the selection / subset process). Geoprocessing for raster layers is available through the GDAL plugin.

3.3 Joining and Mapping Attribute Data

In this section you'll learn how to join an attribute table to a shapefile and map the attributes in that table. Now that the ZCTA boundaries are ready, we need to associate them with census demographic data for those ZCTAs in order to select the optimal neighborhoods for locating our shop.

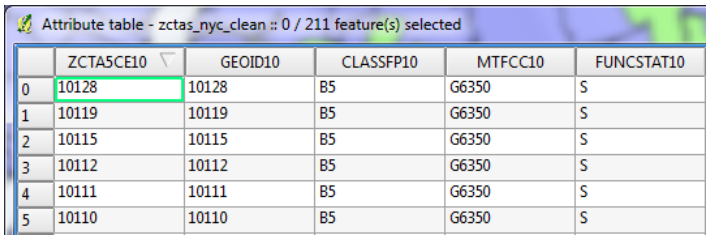
3.3.1 Steps

1. *Open the data file.* Minimize (don't exit) QGIS for the moment. Using your file manager, browse to the data folder for part 3. Look for a file called DEMOG_DATA.DBF. A dbf is a dbase file, used for storing data. This is a stand-alone dbf that is not associated with a shapefile. Depending on what operating system you're using, open this file with a spreadsheet package like Excel or Calc (if you're in Windows select the file, right click, select Open With, and then choose the option to select a program from the list. Choose Excel, hit OK, and open the file).

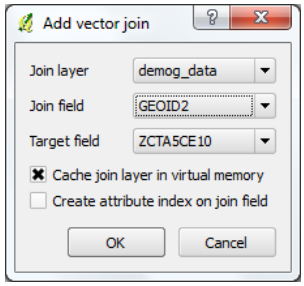
2. *Examine the data file.* The data file contains one row for each ZCTA in NYC and several columns of attributes. The first three columns contain identifiers for each ZCTA; the column GEOID2 is a FIPS code that we'll use to join this table to the shapefile. The first three data columns are from the 2010 Census and contain the total population, women aged 18-49, and percentage of the total population in that age and gender group. The last two columns are from the American Community Survey (ACS) and represent an estimate of median household income and a margin of error for that estimate. You would interpret the estimates thusly: For ZCTA 10001 we're 90% confident (that's the confidence interval for the ACS) that median household income was $67,795 between 2007-2011, plus or minus $6,489.

	A	B	C	D	E	F	G	H
1	GEOID	GEOID2	GEOLABEL	TOTPOP	FEM18_49	PCTFEM	HSHD_INC	MARGIN_E
2	8600000US10001	10001	ZCTA5 10001	21102	6810	32.3	67795	6489
3	8600000US10002	10002	ZCTA5 10002	81410	19583	24.1	32407	2484
4	8600000US10003	10003	ZCTA5 10003	56024	20058	35.8	88601	5397
5	8600000US10004	10004	ZCTA5 10004	3089	1083	35.1	127448	30284
6	8600000US10005	10005	ZCTA5 10005	7135	3272	45.9	117885	13183
7	8600000US10006	10006	ZCTA5 10006	3011	1317	43.7	111617	14445
8	8600000US10007	10007	ZCTA5 10007	6988	2033	29.1	191900	47204

3. *Examine the attribute table of the ZCTAs.* Close the dbf file, exit your spreadsheet software and maximize QGIS. Select the ZCTA layer in the ML, right click and open the attribute table. In the table, note the column labeled ZCTA5CE10. It contains the same FIPS code (the ZCTA number) that was stored in the GEOID2 column in the data table. Since these columns are the same, we can use them to join the two files. Close the table.

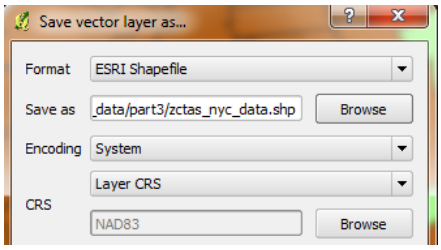


4. *Join data table to shapefile.* Use the browser to browse to your part3 data folder. Select the DEMOG_DATA.DBF data table and drag it into your project, or right click and choose Add Layer. It should appear in the ML. You can select it in the ML and hit the  open table button to verify that the table displays correctly. If all looks good, close the table, and double click on the ZCTAS_NYC_CLEAN layer to open its properties menu. Hit the Joins tab. Hit the green plus button to add a join. The join layer will be the data table DEMOG_DATA. The Join field in that table is GEOID2. The Target field in the ZCTA layer is ZCTA5CE10. Hit OK. Close the properties menu. Right click on ZCTAS_NYC_CLEAN in the ML and open the attribute table. Scroll over to the right, and you'll see all of the layers attributes and the data that is stored in the dbf file. Close the attribute table.



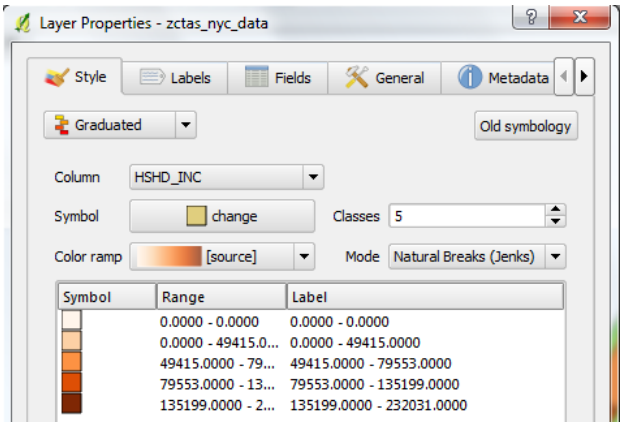
5. *Work-around for joining tables.* Dynamic joins, where layers can be loosely coupled to data tables in a join, were introduced in QGIS 1.7. Unfortunately this feature has a bug; even though the data is joined successfully, we

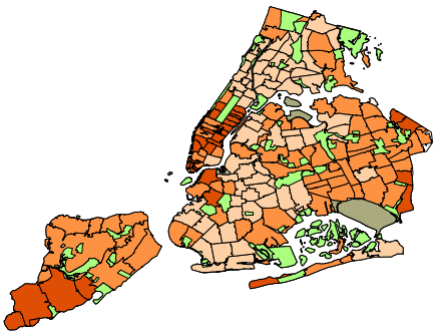
won't be able to classify the data in the table properly in order to symbolize or map it. In order to get this to work we're going to have to create a new shapefile where the data from the table becomes permanently fused to the shapefile. So, select ZCTAS_NYC_CLEAN in the ML, right click and choose Save As. Browse to the part 3 data folder and Save the layers as ZCTAS_NYC_DATA. Leave the encoding and the CRS alone, the latter will be saved as NAD 83. Hit OK. Once it's been saved hit the Refresh refresh button above the browser and use the browser to add the new ZCTAS_NYC_DATA layer to the project.



6. *Reorder the layers.* Select the ZCTAS_NYC_CLEAN layer in the ML, right click and remove it. Also, remove the data table DEMOG_DATA. Drag the new ZCTAS_NYC_DATA layer to the bottom of the ML. Save your project.

7. *Map the income data.* Now that the data is joined to the boundaries, we can map it. Double click the ZCTAS_NYC_DATA layer in the ML and go to the Style tab. Change the Legend type dropdown from Single symbol to Graduated. Change the Classification field to HSHD_INC. Change the mode from Equal Interval to Natural Breaks. In the Color ramp drop down select a scheme that has a range of single-color values that go from light to dark. Hit the Classify button, and then hit OK. You should now have a choropleth (shaded area) map that shows median household income for each ZCTA, classified by natural breaks (divides data into categories based on gaps in values). We'll discuss color and classification schemes in more detail later on. Turn the GREENSPACE and FACILITIES layers on to cover up areas devoid of ZCTAs or portions of ZCTAs that are non-residential. Save your project.





3.3.2 Commentary

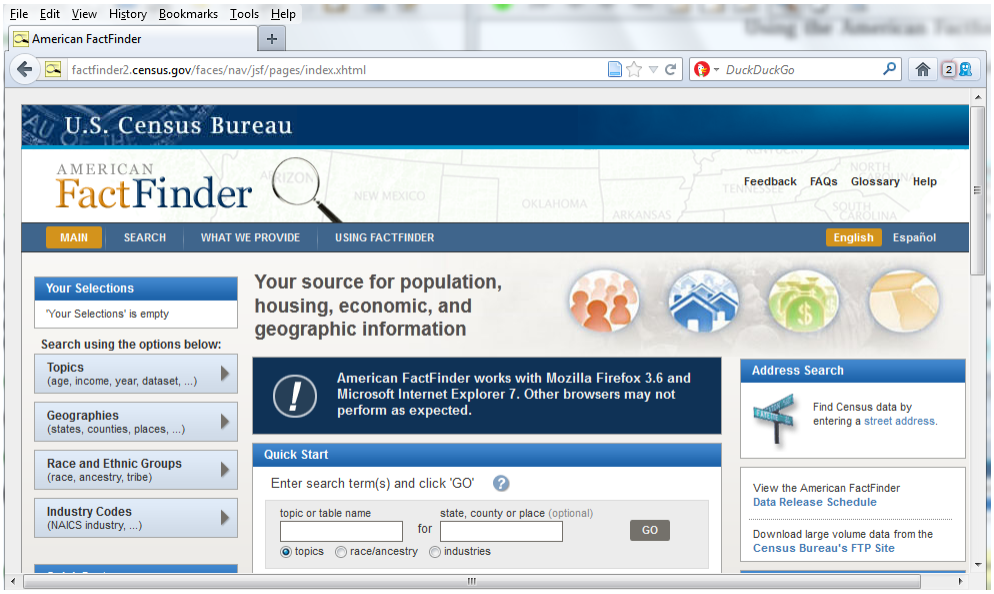
Census Data

The demographic data used in this exercise comes from two US Census Bureau datasets: the 2010 Census (for data on age and gender) and the American Community Survey (ACS - for data on income). Most people are familiar with

the ten-year census, which is a 100% count of the population mandated by law to reapportion seats in Congress. The ACS is an annual sample-survey of population characteristics. Each year the census publishes annual ACS estimates for all geographic areas in the US that have at least 65,000 people. Estimates are at a 90% confidence interval and are published with a margin or error. Since the survey results for areas with smaller populations are often not statistically significant, the bureau averages data over several years for smaller areas. Data for areas that have at least 20,000 people is averaged for a 3-year period, and areas with less than 20,000 people down to the census tract level are averaged for a 5-year period. Each year the bureau releases a new annual data set and updates the two averaged series by adding the latest year of data and dropping the oldest one. For our exercise, we are using 5-year average data from 2007-2011, as that's the only ACS series that is published at the ZCTA level.

The American Community Survey was designed to provide data on a frequent basis and to replace the form on the decennial census that collected detailed socio-economic characteristics of the population. Beginning with the 2010 Census, the decennial census only provides basic demographic indicators of the population such as age, gender, race, and the total number of households and housing units. The decennial census is a count (not a survey) of the population and continues to be useful for making historical comparisons, providing a baseline for creating estimates, for doing analysis below the census tract level, and for providing exact counts when estimates aren't suitable.

A third data product, Population Estimates, is published annually and is created using demographic calculations (as opposed to a count or survey) based on births, deaths, and migration. Basic estimates (total population, age, gender, race, and housing units) are published for states, counties, incorporated places, and metropolitan areas.



All the datasets from the US Census are available for download from the bureau's American Factfinder data portal at `http://factfinder2.census.gov`. All of the data is free and in the public domain. When you download the data you may have to process it to aggregate certain variables before you can use it. The data that we are using in this exercise has been preprocessed to aggregate certain columns and delete unnecessary ones.

Census data from other countries may be more difficult to obtain, as is may not be free or in the public domain, may not be documented in English, and may not be available in a digital format. You can check the website of the statistical agency for an individual country to see what is available, or you can visit the websites of international organizations like the United Nations or World Bank to obtain basic population data for all countries. The US Census Bureau also publishes the International Data Base, which has a variety of demographic indicators for each country.

The decision of which census variables to examine in this study was made by consulting psychographic data and market research reports. This data is generated by marketing surveys to determine which groups of people are interested in products or activities relative to other groups based on age, gender, race, occupation, education level, and geographic location. The census data for this exercise was chosen based on statistics from the Mediamark Reporter; a series of psychographic reports published in a database called MRI+. This data is not freely or publicly available - you would have to access it through an organization that subscribes to the database.

Identifiers

The ability to join data tables in a database or a data table to a shapefile is made possible by the use of identifiers, which are codes used to uniquely identify features. If features in two separate data tables share the same identifier, those data tables can be matched or joined together based on that common identifier, allowing you to create new data or to map data in a table.

There are several standard codes for identifying features. In the United States, ANSI / FIPS (Federal Information Processing Standards) codes are a classification system for identifying all legal, administrative, and statistical areas in the country. For example, FIPS 36061 is the FIPS code for New York County (Manhattan). The first two digits are the code for New York State, while the last three digits are the unique code within New York State for New York County. In an attribute table these codes may appear in separate columns (state, county) or in a single column as one string.

A list of US ANSI / FIPS codes for states and territories is available in the appendix of this tutorial, and the US Census Bureau maintains lists of codes on its website: `http://www.census.gov/geo/www/ansi/ansi.html`

The US government has also created two-letter alpha FIPS codes for each of the world's countries and uses them for international data published by various agencies. However, international data is more commonly coded with ISO codes (ISO 3166) which are available in a two-letter alpha format, a three letter alpha format, and a three-digit numeric format.

Sample Country Codes

Country	FIPS 10	ISO 3166		
Denmark	DA	DK	DNK	208
Djibouti	DJ	DJ	DJI	262
Dominica	DO	DM	DMA	212
Dominican Republic	DR	DO	DOM	214

It is generally best practice to store ID codes as text and not as numbers since they don't represent quantities. Storing ID codes as numbers can result in data loss and misidentification. If codes begin with a value of zero and the ID is stored as a number, the zero will be dropped and the code will be incorrect (i.e. imagine you have a file with US ZIP codes and all ZIP codes that begin with zero are truncated).

In order to join two tables together based on an identifier, you need to be sure that each field is stored in the same data format; if one is stored as text and the other is numeric, the join will fail. Furthermore, you need to insure that each record is unique because one to many joins are not allowed; if you have a data table that has multiple records for one country, only one of those records will be joined to a shapefile and the others will be dropped. Finally, you should never use place names as identifiers or join fields because there are often many inconsistencies (imagine the number of different ways for spelling or abbreviating country names like the United States or South Korea).

Adding or appending identifiers to tabular data that lack this information is a common data processing task that you'll likely have to perform.

Tabular Data: DBF Files

DBF files are an old data table file format from a database system called dBase that was once common in several database systems. While many of these systems are no longer widely used the file format has survived, in part because dbf files are a component of shapefiles that store all of the attributes of features. QGIS is able to take data stored in standalone dbf files and join them to dbfs affiliated with shapefiles based on a common ID code, using basic relational database techniques (a SQL join statement).

Important things to note about DBFs:

- You can view and create DBF files in spreadsheet programs such as any version of LibreOffice or OpenOffice Calc and versions of Microsoft Excel between Office 97 and Office 2003. You can save text files and spreadsheets as DBF in these programs by using Save As and selecting DBF as the option.
- You can open or import DBF files with Microsoft Office 2007 and 2010, but you cannot save changes or create new files because Microsoft decided to deprecate the DBF format. However, several plugins have been developed that allow you to work with DBFs in the newer versions of Office, and you can download these from the web.
- DBF files are VERY particular - names for columns must be kept short (less than 10 characters), should contain no spaces or punctuation (except underscores), and cannot begin with numbers.
- Unlike plain text files, columns in a DBF table have a specific data type associated with them (text, integers, real numbers, etc). In order for joins between DBF files and shapefiles to work, the ID fields must be in the same format - text or numbers - IDs should normally be stored as text.
- You can open and edit DBF files that are associated with shapefiles. However - you should NEVER EVER re-sort the data in a DBF file that is associated with a shapefile - if you do, the data will become misaligned with the features in the shapefile and will no longer match. You also CANNOT add new rows to the DBF, since there will be no geometry in the shapefile to match it. You can edit existing values, add new columns, and delete columns (as long as you don't delete the ID fields at the beginning of the sheet!)
- If you need to do substantial editing of a stand-alone DBF file that is NOT part of a shapefile, it is best to copy all of the data in the dbf and paste it into a new, blank workbook in Excel or Calc format (xls or odt). For example, if you want to create a calculated field with percent change or do ANY work that involves formulas, create a new blank workbook - DO NOT work in the dbf file and do not create a second worksheet within the DBF - DBF can only support single worksheets. Once you finished doing the work in the spreadsheet file, do a copy and paste special in another workbook, pasting only values - no formulas or formatting. Then you can save that sheet as a new DBF file.

CSV files are the other officially supported alternative for getting stand-alone data tables into QGIS. We'll cover these later in the tutorial.

Tabular Data: Spreadsheet Files

Although they are not officially supported, the latest builds of QGIS 1.8 allow you to add individual sheets from spreadsheets as tabular data that you can join to shapefiles and manipulate within the software. If you use the  Add vector data button and change the filter drop-down to view all files in a folder, you can add data from older Excel files (that end with .xls) or LibreOffice Calc files (.ods). You select the spreadsheet, and are then prompted to select an indivdual sheet within the workbook to add to the project. In all likelihood these formats will be officially supported in QGIS 2.0.

3.4 Plotting Coordinate Data

In this section you'll learn how to take a text file with coordinate data, plot the data in GIS, and convert it to a shapefile. It's often difficult to find pre-existing shapefiles of buildings, particularly businesses and residences. But you can create your own point layers if you have the coordinates of the places you wish to plot. In this exercise you'll create a layer of coffee shops from a text file that lists each store with its latitude and longitude coordinates.

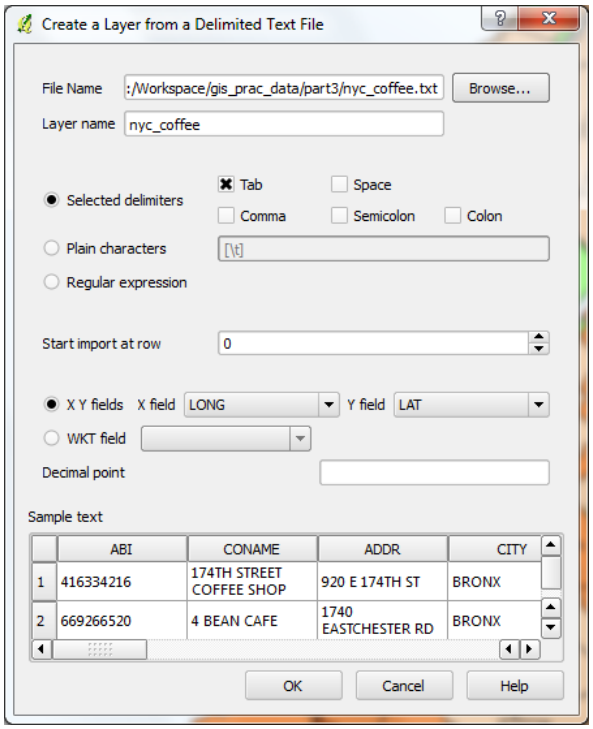
The coordinate data for the coffee shops was downloaded from a database called ReferenceUSA and processed so that it was ready for plotting. Please note that this data is from December 2012 and is used as a teaching example; it should not be used for any commercial purpose.

3.4.1 Steps

1. *Inspect the text file.* Go to your data folder for part 3, open the file NYC_COFFEE.TXT in a text editor (like Notepad on MS Windows) and examine it. This is a tab-delimited text file with data for coffee shops in NYC; each record represents one store and each attribute column is separated by a tab. Close the file when you're finished.

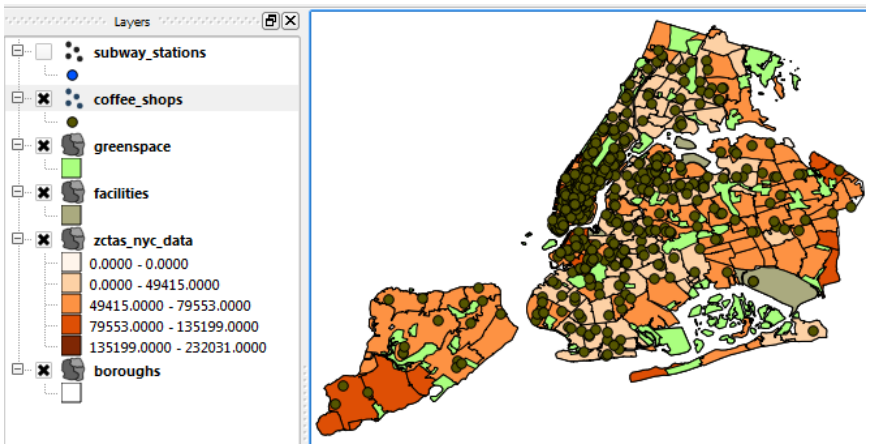
2. *Activate the delimited text plugin.* In QGIS, make sure that this plugin is activated under Plugins > Manage Plugins by checking it off in the list, and that the plugin toolbar is visible by right-clicking an empty area of the toolbar and checking the plugin box.

3. *Launch the delimited text plugin* Click the  add delimited text button or launch it from the Plugins menu. For the delimited text layer browse to the part 3 data folder and select NYC_COFFEE.TXT. Accept the default layer name. Choose the Selected delimiters radio button and the Tab checkbox. Make sure that by XY Fields the X field is Long and the Y field is Lat. Hit OK. When prompted for a CRS, keep NAD 83 as the default and hit OK.



4. *Convert the plot to a shapefile.* Even though the points have been plotted, it isn't a shapefile yet. To convert it, select and right click on NYC_COFFEE.TXT in the ML and choose Save As. Save it as an ESRI shapefile in your part 3 data folder and call it COFFEE_SHOPS. Keep NAD 83 as the default for the CRS.

5. *Add the new coffee layer.* Hit the Refresh refresh button above the browser and add the new COFFEE_SHOPS shapefile to your project. Drag it to the top of the ML. Then select the original text file in the ML, right click and remove it. Save your project.



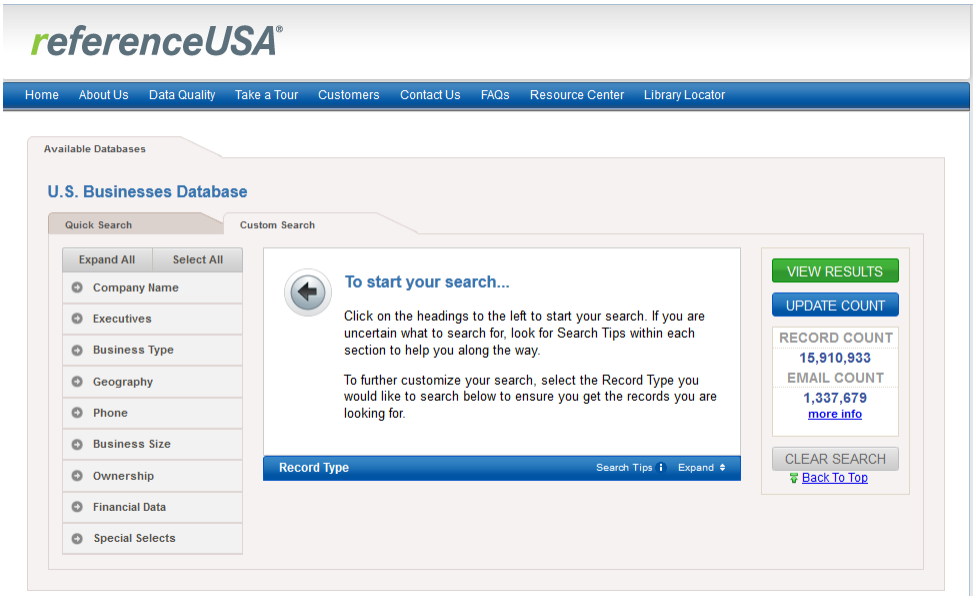
6. *View the attribute table.* Select the COFFEE_SHOPS layer in the ML, right click and open the attribute table, to take a look at what's there. You should see all of the data that's affiliated with the coffee shops. Close the table when you're finished.

3.4.2 Commentary

Coordinate Data Sources

While government agencies often create and provide geographic data for boundaries and physical features, private features like businesses are usually not captured. These datasets must often be purchased or created from address or coordinate data. ReferenceUSA is not a freely available resource, but it is commonly held by many academic and public libraries. You can search for businesses by name, industrial classification code, and geography and download the data in spreadsheet format; although the number of records you can access in one download is limited. They provide comprehensive business, health care, and residence data for the US and Canada. The inclusion of XY coordinates (longitude and latitude) for each record makes it possible to plot the data in GIS.

Ultimately the outcome of this exercise is only as good as the input; when downloading this type of data you must scrutinize it to make sure that you capture as many records that meet your criteria as possible, while removing ones that do not. Don't accept the data as is - consider it as being raw data that you must analyze and clean before bringing it into GIS. For example, in assembling the coffee shop dataset for this exercise many businesses self-identified (based on SIC or NAICS codes) as coffee shops, but based on the name of the business they were actually cafes or diners; in New York the term coffee shop is often synonymous with diner (think of the "coffee shop" in Seinfeld episodes). As a diner is not what we had in mind, these records had to be removed. At the same time, a popular local chain of coffee shops was missing in the initial set of records - subsequent investigation revealed that they identified themselves primarily as coffee roasters and not as retail shops, even though they engage in both activities. The records for these stores were subsequently added.



There are also free, public sources for downloading coordinate data that you can use to create features for natural (lakes, mountain peaks, parks, etc.) and human-made (cities, airports, schools, cemeteries, etc.) features, such as the USGS Geographic Names Information System (for US features) and the NGA's GEOnet Names Server (for international features). The subway stations layer was created from coordinate data provided by the NYC MTA. If you have batches of addresses, you can look-up or assign coordinates to them by using a geocoding service such as the Geocoding Services at the Texas A&M GIS lab at `http://geoservices.tamu.edu/`.

Delimited Text Files

A text file is a plain document format that is often used for storing and sharing data. Since it is relatively simple and contains no formatting it is cross platform and historically stable. The attributes of each record are separated by a delimiter to indicate different fields. This allows spreadsheet and database programs to parse the text file into columns when you open or import it into that software. Common delimiters include commas, tabs, and pipes. The disadvantage of text files is that the fields are not associated with a specific data type; unlike a DBF file where a field can be designated as a string, integer, real, or other type. When importing text files you need to be careful that columns are designated correctly during the import process; strings inadvertently stored as numbers may have zeros dropped, while numbers inadvertently stored as strings cannot be treated mathematically. Depending on the source of the text files, fields that are intended to be strings may be surrounded by quotes, so that software can recognize and import those fields correctly.

3.5 Running Statistics and Querying Attributes

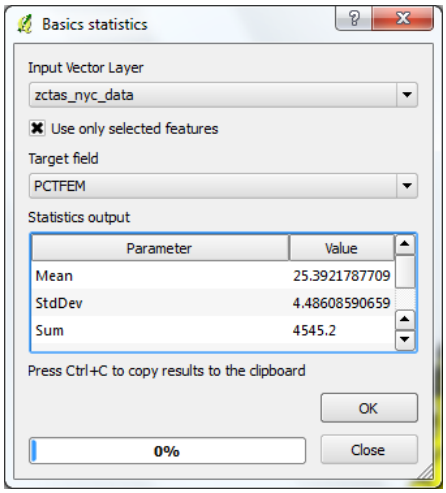
In this section you'll learn to calculate basic statistics for attributes and use some of the advanced query features. Now that all of the data is in place, we can begin to remove ZCTAs that don't meet our site selection criteria. We want to target areas that don't have a large number of existing stores, that have a high percentage of women aged 18 to 49, and that are middle-income.

3.5.1 Steps

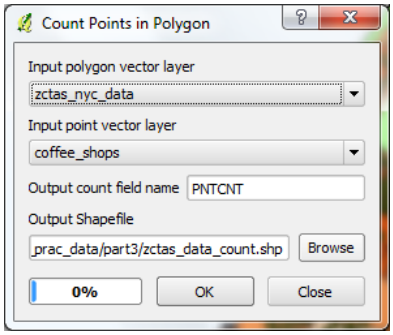
1. *Examine the age distribution.* Open the attribute table for the ZCTAS_NYC_DATA layer. Click on the PCTFEM column heading to sort the data, then examine it. Besides the large number of non-residential ZCTAs where the population is zero, the distribution seems relatively normal. Before we run any statistics we will want to exclude

the non-residential areas. Sort the data by the TOTPOP column. Click on the record where the population is 1,685 by clicking on its record number on the far left-hand side of the screen (number 32). Scroll down to the bottom of the table and **while holding down the SHIFT key** select the last record in the table (number 210). This will select all the records in-between.

2. *Run some basic statistics.* Close the attribute table. On the menu bar select Vector > Analysis Tools > Basic Statistics. Choose ZCTAS_NYC_DATA as the input vector layer. Make sure the box that says Use only selected features is checked. Change the target field to PCTFEM. Hit OK. You'll see that the mean percentage is approximately 25.4% and if you scroll to the bottom, you'll see the median is 24.8%. For the purpose of our example, we'll use 25% as our cut-off; ZCTAs where 18 to 49 year old women make up 25% or more of the population will be included, while any with less than that number will be excluded. Close the stats menu. Hit the  Clear selected features button.



3. *Count stores by neighborhood.* We should exclude ZCTAs that already have a very large number of coffee shops. On the menu bar go to Vector > Analysis Tools > Points in Polygons. Specify ZCTAS_NYC_DATA as the Input polygon layer and COFFEE_SHOPS as the Input point layer. Keep the output count field name as PNTCNT. Browse to your part 3 data folder and save the output as ZCTAS_DATA_COUNT. Hit OK to create the new shapefile. Say Yes and add the file to the project. Close the Point to Polygon menu.



4. *Swap your layers.* Select the ZCTAS_NYC_DATA layer in the ML, right click and remove it. Drag the new ZCTAS_DATA_COUNT layer to the bottom of the ML. Don't worry about symbolizing the new layer.

5. *View the table for the new layer.* Select ZCTAS_DATA_COUNT in the ML, right click and open the attribute table. Scroll the table all the way to the right. You'll see the new PNTCNT field, which shows the number of coffee

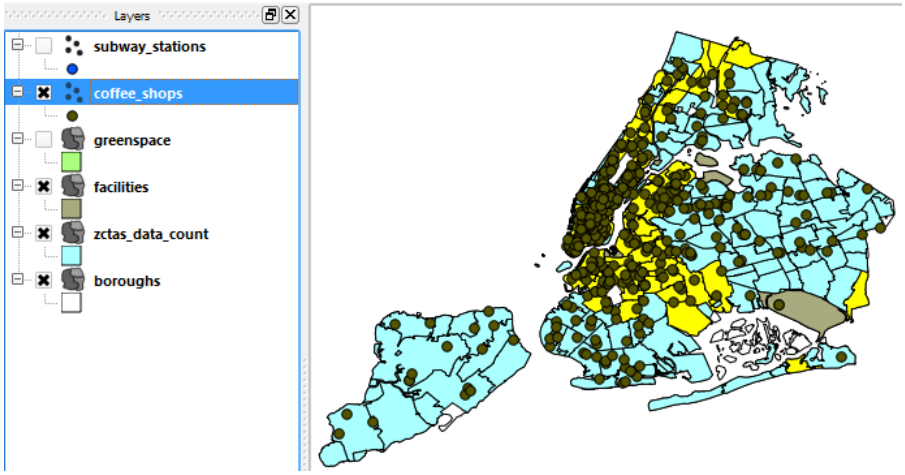
shops in each ZCTA. Click on the PNTCNT column heading to sort the table by that field. You'll see there are a number of ZCTAs that have many coffee shops, but there is a small gap at the top of the distribution between 17 and 14 stores; we can use this as a cut off.

TOTPOP	FEM18_49	PCTFEM	HSHD_INC	MARGIN_E	PNTCNT
16575	5281	31.9	105974	4474	27
56024	20058	35.8	88601	5397	25
21102	6810	32.3	67795	6489	23
42870	12897	30.1	79508	5099	20
24711	6931	28	62997	6288	19
31924	8193	25.7	109625	8945	17
50984	14946	29.3	99700	3163	14

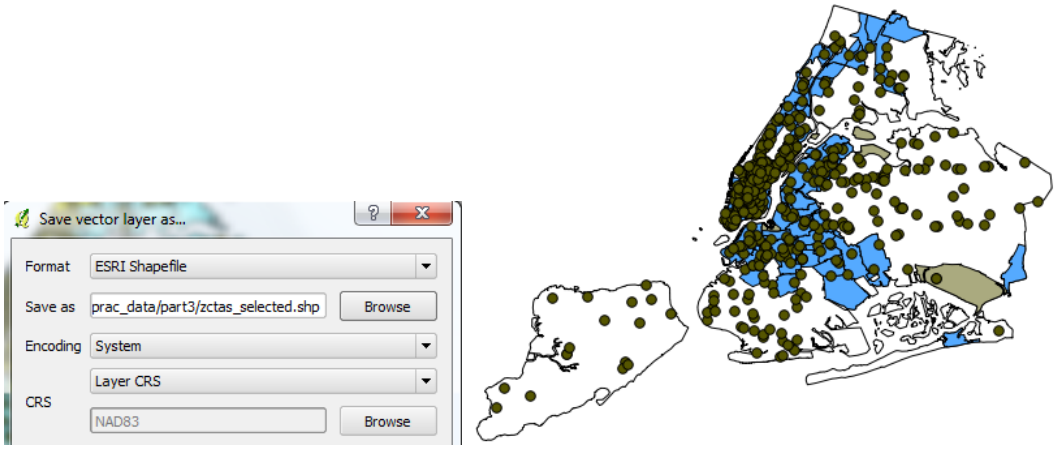
6. *Build an advanced query.* Hit the Advanced Search button in the lower right-hand corner of the attribute table menu. In the Fields box scroll down and doubleclick PCTFEM to add it to the expression. In the Operators box click greater than or equal to (>=). In the SQL where clause box type in the value 25. Hit the AND button in the Operators box. Double-click the HSHD_INC field in the fields box. Hit the greater than or equal to button (>=) in the operators box. Type the value 25000 directly in the SQL Clause box. Follow the same steps to add two more conditions: exclude household income greater than $99,999 and exclude areas where the PNTCNT field is greater than 14. Your final statement should read:

   PCTFEM >= 25 AND HSHD_INC >= 25000 AND HSHD_INC < 100000 AND PNTCNT < 15.

   The Census Bureau classifies income into specific brackets when it publishes data; the values we have used exclude the three lowest and three highest brackets. Hit the Test button to test your statement - you should have 55 features selected as a result. Hit OK. Close the attribute table to view your selections in the map view.



7. *Save your selection as a new shapefile.* Select ZCTAS_DATA_COUNT in the map legend (ML). Right click and choose the Save selection as option. Browse to your part 3 folder and save the selection as ZCTAS_SELECTED. Keep the defaults. Hit OK to save it. Hit the Refresh refresh button above the browser and add the new ZCTAS_SELECTED layer to your map. Select the old ZCTAS_DATA_COUNT in the ML, right click and remove it. Drag the new ZCTAS_SELECTED layer to the bottom of the ML. Save your project.



3.5.2 Commentary

Selection Criteria

Since the goal of our exercise is to demonstrate the capabilities and possible uses of GIS, we're not adhering to strict criteria in our site selection process; the example is merely illustrative. Is a cut off of 25% of the population for women aged 18-49 reasonable? It really depends on your goals, and whether you would prefer to have a focused, narrow selection of places or a more expansive one. Does it make sense to omit a ZCTA that is only a tenth of a decimal place below 25%? These are the kinds of decisions you'll have to make for each project you do. You may decide that a line has to be drawn somewhere and that's it, or you may wish to allow an exception within a few decimal places or to round your numbers. You also could decide to make a qualitative decision - based on what you know about the neighborhood that's near the dividing line, should you include it or exclude it?

You have a few tools at your disposal for making these decisions; the basic statistics for determining mean, median, range, and standard deviation to establish a baseline are helpful. The data classification tools for symbolizing your data based on quantiles or equal intervals can also aid your decision (we'll discuss these later on). Regardless of what you do, look at the attribute table and make sure to examine your data to see what the distribution looks like. It also helps to become familiar with the places you are studying, so you can draw on your more qualitative experiences to make decisions and perform a "reality check" on your observations.

Some Basic SQL

The advanced selection menu under the attribute table allows you to build complex queries for selecting features. QGIS, and most GIS packages, use the Standard Query Language (SQL) that's used when working with databases. Some tips:

- The boolean operator AND is exclusive; use it to select features that meet all of the criteria; the statement PCTFEM >= 25 AND PNTCNT < 15 will only select features where both criteria are met.
- The boolean operator OR is inclusive; use it to select features that meet one of the criteria; the statement PCTFEM >= 25 OR PNTCNT < 15 will select features that meet the first criteria, or the second one, or both.
- Your statements must be explicit; for every operation you must include the field that is part of the operation: PNTCNT > 14 AND PNTCNT < 16 is a correct statement. PNTCNT > 14 AND < 16 will yield an error, because you didn't specify the field for the second operator.
- Statements can be written more than one way. In our example above, PNTCNT > 14 and PNTCNT >= 15 would yield the same result, since the number of coffee shops is saved as an integer.

- If your query includes text rather than numbers, all text must be surrounded by 'quotes', otherwise you'll get an error. BCODE='36081' will return all ZCTAs in Queens. You can also use wildcards. BCODE LIKE '3608*.*' will return all the ZCTAs in Queens ('36081') and Staten Island ('36085').

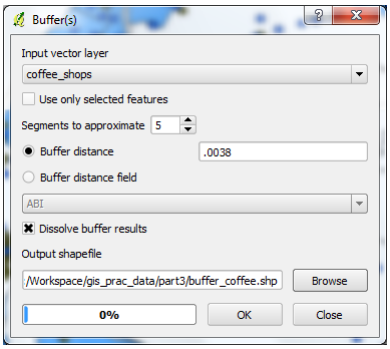
3.6 Drawing Buffers and Making Selections

One of the primary strengths of GIS is the ability to layer different features and to combine or extract information to create new features. In this section you'll learn how to create buffers around features and to deduct areas from selections. For our example, we want to target areas that are near subway stations since these represent traffic and commercial activity, while avoiding areas where competitors exist. We'll identify these optimal areas within the ZCTAs that met our conditions.

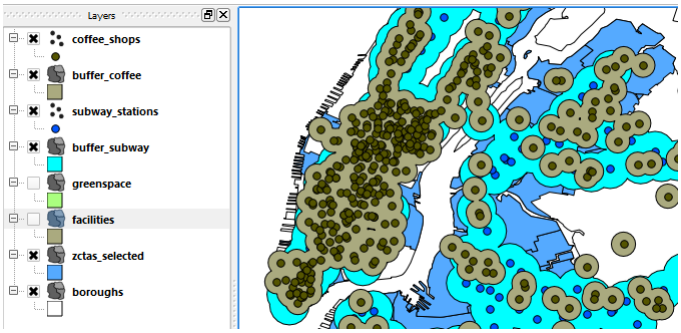
3.6.1 Steps

1. *Activate and de-activate layers.* Hit the check boxes beside the SUBWAY_STATIONS and BOROUGH layer to turn them on, and drag the BOROUGH layer to the bottom of the ML. Uncheck the GREENSPACE AND FACILITIES layers to turn them off. Make sure the subway stations and coffee shops have a different color so they can be easily distinguished.

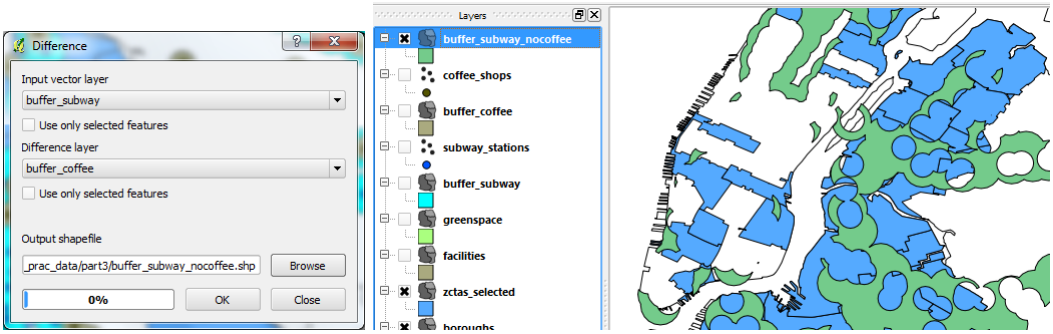
2. *Create buffers.* On the menu bar, go to Vector > Geoprocessing tools > Buffers. Specify the coffee shop layer, COFFEE_SHOPS as the input vector layer. For the buffer distance, type .0038 (this is in degrees and represents approx 1/5 mile; see commentary below for explanation). Check the box that says Dissolve buffer results. Hit the browse button to save the new shapefile in your part 3 folder as BUFFER_COFFEE. Hit OK. Click Yes to add the new layer. Repeat the process for the SUBWAY_STATIONS, but specify a buffer distance of .0062 (1/3 mile) and save the new layer as BUFFER_SUBWAY. Close the buffer menu when you're finished.



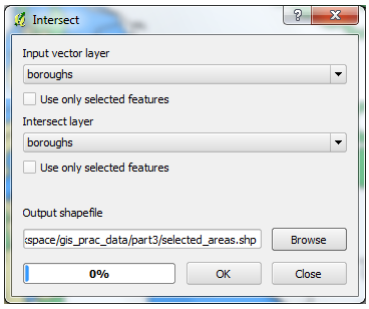
3. *Re-arrange layers.* To see everything more clearly arrange the layers in the ML in this order from the top: COFFEE_SHOPS, BUFFER_COFFEE, SUBWAY_STATIONS, BUFFER_SUBWAY. The two layers at the bottom of the ML should be ZCTAS_SELECTED and BOROUGHS, and the layers in-between should be unchecked and off. You may want to assign different colors to the buffers to make them more visible. Explore the map; you'll see circular zones in a 1/3 mile radius around each subway station and 1/5 mile radius around each coffee shop. The boundaries between each buffer zone are merged where zones intersect (as a result of checking the dissolve results box).



4. *Subtract coffee areas from subway areas.* We are interested in areas that are close to subway stations, unless those areas happen to be near existing coffee shops. To isolate these areas we'll use the Difference tool. On the menu bar go to Vector > Geoprocessing Tools > Difference. Select BUFFER_SUBWAY as the Input vector layer, BUFFER_COFFEE as the Difference layer, and Browse and save the new file in your part 3 data folder as BUFFER_SUBWAY_NOCOFFEE. Hit OK. When prompted to add the layer to the project, say Yes. Close the difference menu.

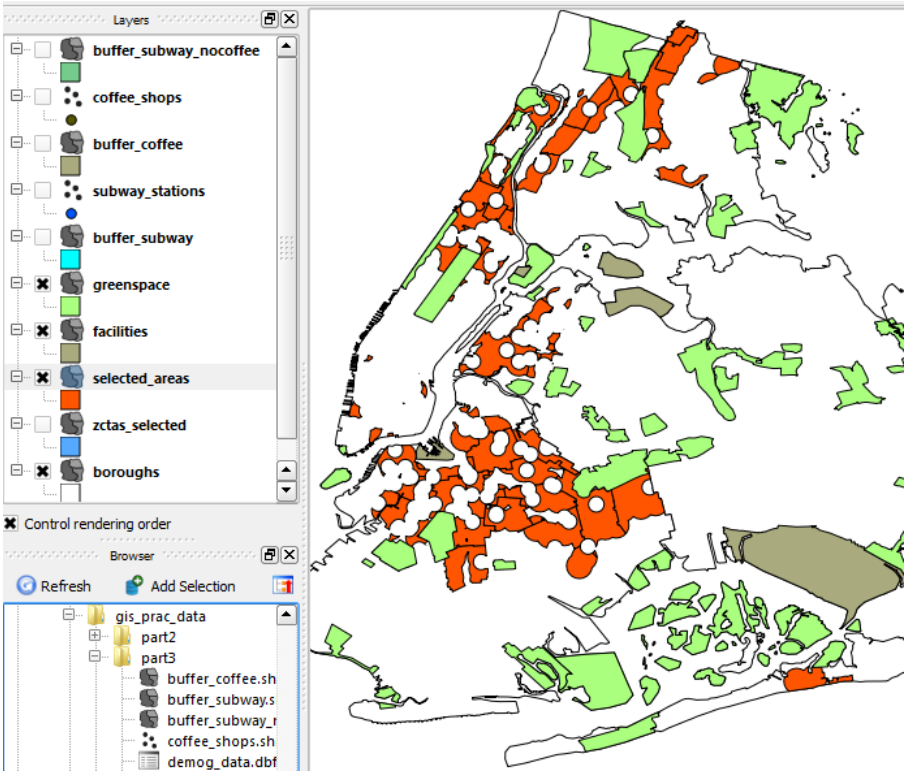


5. *Isolate areas within selected ZCTAs.* Drag the new BUFFER_SUBWAY_NOCOFFEE layer to the top of the ML, and turn off the SUBWAYS, COFFEE SHOPS, AND BOTH BUFFER layers so you can clearly see the result. Ultimately we are interested in the areas around subway stations that don't have coffee shops that are within the neighborhoods that met our criteria. To isolate the former within the latter, we'll use the Intersect tool. On the menu bar, go to Vector > Geoprocessing tools > Intersect. Choose ZCTAS_SELECTED as the input vector layer. Choose BUFFERS_SUBWAY_NOCOFFEE as the intersect layer. Browse and save the new result to your part 3 data folder as SELECTED_AREAS. Hit OK. Close the Intersect menu.



6. *Clean up your map.* Drag the SELECTED_AREAS layer so that it is directly above the BOROUGHS layer in the ML. Check the NYC_FACILITIES AND NYC_GREENSPACE layers to turn them back on, and uncheck the BUFFER_SUBWAY_NOCOFFEE AND ZCTAS_SELECTED layers to turn them off, so you can see the end result more clearly. We could refine our analysis a bit more by subtracting the greenspace and facilities areas that intersect our areas of interest, since we couldn't

build a store on this land. For now, overlaying these land uses on top of our areas of interest should suffice. Select the SELECTED_AREAS layer in the ML, then hit the  zoom to layer button so our areas of interest are maximized within the map window. The new SELECTED_AREAS layer shows you the areas to consider targeting: areas within a 1/3 mile of a subway station but not within 1/5 of a mile from an coffee shop that are located in middle-income ZCTAs where women aged 18-49 represent 25% or more of the total population and there aren't a large number of existing shops (more than 14).



7. *Identify areas.* Through the selection process, the attributes of our previous layers have been preserved in our new layers. Select the SELECTED_AREAS layer in the ML. Use the  identify button and click on one of the areas. You'll see the attributes from our earlier ZCTA layer. While the identifying information is useful, many of the other attributes are now incorrect. The population figures represent the entire ZCTA and not the small subset we've selected. If we were going to save these layers for future analysis or projects, we would want to delete the attributes that are no longer necessary or that are incorrect.  Save your project.

3.6.2 Commentary

Buffers and Distance Measurement

Since the coordinate system of our layers is NAD 83 and it uses degrees of latitude and longitude, we have to specify units for measuring the distance of our buffers in degrees. This is difficult for a number of reasons; it's much easier for us to conceive how large a kilometer or mile is relative to a degree. A thornier issue is that the length of a degree

isn't constant - the distance between degrees of longitude decreases as we move from the equator to the poles. The distance between degrees of latitude is relatively consistent, but is also not equal to a degree of longitude, which requires us (or software) to make complex calculations to transform degrees into simple distance measurements. Here are some ways to get around this problem when creating buffers:

- Do some math and estimate. A degree of latitude at 40 degrees latitude is approximately 53 miles. If we want to draw a 1/3 mile buffer, divide 53 by .33 to get 160.6, then divide 1 degree by 160.6 to get approximately .0062 degrees. So for our example a third of a mile is approximately .0062 degrees (tables for converting degrees to miles and kilometers are available in the appendix of this tutorial).
- Use trial and error. Create a buffer in some unit of degrees, then measure the buffer using the measuring tool to see what it is in miles or kilometers. Then try drawing the buffer again, and base the number of degrees on your previous observation.
- Transform your layers to another coordinate system. Some coordinate systems use units like meters or feet instead of degrees, which would allow you to skip the unit conversion process all together. You can use the appropriate Universal Transverse Mercator (UTM) zone for the area you're studying, and the units will be in meters. In the United States you can also use a State Plane system which is in meters or feet. We'll cover projections and coordinate systems in the next part of this tutorial.

Transforming the layers to an appropriate coordinate system that uses meters or feet is the best approach when drawing buffers or performing any distance calculations. For this example we stuck with NAD 83 for simplicity's sake; it was the original coordinate system for the layers, and we needed to use a CRS that used degrees in order to plot the coffee shop data (which was stored in degrees of latitude and longitude). But ideally, the best CRS for mapping our area and measuring distances would be the local State Plane zone, NAD 83 / New York Long Island (ft US).

In our example we chose to dissolve the boundaries of the buffers where they intersected because we were interested in the total area within 1/3 mile of any subway station. The resulting shapefile consisted of a single feature - the entire buffer. What if we wanted to preserve the individual boundaries of each buffer? We would leave that Dissolve box unchecked. The resulting shapefile would consist of several features, one buffer for each station, AND each feature would take the attributes of the station it surrounds (i.e. the station id, name, trains, etc).

Site Selection

Site selection theories and land use analysis can be traced back to the early 19th century with the introduction of Von Thunen's land rent gradient. Subsequent work that included Weber's median location, Hotelling's competitive location problem, Christaller's Central Place Theory, and Tobler's Laws of Geography have provided a framework for the science (and art) of optimal site selection. Optimal site selection is studied within the fields of geography, location science, and operations management, and has expanded with the introduction and evolution of GIS. The three laws of location science, as summarized by Church and Murray (*Business site selection, location analysis, and GIS, 2009*) are:

- Some locations are better than others for a given purpose
- Spatial context can alter site efficiencies (the unique circumstances of a given area can alter whether or not a site is optimal)
- Sites of an optimal multi-site pattern must be selected simultaneously rather than independently, one at a time (if you're planning to open several franchises you should do the planning all at once; as each site you open can impact another)

It's also important to understand the unique spatial patterns of each type of business or industry; a phenomena that economic and urban geographers have been studying for many decades. Products or services classified as low ordered goods tend to be located in most environments, and there will be more of these businesses in places with higher population densities. High order goods tend to require a higher population density and will be present in fewer

locations. For example, businesses like gas stations, dry cleaners, and family doctor's offices will be located in most areas, while office towers, specialty retail, and major hospitals will be located in fewer places, spaced further apart. Businesses like gas stations and convenience stores tend to cluster around major transportation intersections, while car dealerships and hotels tend to cluster around each other in districts. Movie theaters and large shopping malls on the other hand tend not to cluster together; they are spaced apart to serve different populations.

The location of non-retail or non-service industries is also distinct. Manufacturing industries often depend on the availability of raw materials and inputs and the distance for finished products to reach transportation and markets, while hi-tech industries tend to locate near pools of highly educated labor. Agricultural uses often appear where other land uses are not present and where land is inexpensive. The types of crops or livestock they produce will vary based on environmental factors like climate or soil.

We worked with coffee shops in our exercise as they are an interesting example of a low-order good: they are small businesses that sell basic, low-cost products. They have a relatively small footprint for attracting customers and can be located almost anywhere, in the hopes of grabbing foot traffic (coffee drinkers who want to grab something to go). But in addition to this large, general demographic they also appeal to particular groups who are seeking community space, a certain atmosphere, and better than average coffee. These quality and place-centric aspects of the business means they can't be entirely co-opted by other food services that simply sell coffee (like fast food retailers or donut shops), large retailers, or the Internet.

The bottom line: if you are going to conduct a site selection analysis, you must understand the context: study the industry or business you are interested in, do some market research, make sure you're familiar with the geographic environment you're working with, and choose your geographic units of analysis and indicators carefully.

3.7 Screen captures

In this brief section you'll learn how to create a screen shot of your map that you can easily share with others. You'll learn how to make a presentation quality map in the next part.

3.7.1 Steps

1. *Zoom to layer.* With the SELECTED_AREAS layer selected in the ML, hit the  zoom to layer button. Use the  hand tool to center the map view. If you want to be fancier, you could activate some plugins under Plugins > Manage Plugins and add a north arrow, scale bar, and copyright info to the screen, and the  text annotation button to add a title.

2. *Save the map view screen.* On the menu bar, go to File > Save as Image. Browse to your data folder for part 3 and save the image there as MAP_SCREEN. Change the Files of Type dropdown to PNG file. Click Save.

3. *View your map.*  Save your project and then close QGIS. Navigate to your data folder for part 3. Look for the file MAP_SCREEN.PNG. Double-click it to open the file in your computer's default photo viewing program, and you'll see your map view. This is a quick way to save and share your map content. This is a simple, static image file that is not connected to your project or data files. You can easily email or text this file to anyone.

3.7.2 Commentary

Considerations and Next Steps

Based on our results, what would you do next? How would you decide where to locate the shop? What else would you investigate? Is there anything that we've done in this exercise that you would do differently, if you had to conduct an analysis like this for an actual project?

For more practice, some things to try:

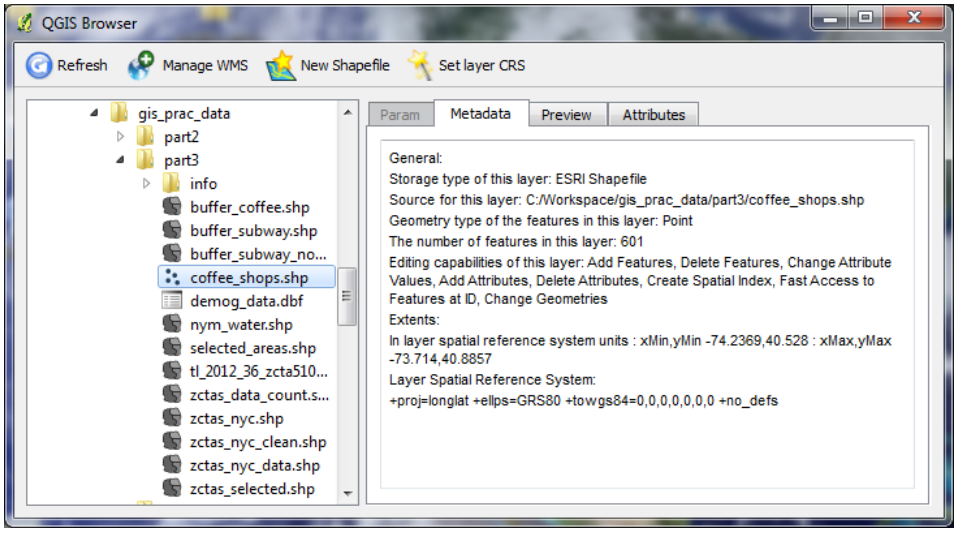
- Expand the selection areas to include ZCTAs where 24% or more of the population are women aged 18 to 49.
- Shrink the selection areas by removing the greenspace and facilities from the final areas (rather tha simply overlaying them on the selection areas).

3.8 QGIS Desktop Browser

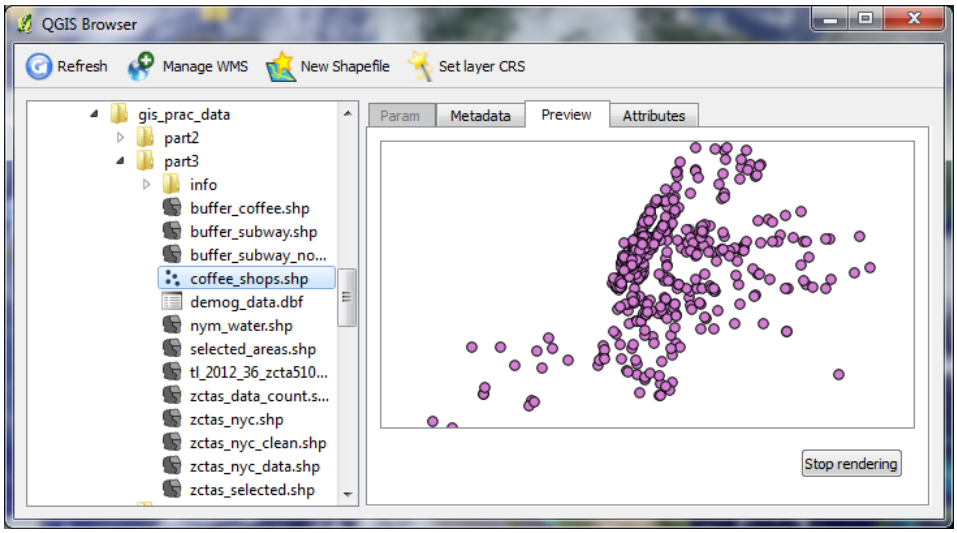
The data browser is a new feature in QGIS 1.8 that makes it easier to manage your files and add them to your projects. In addition to the browser that's embedded in the map interface, there is also a separate browser application, the Quantum GIS Browser, which offers additional features. It is a convenient application for previewing your layers. This section will give you a brief overview by taking a look at the various files we've created throughout this chapter.

3.8.1 Steps

1. *Launch QGIS Browser.* Minimize QGIS Desktop for now and launch the QGIS Browser. If you're using Microsoft Windows, look under the Start Menu > Program Files > Quantum GIS > Quantum GIS Browser. For Mac users, the icon to launch it isn't easy to find. Select QGIS in Applications, and CTRL-Click to browse through the folder contents - the application appears under Contents/MacOS/bin. You will need to create a symlink or Finder alias to make it simpler to launch. A better approach should be implemented in a future QGIS release.
2. *Drill down to the Part 3 Folder.* The folder tree here is similar to the browser that is in QGIS Desktop. You can drill down through folders in your file system or you can view the contents of spatial databases or web services. Drill down to the Part 3 folder to see all of the files we used or created for this chapter.
3. *View the metadata for coffee_shops.* Select the COFFEE_SHOPS layer in the folder tree. The default tab on the right will show you basic metadata for the file, including the type and number of features (601 point features), the spatial extent of the layer in coordinates, and the definition for the CRS.



4. *Preview the coffee_shops.* Select the Attributes tab in the browser to preview the attribute table of the layer. Then, select the Preview tab to preview the geography of the layer. If the geography doesn't render right away, just switch the tab from Attributes to Preview to get the layer to draw.



5. *Experiment with viewing other files.* Select some of the other layers to preview them. The attributes of tabular files like dbf or csv can be previewed, but there is no geometry to display. Clicking on a folder in the browser will display the full contents of the folder. When you're finished exploring, close the browser.

3.8.2 Commentary

File Management

As we've moved through this exercise, we've created many shapefiles along the way; every time we made a selection or performed a geoprocessing function we ended up with a new file. There are two things we should note here.

First, this can get pretty confusing. With each new file you create, it's easy to lose track of what each one represents. You can mitigate this by giving your files names that clearly indicate what they are. Documenting your progress in a logbook, whether it's on paper or in a simple text file, can help you keep things straight. You may also decide to delete files that were created during the middle of the process. This is fine as long as you think you won't need to go back and re-do a step, either because the parameters of your project have changed or you've spotted an error.

Second, it's not always necessary to create a new file with every single processing step. Some menus will give you the option to select features or perform operations on features that are ALREADY selected. This allows you to work with just the features you need from one layer to create a new one, skipping the interim step of creating a new shapefile of just the features you want to work with.

When we've created new layers, we have used underscores instead of spaces when naming files, i.e. ZCTAS_NYC_DATA.SHP. When naming files it's best practice to use underscores instead of spaces and to avoid using any punctuation in file names. This helps to insure compatibility of data across operating systems and to prevent

possible errors when loading or reading data in the software. You should follow the same rules when creating folders to store data. The name of your file should reflect what it contains; you could include the geographic area it covers, the type of feature, and possibly a date or number to indicate different iterations of the data.

The QGIS Browser makes managing and working with your files a bit easier. It filters files in your folders so that only GIS usable files are visible. It also collapses shapefiles to single entries so that it's easier to see what you have. If you have a folder with layers and you're not sure what they are, you can easily use the browser to preview them, rather than having to add and layer them all in the map interface.

Chapter 4

Thematic Mapping

The goal of this chapter is to introduce you to map layout and design, as well as to some additional data processing techniques. You will also grapple with coordinate systems and map projections, which are central components underlying GIS. You'll learn about cartographic representation and design and the practical implications of choosing how to classify and represent your data.

The goal of this particular exercise is to create a stand-alone thematic map to show the distribution of employment in the health care sector by state in the United States. The data we'll use covers both public and private sector employment from 2010 and comes from the US Bureau of Labor Statistics Quarterly Census of Employment and Wages (QCEW) at `http://www.bls.gov/data/`.

4.1 Transforming Map Projections

This section will show you how to transform a file from one map projection to another, and how to define a custom projection. Choosing a coordinate system and map projection for your layers is of critical importance; all layers in a project need to share the same system in order to work together, and the choice of a system is influenced by the type of analysis you're doing and what your final map will depict.

4.1.1 Steps

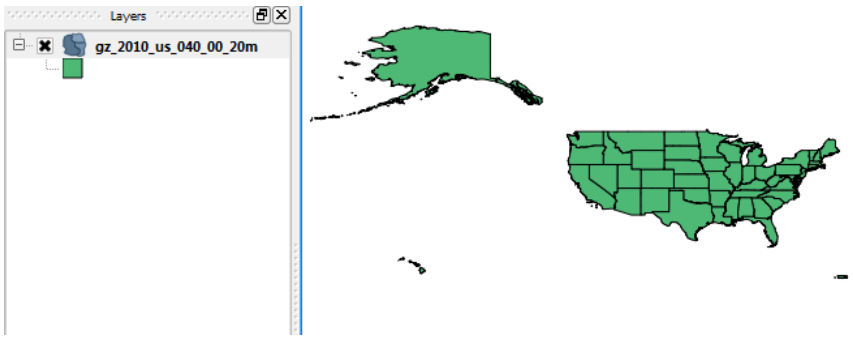
1. *Create a new project.* Open QGIS to an empty, blank project. Hit the  save as button. Browse to your data folder for part 4 and save the project as PART4.QGS. We'll be working with this project throughout this part of the tutorial.

2. *Check the shapefile's CRS.* Minimize QGIS, and use your file browser to browse through the data folder for part 4. You'll see there's a shapefile in the folder called GZ_2010_US_040_00_20M, which represents the states of the United States. It has a .dbf, .shx, .xml, and a .prj associated with it. Open the .prj file in a text editor (if using Windows, select the file, right click, choose the option to select a program from the list, select Notepad and click OK). You will see the projection information stored in the file:

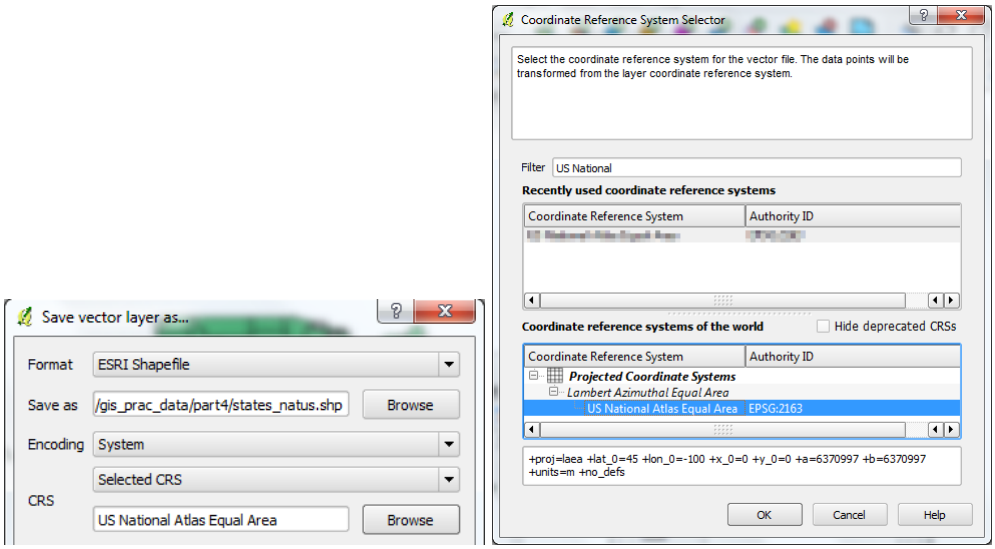
   ```
   GEOGCS["GCS_North_American_1983",
       DATUM["D_North_American_1983",
           SPHEROID["GRS 1980",6378137,298.257222101]],
       PRIMEM["Greenwich",0],
       UNIT["Degree",0.017453292519943295]]
   ```

   This file tells us that the shapefile is projected in the North American Datum of 1983 (NAD 83). Close the file when you're finished.

3. *Add the states shapefile.* Maximize QGIS. Use the browser to browse to the part 4 folder and add the GZ_2010_US_040_00_20M shapefile (select it in the browser and drag it into the project, or select, right click, and choose Add to project). Use the  zoom in button, draw a box around the US and zoom in.



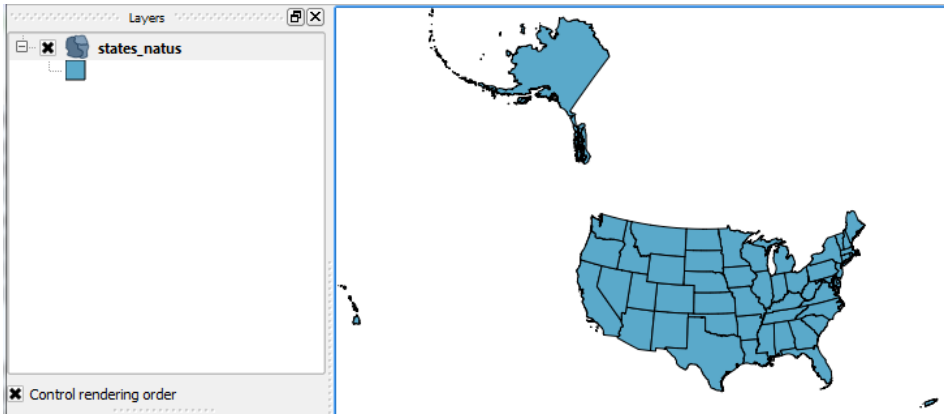
4. *Transform the projection.* Now we need to transform our layer to a different CRS that's more suitable for a thematic map of the US. Instead of using NAD 83, which is a basic geographic coordinate system (GCS), we are going to use a projected coordinate system (PCS). Select the GZ_2010_US_040_00_20M layer in the ML. Right click and hit Save As. Hit the Browse button beside the CRS entry. In the CRS Selector window type in US National in the Filter Box at the top. This filters the entire CRS database by name and we'll see the US National Atlas Equal Area CRS (EPSG: 2163) appear in the bottom window. Select it and hit OK. Back on the On the Save As menu browse and save the file in your part 4 folder as STATES_USNAT.SHP. Hit OK.



5. *Add the new layer.* Hit the Refresh refresh button above the browser, and add the new STATES_USNAT.SHP file to your project. When you do, your screen will fill up with a single color, obscuring everything. What's wrong? The two files we have in our view have different map projections and are incompatible.

6. *Reset the projection for the project.* Select the GZ_2010_US_040_00_20M layer in the ML, right click, and remove it. Select the new STATES_USNAT layer and hit the  zoom to layer button. Notice in the lower right-hand corner of the screen, the projection code EPSG: 4269 is still the old code for NAD 83. Select the STATES_USNAT layer in the ML, right click, and choose the Set Project CRS from Layer option. Now our layer has been reprojected and our project has been set to match that projection - you can see the EPSG Code 2163 in the lower right hand

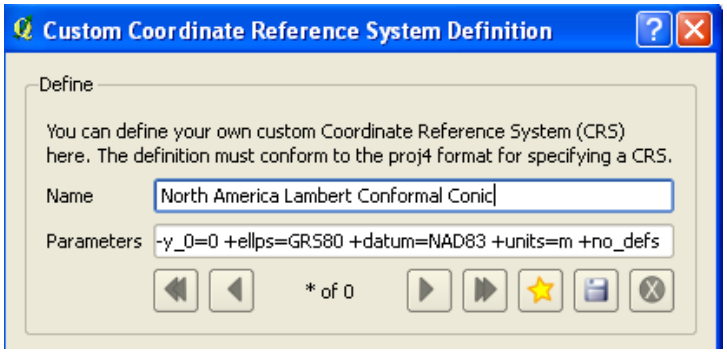
corner, and if you hover over it with the mouse it will say this represents the US National Atlas Equal Area Projection. Also notice that the coordinates in the status bar are no longer in degrees of latitude or longitude, but are in meters, which is the unit of measurement for this CRS. Save your project.



7. *Define a Custom Projection.* While QGIS has access to a large number of GCS's in its library, it doesn't have many PCS's that are suitable for continental or global projections. We do have the ability to add custom projections to the QGIS library using standard definitions. Lambert Conformal Conic is a common PCS for continents. Minimize QGIS, and in the data folder for part 4, open the file called LCC_NA_PROJ4.TXT. Copy the definition from that file:

```
+proj=lcc +lat_1=20 +lat_2=60 +lat_0=40 +lon_0=-96 +x_0=0 +y_0=0 +ellps=GRS80
+datum=NAD83 +units=m +no_defs
```

Maximize QGIS. On the menu bar go to Settings > Custom CRS. Paste the projection information from the file into the Parameters box. Make sure that there is no blank space at the end of the definition. In the Name box, type North America Lambert Conformal Conic. Hit the Save Button, then hit OK to close the window. This definition is now part of the CRS library on this machine, and we can use this definition to transform and reproject layers. But for now, we'll keep our current state layer in US National Atlas Equal Area.



8. *Avoid the Define Current Projection pitfall.* When transforming projections a common mistake is to use the Define current projection tool under Vector > Data Management Tools. You should NEVER use this tool to transform projections; its purpose is to define a projection for layers that are missing CRS data. See the commentary for details.

4.1.2 Commentary

Understanding Coordinate Reference Systems

All GIS layers are created using a specific coordinate reference system (CRS). The reason that we can take data from different sources and overlay them in GIS is because they share the same system; likewise, we can plot coordinate data and create layers because there's a coordinate system under the hood of our map window. In order for everything to work, your layers must share the same system and the map window must be defined to use that system. GIS software can be used to transform layers from one system to another. Each CRS is composed of at least three or four parts:

Spheroid or Ellipsoid: We typically imagine the earth as a perfectly round sphere, but in reality the earth is rather lumpy and uneven, with protrusions in some areas and indentations in others. The shape of the earth is approximated using spheroids, round three dimensional models of the earth, and ellipsoids, which represent the earth as being more oval than sphere-like in nature.

Coordinate System: This is the reference grid used for locating places on the earth and measuring distances. Latitude and longitude is the most common system, but there are other systems with different grid cells and units of measure; for example, the Universal Transverse Mercator (UTM) system uses a unique grid.

Datum: When you apply a coordinate system like latitude and longitude to different spheroids or ellipsoids, there needs to be a method for creating the grid and attaching it to the earth's surface. Mathematically, where does one draw the prime meridian and equator on a particular spheroid in order to accurately represent their location? The frame of reference for drawing these lines and measuring locations on the surface of the earth is called a datum.

Collectively, when you have these three elements: a spheroid or ellipsoid, a datum, and a coordinate system, you have something called a Geographic Coordinate System (GCS), which uses a three-dimensional spherical surface to define locations on the earth. The terminology is confusing, as a coordinate system is one part of a geographic coordinate system, and some systems are named based on the datum they use. For example, WGS 84 (World Geodetic System of 1984) is the most common GCS and uses the WGS 84 spheroid, WGS 84 as a datum, and latitude and longitude as a coordinate system. WGS 84 is used by the Global Positioning System of satellites and thus by individual GPS units as a default, and is commonly used by online mapping applications. It is so common that it is often referred to a THE Geographic Coordinate System. There are other systems; in North America NAD 83 (North American Datum of 1983) is widely used, particularly by government agencies. It uses GRS 1980 as a spheroid, NAD 83 as the datum, and lat and long as the coordinate system.

If you add a map projection as the fourth element to the spheroid/ellipsoid, datum, coordinate system trio, you have a projected coordinate system (PCS), which is defined on a flat two-dimensional surface:

Projection: Map Projections are mathematical systems for taking the three dimensional earth and transforming it to a flat two dimensional surface. There is no way to take a 3D shape and accurately represent it on a 2D surface, so map projections are designed to preserve one quality of the earth - area, shape, or distance/direction, or are created as a compromise to make the earth appear the way we expect it to appear on a flat surface.

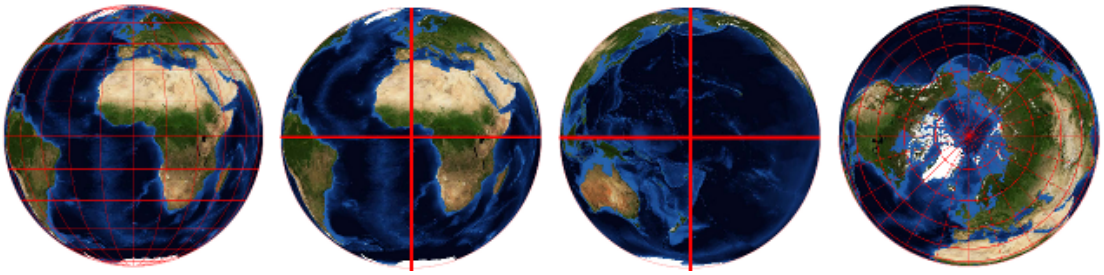
We defined the layer in the previous example as the US National Atlas Equal Area Projection, a PCS more commonly known as the Lambert Azimuthal Equal Area Projection. This projection preserves equal areas and true direction from the center of the map. It uses Clarke 1866 as the spheroid, uses a datum based on Clarke 1866, and the coordinate system is in meters.

In most GIS software, libraries of GCS and PCS system definitions are stored or organized separately, under their own menus or tabs.

Latitude and Longitude

The most common coordinate system is latitude and longitude, a grid system that covers the earth and uses a unit of measurement called a degree. Lines of latitude, called parallels, run east-west. The origin of latitude is the equator, which is zero degrees latitude. The equator bisects the earth and along this line there are twelve hours of daylight and twelve hours of darkness each day, throughout the year. Lines of latitude run 90 degrees to the north pole and 90 degrees to the south pole. One degree of latitude is equal to approximately sixty-nine miles, and since they are parallel lines they never converge.

Lines of longitude, called meridians, run north-south. Unlike the equator, which is the defacto line of latitude based on natural phenomena, the selection of an origin for longitude is arbitrary. The Prime Meridian, zero degrees longitude, was designated as the origin parallel in the 19th century. It runs through the center of the astronomical observatory in Greenwich, UK. There are 180 degrees of longitude to the east and to the west of the prime meridian. The meridian that is opposite the prime meridian on the far side of the globe, 180 degrees longitude, is the International Date Line (approximately). Unlike latitude, longitude converges at the poles to a single point at zero degrees. Since lines of longitude converge there isn't a uniform distance between them - the distance decreases as you move away from the equator. At the equator one degree of longitude is approximately 69 miles across, while at the poles it is zero miles.



There are two conventions for recording coordinates: in degrees, minutes, and seconds (DMS) or as decimal degrees (DEC). Take a look at the following coordinates for Philadelphia, PA from the USGS GNIS gazetteer:

39 deg 57' 08" N 75 deg 9' 50" W (DMS)

39.952335, -75.163789 (DEC)

The DMS notation is similar to the notation for telling time - there are 60 minutes in one degree and 60 seconds in one minute. DEC notation is preferable for computer processing; if you're plotting coordinates in GIS they should be in DEC. In DEC, latitudes south of the equator and longitude west from the prime meridian to the international date line are recorded as negative numbers. It is crucial that DEC coordinates indicate direction, otherwise you'll be confusing your point with a different place:

39.952335, -75.163789 is Philadelphia, PA USA

39.952335, 75.163789 is a remote area in western China near the Kyrgyzstan border

Map Projections

Most people today would agree that the earth is round. Most maps, whether they're on paper or a computer screen, are flat. When you take a three dimensional sphere and flatten it to two dimensions, you get fair amount of distortion. Imagine removing the peel from an orange and laying it out flat - you can't do it without tearing the peel. A map projection is a method for taking the three dimensional earth and transforming it to a flat surface.

For a nice overview, visit `http://www.radicalcartography.net/?projectionref` - Radical Cartography's projection page and note the common projections (marked in pink). Projections can be classified based on how the grid is applied to the earth's surface - a grid laid flat on top (azimuthal), wrapped as a cone on the top half of the earth (conical), wrapped around the earth as a cylinder (cylindrical), etc. They can also be organized based on which property they preserve:

Area (Equal-Area) - areas that are the same size on the globe appear as the same size on a map. Examples: Mollweide projection for the earth, Albers Equal Area for continents.

Shape (Conformal) - preserves angular relationships and shapes for small to medium areas (but distortion of shape occurs for larger areas). Examples: Mercator for the world, Lambert Conformal for continents.

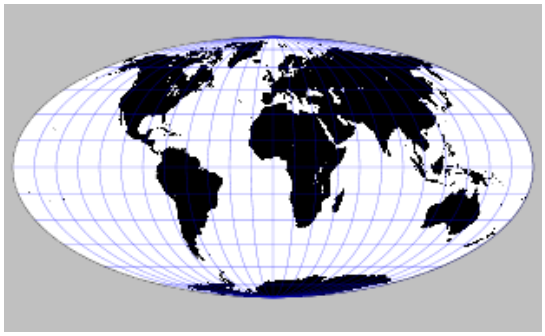
Distance (Equidistant) - maintains accurate distances from the center of the projection along specific lines; a straight line on the map will give you the shortest distance between two points, the same distance as a great circle on a globe. The Geographic Projection, also known as Plate Carree or Equirectangular, is the most common.

Direction (Azimuthal) - maintains accurate directions (and thus angular relationships) from a given central point. Azimuthal Equidistant and Gnomic are examples.

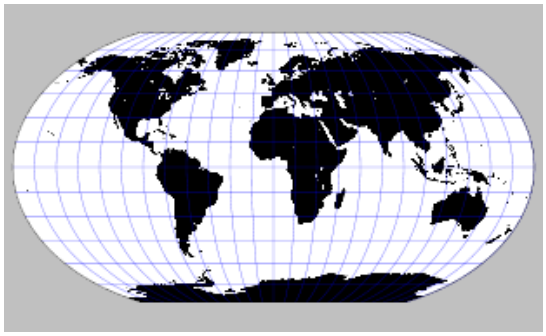
Other projections:

Interruptions - these projections show tears in the earth's surface and try to mitigate them to create something readable. Goode's Homolosine is good for showing land areas, but poor for showing oceans (as these are interrupted).

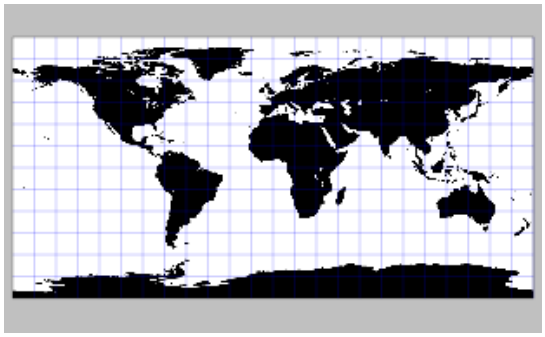
Compromises - these projections don't preserve any quality of the earth exactly, but they compromise to make a map of the earth that "looks right". Good compromise projections of the earth include Robinson and Winkel Tripel.



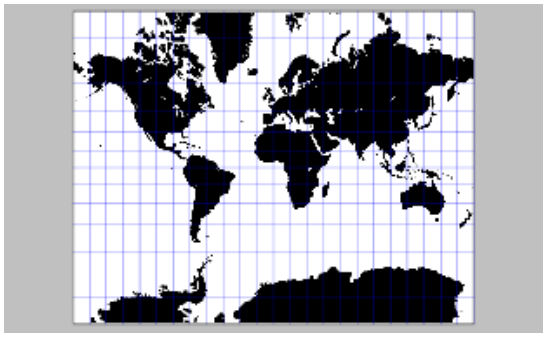
Mollweide



Robinson



GCS (Equirectangular)



Mercator

You can compare maps that use different projections to get a sense for how they distort different areas (in particular, observe Greenland). Common map projections for the world for general reference or thematic use include Robinson, Mollweide, Goode Homolosine, and Winkel Tripel (the first two have proj4 definitions that can be custom defined in QGIS). In general, projections that appear oval-like, showing the curvature of the earth at the edges, are best for general or thematic use.

Every continent and country has a preferred map projection or set of projections that is appropriate for each area based on its size and shape. Look at atlases or pre-existing maps to get an idea of what these are. Albers Equal Area, Lambert Equal Area, and Lambert Conformal are common and are adjusted to focus on specific continents or countries. Orthographic projections are used to map polar areas.

CRS Definitions

Several formats have been created for recording the definition of projections. There's the Open Geospatial Consortium's Well-Known Text Format (OGC WKT) as seen in the example we worked through, the Proj4 format, which we used to define a custom CRS in QGIS, and .prj file format created by ESRI. To look up CRS information, you can use the Spatial Reference website at `http://spatialreference.org/`. Use that site to get the proj4 format for creating custom projections in QGIS. When you open a .prj file and look at the definition, you'll see the elements that make up the GCS (projection, datum, spheroid) as well as units of measurement and origin information:

```
PROJCS["North_America_Lambert_Conformal_Conic",
    GEOGCS["GCS_North_American_1983",
        DATUM["North_American_Datum_1983",
            SPHEROID["GRS_1980",6378137,298.257222101]],
        PRIMEM["Greenwich",0],
        UNIT["Degree",0.0174532925199433295]],
    PROJECTION["Lambert_Conformal_Conic_2SP"],
    PARAMETER["False_Easting",0],
    PARAMETER["False_Northing",0],
    PARAMETER["Central_Meridian",-96],
    PARAMETER["Standard_Parallel_1",20],
    PARAMETER["Standard_Parallel_2",60],
    PARAMETER["Latitude_Of_Origin",40],
    UNIT["Meter",1],
    AUTHORITY["EPSG","102009"]]
```

From this definition, we can see that North America Lambert Conformal Conic projection uses GRS 1980 as a spheroid, NAD 83 as the datum, and meters as the unit of measurement for the coordinate system. As a conformal projection it preserves angular relationships.

Geographic reference systems have also been classified with codes, which makes them easier to identify and retrieve. QGIS uses a CRS library called the European Petroleum Services Group (EPSG). This library contains most of the primary GCS systems, such as WGS84 and NAD83, and local PCS systems like State Plane. For example, EPSG 4269 is the code for NAD 83, and EPSG 4326 is the code for WGS 84. The advantage of the codes is clearer when you're working with longer names: NAD 83 NY State Plane Long Island (feet) is abbreviated to EPSG 2263. The EPSG library lacks most of the PCS systems for continental and global map projections, which is why these are not available in QGIS; search Spatial Reference to find the proj4 definitions for these projections in order to custom define them. US National Atlas Equal Area EPSG 2163 is one exception, and is fine for making basic thematic maps. A brief list of common projections and definitions is included in the appendix of this tutorial.

Defining Undefined Projections

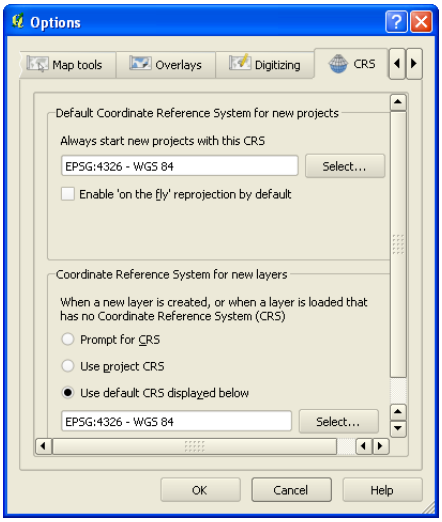
All shapefiles have a CRS and were created based on a particular one, but in some cases you may download or come across a file where the projection information for the shapefile, the .prj, is missing. In order to use the shapefile you will have to define the projection and create a .prj for it, so that the software will know how to render and layer it properly. To do this you'll have to go back to the website or source and look for some metadata that will tell you what CRS the file is in. The metadata could be listed on the download website, in a README or narrative file that accompanies the shapefile, or in an XML file accompanying the shapefile that was written based on metadata standards.

Once you know what the projection is, you can go to Vector > Data Management Tools > Define current projection. You can assign the projection from the QGIS databases of projections or you can import it from an existing shapefile that has the proper projection.

Note that defining a projection is DIFFERENT from transforming one. You DEFINE projections for shapefiles that are undefined, in order to tell the software what projection it is in. Use the Define current projection tool for that purpose. You TRANSFORM projections for shapefiles that are defined and have a projection, in order to convert them from one projection to another for a specific purpose. Select the shapefile in the Map Legend and do a Save As to convert the shapefile from one projection to another.

QGIS Projection Handling

In QGIS 1.8, when you open a new, blank project the default CRS is WGS 84. Then when you add your first layer, your project automatically takes the CRS from that layer, provided that it recognizes the definition. This is different from previous versions of QGIS where you were required to explicitly define the project window. If you add subsequent layers that don't share the same projection you'll have to transform them so they match, and reset the project window to the new CRS if your new files don't match the project window - reset it by selecting a layer in the ML, right click, and choose the option to Set Project CRS From Layer.



You can change the projection options under Settings > Options > CRS tab. You can change the default projection from WGS 84 to something else, if you know that you'll usually be working in another projection. Below this is a checkbox for enabling on the fly projection. If you enable this, QGIS will attempt to redraw layers if they don't match the projection of the window or other layers. This makes the software easier to use, but could lead to problems later. In this tutorial, and in general, I suggest that you know what CRS your layers are in and make sure all of the files you're using share the same CRS - don't use the on the fly option. I believe that this cuts down on confusion and

helps avoid errors caused by mis-aligning data layers and using systems of measurement that don't match.

You also have the option of setting the default projection for new layers that are created or are added to the window without a projection. The default setting is one projection that you select, and it's WGS 84. Once again, if you know you'll be working with a particular projection constantly you can select it here. Alternatively, you can choose one of the other radio button options - using the project window's CRS is a safe bet, if you're following the practice of keeping your window and all of your files in the same CRS.

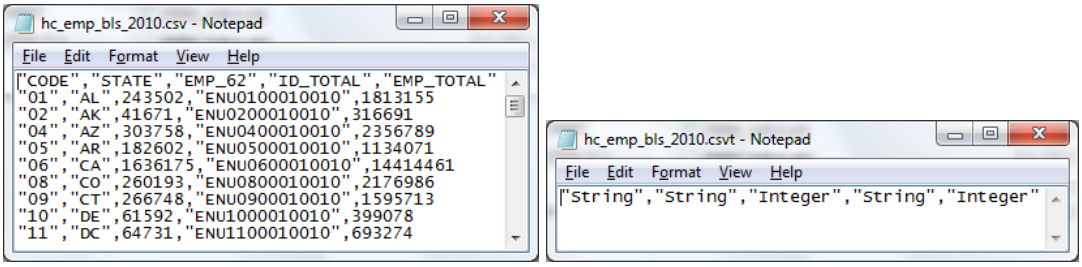
4.2 More Geoprocessing and Joining

This section will demonstrate a few more geoprocessing techniques that you're likely to need. You'll do another table join, and will learn how to edit a layer in order to delete individual features.

4.2.1 Steps

1. *Count features for your layer.* Select the STATES_NATUS layer in the ML, right click and check the Show features count box. It tells us there are 52 features. That's 50 states plus two (DC and Puerto Rico).

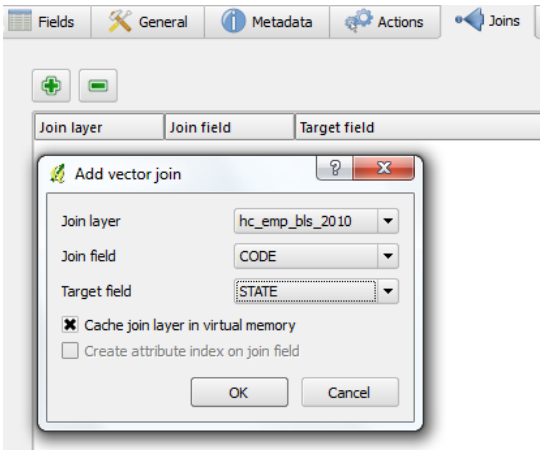
2. *Examine the employment data table.* Minimize QGIS. Use your file browser to go to the part 4 data folder. Find the file called HC_EMP_BLS_2010.CSV. Right click on the file and open it with a text editor - DO NOT open it in Excel (Windows users should use Notepad). This data is stored in a plain text, comma delimited format. Each column or field is separated by a comma, and each record (one for each state) is stored on a separate line. The data is from the US Bureau of Labor Statistics (BLS) and there are 51 records, one for each state and DC. The CODE field is a state FIPS code we can use for joining (a list of these is included in the appendix), EMP_62 is the number of people who are employed in the Health Care and Social Assistance sector, as defined by the North American Industrial Classification System (NAICS), and TOTAL_EMP is the total number of people in the labor force in 2010. Close the file. Now look for the file HC_EMP_BLS_2010.CSVT and open it. This file contains a single line and has one entry for each column in our csv file. It specifies the type of data stored in each column - text (string) or number (integer in this case). QGIS will reference this file to store and display our csv data correctly when we open it in QGIS. Close the file.



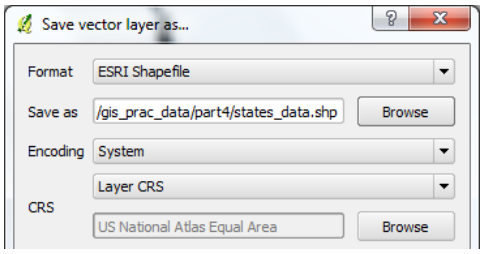
3. *Add the CSV table to your project.* Maximize QGIS. Use the browser to add HC_EMP_BLS_2010.CSV from your part 4 folder to the project. Select it in the ML, right click, open the attribute table. If the file imported correctly the text-based columns should be left-centered and numeric columns should be right-centered (if this isn't the case then there's a problem with your csvt file). Notice the CODE field - this is the two digit FIPS code that identifies each state. We'll be able to match this to the FIPS code stored in our shapefile in the STATE field. Close the table.

4. *Join the data to the shapefile.* Select STATES_NATUS in the ML, double click, and open the Joins tab in the properties menu. Hit the green plus sign to add a new join. HC_EMP_BLS_2010 is the join layer, CODE is the join field in that layer, and STATE is the target field in our shapefile. Click OK. Close the properties menu. Select STATES_NATUS, right click and open the attribute table. You'll see that all of the data has been added. However, if you look at

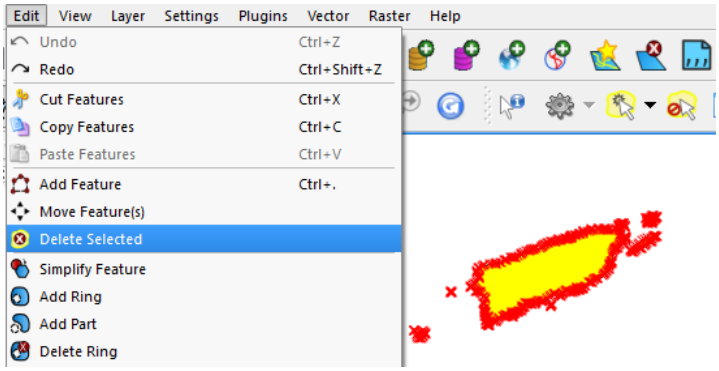
the record for Puerto Rico, you'll see that the values from the data table are NULL; this is because we had a feature for Puerto Rico in our shapefile, but no record in the data table. Close the table. Save your project.



5. *Work around for joins* Because of a bug in versions 1.7 & 1.8 we'll have to permanently fuse our data table to our shapefile in order to categorize and symbolize the data properly. Select STATES_NATUS, right click, and select Save as. Save the new layer as STATES_DATA in your part 4 data folder. Keep the CRS as US National Atlas Equal Area. Hit OK to save the new layer. Hit the Refresh refresh button above the browser, and add the new STATES_DATA layer to the project. Remove the old STATES_NATUS layer and the HC_EMP_BLS_2010 table. If you open the attribute table for STATES_DATA all of the data for the 50 states and DC should be there. Save your project.



6. *Edit the layer.* Since the values for PR are null and we have nothing to map for it, we'll modify the shapefile to delete the feature. Select STATES_DATA in the ML, right click, and choose the option to Toggle Editing. Each feature will be outlined in red; the red X's represent the individual vertices that each polygon is composed of. Zoom in to the area around Puerto Rico. Click the Select features button in the toolbar. Click on Puerto Rico to select it. Go to Edit > Delete Selected. Confirm the deletion. Zoom back out to see the rest of the US. Select STATES_DATA in the ML, right click and hit Toggle Editing. Save your edits.



7. *Inspect the layer.* Zoom in to the northeastern US, to the area around New York City. You'll notice that, unlike the previous census file we worked with from TIGER, this file has already been modified to remove bodies of water from state boundaries. But if you look at the NYC area, you'll see that Manhattan and Long Island appear joined to the mainland. This shapefile is from the Census Cartographic Boundary Files; they are TIGER files that have had their boundaries simplified so they appear less jagged at small scales (viewing the US as a whole) but are not appropriate for large scale maps (viewing a small area like the NYC metro).  Zoom back out.

4.2.2 Commentary

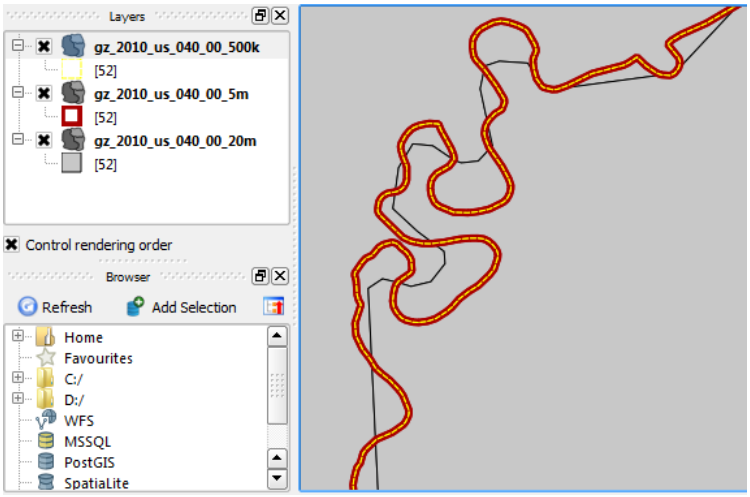
Generalization and Scale

The Census Cartographic Boundary Files (`http://www.census.gov/geo/maps-data/data/tiger-cart-boundary.html`) that we are using in this part of the tutorial were designed for creating maps of the US at a national or regional scale. According to the Census Bureau, "The cartographic boundary files are primarily designed for small scale, thematic mapping applications at a target scale range of 1:500,000 to 1:20,000,000." The file we're using in this exercise is the most generalized national file available, at 1:20,000,000. Boundaries have been generalized to depict land areas, to smooth coastlines and borders, and to remove small islands. This makes the boundaries appear smoother and cleaner at these smaller scales, while sacrificing geographic accuracy that wouldn't be visible.

When choosing vector files for thematic mapping you will need to make sure that the generalization for the file is appropriate for the scale you're working at. If you were creating a map of the NYC metro area, you would not want to use these boundary files as the generalizations become apparent at this larger scale and will make your maps appear inaccurate. You can identify whether a layer is appropriate by looking at the metadata and seeing if an optimal scale is indicated. Scale is a proportion of units of measurement on the map versus the actual distance in reality. A scale of 1:20,000,000 indicates that one measurement unit on the map represents 20,000,000 units in reality. Small scale

maps cover large areas while large scale maps cover small areas; this may seem counter-intuitive, but remember that scales represent fractions: 1/20,000 is a larger number (and thus larger scale) than 1/20,000,000. Most GIS software have tools for generalizing boundaries if you need them to be more simplified.

The screenshot below illustrates differences in generalization in scale using three different files from the 2010 Census Cartographic Boundary Files, from least to most generalized: 1:500,000, 1:5,000,000, and 1:20,000,000. The image depicts a portion of the boundary between Mississippi and Arkansas. The thin black line is the boundary from the most generalized file (1:20mil) that we are using; notice that most of the curves in the boundary have been straightened out relative to the medium (1:5mil dotted line) and least (1:500k thick line) generalized file.



Tabular Data: CSV Files

CSV files (comma separated values) are an alternative, stand-alone data format to dbf files that you can use in QGIS to match data tables to shapefiles. In Part 3 of this tutorial we worked with a tab delimited text file to get coordinate data into QGIS. CSV files are essentially the same format; they are plain text files where fields are separated by commas (as opposed to other delimiters like tabs or pipes) and records are separated on different lines. Compared to dbf, csv is a much more common format that you can create in any text editor or spreadsheet program, and when you download attribute data from a website or digital repository csv is almost always an option.

Unlike dbf, csv files do not contain any embedded information that specifies the type of data stored in each field. QGIS automatically imports all fields in csv files as strings. This is problematic, as numbers imported as strings cannot be treated as numbers (grouped into graduated categories or operated on mathematically). CSVT files are used to overcome this. You must create these by hand in a text editor and provide a data type for every column in your csv file. The names of the data types are placed in quotes and separated by commas. The file must have the same name as the csv file, must be saved with the extension .csvt, and must be stored in the same directory as the csv. The following data types are supported:

- Integer (for whole numbers)
- Real (for decimal numbers)
- String (for text)
- Date (YYYY-MM-DD)
- Time (HH:MM:SS+nn)
- DateTime (YYYY-MM-DD HH:MM:SS+nn)

You should be careful when opening csv files, or any delimited text files, in Microsoft Excel. Excel imports the csv and automatically saves any value that looks like a number as a number. This has the unintended effect of rendering identifiers like FIPS codes and ZIP Codes useless, as zeros are dropped from preceding values. Even if you open a csv file in Excel and don't save it, the file is still altered. In order to convert values back to their original form you would have to use the concatenate formula on any values that are less than their expected length and pad them with zeros. You can avoid these problems by importing the data using the From Text tool on the Data ribbon in Excel (rather than clicking on the file or opening it within Excel), working with csv in text editors, or by using other spreadsheet programs. For example, when you open a csv file in LibreOffice or OpenOffice Calc you are prompted to designate the data type for each field. Designate your identifiers as text / strings and they will be preserved.

If you would rather not bother with csvt files, you can import the csv and change the columns from text to numbers after the fact. You can use the field calculator, or a suite of third party tools called mmqgis that includes a few tools for working with csv files. You can activate mmqgis by activating the Plugin Installer under Plugins > Manage Plugins, and then going to Plugins > Fetch Python Plugins to search for and install mmqgis from the repositories.

4.3 Creating Calculated Fields

This section will show you how to add new calculated fields to a shapefile in QGIS. In many instances mapping numbers that represent whole values may not make sense and you'll want to create derived values. In this section you'll create a percent total to show the concentration of employment in the public health sector across different states.

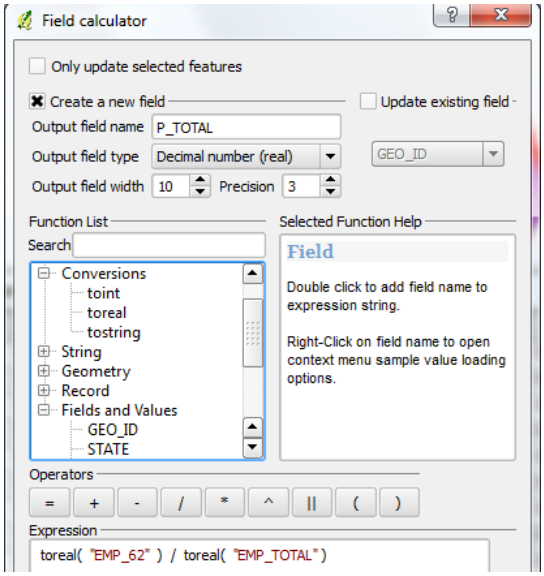
4.3.1 Steps

1. *Enter the edit mode.* Select the STATES_DATA layer in the ML, right click and open the attribute table. Hit the  edit button below the table to enter the edit mode. Since we are making changes to the actual shapefile we need to do that from an edit mode.

2. *Launch the field calculator.* Hit the field calculator button that's a few buttons to the right of the edit button. This opens the Field Calculator window. Under New Field, type P_TOTAL as the output field. Change the field type to a Decimal number (real). Keep the output field width to 10 (default width setting in the table window) but change the precision to 3 (number of places right of the decimal point). In the Function List expand the Conversions menu and double click toreal. Expand the Fields and Values menu to view the columns in the attribute table. Double click EMP_62 to add it to the expression box. In the expression box close the parentheses. Click the divisor symbol under the operators. Double click toreal again, and then double click EMP_TOTAL from the Fields menu. Close the parentheses. Your field expression should read:

   toreal("EMP_62") / toreal("EMP_TOTAL")

   The toreal operator converts the employment fields from integers to decimals, so the output can be saved as decimals. Hit OK. Back on the attribute table screen, hit the  edit button to stop editing and save your edits. You'll see the new percent total field appended on the right.



3. *Examine your data.* Click on the P_TOTAL column to sort the data. When you're finished examining the data close the table. You can Save your project at this point, but the edits to your shapefile have already been saved, since we were working on the shapefile directly and saved it after the edit mode.

NAME	LSAD	CENSUSAREA	TATE_	EMP_62	ID_TOTAL	EMP_TOTAL	P_TOTAL
Nevada	NULL	109781.18	NULL	102240	ENU3200010010	1108238	0.092
District of Columbia	NULL	61.048	NULL	64731	ENU1100010010	693274	0.093
California	NULL	155779.22	NULL	1636175	ENU0600010010	14414461	0.114
Utah	NULL	82169.62	NULL	131066	ENU4900010010	1150737	0.114
Colorado	NULL	103641.888	NULL	260193	ENU0800010010	2176986	0.12

4.3.2 Commentary

Representing and Calculating Values

In some circumstances it may make sense to map values as whole numbers - cities by number of crimes, states by total population, counties by number of renter-occupied housing units, etc. But in each of these examples a particular place could have a higher value simply because it has more people or is a larger place. In order to make more meaningful comparisons it's often necessary to do a little math:

Percentage - (value of subset / total value)*100: (3,000 renter units / 10,000 renter units)*100 = 30% units are rentals

Rate - (value / total value) * multiplier: (400 robberies / 50,000 people)*100,000 people = 800 robberies per 100,000 people

Ratio - (value 1 / value 2): (4000 cars / 3000 people) = 1.33 cars per person

Density - (value / land area): (800,000 people / 2500 sq miles) = 320 people per sq mile

Percent Change - [(recent value / older value)-1] * 100: [(10,000 people / 9,000 people)-1] * 100 = 11.1% change

Location Quotient - (employment in industry in local economy / total employment in local economy) / (employment in industry in national economy / total employment in national economy)

A location quotient is a common indicator used in economic base analysis. It compares a local economy to the greater economy in order to measure how specialized a local economy is for particular industries. The result is a ratio that measures the concentration of economic activity, where a value > 1.0 indicates that the local economy is more specialized in a particular industry relative to the nation while a value < 1.0 indicates that the local economy is less specialized. You can use the EMP_62 and TOTAL_EMP fields in our data table, along with the total labor force values for the US, to calculate location quotients for the health care sector for every state:

(toreal(EMP_62) / 18074841) / (toreal(TOTAL_EMP) / 127820443)

The field calculator was revamped in QGIS 1.8. If you are dividing two integers and want the output to be in a decimal format, you have to use the toreal function to convert each integer to a decimal before doing the division. Likewise, if you had two reals and wanted the output as whole numbers, you would convert them using toint. The tostring function can be used to convert numeric values to text.

Industrial Classification: NAICS

The North American Industrial Classification System (NAICS) is a hierarchical system of codes used to classify businesses into industries in the US, Canada, and Mexico. It was created in the mid 1990s and replaced the older Standard Industrial Classification (SIC) system. The NAICS system consists of broad industrial sectors defined with two digits that can be broken down into more specific subsectors with additional digits.

In our example we are studying the labor force of NAICS 62, the Health Care and Social Assistance Sector. Establishments in NAICS 62 provide health care and social assistance for individuals. The services provided by establishments in this sector are delivered by trained professionals; health practitioners or social workers with the requisite expertise. NAICS 62 can be broken down into more specific subsectors that include:

- 621 Ambulatory Health Care Services
- 622 Hospitals
- 623 Nursing and Residential Care Facilities
- 624 Social Assistance

Each of these 3 digit subsectors can be broken down into 4 digit groups. For example, subsector 621 Ambulatory Health Care Services can be broken down to:

- 6211 Offices of Physicians
- 6212 Offices of Dentists
- 6213 Offices of Other Health Practitioners
- 6214 Outpatient Care Centers
- 6215 Medical and Diagnostic Laboratories
- 6216 Home Health Care Services
- 6219 Other Ambulatory Health Care Services

4 digit groups can be broken down further to 5 digit industries (6124 Outpatient Care Centers breaks down to 62141 Family Planning Centers, 62142 Outpatient Mental Health and Substance Abuse Centers, 62149 Other Outpatient Care Centers), and 5 digit industries can be broken down to 6 digit national industries (62149 breaks down to 621491 HMO Medical Centers, 621492 Kidney Dialysis Centers, 621493 Freestanding Ambulatory Surgical and Emergency Centers, and 621498 All Other Outpatient Care Centers).

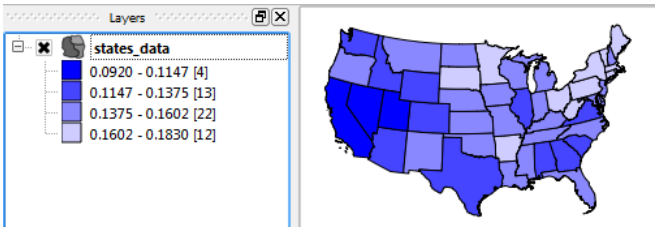
You can browse and download the codes at `http://www.census.gov/eos/www/naics/`. They are widely used by government agencies that produce data for industries (US Bureau of Labor Statistics, US Census Bureau, Statistics Canada, National Institute of Statistics Mexico) as well as private companies that produce databases or information retrieval systems that focus on industrial research. The NAICS system is largely compatible with the UN Statistics Division's International Standard Industrial Classification (ISIC) codes.

4.4 Classifying and Symbolizing Data

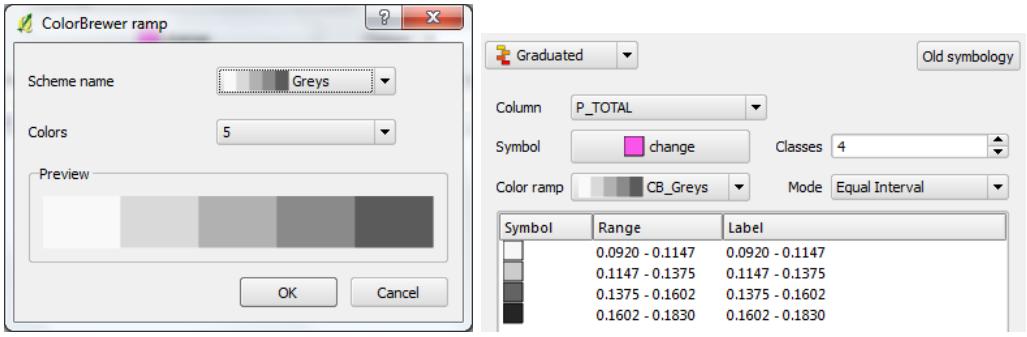
In this section you'll learn about the different methods for classifying data and the best approach for choosing color schemes to symbolize your data. These are important concepts to grasp, as they have a direct impact on how successful your map will be in communicating your data.

4.4.1 Steps

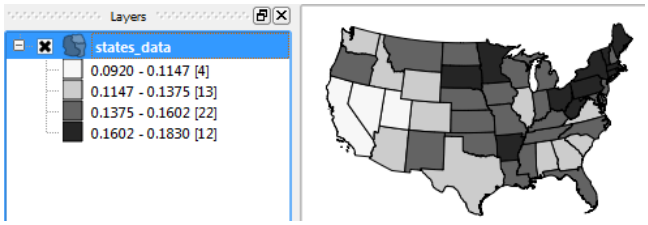
1. *Classify your data.* Select STATES_DATA in the ML and double-click to open the Properties menu. Go to the Style tab. If it's not currently active, click the New Symbology button to apply the new symbols for this layer. In the New Symbology change the classification drop down from Single Symbol to Graduated. Change the Column (the field you're classifying) from CENSUSAREA to P_TOTAL. Change the number of classes from 5 to 4. Keep the mode as Equal Interval. Choose one of the default color ramps and hit the Classify button at the bottom of the menu.



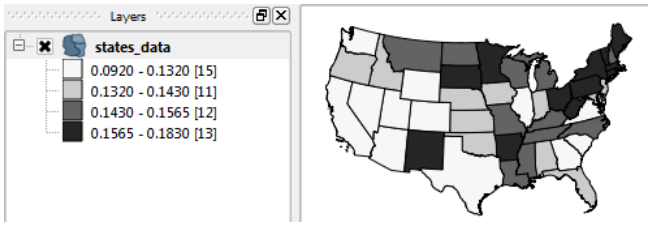
2. *Select a new color scheme.* You have the ability to access a number of color ramps rather than using the defaults. Hit the color ramp dropdown and choose New color ramp at the bottom. On the color ramp type screen choose ColorBrewer and hit OK. On the Colorbrewer ramp choose a scheme - for quantitative data with only positive values you should choose a color scheme that uses a single color value from light to dark - DO NOT choose a multi-color or random scheme. Hit OK once you've made your choice, and then give your color layer a name (like CB_greens or CB_oranges). Hit OK. Back at the Style menu, choose your newly added color scheme from the dropdown and hit Classify to reclass your data with the new colors. Keep Equal Intervals as the classification scheme. Hit OK to map your data.



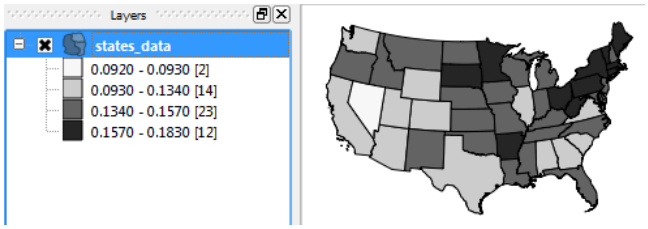
3. *Examine the Equal Intervals map.* In the ML, expand the menu for STATES_DATA to see the classes. Equal Intervals is the default classification scheme; it took our four classes of data and divided it so that each class has an equal range of values; with a min value of .092 and a max value of .183 our data has a range of .091 - divide by four and each class covers a range of .2275, sorted from lowest to highest. Remember that these are percentages in decimal format (.092 is 9.2% and .183 is 18.3%, etc). Right click on STATES_DATA in the ML and check the Show feature count option. You'll see the number of states in each class varies, but the range of values in each class is constant (2.275 percentage points in each class).



4. *Map data using Quantiles.* However, we could use an alternate classification method called Quantiles. Double click on the STATES_DATA layer to go back to the Style tab under the Properties menu. Change the classification mode to Quantiles and hit Classify. Hit OK to re-map your data in this scheme, and take a look at the result. Compared to the equal intervals map, quantiles show us a greater range of colors since each class has the same number of features. Quantiles divides our data into classes that have an equal number of data points. Since we have 51 data points we have about 13 states in each class sorted from low to high, as you can see in the feature count.



5. *Map data using Natural Breaks.* We have another option. Double click on the STATES_DATA layer to go back to the Style tab under the Properties menu. Change the classification mode to Natural Breaks (Jenks) and hit Classify. Hit OK to re-map your data in this scheme. The natural breaks method classifies data based on the location of gaps or breaks in the data range, which is less arbitrary than equal intervals or quantiles. Notice how there are only two states in the lowest category. If you select STATES_DATA in the ML, open the attribute table, and sort by P_TOTAL, you'll see Nevada and DC are in this class. After DC, there's a large gap of 2% points between DC and the state in the next class, California, large enough that the natural breaks formula created a class break here.



6. *Save your project.* At this point save your project. For our map we'll stick with the natural breaks method, but read the commentary below for an explanation of each method and it's advantages and disadvantages.

4.4.2 Commentary

Data Classification and Color Schemes

The purpose of a thematic map is to communicate a message about the data. If a map uses too few classes, then the data is too generalized and meaningful patterns can be hidden. If a map uses too many classes, then a pattern becomes difficult to detect because there is too much detail. It is difficult for the human eye to distinguish between too many colors or variations of color. Generally speaking, it is a good idea to use 3 to 6 classes, and ideally 4 or 5. When choosing the number of classes you should consider the number of data points, the range of the data, the purpose of the map, and the color choice based on the output. While a certain number and range of colors may look good on a color printed map, they may appear washed out if the map is shown on a projector or blurred together if photocopied in black and white. You should design with the final output in mind.

After ranking the data from lowest to highest values, there are a number of classification methods:

Equal Interval - each class has the same range of data values. Easily understood by map readers, but does not account for data distribution and can result in categories with few or even no values.

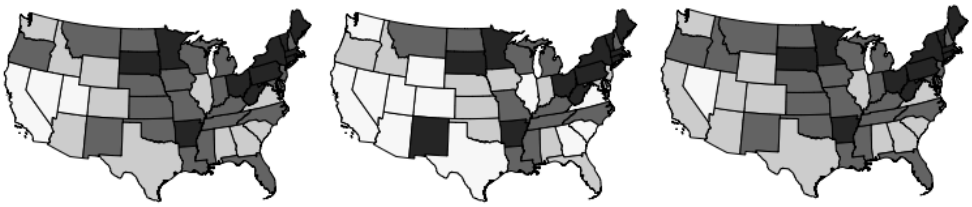
Quantiles - each class has the same number of data points. Always produces distinct map patterns, but can often create categories that have an inconsistent range of values.

Natural Breaks - classes are created based on the location of gaps in the data. Since the data is divided based on its distribution it is good for distinguishing patterns, but creating the classes manually is more labor intensive than other methods.

Unique / Manual - classes created based on some external criteria. Should only be used when justified, otherwise the classification is completely arbitrary.

It's often necessary to make some common sense adjustments to any classification scheme, such as creating unique classes for values of zero or missing values, and adjusting classes so they don't contain a mix of negative and positive values. In QGIS you have the ability to adjust classes or create manual classes. To do this, you classify the data using one of the standard methods in the Style tab for the layer, then select the class that you want to change and double click on the range. You'll be able to type the values in by hand.

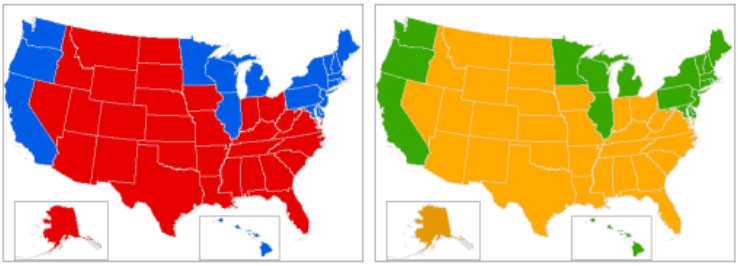
Take a look at the maps and data for this project below to compare how the different classification methods group the data. In this particular case the equal intervals and natural breaks method yield similar results; this is coincidental, as the data we're examining is rather evenly spaced around the mean. In other cases these classification methods will yield quite different results.



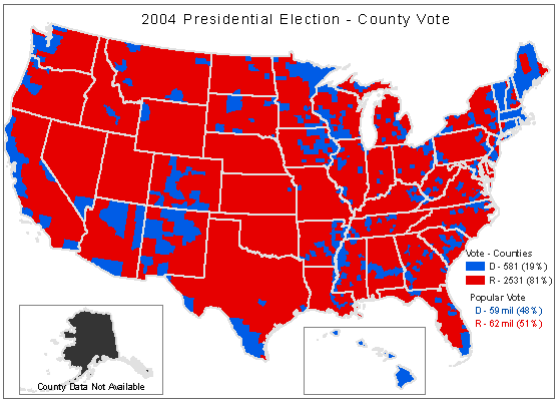
Data Classification for % of Labor Force in Healthcare

	Equal Intervals		Quantiles		Natural Breaks	
Class	Range	Count	Range	Count	Range	Count
1st	.092 to .114 NV to UT	4	.092 to .132 NV to TX	15	.092 to .093 NV to DC	2
2nd	.120 to .136 CO to ID	13	.134 to .143 AL to NJ	11	.114 to .134 CA to AL	14
3rd	.138 to .157 OK to NM	22	.144 to .156 WI to LA	12	. 136 to .157 ID to NM	23
4th	.161 to .183 AR to WV	12	.157 to .183 NM to WV	13	.161 to .183 AR to WV	12

Color schemes for displaying quantitative values on choropleth (shaded area) maps should show a logical progression of color values. The progression from light to dark helps convey the change in data values from low to high, and most map readers can infer this without even looking at the map legend. Creating a mixed, fruit salad of colors will defeat this natural inference and will confuse the map reader - so don't do it. When comparing qualitative values (categorical data instead of ranges of values), a map should use colors that reflect those values. For example, it makes sense to use reds and blues to show which political party a state voted for, as these colors have become associated with the US political process. Without even looking at a legend or description, the average American will instantly understand what this map is about. Depicting the same data with greens and yellows doesn't make much sense, and results in confusion.



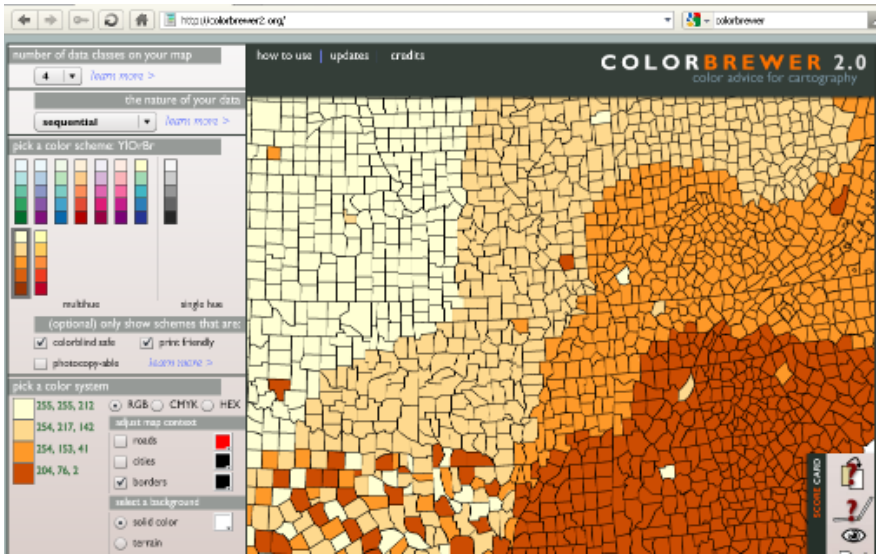
While we're not considering it for this exercise, the unit of geography used to map phenomena can profoundly affect the interpretation of a distribution or pattern and the ultimate message that your map sends. Mapping populations of US states or Canadian provinces is fine if you are interested in seeing which ones have the most people. But these maps tell you very little about how the population is distributed across these countries, since there is considerable variation in the concentration of people in each state / province. Using a smaller unit of geography can give you a better idea of the distribution of the population. For example, compare the state-level election map above with a county-level map that depicts majority votes by political party:



Oftentimes you'll be limited to using certain geographic units based on the availability of the data, making it necessary to compromise.

Colorbrewer

Colorbrewer is an online tool for choosing good color schemes for thematic maps. Recent versions of QGIS have integrated many of the color schemes from this tool in the New Symbology tools, but it's still worth visiting the site at `http://colorbrewer2.org/` for color selection advice. The tool let's you choose the number of classes and class options like sequential (for quantitaive data we've used in our example), categorical (for nominal or qualitative data), and others. You also have the ability to filter color schemes based on desired output. In the lower-right hand corner of the map, you can click on a scorecard that shows whether your choice is ideal for the color blind, color printing, photocopying, and viewing on an LCD screen. You should always choose color schemes based on what your final output format will be. Colorbrewer gives you the option to export your color choices out as text, where the text is some notation for representing color such as RGB or hexadecimal (used in HTML for identifying colors).

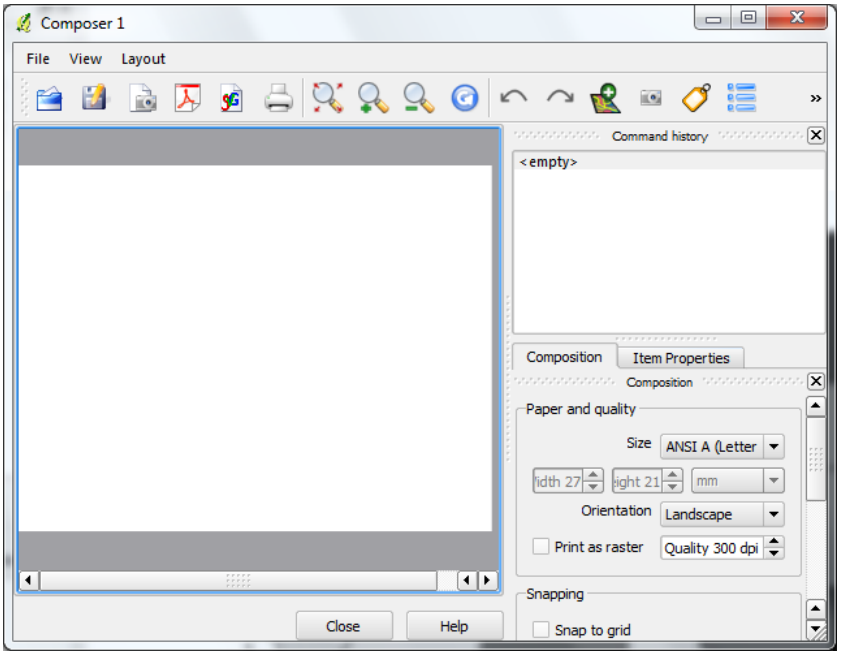


4.5 Designing Maps

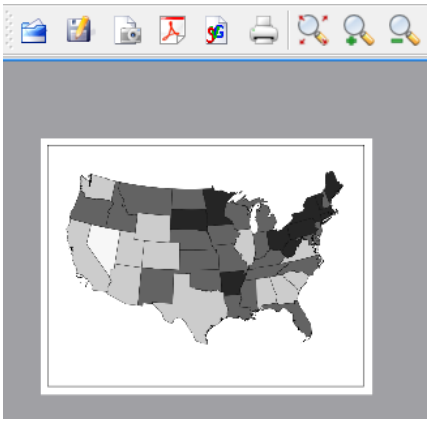
In this section you'll learn how to create a finished map that includes typical map elements: legend, title, scale bar, and source information.

4.5.1 Steps

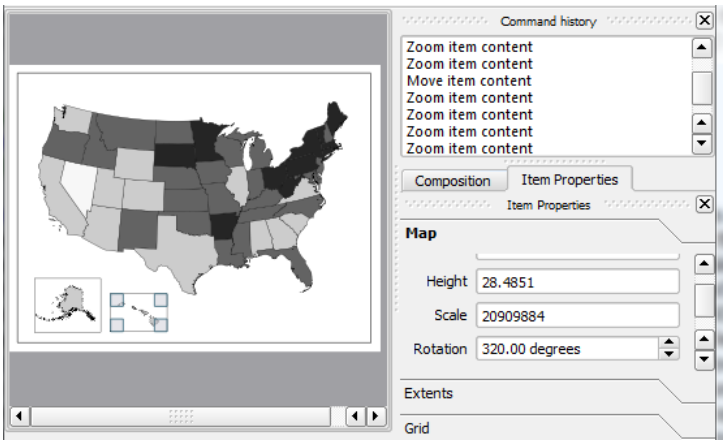
1. *Set the environment for the print layout.* Hit the print new button to enter the print layout screen. On the General tab in the Paper and quality menu on the right-hand side change the paper size from A4 to ANSI A (letter 8 1/2 by 11). The composition tab provides you with options for the map canvas as a whole. Once you add individual items (a map, label, legend, etc) the item tab will become active, and if you have the item selected in the canvas you'll be able to edit its attributes. Each tab has collapsible menus for editing various elements.



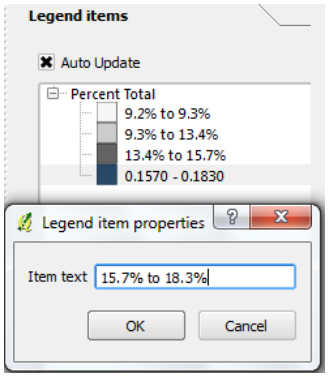
2. *Add your map and configure zoom.* Hit the add map button in the toolbar. Then draw a box on the map canvas (click on upper-left hand corner, hold down left mouse button, drag box), leaving an even amount of space on each side so there is a gap between the map and the edge of the canvas. If you don't get it right on the first try, you can always hover over an edge of the map, hold down the left mouse, and drag the edge to change the size. Or, to shift the entire map on the page, use the select move button. This button moves the entire map box. To shift the geography *inside* the map box, use the adjacent move item button. Move the map around so that the lower 48 states are roughly centered in the box. With the move item button selected, you can also change the zoom of the map by using the mouse wheel, or by clicking on the item tab on the right and experimenting with the scale in the Map menu under the Item properties tab. The regular zoom buttons on the toolbar will NOT effect the zoom of the geography; these zoom buttons just zoom you closer and further from the map canvas, similar to taking a piece of paper and holding it closer or further from your face. Experiment with them and see. If your map looks blurry from resizing a window, just hit refresh button. When you're finished, with the map selected go to the Item tab, and under General option increase the outline width of your map from .3 to .5.



3. *Add additional maps for Alaska and Hawaii.* Given the vast distances between the lower 48 states, Alaska, and Hawaii, it doesn't make sense to include them in the same map window at the same scale; look at most maps of the US and Alaska and Hawaii appear in separate maps or boxes so that optimal scale can be achieved for all three areas; we'll do the same with our map. Hit the add map button and draw a smaller box in the lower left hand corner. Use the move item button to shift the focus of the map to Alaska, and with this button selected use the map wheel to change the zoom. If you have trouble getting the zoom "right", open the Map menu on the Item tab on the right, watch how the scale changes as you zoom in and out with the mouse wheel, type in an estimated scale that's somewhere in-between. Right below the scale in the menu is rotation, which is currently set to 0. You can type values here to rotate the items in the map from 0 to 359 degrees clockwise. Since Alaska looks a little skewed (since we're using a map projection for the whole continent and AK is on the edge) change the rotation to 325 to straighten Alaska out. Once you're finished, repeat the same step for Hawaii: add another map, zoom in to focus on the main eight islands, and rotate it by 320.



4. *Add a legend.* Hit the add vector legend button and click on the lower right-hand corner of the map. With the legend selected, go to the Item Properties tab and the General Menu. Change the generic "Legend" title to NAICS 62 Employment. Hit the Title Font button that's directly underneath the title and change the font to 12. Next, go to the Legend Items menu. Select STATES_DATA in the list, hit the edit legend button, and change the name to Percent Total. You should also edit each data range to change the label to change our percentages to whole numbers. Open the Percent Total dropdown, select each range, hit the edit button, and type in the percentage values. The final step is to move the legend to an ideal position in the corner of the map (which may require you to shift the map around a bit).



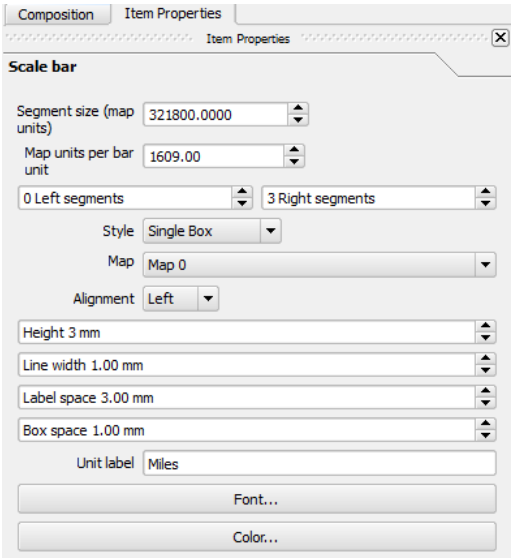
5. *Add a title.* Hit the add label button. Click on the top of the map, and a generic label is added. In the Label menu under the Item tab, change the default Quantum GIS label to Employment in the Health Care Sector in 2010. Change the font to 18 using the font button. On the General options menu uncheck the option that says Show Frame. This will turn off the label outline. Click on the label in the map, and using the select move button, move the label to the top center of the map, and expand the size of the label box so the title appears on one line.

6. *Add a label with source information.* Hit the add label button. Click on the bottom of the map to add the generic label. In the Label menu on the Item tab, change the label to read: Source: US Bureau of Labor Statistics QCEW 2010. Change the font to size 8. On the General options menu uncheck the option that says Show Frame. Click on the label in the map, and using the select move button, move the label to the bottom center of the map, and expand the size of the label box so the text appears on one line.

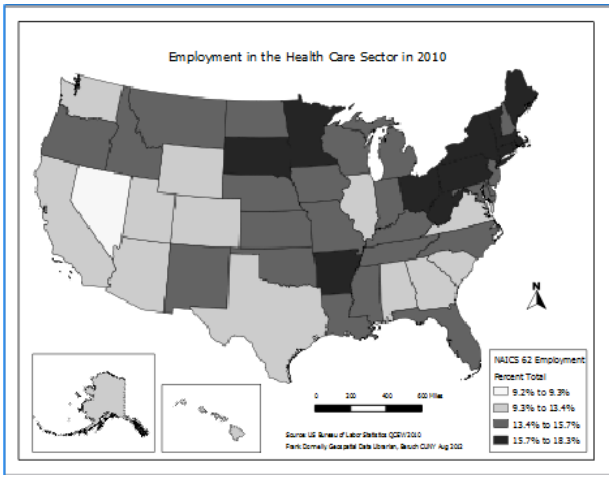
7. *Add a label with author information.* Repeat the same step above to add a label with your information - Map created by [insert your name / organization] [insert date]. Move this label underneath the source label.

8. *Add a north arrow.* Hit the add image button. Click somewhere to the right of the US in the map, above the legend. There may be a momentary pause while QGIS loads the images. Scroll through the picture options in the item properties and select a simple north arrow. In the Item tab for the image, go to the General options and turn the frame for the arrow off. Move the arrow around on the map to get it centered, and resize it to make it a bit smaller.

9. *Add a scale bar.* Hit the add scale bar button. Click below the map of the US to add the scale bar. The result doesn't look promising, but we'll fix that. In the Item tab for the scale bar go to the Scale bar menu. Change the segment size to 321,800. Change Map units per bar to 1609. This gives us a scale bar with segments of 200 miles; our map units are in meters so we had to do some conversion (see the commentary). Increase the number of segments from 2 to 3. Change the Map dropdown to Map 0 (0 is the first map we added, for the entire US, 1 is AK and 2 is HI). Change the height of the bar from 5 mm to 3 mm. In the Unit labels box type Miles. Change the Font size to 8. Lastly, go to the General options menu at the bottom and uncheck the Show frame box to turn it off. Use the select move button to position the scale bar on your map.



10. *Balance your map elements.* At this point you should have all of your map elements in place. You may need to resize and shift elements around in order for the map to appear balanced. If you want to insure that boxes are lined up properly, you can hit the select move button and click on individual features while holding down the CTRL key to select multiple items. You can use the various align buttons to arrange elements in a certain way, and you can use the group button to bind several features together so you can move them in unison.



11. *Close the composer and save.* Oddly, there is no save button within the composer (the one on the toolbar is for saving a template of your map, and not the map itself). Close the composer window, and back out at your map view save your project. This will save the map you just created. It's IMPORTANT that you save your map prior to printing or exporting it - this insures that if the export or print goes wrong or crashes, you won't lose your map. Once you save, hit the print button, select the first composer from the list and hit show, and you'll be back to your finished map. If your map looks grainy or out of focus, don't worry - it's really ok. To assuage any worries, you can hit the refresh button.

12. *Print to PDF.* PDFs are good stand-alone maps. Before you export, make sure you don't have any map elements selected and return to the Composition tab for the map. Hit the export to PDF button and save your map as a PDF file, HC_EMP_2010.PDF, in your part 4 data folder. The program may hang for several seconds while the map is being exported. After a few moments you can click on the composer to reactivate it, or minimize and maximize QGIS to get back to the composer.

13. *Export as PNG.* You can also save your map as an image file like a jpg or png. Normally we would want to design the map to be the size of the desired image, and we'd want to adjust the DPI quality (just above the Save as Raster checkbox in the Composition tab) to reduce it's size. Hit the save as image button. Browse to your data folder for part 4 and save the map there as HC_EMP_2010.PNG. Make sure that you specify png as the file type. After you hit save, QGIS will hang for a moment while it exports the file - just wait for a few seconds and the export will be finished.

14. *Take a look at your maps.* Minimize QGIS and use your file browser to go to your part 4 data folder. Double click on the PDF file to open it in Adobe or your PDF viewing software. Double-click on your png file to open it in your default image viewing program (or open it with your web browser). Congratulations on creating a finished map!

4.5.2 Commentary

QGIS Map Composer: Scale Bars and Other Details

In some GIS software packages the current view in the map window and the print layout are dynamically linked, and a change in one (such as adjusting the zoom) affects the other. This isn't the case with QGIS; the two are separate. If you do change something in the map view, such as reclassifying the data, you can update the map composer under the item tab for the map by hitting the Update Preview button. Changes in focus or zoom between the view and the composer are not connected at all, which relieves a lot of potential headaches.

The print composer allows you to customize minute details of the canvas, map, and legend, more so than other open source packages. The composer also gives you the ability to draw shapes or add portions of an attribute table directly to a map. You can also store more than one map in a single project. From the map view, you can use the print new button to create new, individual maps, and the print composer button to manage your maps and choose a particular one to show or edit.

The scale bar feature in QGIS automatically takes the units from the layers' CRS. The US National Atlas Equal Area projection is in meters, so by default the units in the scale bar are in meters. In most cases you will have to convert measurements to larger units that make better sense. To do this, decide how many units you want an individual box in the scale bar to represent, and then do the conversion. A simple example: if we want the individual segments of the scale bar to represent 300 km, we would enter 1000 in the the Map Units Per Bar (as 1000 meters = 1 km), and then multiply the conversion factor by the segments we want (300 * 1000 = 300,000), and enter the result in the Segment Size Box. To make sure we did the math correctly compare one segment of the scale bar to the length of a known feature on the map. For example, Colorado is just over 600km in width, so you can hover the scale bar over the state to see if it's approximately correct.

Using kilometers on a map of the US would be heretical, so we need to use a different conversion factor. If we want the individual segments of the scale bar to represent 200 miles, we would enter 1609 as the the Map Units Per Bar (as 1609 meters = 1 mile), and then multiply the conversion factor by the segments we want (200 * 1609 =

321,800), and enter the result in the Segment Size Box. We can compare our scale bar to Illinois, which is just over 200 miles across at its widest point, to make sure we have the math right.

Most continental and global projections are in meters and degrees. Converting degrees to other units of measurement, particularly at this scale, is complicated and should be avoided. Use a projection in meters, and convert to kilometers or miles. For regional and local projections like UTM and State Plane, US mappers will have an option between meters or feet.

In our map we created a scale bar that's just for the US; conventional practice would require us to create separate bars for Alaska and Hawaii since they are not at the same scale. On the other hand scale bars and north arrows are only crucial on reference maps (street maps, property maps, topographic maps, etc.), where the emphasis is on depicting direction or distance; for many thematic maps they can be considered optional.

General Map Design

When creating maps you need to design with the end use, format, and audience in mind. If you're designing a map that you're going to embed as an image in a document or web page, you should change the size of the canvas and design the map to the specifications for the document. Creating a full size 8 1/2 by 11 map and scaling or cropping the final image is a bad idea; you'll introduce distortion into the map and text will become illegible. You also need to think about page orientation; it's appropriate to map the United States using a landscape page layout, but if you were mapping an area that was taller rather than wider (South America) you'd want to flip the page to portrait.

Individual map elements (maps, title, arrow, legend, source text) should be balanced on the page to achieve some harmony; avoid lumping too many elements together or having large areas of white space. The title and legend should concisely and accurately describe what the map is about and what you are mapping. The amount of detail you provide and the terminology you use should vary with your audience; for example if we were going to circulate this map to the general public we may want to include a brief definition of NAICS 62 and what is included in it. You should always include the source of your data in the map. The fonts, north arrows, and other elements should also be tailored to the map content; a title in calligraphy font and an ornate compass rose may look good if you're recreating one of Christopher Columbus' charts, but it would look rather silly on our US health care employment map.

Maps are a form of communication, designed to send a message. Like a book or article that is poorly written, maps that are poorly designed will fail because they do not effectively communicate their message to their audience. Some reasons why maps can flop:

- Poor layout - map elements (map, legend, title, text) arranged in an uneven or sloppy way
- Poor use of symbols - circles too big or small, not enough dots per person, etc
- Improper data classification - too many or few classes that obscure patterns, illogical scheme for dividing data
- Violation of basic cartographic convention - improper conventions for labels and color
- Poor figure-ground relationship - inability to clearly distinguish land from water or foreground from background
- Poor color scheme - random schemes for quantitative data, color that's improper for final format (color print, photocopy, screen projection, etc.)
- Information overload - too much information (several variables or map elements) or noise (unnecessary information)
- "Chartjunk" - concept defined by the graphic designer Edward Tufte, refers to kitschy or gimmicky elements that add nothing to the message of a map or graphic
- Factual errors - mistakes with labels, data, or geography

- Violates expectation of the user - simplification or generalization is too much for the user to accept
- Offends culture of the user - the message or how the message is communicated (text, colors) violates taboos that a user or group cannot accept

Output Formats

PDFs are a good format for creating stand-alone documents. PDFs are also a vector-based file, meaning that the geometry of every shape (point, lines, and polygons) is stored as a series of coordinates. If you're working with vector features to begin with, the output in the PDF should be fairly smooth, and if you zoom in to the document you should see all of the detail stored in the original file. If the PDF takes too long to open or draw or the PDF file is too large, you may want to consider checking the option to save the map as a raster within the PDF. The problem with PDFs is they are stand-alone; SVG files are a vector format that can be embedded in other documents, but support for them is uneven, and the SVG export in QGIS is still a work in progress.

Image formats are raster-based, meaning that the image is composed of individual pixels or grid cells. Rasters are designed for a specific scale; zoom in too close and the image quality deteriorates as each individual cell becomes more distinct. Rasters can stand alone or can be embedded in documents. PNG files are an open format, compressed raster. They're a good alternative to jpgs; they have better image quality and are widely supported. Tif files are a lossless, uncompressed format - use these only if you need to preserve the image at its highest quality (these files get pretty big). When exporting to a raster, be sure to adjust the dpi (dots per inch) setting, which will adjust the resolution of the image (and affect it's size and quality).

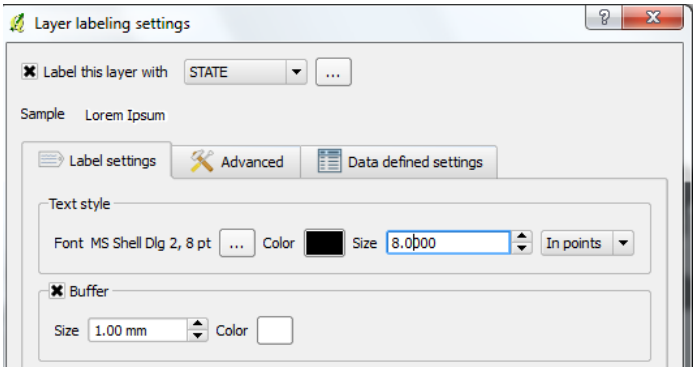
When printing hard copy maps, what you see on the screen is not exactly what you'll get on paper, so be prepared to print test copies and go back and revise. Because there are different screen resolutions and different printers (in terms of print method and quality) colors and outlines will vary.

4.6 Adding Labels

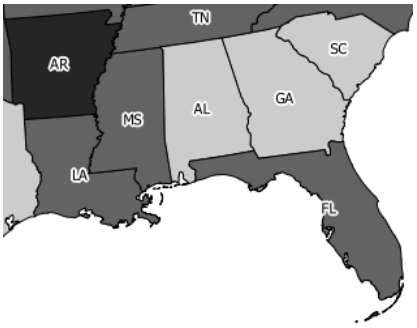
In this section we'll go back and add some labels to our map. Like symbology, two different labeling systems (an old stable one and a new experimental one) have existed side-by-side in QGIS for several versions now, and it's likely in the near future that the new version will replace the old entirely. The old labeling system is available in the Labels tab under the Properties menu for a particular layer. The new system is available via  labels button on the toolbar. We'll use the new system.

4.6.1 Steps

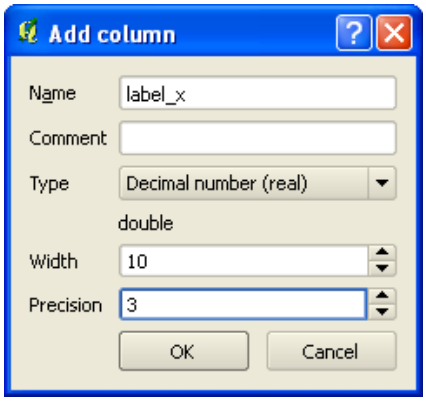
1. *Turn labels on.* Close the print composer and go back to your QGIS map view. Select STATES_DATA in the ML and hit the  labels button on the toolbar. On Label Settings check the box to Label this layer. In the Fields with labels dropdown choose STATE as the label field (this field has the two-letter postal code for each state). Change the size of the text from 8.5 to 8. Hit the Advanced tab, and on that tab click the Over Centroid radio button, and under the buttons move the Priority slider all the way over to High. Hit OK to apply the label settings.



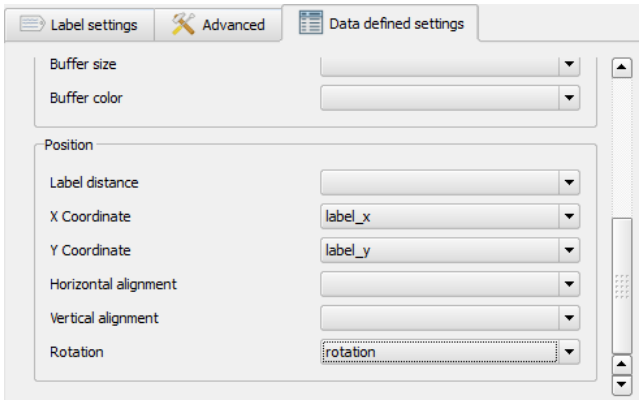
2. *Inspect the labels.* At first glance the label placement looks pretty good, and is a vast improvement over previous versions of QGIS. There are a few small issues; the labels for Florida and Louisiana look a little off center. And if you're zoomed out so the contiguous 48 states fill the screen, the label for Washington DC is omitted, as it overlaps with labels of neighboring states. With a little extra work we can fix that.



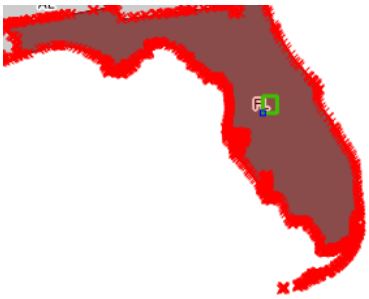
3. *Add new columns to the attribute table.* The labels are automatically placed in the center of the state. In order to define and store a specific position for them, we have to add some new columns to the attribute table. And to do that, we'll have to enter an edit mode so we can actually modify our file. Open the attribute table for STATES_DATA. Hit the edit button at the bottom of the table. Hit the New column button. In the Add Column window name the new field label_x. Assign it a Decimal number type. Give it a width (number of characters) of 10 and a precision (number of decimal places) of 3. Hit OK, and the new column gets tacked on at the end of the table. The label_x column will hold the X (longitude) coordinates for our label. But we need a second column to hold our Y coordinates (latitude). Repeat the previous step to add a second column called label_y. Finally, add a third column called rotation, and give it the same attributes. Once you've added it, hit the edit button to save the changes, and the columns become permanent. Close the attribute table.



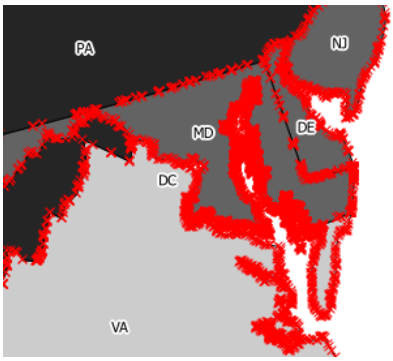
4. *Update label menu settings.* Before we can start moving labels we have to tell QGIS to store the positions for our labels in these new fields. Hit the labels button and on the labels menu go to the Data defined settings tab. Scroll down in this window to the Position box. In the dropdown for X Coordinate, select the x_labels field. In the dropdown for Y coordinate, select the y_labels field. For Rotation, select the rotation field. Hit OK.



5. *Move the label for Florida.* With STATES_DATA selected in the ML, right click on it and hit the edit button to enter an edit mode. You'll see each state outlined with little x's; these are the individual nodes that make up the points of each polygon, and this is your clue that you're in an editing mode. You'll also see that the move label button on the toolbar is now active. Hit the button, and you'll see a crosshairs as you move across the map. Adjust your map so that Florida (FL) is visible and centered. Move the crosshairs over the FL label, hold down the left mouse button, drag the label to the center of the state, and release.



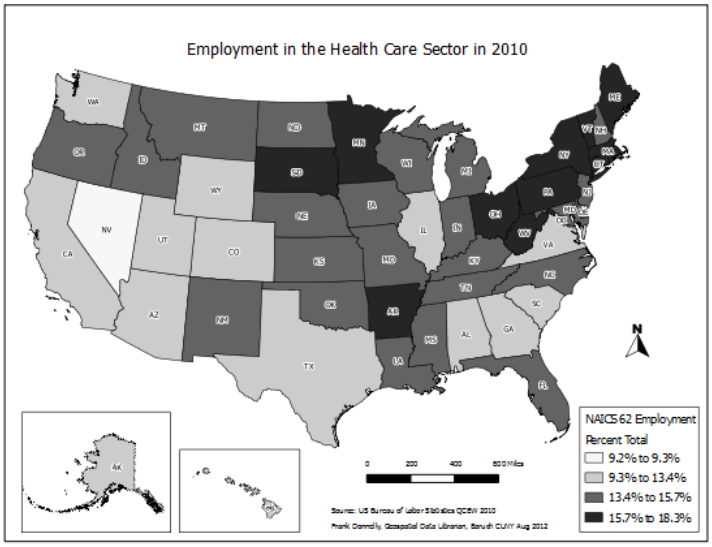
6. *Adjust additional labels.* Do the same to move the label in nearby Louisiana (LA). Then, use the pan tool to move to the northeastern US, then reactive the move label button. Move the label for Maryland (MD) to the north and the label for DC to the south so that both will be visible. The labels aren't going to look right at this scale, so zoom out back to the continental US to make sure the labels look OK at that scale. Once you're satisfied, hit the edit button for the layer to stop editing and save your edits. You may have to enter the edit mode, move labels, and exit a few times until you get the labels right (as it may be difficult to see their placement in the edit mode). When you're finished, you can open the attribute table for the layer, scroll to the right, and you'll see that coordinates are stored in the x_label and y_label field for the labels you moved. Close the table.



7. *Adjust rotation for AK and HI labels.* Even though they may look fine in our map view, our labels for Alaska and Hawaii are going to look askew when we re-open our Map Composer. This is because we rotated the maps of AK and HI so that they appeared "normal" in orientation relative to the rest of the country. So, we also have to alter the rotation for the labels to match. Enter an edit mode. Hit the change label button. Zoom up to Alaska and click on the AK label. At the bottom of the Labels properties box type 325 in rotation and hit OK. (325 is the number of degrees we rotated Alaska in the map composer - you could go back into the composer to find this info). Repeat the same step for Hawaii, but specify a rotation of 320. Exit the edit mode and save the changes.

8. *Update your map composer.* Hit the print button and show Composer 1. Hit the refresh button. You should see all your map labels - don't worry if they appear overlapped; they should turn out fine in the export. If you don't see the labels, select each map and under the Item tab and hit Update Preview.

9. *Save and export.* Close the map composer and back in the map view hit the save button. Then go back to the map composer and export your map (remember - we exit, save, and return just in case the export crashes). Print your map out as a PDF or save it as an image. Save it in your part 4 data folder as HC_EMP_2010_LABELS.PNG (or .pdf). Minimize QGIS, go to your part 4 data folder, and take a look at your final map.



4.6.2 Commentary

Labeling in QGIS

Automatic labeling placement in QGIS, and the ability to move labels and customize them, has vastly improved in the latest versions of QGIS. There are some other options at your disposal:

- The new, experimental labeling tool is available in the map view on the toolbar, and will eventually become the default for labels. You can also add columns to your attribute table that allow you to specify label details for each feature such as font type, size, color, placement, and rotation. The old labeling engine is still available on the labels tab under the properties menu for individual layers, but at this stage it's best to stick with the new label options.
- The text annotation tool allows you to add call out boxes directly in the map view. This is practical if you only need to place a few labels.
- You can also use the add label feature within the map composer. This can be a little cumbersome since you cannot copy and paste labels, but must create each one from scratch; ok if you only need to add a few labels.

Generally, features can be displayed and differentiated from each other using text. For example, the standard cartographic convention for labeling bodies of water is to use an italic font and, when possible, a dark blue color. The size of a label indicates the hierarchy of the feature - oceans have larger fonts than seas, which have larger fonts then rivers, larger than streams, etc. Land features are labeled in black, or anything that isn't blue, and are never written in italics. Larger features, land or water, may be written in all capital letters, while smaller features are in lower case.

ATLANTIC OCEAN *GULF OF MEXICO* *Lake Ontario* *Hudson River*

UNITED STATES NEW JERSEY Philadelphia Trenton

Thematic Maps and Symbols

In this tutorial we worked through an example for creating a shaded area or choropleth map. However, there are a number of other techniques that you can use to create a thematic map. QGIS also supports graduated symbols for point and line layers, where the relative size of the symbol (a circle, square, line, or image) represents a value (if you look at the style tab for a point layer, you can change the legend type to graduated symbols). If you have a polygon layer that you'd rather map as graduated circles (instead of shaded areas) you have to convert it to a point layer first (you can do this under Vector > Geometry Tools > Polygon Centroids).

Symbols are used to show qualitative data (name or feature type) or quantitative data (proportions or numbers) and are often divided into four types:

Nominal - qualitative measurements like the name or type of feature, shown using unique symbols.

Ordinal - quantitative measurements with a general order of size, like small, medium, or large, shown using symbols of different sizes or colors.

Interval - quantitative measurements with a specific beginning point and range of specific values (distance, temperature, elevation), shown using a variety of symbols (isolines, shaded areas, graduated symbols).

Ratio - a type of interval measurement that shows the relationship between the area and some phenomena (time to cover a distance, population density).

Symbols are often designed to mimic the features they represent, i.e. airplanes for airports, little buildings with flags to represent schools, etc (these are all examples of nominal symbols). In some cases, features may be represented with geometric shapes (circles, squares, triangles) that can be easily distinguished on small scale maps. Some features may be represented using a standard convention for classifying them, i.e. mining maps may label minerals based on their abbreviation in the periodic table - Sn for tin, Pb for lead, Cu for copper, etc.

A single symbol can be used to identify a feature. Varying the size or color of the symbol can indicate quantity. The width and color of roads on a map is highly standardized to show the type of road and volume - thick blue roads are interstate highways, thick green roads are toll highways, thinner red roads are US highways and thinner black roads are state or local roads (all ordinal symbols).

Considerations and Next Steps

Now that we have mapped this data - what does it mean? How would you interpret this map? Are there any spatial patterns to the data (clustering) or does it appear more or less random? Maps have the ability to answer questions but also raise new ones. In order to understand what's going on, we have to become familiar with the underlying dataset. What kinds of occupations are included in the Health Care Sector, and how might that explain the distribution across different states?

For more practice, some things to try:

- In addition to shaded areas, we can also create graduated circle maps. Convert the STATES_DATA polygon layer to a point layer (using Vector > Geometry Tools > Polygon Centroids) and symbolize the point layer based on the total number of jobs in the health care sector by state. Hop into your map layout and create a bi-variate (two variable) map that shows the number of jobs (as circles) and percent total (as shaded areas). You'll have to turn the labels off, as the map will look too busy.
- Instead of mapping percent totals, try calculating and mapping location quotients instead. Use the formula from the commentary in section 4.3 to populate the column.

- Now that you're familiar with map projections, go back to some of your layers from Chapter 3 and reproject them. Open a blank project and add the SUBWAYS, COFFEE SHOPS, AND ZCTAS_DATA_COUNT layers. Transform them from NAD 83 to NAD 83 / New York Long Island (ftUS) - this is EPSG 2263, which is the NY State Plan zone for NYC. After you transform the layers remember to set the CRS of the project to match them. Since this CRS is in feet, you can experiment with some of the distance tools to create meaningful ouput; under Vector > Analysis Tools you can use Distance matrix to measure the distance from each coffee shop to each subway station, or you can try measuring the distance to the three closest coffee shops for each station.

Chapter 5

Going Further

This tutorial has provided you with a basic introduction to GIS concepts and applications using QGIS. This chapter will cover the next steps you can take on your own.

5.1 Finding Data

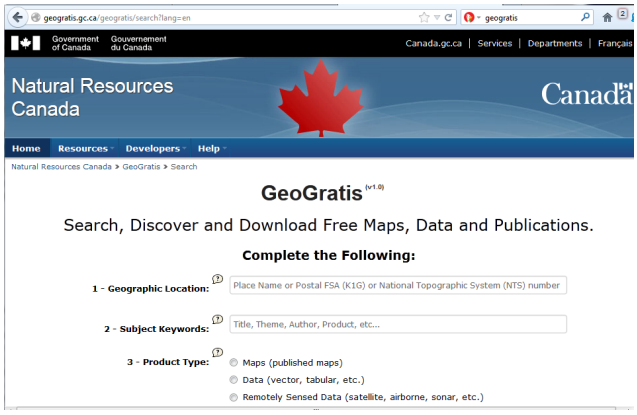
Throughout this tutorial you've been provided with data that you've used to work through various exercises. Once you're working on your own projects, you'll need to find or create the data you need. There is a lot of free GIS data available on the web, created by various government agencies, academic and non-profit organizations, and private companies. You can try a search engine or look at an academic map / GIS library website for a list of helpful links (a list of suggestions is included in the following section). To be strategic about your search, it helps to understand who creates and provides the data:

- *Global / international:* Look at supra-national agencies, like the United Nations (in particular, the UN's Environment Programme has a good site) or academic / non-profit organizations who have enhanced and updated public domain data such as the Global Administrative Areas (GADM) site, the DIVA GIS data page, and the Natural Earth project. If you need satellite imagery the best sites to visit are the USGS and NASA.
- *Country level:* In some cases you'll want to visit a few of the international sites, like DIVA GIS and Natural Earth, to get basic country-level datasets like state or provincial boundaries. But in many instances you may want to visit a mapping agency website or data depository for the specific country you're interested in; you'll find more country specific layers and they will be processed in a way that is readily compatible for mapping attribute data from that country. Most countries have one or two agencies that will provide the bulk of the country's GIS data - a statistical agency responsible for the census, or a mapping agency responsible for surveying. In the US you could go directly to the US Census Bureau or the USGS to download data, or you could visit the central data.gov repository. In Canada, you could visit Statistics Canada directly or visit the Geogratis repository. Some countries provide one central source (Australia), whereas other countries may provide limited or no data.
- *State / Provincial:* You may be able to visit a country level source, like the US Census Bureau to get state, county, or zip code boundaries for the entire state, or you can visit a state level agency to get more specialized datasets for that state. Some states will have state government portals where you can access all data for a state, others may cooperate with a college or university located in that state to provide data via the university's portal. In addition to centralized portals, individual departments or agencies may also provide data directly; road and transportation layers may be provided by a state department of transportation or may be provided through the state's central portal. State agencies are also the most likely source for aerial photography.
- *County / City / Local:* Local governments may have portals where they provide administrative boundaries, transportation data, and real estate or tax parcels, and datasets that would be of local interest (such as

neighborhood boundaries that may not be formally defined elsewhere). You can also look at the geography one step above (state level) to see if data is available for the local area.

- *Gazetteers and Geocoding:* if you can't find an existing GIS dataset, you can always try to create one from an online gazetteer that provides latitude and longitude coordinates for point-based features; the USGS has a US level gazetteer, while the NGA has an international gazetteer. Do you have a list of addresses but no coordinates? Try uploading them to a free geocoding service like the one at Texas A&M GIS Laboratory, which will translate your addresses into coordinates.

- In some cases you may find university or non-profit sites that provide data within a specialized area of interest. While universities typically provide data for the geographic areas where they reside, there may be special labs or research groups that provide data beyond that area; the CIESN (Center for International Earth Science Information Network) site at Columbia University and the NHGIS (National Historic GIS) at the University of Minnesota are two examples.



Regardless of where you download your data, you'll want to examine the metadata for the layers. Metadata can be formally or informally described on the website where you downloaded your files, in narrative documentation that is included with the files you downloaded, or in special XML files that accompany each of your GIS files. There are a few well-defined standards such as the FGDC and ISO 19139 that data creators use to document data, and include elements that explain who created the data, when it was last updated, what the file contains, what the intended purpose of the file is, if it was created for a specific optimal scale, the coordinate system and map projection it was created in, and copyright and use restrictions. You'll want to check the metadata to verify that the data is going to meet your needs and that you can use it for your intended purpose. For example, you wouldn't want to use a generalized boundary file if you're mapping at a large, local scale, and if you are going to use the data for a commercial purpose you need to verify that that's permitted. In any event, you should cite the source of your data in any maps, tables, or reports you create from it.

If you are looking for a particular GIS file and it's provided by several sources, which source should you use? For example, if we wanted census tracts for a particular city, we could download them from the city's GIS page, from a state-based site, from one of ESRI's pages, or from the Census Bureau itself, via the TIGER page or the generalized boundary page. To answer this question, you'll have to examine the download page, and even download the files to view them and their metadata. Here are some things to consider:

- How are the files packaged for download? Do I have to download them one place at a time, or could I get the entire area in one download?

- Who created the files originally? Is it better to go with the original source? Or has a secondary source added some value that makes their files more desirable?

- Can I trust the source? Is there metadata? How did they create the data?

- For vector files, are the layers generalized or not? What scale are they appropriate for?
- For vector files, are the polygons saved as single or multipart layers?
- For vector files, what attributes are available in the attribute table? Are there ID codes that I can readily use to join data? Are there place names that I can readily use as labels?
- For raster files, what is the resolution of the data? Is it appropriate for my intended use?
- What format is the file in? Is it a format I can use, or at least one that I can easily convert?
- Are there any copyright or use restrictions with the data?

Finally, remember that GIS data is often just one piece of the puzzle. It represents the geographic features, but if you need attributes to go with these features (demographic data, weather data, sales data, etc) you'll have to download this data from someplace else (or create it yourself) and process it to make it usable with your GIS data.

5.2 Data Sources

Global

- *DIVA GIS data* `http://www.diva-gis.org/gData`: Country level vector and raster data for every single country in the world. Download individual files or geodatabases. Assembled for the BioGeomancer Project at UC Berkeley and part of the DIVA GIS project. For just global administrative boundaries, you could also visit the GADM database page at `http://www.gadm.org/`.
- *Natural Earth* `http://www.naturalearthdata.com`: Generalized raster and vector data for countries, available at three different scales.
- *United Nations Environment Program* `http://geodata.grid.unep.ch/`: Geodata Portal. Click on "Advanced Search" select "Geospatial Data Sets" under the first drop down box, and hit the red "Search" button. This will take you to a list of global or continental GIS files that you can download.
- *Center for International Earth Science Information Network* `http://www.ciesin.org/download_data.html`: Center for International Earth Science Information Network, hosted by Columbia University, it contains links to datasets for the world, various countries, and the US.

Canada

- *GeoGratis* `http://geogratis.gc.ca/geogratis/search?lang=en`: Canadian government GIS repository provided by the Earth Sciences Sector of Natural Resources Canada.
- *Statistics Canada, maps and Geography* `http://www.statcan.gc.ca/mgeo/geo-eng.htm`: Boundaries, road networks, and place name files from Canada's statistical agency.

United States

- *TIGER Line Shapefiles, U.S. Census Bureau* `http://www.census.gov/geo/maps-data/data/tiger-line.html`: Extracts of the bureau's TIGER Line files for several legal, administrative, and statistical areas in the US, updated annually.
- *Cartographic Boundary Files, U.S. Census Bureau* `http://www.census.gov/geo/maps-data/data/tiger-cart-boundary.html`: Generalized extracts of the bureau's TIGER Line files for several administrative areas (i.e. states, counties, zip codes) and census (i.e. tracts, block groups, metros) areas in the US.

- *National Historical Geographic Information System* `http://www.nhgis.org/`: The NHGIS is a project at the University of Minnesota that compiles and provides historical census boundaries and data for the United States from 1790 to 2000. New users must register, but there is no cost and downloads are free.
- *Data.gov's Geodata Catalog* `http://www.data.gov/catalog/geodata`: Data.gov's Geodata Catalog, a large depository of GIS data from several federal agencies.
- *USGS National Map* `http://nationalmap.gov/viewer.html`:This federal agency provides imagery, digital topographic maps (DRGs), elevation data, and some boundary files.
- *Libre Map Project* `http://libremap.org/`: a non-profit site that provides all of the 24k scale USGS topographic maps (DRGs) for the US.

State of New York

- *CUGIR* `http://cugir.mannlib.cornell.edu`: Cornell University's Geospatial Information Repository. They also compile data at the state, county, and local levels for NY State and they coordinate their activities with NYS GIS.
- *NYS GIS Digital Orthoimagery Direct* `http://gis.ny.gov/gateway/mg/nysdop_download.cfm`: The NYS GIS page for imagery (orthophotos), tiles can be searched by county and year. Imagery for the five boroughs for the most current series is only available by direct, special request. Imagery from the older series is available for all areas.

New York City

- *NYC OpenData* `https://data.cityofnewyork.us/`: this site is a repository of geospatial and attribute data from several city agencies.
- *BYTES of the BIG APPLE* `http://www.nyc.gov/html/dcp/html/bytes/applbyte.shtml`: The NYC Department of City Planning's page has administrative and political boundaries, streets, transportation networks, shorelines, and tax parcels.
- *DoITT Services: GIS* `http://www.nyc.gov/html/doitt/html/citywide/gis.shtml`: The NYC Department of Information Technologies and Telecommunications has transportation networks, survey points, water bodies, building footprints, and open spaces.

Baruch Geoportal

This is Baruch's GIS data repository at `http://www.baruch.cuny.edu/geoportal/`; it includes a mix of public and Baruch-only datasets. Some can be downloaded directly from the web while others can only be accessed by making arrangements with the geospatial data librarian.

- *DRG's for NYC Metro*: scanned and georeferenced USGS topographic maps for the NYC metro area (public)
- *ESRI International data*: features for the world, Canada, Mexico, and Europe (CUNY only)
- *ESRI USA data*: features for the United States (CUNY only)
- *MapPLUTO*: 2008 tax parcel and real estate datasets for NYC (Baruch only)
- *NYC Geodatabase*: geodatabases of NYC neighborhood features and census data (public)
- *NYC Transportation Data*: city and metro area transportation features including: buses, subways, trains, and truck routes (Baruch only)

5.3 Additional Concepts and Applications

In this tutorial you've learned what GIS is, what it looks like, and generally how it works. You've learned how to work with vector-based GIS data to do some basic geoprocessing and analysis, and you've learned the basics of thematic mapping and map design. Here are some things that we didn't cover that you may wish to explore next:

- *Geodatabases.* Instead of storing all of your features in individual shapefiles and your attribute data in several DBFs, store everything in a single database file. Use the database software to organize your data to run spatial and non-spatial queries. QGIS can directly connect to the desktop Spatialite database or the network-based PostGIS database.
- *Working with rasters.* The GDAL plugin allows you to do more interesting things with rasters.
- *Creating and editing vector layers.* QGIS has an entire suite of tools that allow you to edit files point by point, line by line, feature by feature, and to create files from scratch.
- *Georeferencing.* The georeferencing plugin gives you the ability to take non-GIS raster files (a map or chart in a jpg or basic image file that lacks coordinates) and transform it into a GIS layer.
- *Plugins.* Many developers have taken advantage of QGIS' extensible architecture to build plugins that offer a variety of additional features. The officially supported plugins are available under Plugins > Manage Plugins. If you activate the Plugin Installer in that menu you can go to Plugins > Fetch Python Plugins to access and search a database of third-party plugins. The mmqgis plugin is stable and consistently developed, and offers a suite of useful tools.
- *WMS and WFS.* Tap into server and web-based datasets without downloading anything. Data provided in a WMS (web mapping service) format can be displayed as a raster in QGIS, while WFS (web feature service) layers can be viewed as vectors via a plugin.
- *Learn command line tools.* Need to export data from one format to another? Or reproject files? Or rename them? Do you have large batches of files to change or transform? The GDAL / OGR tools, many of which are embedded in QGIS, are also available via the command line or shell and can make your life a little easier.
- *Need more analytical capabilities?* There are a number of other analysis tools that are avaialble under the ftools menu and via plugins. You can also try the QGIS GRASS plugin and learn how to use the powerful GRASS GIS software. The learning curve is steeper, but with the GRASS tools you'll have more than enough features to match the major proprietary software. Geodatabases like PostGIS also give you the capability to perform spatial queries and geoprocessing operations.

The QGIS website and the OSGeo foundation have links to additional manuals and tutorials for learning QGIS and GRASS (see `http://www.osgeo.org/educational_content`). Online, Harvard has a concise and graphics-rich QGIS tutorial at `http://maps.cga.harvard.edu/qgis/` and the Quantum GIS (QGIS) Tutorials blog (not affiliated with the QGIS project) has detailed tutorials for individual tasks at `http://qgis.spatialthoughts.com/`.

In print, Sherman and Mitchell's *The Geospatial Desktop* is great for delving deeper into QGIS and for providing a crash course in GRASS, PostGIS, and the GDAL OGR command line tools. *Open Source GIS: A GRASS GIS Approach* by Neteler and Mitasova is the definitive source for learning about GRASS, and *PostGIS in Action* by Obe and Hsu has become the text for learning PostGIS (although it's easier to grasp if you have general database experience first). In addition to QGIS and GRASS there are a number of other open source GIS products bouncing around that are worth a look. gvSIG, an open source desktop GIS package created by local government agencies in Spain, is a notable alternative.

If you think you're going to become deeply involved in GIS, you may want to consider trying the major proprietary packages in the industry such as ESRI's ArcGIS or Pitney Bowes MapInfo. If you're a current Baruch student, faculty, or staff member you can sign up to take free, self-paced, online courses in ArcGIS as part of the ESRI

Virtual Campus program. Visit the ESRI VC page under the Tutorials and Courses tab on Baruch GIS Guide at `http://guides.newman.baruch.cuny.edu/gis` for information on how to sign up. ArcGIS is available in several computer labs on campus. CUNY affilaites outside of Baruch should contact the site license administrator of ArcGIS on your campus to see who administers the courses to gain access. Once you're familiar with QGIS, the leap to one of the proprietary packages isn't too great because they use a similar interface and operate under the same basic principles. ArcGIS is well documented; there are many books and online tutorials. On the flip side, the software is more resource intensive, is only available for the Windows operating system, and is expensive enough that it's not a viable option for an individual user. You can download and sample a basic, freeware version called ArcGIS Explorer from ESRI's website.

Appendices

Appendix A

ID Codes

- *ISO Country Codes:* `http://www.iso.org/iso/country_codes.htm`
- *US ANSI (FIPS) Codes:* `http://www.census.gov/geo/www/ansi/ansi.html`

INCITS 38:2009 ID Codes for US States (formerly FIPS 5-2)

Name	ANSI/FIPS	USPS Code	Name	ANSI/FIPS	USPS Code
Alabama	01	AL	Montana	30	MT
Alaska	02	AK	Nebraska	31	NE
Arizona	04	AZ	Nevada	32	NV
Arkansas	05	AR	New Hampshire	33	NH
California	06	CA	New Jersey	34	NJ
Colorado	08	CO	New Mexico	35	NM
Connecticut	09	CT	New York	36	NY
Delaware	10	DE	North Carolina	37	NC
District of Columbia	11	DC	North Dakota	38	ND
Florida	12	FL	Ohio	39	OH
Georgia	13	GA	Oklahoma	40	OK
Hawaii	15	HI	Oregon	41	OR
Idaho	16	ID	Pennsylvania	42	PA
Illinois	17	IL	Rhode Island	44	RI
Indiana	18	IN	South Carolina	45	SC
Iowa	19	IA	South Dakota	46	SD
Kansas	20	KS	Tennessee	47	TN
Kentucky	21	KY	Texas	48	TX
Louisiana	22	LA	Utah	49	UT
Maine	23	ME	Vermont	50	VT
Maryland	24	MD	Virginia	51	VA
Massachusetts	25	MA	Washington	53	WA
Michigan	26	MI	West Virginia	54	WV
Minnesota	27	MN	Wisconsin	55	WI
Mississippi	28	MS	Wyoming	56	WY
Missouri	29	MO			

INCITS 38:2009 ID Codes for US Territories (formerly FIPS 5-2)

Name	ANSI State Numeric Code	USPS Code
American Samoa	60	AS
Guam	66	GU
Northern Mariana Islands	69	MP
Puerto Rico	72	PR
U.S. Minor Outlying Islands	74	UM
U.S. Virgin Islands	78	VI

SGC Codes for Canadian Provinces and Territories (2011)

Name	SGC Code	Canada Post Code
Alberta	48	AB
British Columbia	59	BC
Manitoba	46	MB
New Brunswick	13	NB
Newfoundland and Labrador	10	NL
Northwest Territories	61	NT
Nova Scotia	12	NS
Nunavut	62	NU
Ontario	35	ON
Prince Edward Island	11	PE
Quebec	24	QC
Saskatchewan	47	SK
Yukon	60	YT

Appendix B

Latitude and Longitude Distances

Length of a Degree (WGS 84 Ellipsoid)

	Latitude			Longitude	
Degrees	Miles	Kilometers	Degrees	Miles	Kilometers
0°	68.71	110.57	0°	69.17	111.32
10°	68.73	110.61	10°	68.13	109.64
20°	68.79	110.70	20°	65.03	104.65
30°	68.88	110.85	30°	59.95	96.49
40°	68.99	111.04	40°	53.06	85.39
50°	69.12	111.23	50°	44.55	71.70
60°	69.23	111.41	60°	34.67	55.80
70°	69.32	111.56	70°	23.73	38.19
80°	69.38	111.66	80°	12.05	19.39
90°	69.40	111.69	90°	0.00	0.00

Appendix C

Some Common CRS Definitions

C.1 Common Definitions Included in the EPSG Library in QGIS

C.1.1 Geographic Coordinate Systems

WGS 84 (EPSG 4326): World Geodetic System of 1984, commonly used by organizations that provide GIS data for the entire globe or many countries and used by most web-based mapping engines.

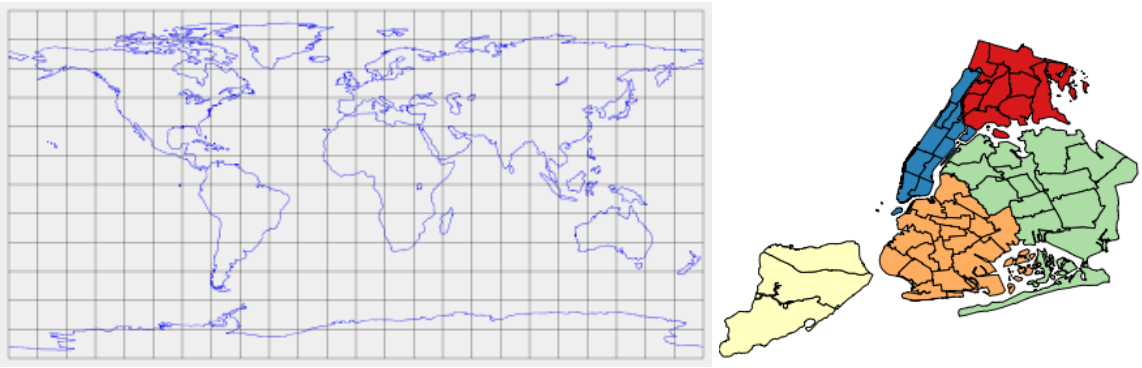
```
+proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs
```

NAD 83 (EPSG 4269): North American Datum of 1983, commonly used by most US and Canadian federal government agencies (the US Census Bureau in particular) that provide GIS data. The definition can be written in two different ways; the first option is more common:

```
+proj=longlat +ellps=GRS80 +datum=NAD83 +no_defs
```

```
+proj=longlat +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +no_defs
```

Since WGS84, NAD83, and all geographic coordinate systems are unprojected they will all look like Equirectangular or "Plate Caree" projections regardless of scale. Global view on the left, zoomed into NYC on the right:



C.1.2 Projected Coordinate Systems for Local Areas

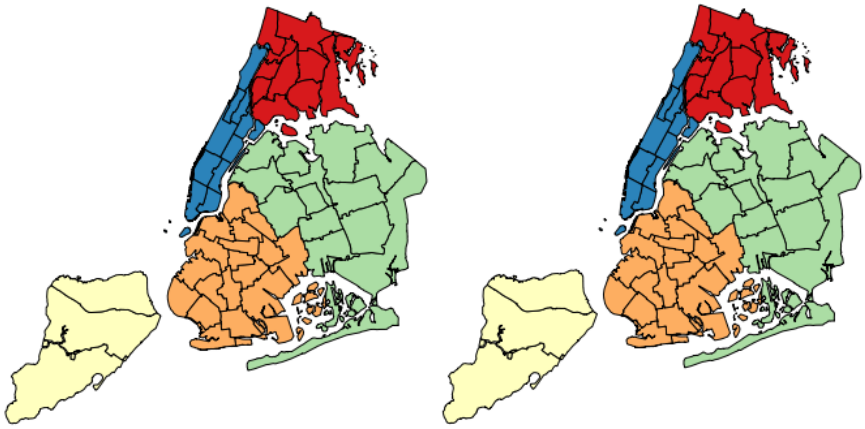
NAD 83 / New York Long Island (ft US) (EPSG 2263): The State Plane zone that covers Long Island and New York City is used by all NYC agencies that produce GIS data. An alternate projection, EPSG 32118, represents the same zone but uses meters instead of feet. Many city, county, and state agencies in the US produce data in their specific state plane zone.

```
+proj=lcc +lat_1=41.03333333333333 +lat_2=40.66666666666666 +lat_0=40.16666666666666
+lon_0=-74 +x_0=300000.0000000001 +y_0=0 +ellps=GRS80 +datum=NAD83
+to_meter=0.3048006096012192 +no_defs
```

NAD 83 / UTM Zone 18N (EPSG 26918): An alternative to State Plane that is better for larger regions and that is applicable outside the US; satellite or ortho imagery is often provided based on the UTM zone where the tile is located. UTM Zone 18N covers much of the east coast of the US. An alternate projection, EPSG 32618, uses WGS 84 as a datum instead of NAD 83.

```
+proj=utm +zone=18 +ellps=GRS80 +datum=NAD83 +units=m +no_defs
```

Visually the difference between State Plane (on the left) and UTM 18 North (on the right) is almost imperceptible when focused on the NYC area, but both are clearly distinct from the basic GCS (WGS 84 / NAD 83):



C.1.3 Continental Projected Coordinate Systems

US National Atlas Equal Area (EPSG 2163): More commonly known as the Lambert Azimuthal Equal-Area projection, this CRS preserves equal areas and true direction from the center point of the map. It's the best CRS in the EPSG library that is appropriate for mapping the North American continent.

```
+proj=laea +lat_0=45 +lon_0=-100 +x_0=0 +y_0=0 +a=6370997 +b=6370997 +units=m +no_defs
```

C.2 Common Definitions That Must be Self-Defined

The following CRS are common continental and global projected coordinate systems that are NOT included in the EPSG library that is part of QGIS; you have to custom define them using the proj4 definitions available at `http://spatialreference.org/`

C.2.1 Continental Projected Coordinate Systems

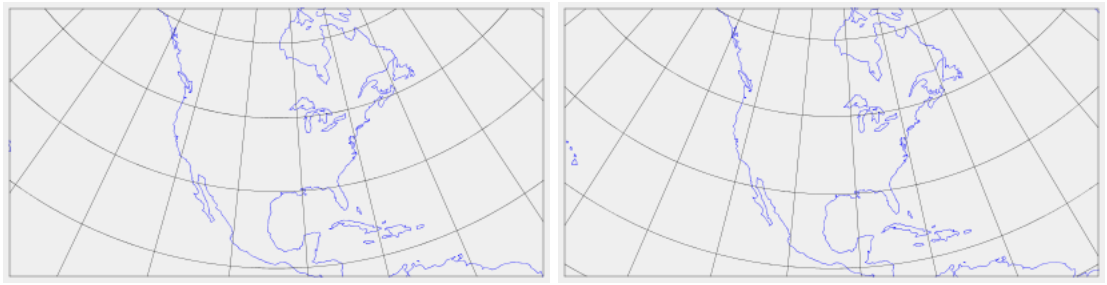
North America Lambert Conformal Conic : Perhaps the most common map projection for North America, a conformal map preserves angles. LCC can be modified for optimally displaying specific countries (i.e. USA and Canada), other continents (i.e. South America, Asia, etc.), or other ellipsoids and datums (WGS 84).

```
+proj=lcc +lat_1=20 +lat_2=60 +lat_0=40 +lon_0=-96 +x_0=0 +y_0=0 +ellps=GRS80
 +datum=NAD83 +units=m +no_defs
```

North America Albers Equal Area Conic : An alternative to LCC, all areas in an AEAC map are proportional to the same areas on the Earth. Can also be modified for specific countries or other continents.

```
+proj=aea +lat_1=20 +lat_2=60 +lat_0=40 +lon_0=-96 +x_0=0 +y_0=0 +ellps=GRS80
 +datum=NAD83 +units=m +no_defs
```

Although difficult to see at this scale, visually Albers Equal Area Conic (on the right) looks more compact east to west versus Lambert Conformal Conic (on the left):



C.2.2 Global Projected Coordinate Systems

Robinson : A global map projection used by National Geographic for many decades. The Robinson map is a compromise projection; it doesn't preserve any aspect of the earth precisely but makes the earth "look right" visually based on our common perceptions.

```
+proj=robin +lon_0=0 +x_0=0 +y_0=0 +ellps=WGS84 +datum=WGS84 +units=m +no_defs
```

Mollweide : A global map projection that preserves areas, often used in the sciences for depicting global distributions on small maps.

```
+proj=moll +lon_0=0 +x_0=0 +y_0=0 +ellps=WGS84 +datum=WGS84 +units=m +no_defs
```

Visually the difference between Robinson (on the left) and Mollweide (on the right) is apparent: