Introduction to Spatial Data Formats

Michele Tobias

2024-09-27

Overview

Description

In this workshop, participants will learn about commonly used spatial data formats - their structures and when to use any given format.

Learning Goals

By the end of this workshop, participants will understand

- The difference between vector and raster data, as well as some niche formats
- The reasons to use different data formats
- The underlying structure of each data type, conceptually

Prerequisites

There are no prerequisites for this workshop and no software to install.

Introduction

Data Models

All data is an abstraction... a sample... a way of representing some aspect of the real world. We cannot capture everything. A data model is a way to conceptualize data. You can think of these as themes in how data is represented.

With spatial data we typically divide data sets into two data models: Vector Data and Raster Data. We'll explore these in greater depth in this workshop.

File Formats

A data model is related to the concept of a file format. Every data model has at least one file format (and usually several file formats) to store data.

For example, consider a nonspatial data model, the text document. This data model is a file containing text characters. You have a number of options for file formats when you save a text document, including a Word Document (.docx), text file (.txt), Google Doc, and Libre Office File. All of these file types store the same critical thing - text - but each may support additional features like text formatting or inclusion of images. Internally, they may store information in different ways, and each file format may differ in human readability (Can you open the file with any text editor or do you need a specific program like Libre Office for it to work?), but they all share a similar way of representing information - text.

Spatial data models are ways of abstracting the world and each has a set of file formats associated with it. For example, vector data may be saved as a shapefile (.shp), geopackage (.gpkg), or as a comma separated variable (.csv), just to name a few. We'll explore this in more detail later as well.

Why?

Why do we need to know about different spatial data models and data formats? Can't our computers tell what the data is and handle it automatically?

When we **use** spatial data, we need to know what model it uses so we can make good decisions about how to analyze it efficiently. Knowing the model and structure gives us intuition about how we can work with the data inside any given file.

When we **make** spatial data, we need to choose a model and file format that will represent the real world with the highest degree of accuracy.

The Big Picture

Spatial Data is usually composed of two parts:

1. Geometry = The location of the data, where it is in space
2. Attributes = Information about the locations

A quick note about the geometry: we are simplifying a bit here. Geometries not only contain locations but also require a coordinate reference system (also known as a projection). Data typically comes with a coordinate reference system already defined, so we'll deal with this concept another day. For more on this topic, see DataLab's Coordinate Reference Systems Workshop.

You're probably already familiar with the idea of locations with attribute information from apps on your phone or other map-based information sources that are so common today. Let's take a look at an example of a website that helps users find gas stations, GasBuddy:

Each gas station location is indicated with a marker showing the price of gas, but you can also click on the marker to learn even more about the station - the name, address, rating, and when the price was last updated.

All spatial data uses this format - location + attribute information - but how the information is structured depends on the data model and the file format you store the data in.

Let's dig into some more specifics of each data model!

Vector

Exercise

Draw a map on a piece of paper. It could be a map of how you got to this workshop, or how to get to your favorite hiking spot, or of all the mochi doughnut locations in downtown. The topic doesn't really matter. Just draw a map of your choosing.

How did you represent your data? Did you use lines and maybe squares or circles? When asked to draw a map on paper, people usually use the *vector data model* to represent their ideas.

Why? Vector data is best for *discrete objects*. The prompt asked you to think about discrete things - buildings, routes, etc. - so you intuitively picked a data model that worked well for this kind of data.

Description

Geometry Representation: Points, Lines, & Polygons

Attributes Representation: Tables

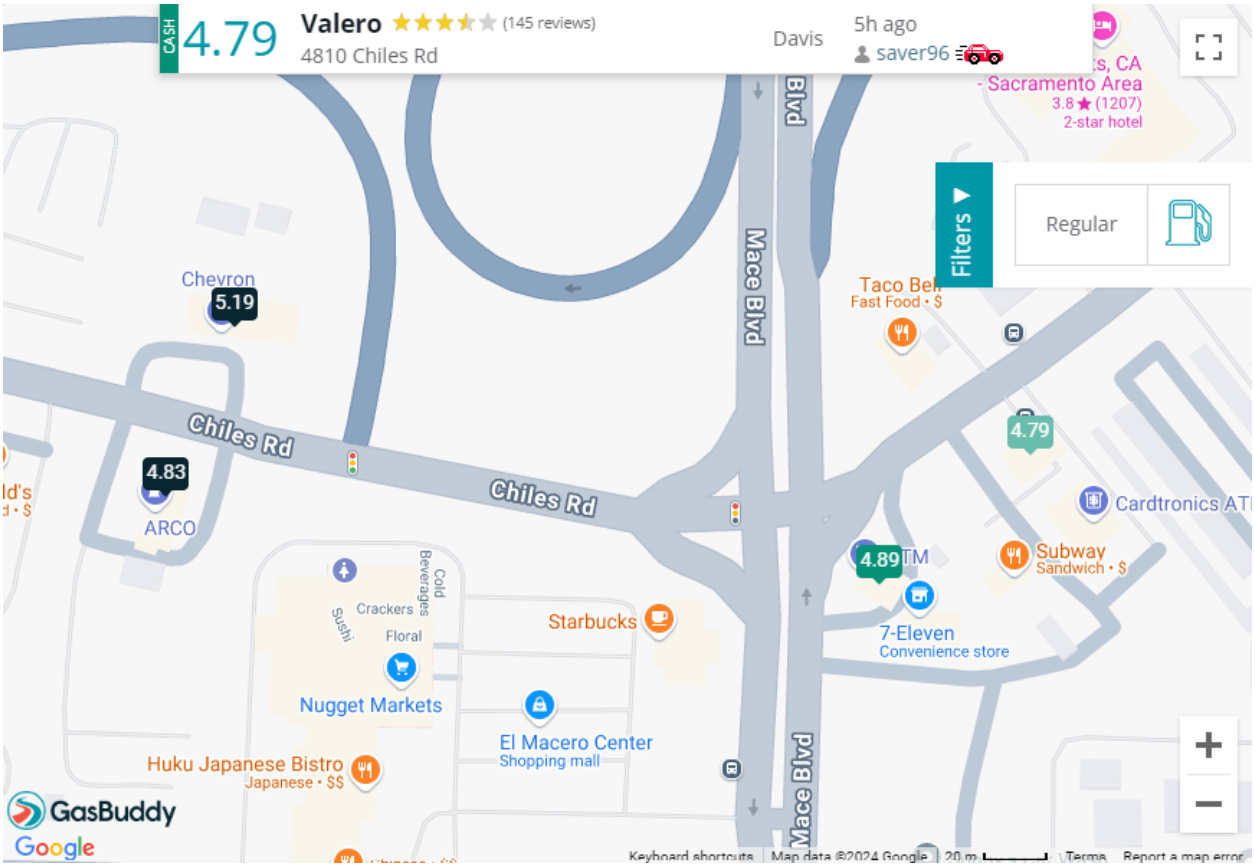

Figure 1: Screenshot of GasBuddy showing 4 gas station locations with their prices

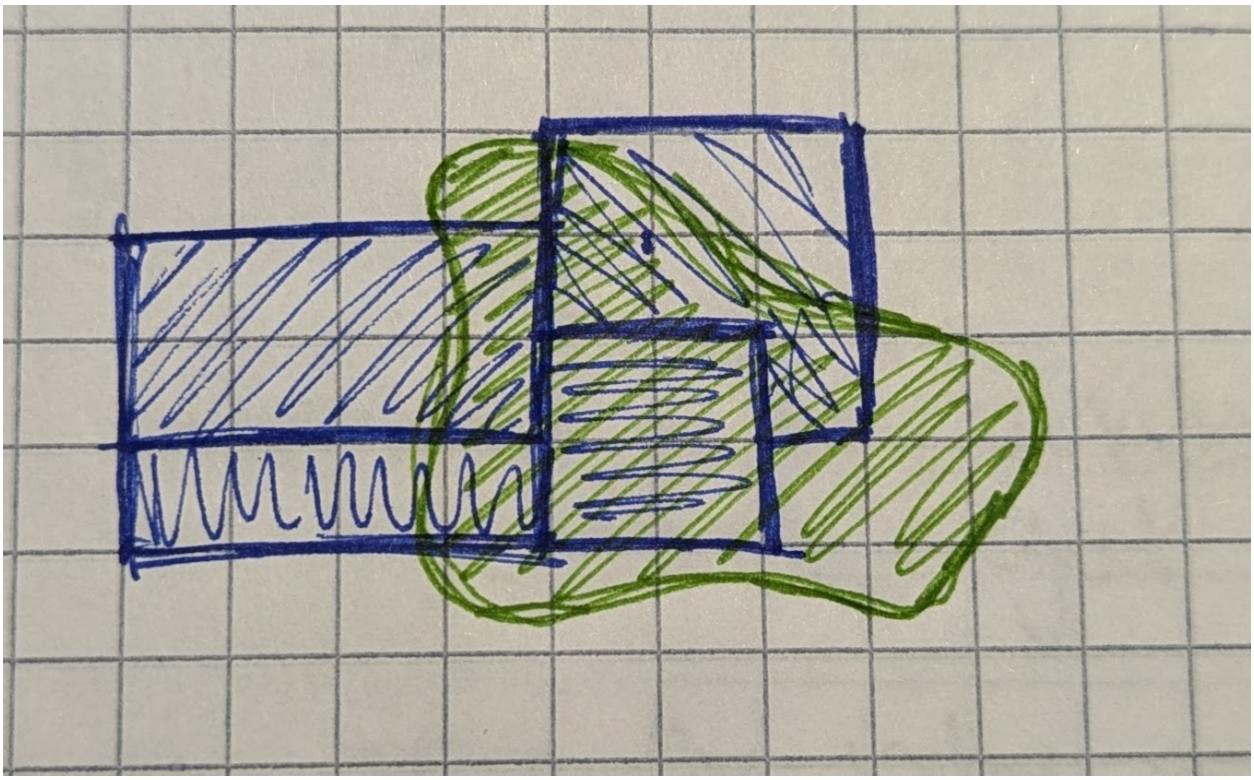

Figure 2: A hand-drawn map

Let's look at some examples of data we can imagine:

Geometry	Example	Attributes
Points	Restaurants	name, region, price range
Lines	Roads	name, speed limit, type
Polygons	Parks	name, year opened, playground (Y/N)

The geometry is usually visually represented at images (points, line, and polygons) but the data is actually stored numerically inside the file as a series of coordinate pairs. The data only covers the place where there is data, not the spaces in between the points, lines, or polygons. Compare this with Raster data in the upcoming sections.

File Formats

Here are some common file formats you'll encounter when working with **vector** data:

Name	File Extension	Notes
Shapefile	.shp (with .shx, .dbf, etc.)	Keep all the "sidecar" files together
Geopackage	.gpkg	Open format, easy organizing
Geojson	.geojson	Open format, human readable

Name	File Extension	Notes
Google Keyhole Markup Language (KML)	.kml, .kmz	Google's spatial data format
GPS eXchange Format	.gpx	A common GPS file type

Raster

Exercise

Open an image or photo on your phone or computer. Zoom in really far. What do you see?

Image files are composed of a grid (imagine graph paper) where each square contains a color. When you zoom out, our brains interpret the image as objects, but it's really a bunch of colored squares. Squares next to each other might be similar in color, but contain small differences in shade or value. A square could contain virtually any color. (Technically, there are limits on the number of colors, but we're not going to get into that today.)

Most photographs you find will be formatted as Raster data.

Why? Raster data is best for gradients.

Description

Geometry Representation: Grid

Attributes Representation: each cell contains one piece of information Can have multiple layers (often called bands)

One piece of information can be stored in each layer or band. This is why it is common to have multiband images. For example, a "color" image is composed on 3 bands - the red, green, and blue components are stored in separate layers and combine by your image viewer software to make it look they way you'd expect.

Coverage is continuous across the dataset, unlike vector data.

File Formats

Here are some common file formats you'll encounter when working with **raster** data:

Name	File Extension	Notes
Geo TIFF	.tiff or .tif	
geopackage	.gpkg	Yes, it stores both raster and vector data
netCDF	.nc	Common in weather & climate data
Hierarchical Data Format (HDF)	.hdf5	Common in weather & climate data

Cross-Over Data

Vectors Representing Gradients

Topo lines, Isotherms, etc.

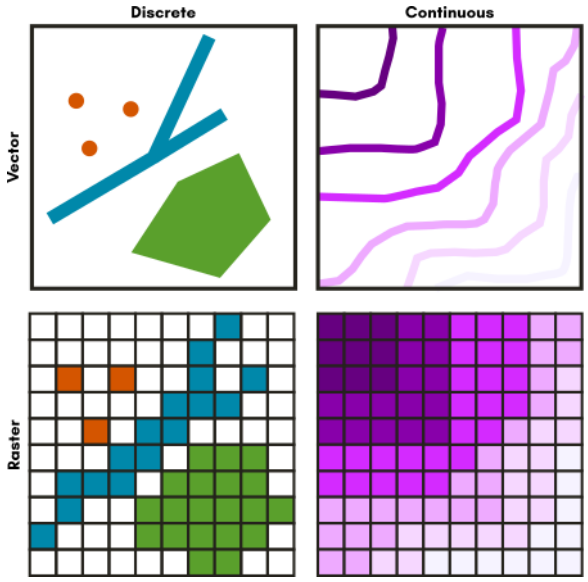

Figure 3: A grid comparing discrete vs. continuous data and raster vs. vector data representations of these two situations.

Raster Representing Discrete Objects

Crop Type Maps, raster masks, scanned maps like USGS topo maps

Other data models

Point Cloud Data

Point cloud data is a special instance of vector data. Points represent the elevation at which a lidar beam intersects with an object to produce a high volume of points that can be used to construct elevation models or 3D models.

The PDAL library (pronounced "pee-dahl" or "poodle") is a common tool for working with and visualizing this kind of data.

Triangular Irregular Network (TIN)

Triangular Irregular Networks (TIN) are useful to represent surfaces. Data you might see stored as a TIN includes slope, elevation, and aspect. The geometry for this data model is combination of points, lines, and polygons, called points, edges, and facets.

Mesh Data

Here are some common file formats you'll encounter when working with **other** data models:

Name	File Extension	Notes

Additional Resources

Workshops & Tutorials

DataLab: Intro to GIS with QGIS

DataLab: Cartography for Map Figures in Academic Journals & Books

DataLab: Spatial SQL

GIS Geography: The Ultimate List of GIS Formats and Geospatial File Extensions

NCAR Climate Data Guide: Common Climate Data Formats: Overview

#maptimeDavis' Workshop Archive - click Archive

Software

QGIS - free and open source desktop GIS with a graphical user interface

ArcGIS Pro - proprietary desktop GIS with a graphical user interface; license fees are negotiated and paid for by the university for UC Davis campus affiliates

Books

Bolstad, Paul, and Steven Manson. *GIS Fundamentals: A First Text on Geographic Information Systems*. 7th edition., Eider Press, 2022.

Citations

This workshop uses the following materials as reference:

Bolstad, Paul. 2019. *GIS Fundamentals: a first text on geographic information systems*. 6th edition, XanEdu.