learning.oreilly.com

1. Introduction to Geospatial Analytics

27-34 minutes

Are you a geographer, geologist, or computer scientist? Impressive if you answered yes, but I am a spatial data analyst. In a nutshell, I am interested in exploring data and integrating location. Now that access to location data and geospatial datasets is fairly ubiquitous, most of us are becoming data curious regardless of professional title or area of study. Appreciating the *where* in our analyses introduces a new dimension of comprehending the impact of a wider variety of features on a particular observation or outcome. I spend a lot of professional time examining public health data and large open source datasets in healthcare. Once you become familiar with geocoding and spatial files, not only can you curate insights across multiple domains but you will also be able to target areas where profound gaps exist.

This is also the age of citizen scientists. The accessibility of open source tools and massive open online courses (MOOCs) empowers a broader range of not only professionals but also the data curious. Perhaps you have a hobby or interest in a certain species of bird and would like to access spatial data to learn about their habitats. Where are they nesting? Where are they traveling from and to? Which habitats support the most species and how is this changing over time? You might be able to create maps of your sightings or other variables of interest. We'll explore this idea a little later when we generate a data question to explore.

Meanwhile, if you are interested in a citizen scientist project look no further than the New York Public Library (NYPL) Map Warper project. As you might imagine, NYPL has a vast amount of historical maps. The challenge is to correct the errors in outdated survey technology by searching for modern matching *ground control points* (GCPs) and warp the image accordingly. This is known as *map rectification*.

The [NYPL Map Warper](https://maps.nypl.org/warper) (shown in [Figure 1-1](#)) is a tool for *rectifying* (digitally aligning) historical maps from the NYPL's collections to match today's precise maps. Visitors can browse already rectified maps or assist the NYPL by aligning a map. Everyone is welcome to participate!

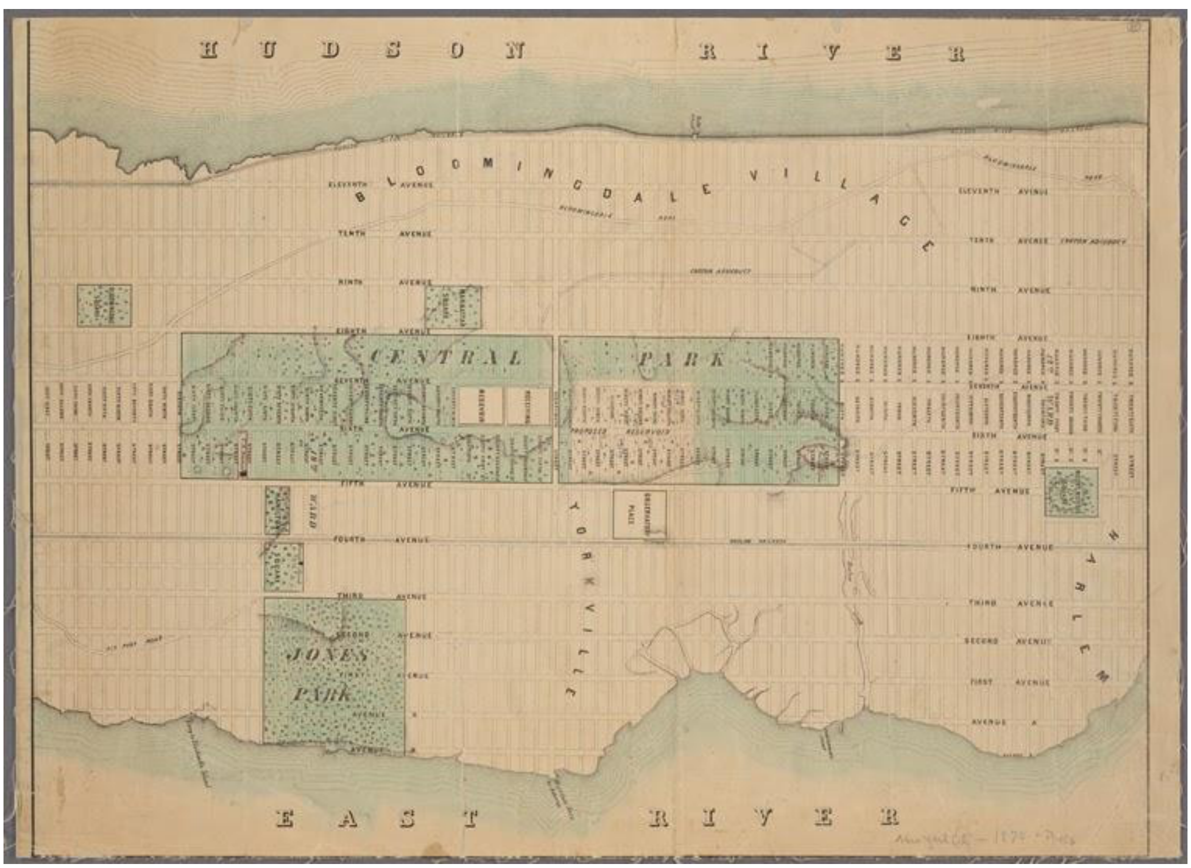


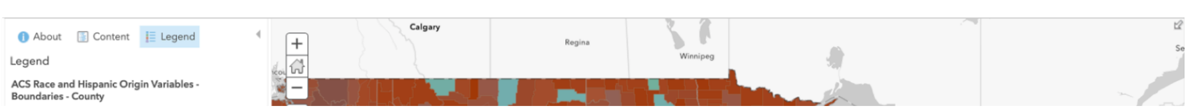
Figure 1-1. Map of Manhattan from 1870

I have used rectified maps to explore development over time in different cities or to accurately reimagine a historical location. How do investments in infrastructure and industrial development impact neighborhoods over time, for example? Figure 1-2 shows a rectified map.

Figure 1-2. Rectified contemporary Manhattan map

There are many opportunities for professionals across multiple industries to include location intelligence in their analytics. Location intelligence describes actionable information derived from exploring geospatial relationships when formulating data questions and evaluating hypotheses. Open source tools are welcoming new end-users, and what we need is a lexicon applicable to a myriad of diverse interests, resources, and learning pedigrees. Yes, enterprise solutions are powerful, but many have limited access to subscription-based applications and tools. There are a variety of options in geographic information systems (GIS) software with pros and cons associated with all of them. I will mention a few when they come up, but although I have access to ArcGIS and QGIS (Quantum GIS), I like to give QGIS the main stage. It is truly open source, meaning that you don't need different levels of licensing for access to all of the available tools. Since this book is intended for a wide level of interests, I want you to be able to explore all of the tools. In my professional work, yes I will move between both ArcGIS and QGIS -- mainly for access to the abundant catalog of GIS datasets in ArcGIS online. I did have to learn the hard way when geocoding in ArcGIS. ArcGIS uses a credit system, and it was easy for me to unknowingly get on the wrong end of it. I didn't realize that the paid service kicked in fairly automatically when I uploaded a CSV file with location data. QGIS, on the other hand, offers two options we will explore later -- both free.

The polygons on the map in Figure 1-3 are different colors corresponding to racial categories. What might we determine about these clusters if we also examine other attributes? Let's sit with these questions for a while. I will give you a hint: the devil is in the details, or should I say, is in the *layers*. The polygons are shaded based on areas where a certain race is the majority population. This is a good place to introduce the expression *defaults should not be the endpoints*. A deeper understanding of GIS will allow you to move away from default settings and create unique and deeper insights.



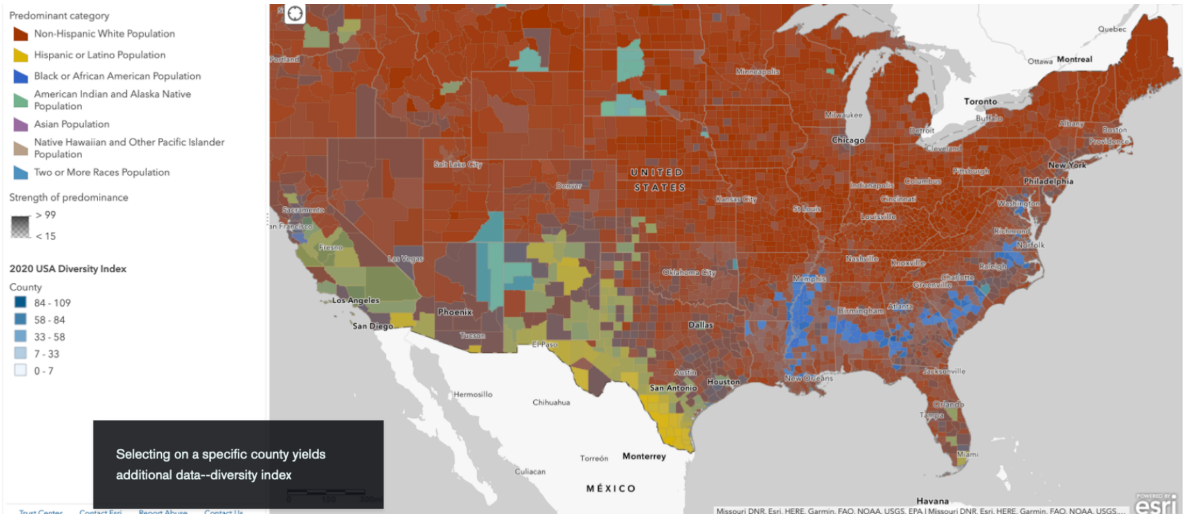


Figure 1-3. Race and diversity (ArcGIS)

Spatial Data Analysis 101 tells us that any analysis requires a defined question. Once you formulate the question you can discover whether there is data that will help you answer or address the hypothesis. I don’t mean cherry-picked data, but all of the data imperative to shape a hypothesis or generate an insight. Should race be treated like a poor biologic proxy for something else or as a social or political construct? We need to understand the variables we have gathered to be able to curate empathy, reveal policy gaps, or address unmet health needs. We may need to reformulate our question if faced with missing data or resources we are unable to access.

When you become familiar with Census data you begin to understand the heavy lifting race has been asked to accomplish. The ability to now examine data on place (housing, employment, transportation, and education to name a few) illuminates the role of spatial data and launched my integration of geospatial applications like ArcGIS and QGIS into my talks about poverty, racial inequity, structural determinants of health, and a wide variety of new questions continuing to emerge. When you rely on spreadsheets or tables of data, I would argue in the absence of spatial considerations you might be missing out on critical insights. Static metrics like where a road might be located, or the coordinates of a specific event, or dynamic measures like the spread of an infectious disease become more powerful when integrated with location intelligence.

A problem uniquely solved by spatial analysis examines the relationships between features identified within a geographic boundary. For example, non-geographic data describes how values are distributed--and we can rely on descriptive statistics. But what if we are curious about the impact spatial relationships might have on these values?

At first glance we can appreciate county level race variables from the American Community Survey. We can see clusters of categorical variables, but do we know anything else? These are predominant categories in each region, but what else might we want to know? This is where geospatial data becomes so valuable.

Note

The annual American Community Survey (ACS) replaced the long form of the decennial census appearing in 2005 and contains a wide variety of questions to identify shifting demographics for gathering information about local communities.

What if we think a little more outside the box? This map you see in Figure 1-4 was created as an exploration of populations where an unexpected emergency would exceed the financial wherewithal of a family. The red squares now tell us that these are the families at the most risk, and bigger squares reflect a larger number of households. To demonstrate the value of geospatial analysis, the green squares are the sites for the summer meal sites in 2020 during the global pandemic. The summer meal sites are part of the Summer Food Service Program (SFSP) serving free meals to low-income children when school is not in session at approved locations.

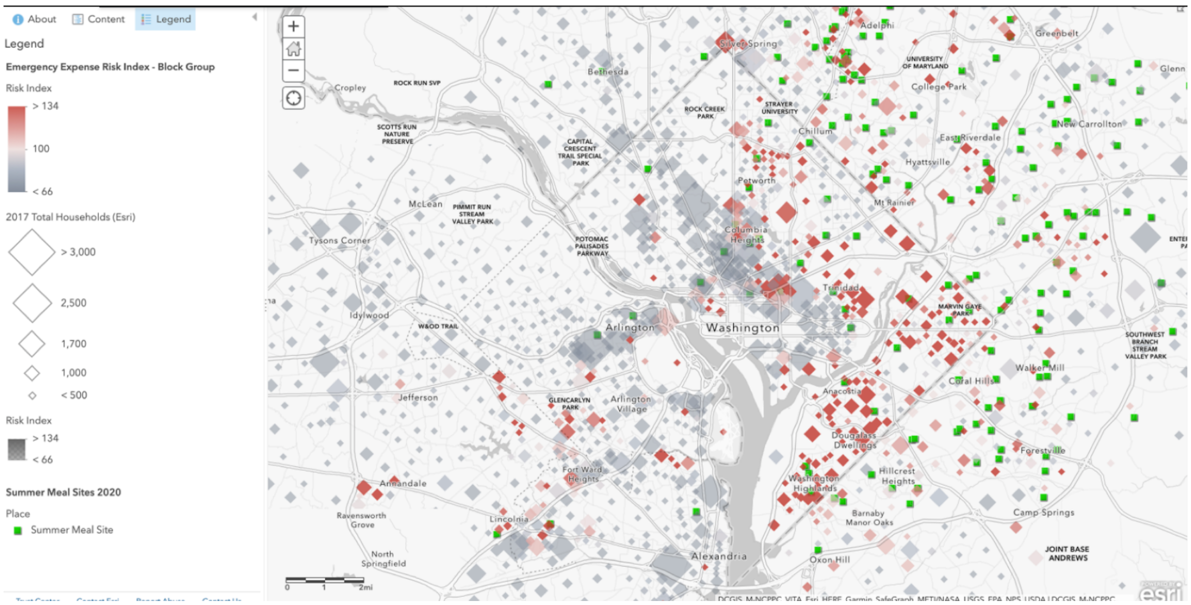


Figure 1-4. Risk Index Summer Meals (ArcGIS)

Open source resources like the Census API yield demographic data accessible for retrieval with a few lines of code in Python. If you aren't familiar with API, it stands for *application programming interface*, a computer interface enabling data transmission between software products. The beauty is in the interface part. The Census API is a wrapper for the United States Census. You can now pull specific data from the annual ACS and the census without downloading entire humongous files of data. Geographies are also captured in census data either at the census website or downloadable from IPUMS, formally known as Integrated Public

Use Microdata Series. The beauty of IPUMS includes person-level data, the harmonization of variables across surveys and the ability to easily download raw data files.

Note

From the IPUMS website:

IPUMS provides census and survey data from around the world integrated across time and space. IPUMS integration and documentation makes it easy to study change, conduct comparative research, merge information across data types, and analyze individuals within family and community context. Data and services are available free of charge.

In my evolution as a data analyst, I began to realize I had bigger and more complex questions to consider, and I needed more resources. With an eye towards working with Census data, I enrolled in an online executive education course in applied analytics sponsored by Columbia School of Engineering. I had worked in the R programming language, so was a bit nervous when I realized the entire course would be taught in Python. I made it through but remember feeling a little whiplashed. After months of completing assignments and a capstone project, I wasn’t sure how to apply these facts to emerging tasks at hand. I learned a lot in the following months that I wish I had been taught in parallel to learning how to code. What are some real-world examples of recursive iteration? Which functions will simplify my work as I wrestle with large datasets? I work independently and was hoping for a solid workflow to guide me through my tasks. My data analytics company is small -- an *n* of 1 to be accurate. The buck starts and stops on my desk.

What I hope to share here is not the complete coding paradigm of Python. But by learning how to write actionable code either within a notebook or a console within QGIS, we can learn by doing. The book contains simple examples, and we will be exploring these concepts in more detail as we move through the chapters. Graphics in earlier chapters are included to familiarize you with how these maps or relationships may be rendered. Later chapters explore the code and platforms empowering Python as a resource for answering geospatial questions.

When formulating a data question, I like to refer to Tobler's First Law of Geography. For example, examining access to public transportation at the community level, understanding how transportation and commute times impact neighborhood home values outside of New York City are important considerations.

> Everything is related to everything else, but near things are more related than distant things.
>
> Waldo Tobler's First Law of Geography

The homes within a certain buffer region of Manhattan are likely to share a wider variety of attributes when compared to outer more distant regions. The ability to map the areas and focus on home values within a certain radius might be confirmatory.

Geospatial data is being collected everywhere. Dynamic sensors with temporal metrics are generating massive amounts of data often being made available to the public. Using open source solutions optimized for geospatial projects is my primary goal. Python is the missing piece for flexibility in manipulating data in both open source and proprietary systems. It is fairly easy to learn, and has a variety of libraries to pivot and reshape tables, merge data, and generate plots either inside or outside of a desktop environment.

Integrating Python into spatial analysis, whether running code in Jupyter notebooks or relying on open source tools like QGIS with the hosted python plug-in, is the focus of this book.

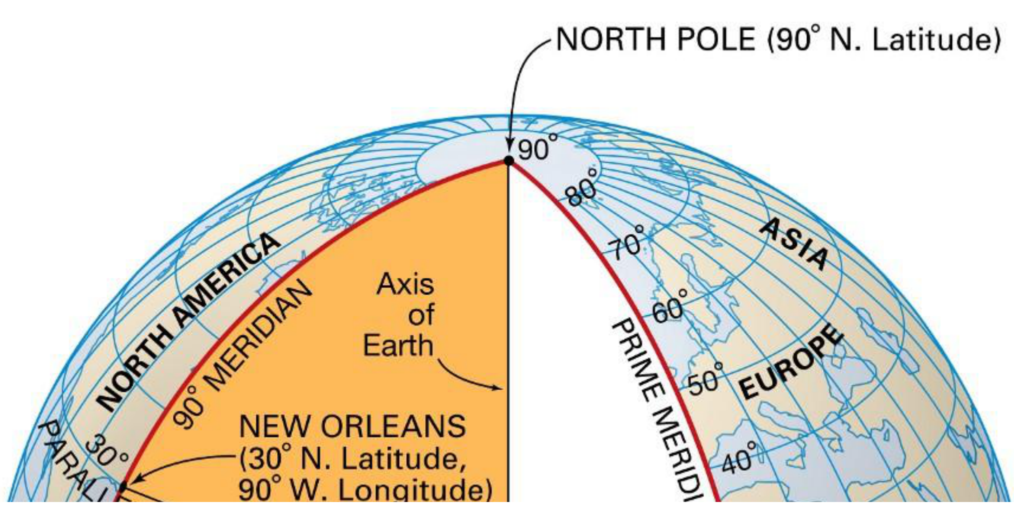
As data visualization continues to become more relevant, think about the most familiar graphics: Maps. Our level of comfort in viewing maps belies the complexity of layers of information, interactivity of multiple datasets, and even how we communicate findings. Accessibility to open source data and platforms extends spatial data science to not only professionals but to hobbyists and individual analysts. But we also need to exercise caution. Familiarity isn't the same as accuracy or competency.

Open source platforms like OpenStreetMap allow us to zoom into a level of detail, revealing structures in the landscape, and add attributes to our analysis. There are Python packages like OSMnx that allow OpenStreetMap downloads into a Jupyter Notebook independent of a specific application or tool (see nearby Note for more).

OSMnx is a Python package that lets you download spatial data from OpenStreetMap and model, project, visualize, and analyze real-world street networks. You can download and model walkable, drivable, or bikeable urban networks with a single line of Python code, and then easily analyze and visualize them. You can just as easily download and work with other infrastructure types, amenities/points of interest, building footprints, elevation data, street bearings/orientations, and speed/travel time.

Conceptual Framework for Spatial Data Science

Planet earth is not perfectly spherical. That makes sense when you think of the chemical nature of the planet, and how the centrifugal force caused by spinning in space would tend to push out the middle, resulting in an oblate spheroid shape. Technically the shape of the earth is an ellipsoid, the circumference around the poles is shorter than the circumference around the equator, almost like the planet has been squished from top to bottom. When we attempt to map the surface, we take a geographic coordinate system and translate it into a projected coordinate system -- for a flat map. The establishment of a *graticule*, the latitude and longitude lines framework of a map, is expedited by QGIS and other GIS applications. Coordinate systems attempt to render the location data in meaningful ways that are close to a single source of truth. These different projections are called *conic*, *azimuthal*, and *cylindrical*. Each one leaves a distortion that must be dealt with. Depending on the chosen coordinate system, we might have distortions in area, distance, direction, and size. You will be glad to know that we don't need to wrestle with these compromises alone -- software manages much of the complicated math. Figure 1-5 shows a geographic coordinate system.



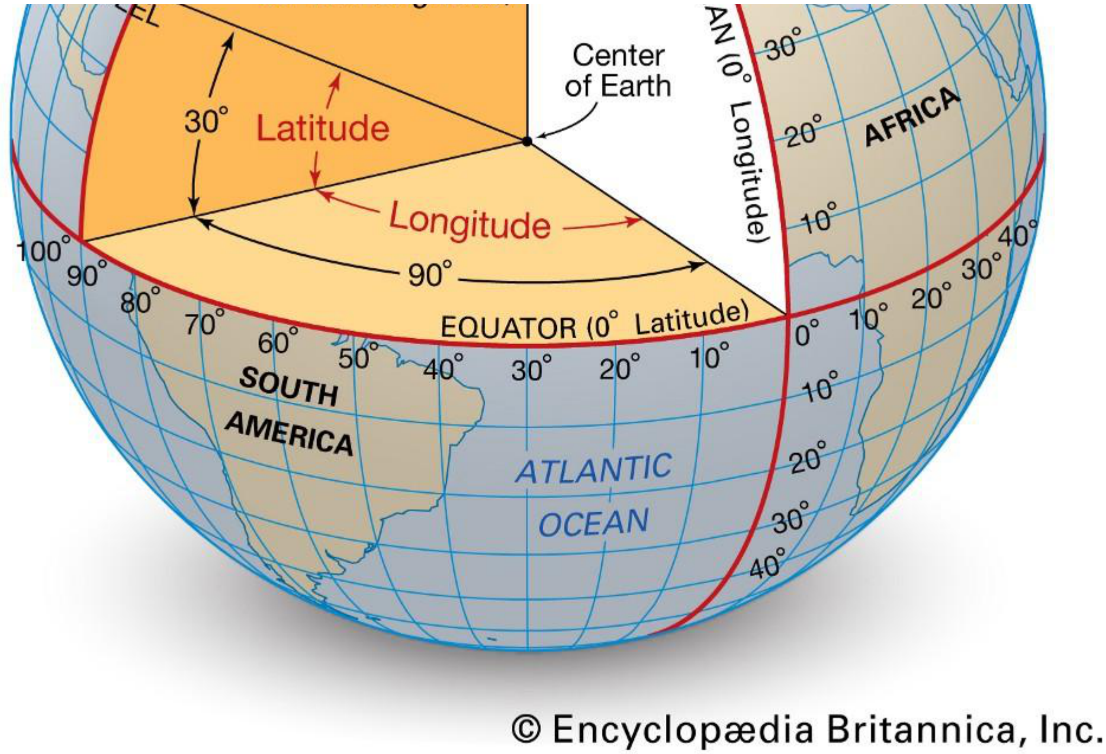
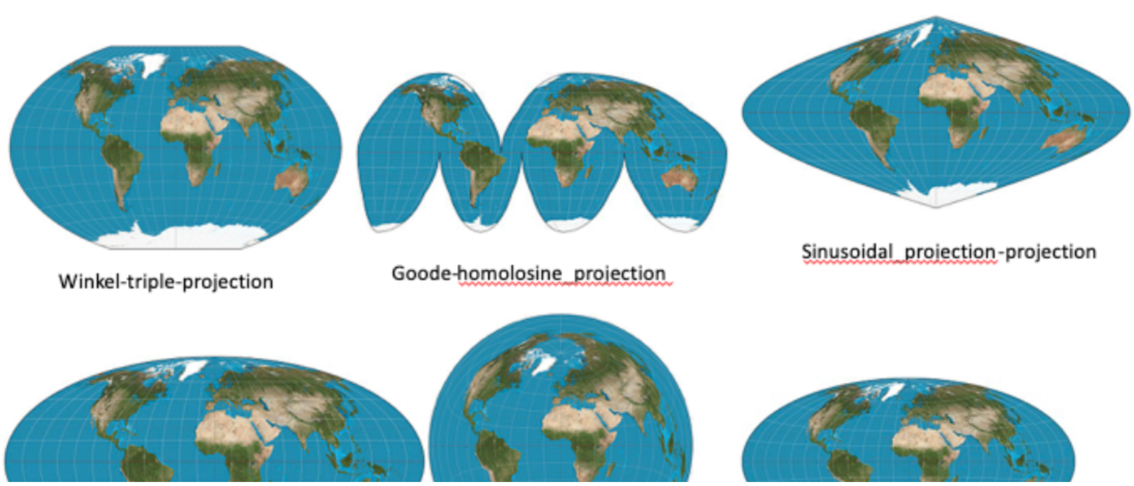


Figure 1-5. Geographic coordinate system.

If you use OpenStreetMap or Google Maps you will be familiar with the Web Mercator coordinate system. Familiarity with variations of these maps will allow you to select the optimal projection as compromises will need to be made. For example, in my work in population health, it is critical that area is maintained on projections. If I'm mapping percentages or raw numbers on a map, and I want to be as impartial as possible, making a small place look too large compared to other places, there's an inherent bias there that affects my interpretation of that map. I'm going to do my best to look at the projection's weaknesses and its strengths and say, "I'm going to choose one that maintains area."

Maps that maintain area are called *equal area projections*. Later chapters discuss this in more detail. I will review QGIS and ArcMap and show how they convert projections for the best cartographic results. Because the earth is not a perfect sphere, we do have to make compromises and accept a little bit of distortion. But if we ensure that the measures most relevant to our visualization are captured in the coordinate system we select, we are most of the way there. Naturally we would like our values to closely match the actual values in the real world. The reality of a "where" question in an interesting dataset is often the impetus for visual communication. For example, spatial phenomena are often iterative instead of linear. Perhaps you are intrigued by an extent or scale of a systematic geography. Where do you begin? The whole world won't fit on a piece of paper or a computer screen--at least not in a visibly available manner for interpretation.

The error or deviation from the accuracy of a globe (remember we are trying to represent the ellipsoid geometry on a flat surface) is typically measured by the Goldberg-Gott error score. The most recent attempt at minimizing the error was researched by J. RIchard Gott and colleagues. For reference the popular Mercator projection (Figure 1-7) has an error score of 8.296 while the Winkel Tripel comes in at 4.563. The lower score reflects fewer errors or compromises in the rendering of the flat map. The new J.Richard Gott map (0.881) has the lowest error score to date when compared to a globe (0.0).



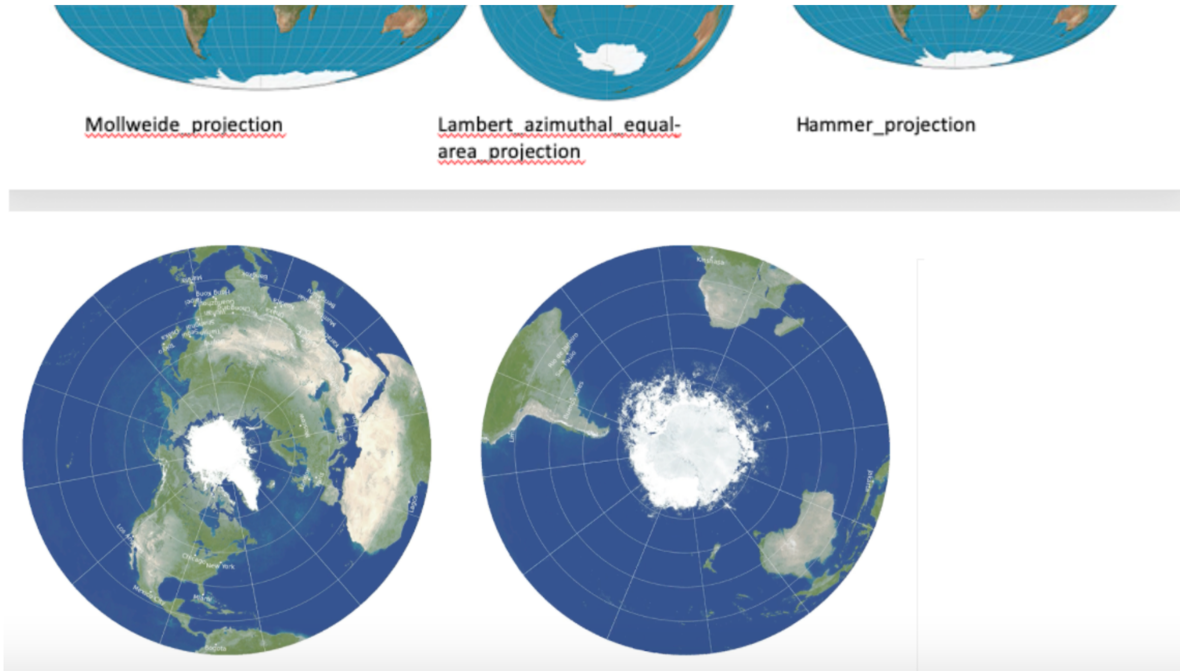
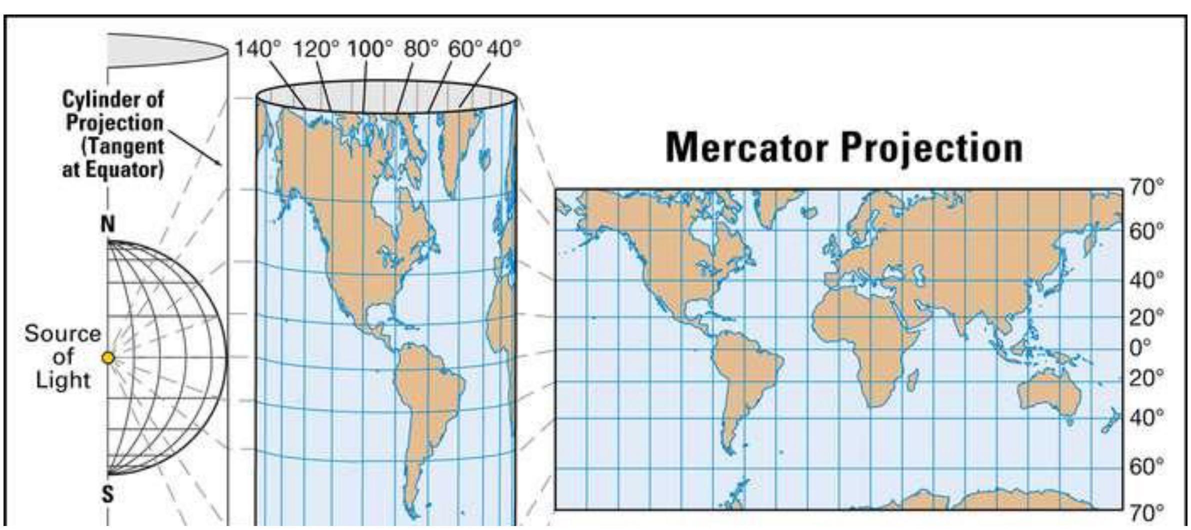


Figure 1-6. Equal area projections

Figure 1-6 shows a few projections, and you can see the impact on our visualization. If you view Greenland in all of these equal-area projections, you can see that it remains the appropriate scale.

Mercator projection maps, on the other hand, like the one shown in Figure 1-7, distort in the area near the poles. We can demonstrate the differences in the maps we create based on which projection we select. This will be evident when we make our selection of coordinate systems in later chapters.



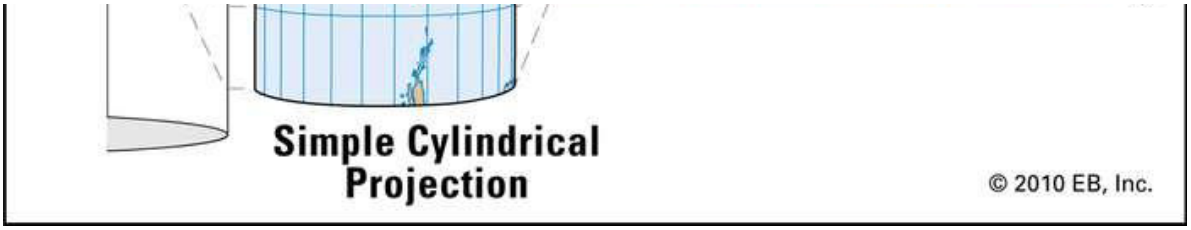


Figure 1-7. Mercator projection

Now that you have been introduced to broad objectives of spatial data analysis, consider challenges facing local communities all the way up to the societal and global level. Think environment, healthcare, biology, geography, economics, history, engineering, local government, urban planning, and supply chain management, for example. Even issues such as access to healthcare, environmental regulation, community planning, and transportation that seem local or regional cross physical and political boundaries, ecological regions, municipalities, and watersheds and possess a spatial component. Since maps are one of the first appreciations we have of data visualization, it makes sense that after interrogating our data we might become curious about location.

Geographic Information Systems (GIS) continuously analyze data and provide real-time insights across a wide variety of industries. Although there may be similarities between spatial and non-spatial analyses, spatial statistics are developed specifically for use with geographic data. Both are associated with geographic features but spatial statistics look specifically at geocoded geographical spatial data. For example, when imagining airport data there are non-spatial statistics for variables such as region or use (military, civilian/public) or on-time arrivals and departures as well as spatial components for analysis (elevation and geographical coordinates). Unlike traditional non-spatial statistical methods, they incorporate space (proximity, area, connectivity, and/or other spatial relationships) directly into their mathematics. Additionally, for those tools written with Python, the source code is available to encourage you to learn from, modify, extend, and/or share these and other analysis tools with others.These complex problems are spatial. Where are these problems occurring and how can we plan for better outcomes in the future?

Note

As already mentioned, I prefer describing what I do as *location intelligence*. It makes sense and is much clearer than saying you work in GIS. Why identify an analysis by the tools? Much better to focus on the outcomes.

The idea of *spatial thinking*, in terms of proximity, overlap containment, adjacency, and the ways of measuring geographic space and the relationship of features and phenomena to one another. This is what we can learn with an introduction to spatial literacy. *Spatial literacy* begins with content knowledge and an understanding of systems and how they interact with the sphere of human influence. The Aspen Global Change Institute (AGCI) identifies six systems of the earth:

1. Atmosphere
2. Cryosphere
3. Hydrosphere
4. Biosphere
5. Geosphere

6. Anthroposphere

   The anthroposphere refers to human presence on earth. Geospatial data allows us to comprehend the interconnectivity of all of these systems, and we have *big data* -- lots of data -- accessible for well-formulated data questions.

   You don't have to become an expert to retain important skills for bigger questions. If you understand at a fundamental level how things work in geospatial data and technology, you are already on your way to more complex ideas. You will learn to formulate a data question and be ready to determine actionable steps to move toward developing your novel application or solution.

   Geospatial problems are complex and change over space and over time and increasingly affect our everyday lives. Look no further than current headlines to find examples of challenges we face in the years to come: racial inequity, climate change, structural determinants of health, criminal justice, safe drinking water, sustainable agriculture practices, ocean acidification, poverty, species endangerment and extinction, and economic strife -- to name a few. How does an individual's location influence their health, well being, or economic opportunity? Questions like these can be examined utilizing GIS by showing patterns between diffusion rates of disease, distance to nearest hospital, roadways, waterways, tree cover and city walkability.

Places as Objects (Points, Lines, and Polygons)

We will dig deeper into how to explore vector data in Python, but first I need to introduce a few concepts that will be useful as we move through the book. *Vector* data is explained by points, lines, and polygons. We will use Python scripts and QGIS integrations to load datasets into a map and examine the structure of vector data. Later chapters will also teach how to customize maps with colors and symbols to improve clarity and accuracy.

In the ArcGIS rendering of Central Park and surrounding buildings we can see the geometry represented as polygons. Additional information about the features might reveal the type of structure, year built, architectural dimensions, and other attributes accessible in an attribute table. The geometry of a feature determines how it is rendered -- either as a point, line, or polygon.

Coverage, shapefiles, and geodatabases are an evolution of ESRI products denoting different generations of their file formats. I won't spend a lot of time discussing the different formats but you may see data files designated as shapefiles (*.shp*) or geodatabase (*.gdb*), so it is good to know their advantages and disadvantages. Compare the file formats of word-processing programs like Microsoft Word to a simple text file. The content (the words on the page) may be the same but the complexity and sophistication is certainly less in the text file. Same thing with the different GIS file formats. Although the content is the same, the elements of functionality are different. If you wanted to share a piece of writing, why would you choose a text file? What if you wanted the document to be read by everyone regardless of software, or what if you wanted ease of portability or a smaller storage format? GIS format files vary in functionality -- shape files do not have a topological or spatial layer, whereas with a geodatabase such a layer is optional. They also vary in simplicity, redundancy, error detection, and storage size. You may not have a choice when working with a dataset that has location data, so it is important to know what is available and what you can say geographically.

Figure 1-8 shows types of buildings in New York City. If you have worked with Census geographic data you are likely familiar with TIGER/Line extracts, or shapefiles. They are grouped as a set with digital files (vector coordinates with a *.shp* extension), an index (*.shx* extension), and dBASE attribute data (*.dbf* extension). There is always talk of their slow disappearance, but because they are prevalent in open source and proprietary systems, they aren't going away any time soon.

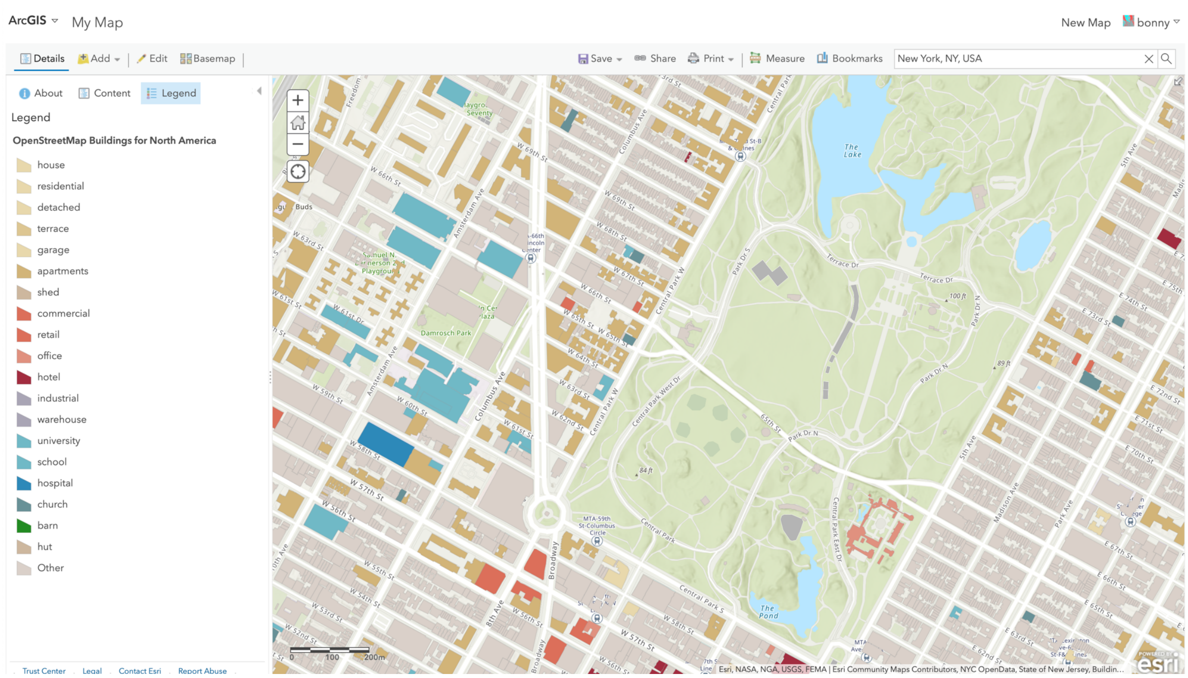


Figure 1-8. Building types in NYC

Geographical systems can work with many types of data. Vector data is what we are referring to when we think of points, lines, or polygons (like a shapefile, for example) and LIDAR (light detection and ranging) surveys. When examining a GIS environment like a landscape, for example, everything within the range of your vision is a feature when rendered in an application like QGIS. Details about the features are captured as attributes.

When location data appears in a spreadsheet, the columns for latitude and longitude create a point. The air quality spreadsheet shown in Figure 1-9 has geographic (*latitude* and *longitude*) and non-geographic data (air quality measurement in the *value* column), allowing a GIS application to add information associated with a particular geographic location. A point feature has an X, Y, and (optionally) Z value. The most familiar coordinate projection system is longitude and latitude. You can be dropped anywhere in the world, and if you provide your longitude (X) and latitude (Y), your location is known -- that's precisely what we mean by a *point* attribute. This coordinate accurately describes where a particular place is on the earth's surface. Point attributes can be quantitative or qualitative descriptions.

A1 *fx* locationId

	A	B	C	D	E	F	G	H	I	J	K
1	locationId	location	city	country	utc	local	parameter	value	unit	latitude	longitude
2	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	27.7	¬µg/m¬≥	34.1324	-118.1834
3	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	25.6	¬µg/m¬≥	34.1324	-118.1834
4	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	26	¬µg/m¬≥	34.1324	-118.1834
5	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	25.4	¬µg/m¬≥	34.1324	-118.1834
6	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	22.8	¬µg/m¬≥	34.1324	-118.1834
7	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	22.6	¬µg/m¬≥	34.1324	-118.1834
8	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	22.2	¬µg/m¬≥	34.1324	-118.1834
9	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	21.5	¬µg/m¬≥	34.1324	-118.1834
10	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	21.5	¬µg/m¬≥	34.1324	-118.1834
11	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	25.4	¬µg/m¬≥	34.1324	-118.1834
12	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	23.1	¬µg/m¬≥	34.1324	-118.1834
13	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	22.5	¬µg/m¬≥	34.1324	-118.1834
14	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	21.2	¬µg/m¬≥	34.1324	-118.1834
15	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	21.8	¬µg/m¬≥	34.1324	-118.1834
16	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	26.6	¬µg/m¬≥	34.1324	-118.1834
17	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	24.4	¬µg/m¬≥	34.1324	-118.1834
18	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	25.2	¬µg/m¬≥	34.1324	-118.1834
19	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	26.1	¬µg/m¬≥	34.1324	-118.1834
20	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	25	¬µg/m¬≥	34.1324	-118.1834
21	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	22.9	¬µg/m¬≥	34.1324	-118.1834
22	62724	Far West Pasadena		US	2021-02-02T	2021-02-01T	pm25	23.5	¬µg/m¬≥	34.1324	-118.1834

Figure 1-9.

Raster data, such as that shown in Figure 1-10, is digital data displayed as a pixelated image, with each pixel corresponding to a specific geographical location. In vector data, instead of a matrix of pixelated data we have points and lines. Both of these types of data will be easier to visualize once we being working with actual data.

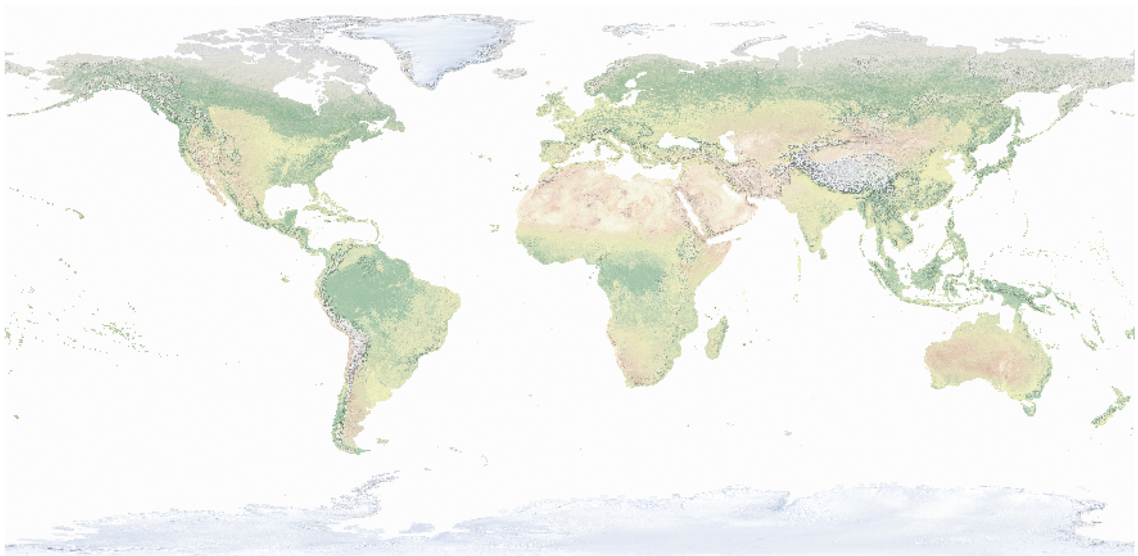


Figure 1-10. Example of raster data (QGIS)

Evaluating and Selecting Data

There are many datasets to explore for use in tutorials for learning a new skill, following along in a new application, or even for launching your own independent geospatial project. Most of the datasets you will see in this book have been vetted and found workable on a wide variety of applications and workflows.

Before selecting your dataset, you need to evaluate the data. The information about your dataset is called *metadata*. You can learn a lot by looking through the metadata, but the most important information about your data includes geographic area, attributes, map projection, scale, and whether there is a fee to use it. You can think of metadata as the label on a can of soup. You want to know what the ingredients might be, and more importantly, whether the soup is good for you. Figure 1-11 shows an example of metadata.

Often there is also a supplemental data file that describes attributes like field headings. This is typically called a *data dictionary*.

Although it can be tempting to explore your own interests and data, I suggest first attempting to follow along with the following suggested data resources. Once you feel confident, then explore and see what you can discover:

- National Geospatial Program
- Geographic Resources Analysis Support System (GRASS)
- OpenGeoportal
- ServirGlobal
- DIVA-GIS
- Natural Earth
- OpenStreetMap
- The National Geospatial DIgital Archive
- National Historical Geographic Information System (NHGIS)
- Geography and American Community Survey (US Census)
- Center of Excellence for Geospatial Information Science (CEGIS)