Urban data science and quantitative methods

Basics of data analysis and visualization

Anand Gajaria & Roman Okhrimchuk

13.05.2025

KSE Kyiv School of Economics

HAMBURG UNITAC

hcu HafenCity Universität Hamburg

UN-HABITAT

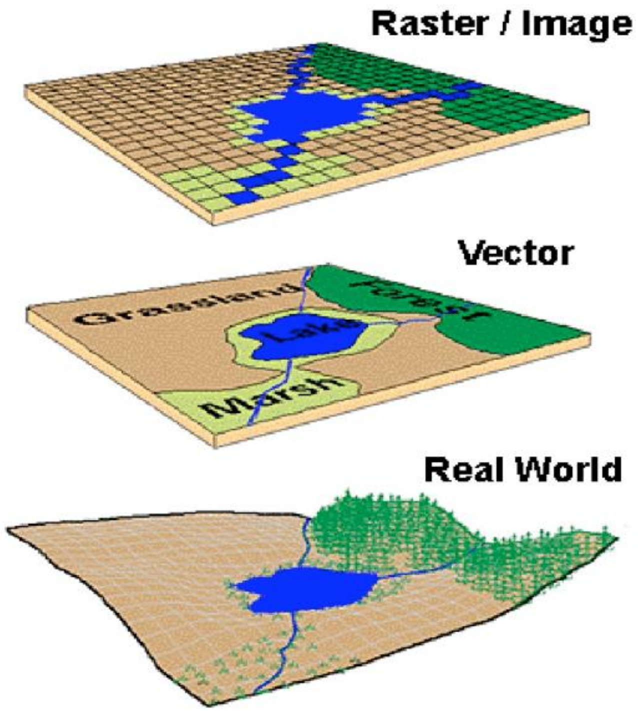
UNITED NATIONS
OFFICE OF INFORMATION AND
COMMUNICATIONS TECHNOLOGY

Presentation Outline

- Introduction to Georeferenced data.
- What is Vector data?
- Types of Geometry.
- Vector data formats and tools to process it.
- Processing Vector data.
- Detecting Anomalies in vector data
- Data Visualization.
- Practical use case presentation.

Georeferenced data

- Geospatial data, also known as **spatial data**, refers to information that is linked to a specific location on Earth.
- It is tied to a **coordinate system** (such as latitude and longitude), allowing precise placement on a map or globe.
- **Primary formats:**
  - Vector data
  - Raster data
- Adds a **spatial dimension** to data.
- Essential for **urban planning**, **infrastructure design**, and **emergency response**.

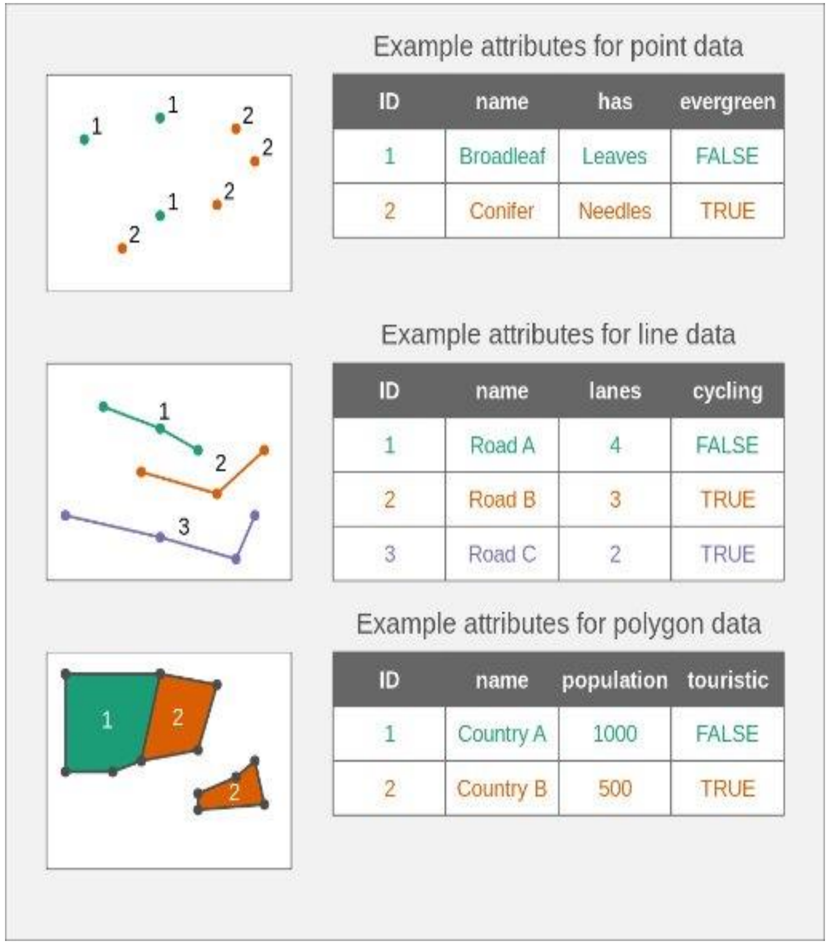


What is Vector data?

- Vector data represents real-world features (e.g., buildings, roads) in a GIS.
- Each feature is something visible on the landscape.
- Every feature has:
  - **Attributes** – Descriptive data (name, population, type).
  - **Geometry** – A shape made up of vertices (coordinates).

Types of Geometry

- **Point**: A single vertex. Used for discrete features.
- **Polyline (Line)**: Two or more vertices, not closed.
- **Polygon**: Three or more vertices, closed shape.



Example attributes for point data

ID	name	has	evergreen
1	Broadleaf	Leaves	FALSE
2	Conifer	Needles	TRUE

Example attributes for line data

ID	name	lanes	cycling
1	Road A	4	FALSE
2	Road B	3	TRUE
3	Road C	2	TRUE

Example attributes for polygon data

ID	name	population	touristic
1	Country A	1000	FALSE
2	Country B	500	TRUE

Vector data formats & Tools to process it.

- **Formats**
  - Shapefiles
  - Geojson data
  - KML/KMZ
  - PostGIS
- **Tools**
  - QGIS
  - Python Programming
  - Spatial database

Processing Vector Data

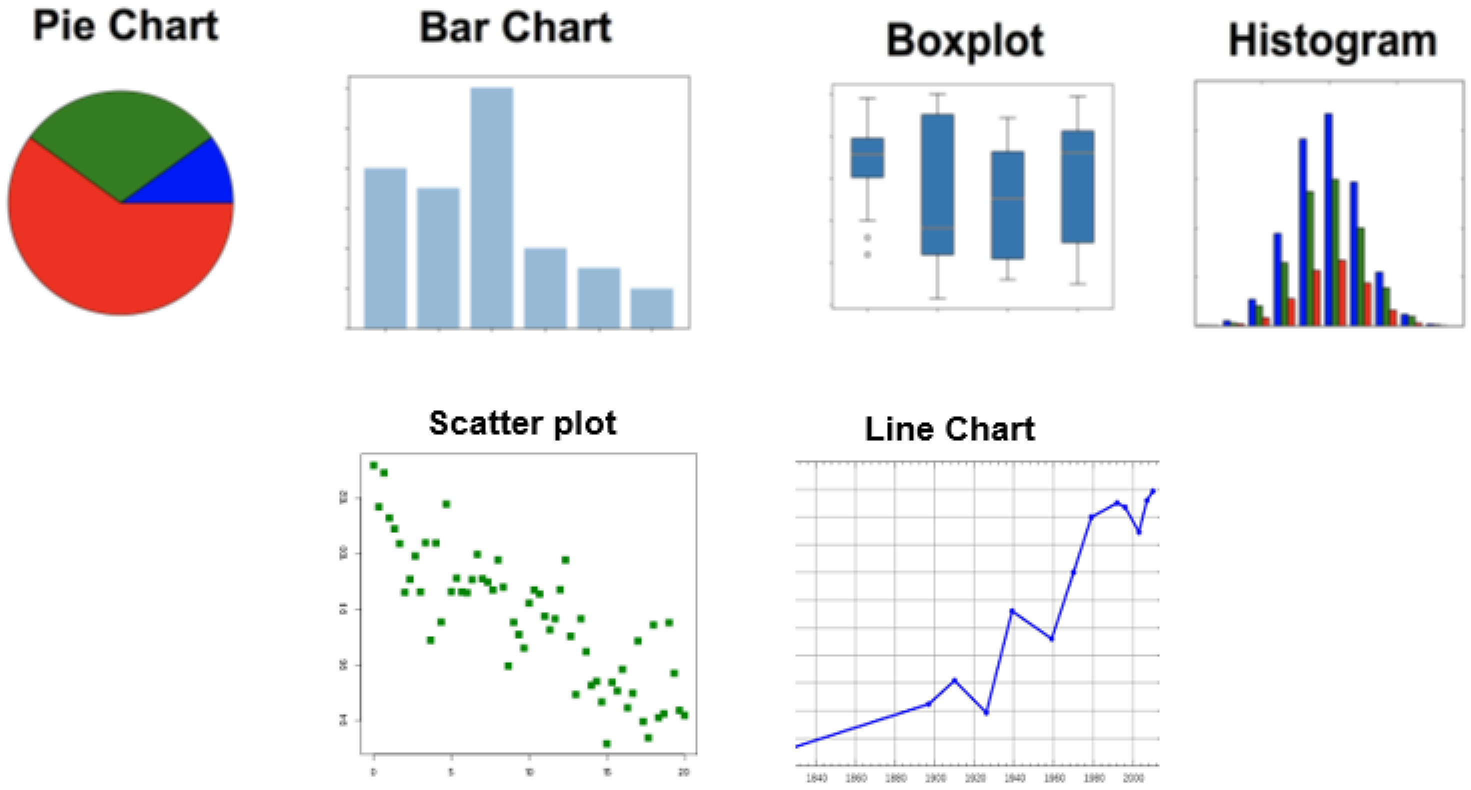
- **Reproject**
  - Converts data into a common coordinate reference system (CRS). *Use it when combining data from multiple sources to ensure alignment and accurate distance/area calculations.*
- **Buffer**
  - Creates zones around features (e.g., 500m radius around a school or hospital).
  - *Use to model walkable access, service areas, or impact zones.*
- **Clip**
  - Cuts features to fit within a specific boundary.
  - *Use when focusing analysis on a specific area like a neighborhood, district, or city.*
- **Spatial Join**
  - Combines data from two layers based on location (e.g., assigning schools to the district they fall within).
  - *Use to transfer attributes across layers for summarization or analysis.*

Detecting Anomalies in Data

- **Geometry Errors**
  - Invalid or empty shapes
  - Self-intersections or disconnected geometries
- **Attribute Inconsistencies**
  - Misspelled values (e.g., “scool” instead of “school”)
  - Missing or inconsistent field names
- **Spatial Outliers**
  - Features located far outside the study area

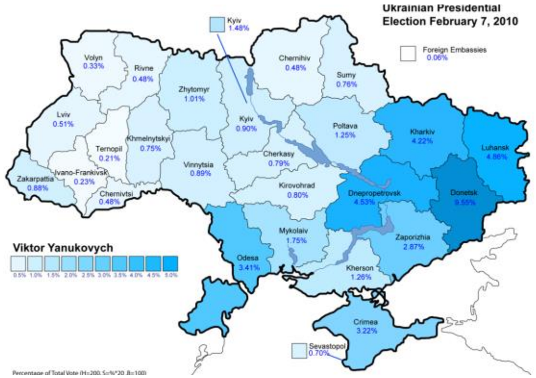
Attribute Data Visualization

-



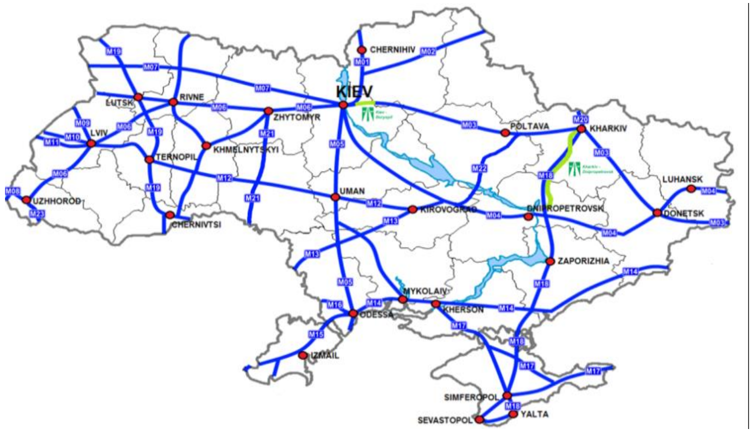
Vector Data Visualization

-



Choropleth Map

Point Map

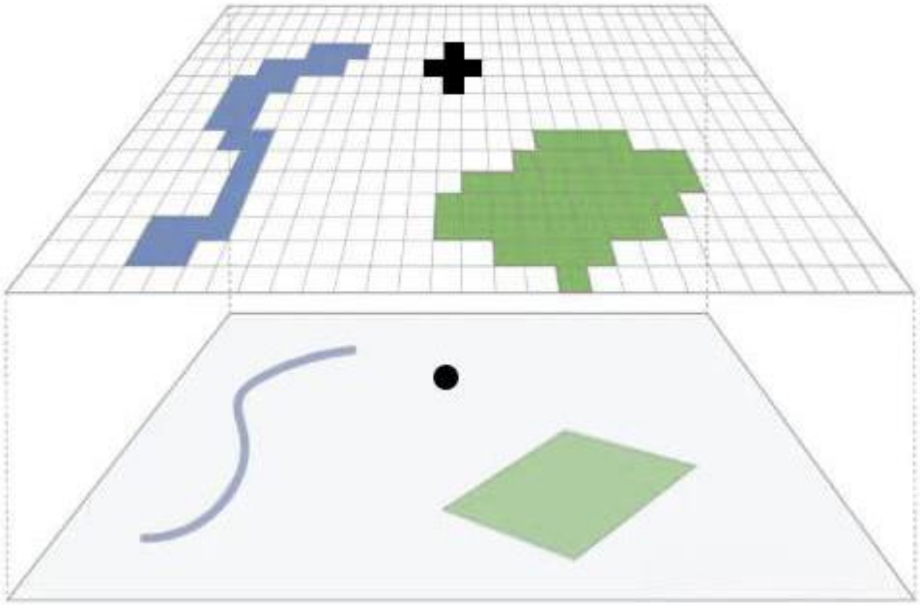


Line and Network Map

What is raster data?

Raster data is a digital representation of spatial information organized as a grid of pixels (cells), where each pixel holds a value representing attributes like color, elevation, temperature, or land cover.

The difference between vector and raster data types



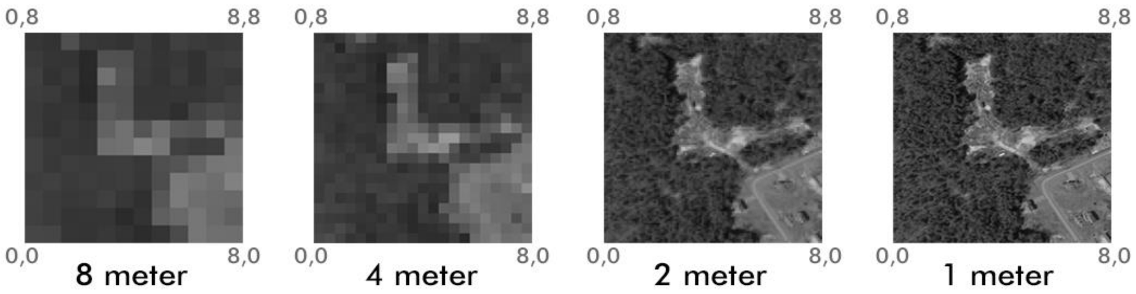
In the bottom part of the image, we see **vector data**:
A **point** (black dot) representing a specific location
A **line** (curved blue line) representing a road or river
A **polygon** (green shape) representing an area like a park or building

In the top part, we see how these features are **converted to raster format**:
The space is divided into a **regular grid of square cells**
Each vector feature is filled into the grid based on its shape and location
Each cell is assigned a value (e.g., 1 for road, 2 for forest)

This process is called **rasterisation,** it allows vector data to be used in raster-based analysis.

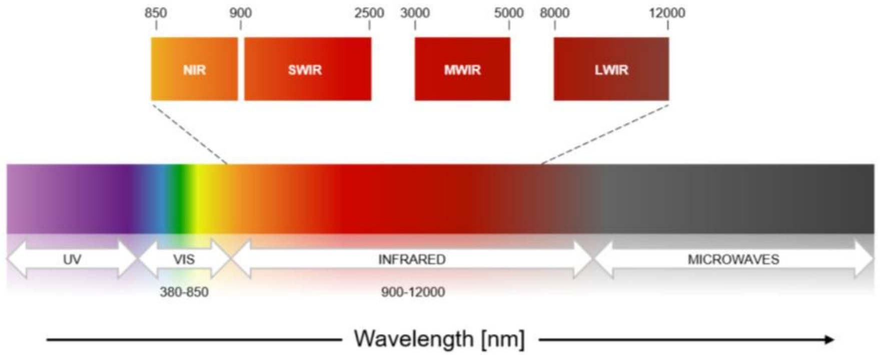
What the term "resolution" means in raster data and why you should know It

Spatial



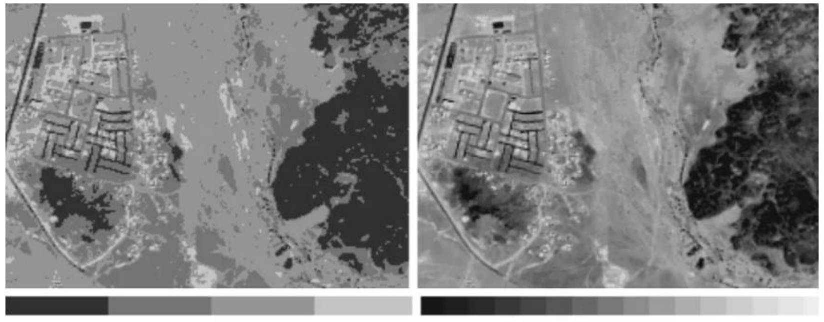
Source: National Ecological Observatory Network (NEON)

Spectral



Source: https://sunex.com/2021/02/17/swir/

Radiometric

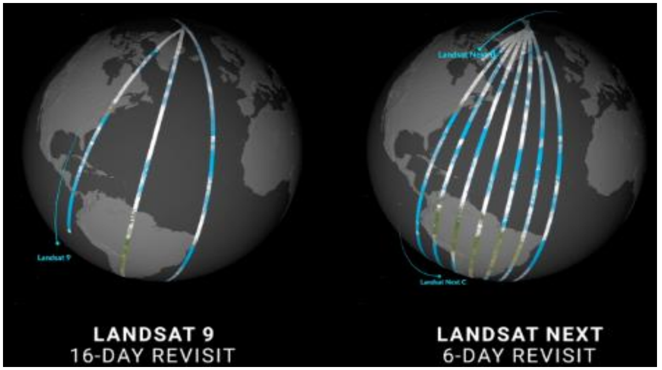


2 bits (4 Values)

4 bits (16 Values)

Source: https://www.earthdata.nasa.gov/

Temporal



Source: https://svs.gsfc.nasa.gov/14379

Geographic vs projected coordinate systems in raster data

Geographic coordinate systems (GCS) represent locations on the Earth's surface using latitude and longitude expressed in degrees. While suitable for global referencing and data integration, GCS is not ideal for spatial analysis, as degrees are angular units and do not represent consistent distances. As a result, area, distance, and other spatial measurements can be highly inaccurate, particularly at higher latitudes.

In contrast, **projected coordinate systems (PCS)** convert the curved surface of the Earth into a flat, map-like surface using mathematical formulas called projections. PCS use linear units such as meters or feet, allowing for accurate calculations of distance, area, and spatial relationships. This makes projected systems essential for raster analysis, map algebra, buffering, and other geoprocessing tasks that require spatial precision.

To ensure meaningful and accurate results, raster data should be reprojected into an appropriate projected coordinate system before conducting any analytical operations.

Raster data formats with georeferencing

GeoTIFF (Georeferenced Tagged Image File Format) - a TIFF-based format with embedded geospatial metadata (coordinates, projection) for maps, satellite imagery, and GIS applications;

COG (Cloud Optimized GeoTIFF) - a GeoTIFF variant optimized for cloud storage, enabling efficient partial data access;

NetCDF (Network Common Data Form) - a flexible, self-describing format for multidimensional scientific data (e.g., climate models, oceanography), supporting metadata and large datasets;

JPG (Joint Photographic Experts Group) - used for georeferencing scanned maps in GIS software (e.g., QGIS) by adding control points and generating a world file (.jgw);

PNG (Portable Network Graphics) - used for georeferencing scanned maps in GIS software (e.g., QGIS) by adding control points and generating a world file (.pgw);

GeoPDF (Geospatial Portable Document Format) - a geospatial PDF standard that embeds coordinates, projection, and vector layers directly into a PDF.

Software for processing and visualizing raster data

QGIS

The most popular open-source GIS software for working with both raster and vector data. Free and powerful, QGIS supports visualization, reprojection, raster calculations, and more, all through a rich set of built-in tools and community-developed plugins.

ArcGIS PRO

ArcGIS Pro is a powerful professional GIS software developed by Esri. The software supports advanced spatial analysis, visualization, and high-quality map production. It also integrates well with cloud-based services and the broader ArcGIS ecosystem.

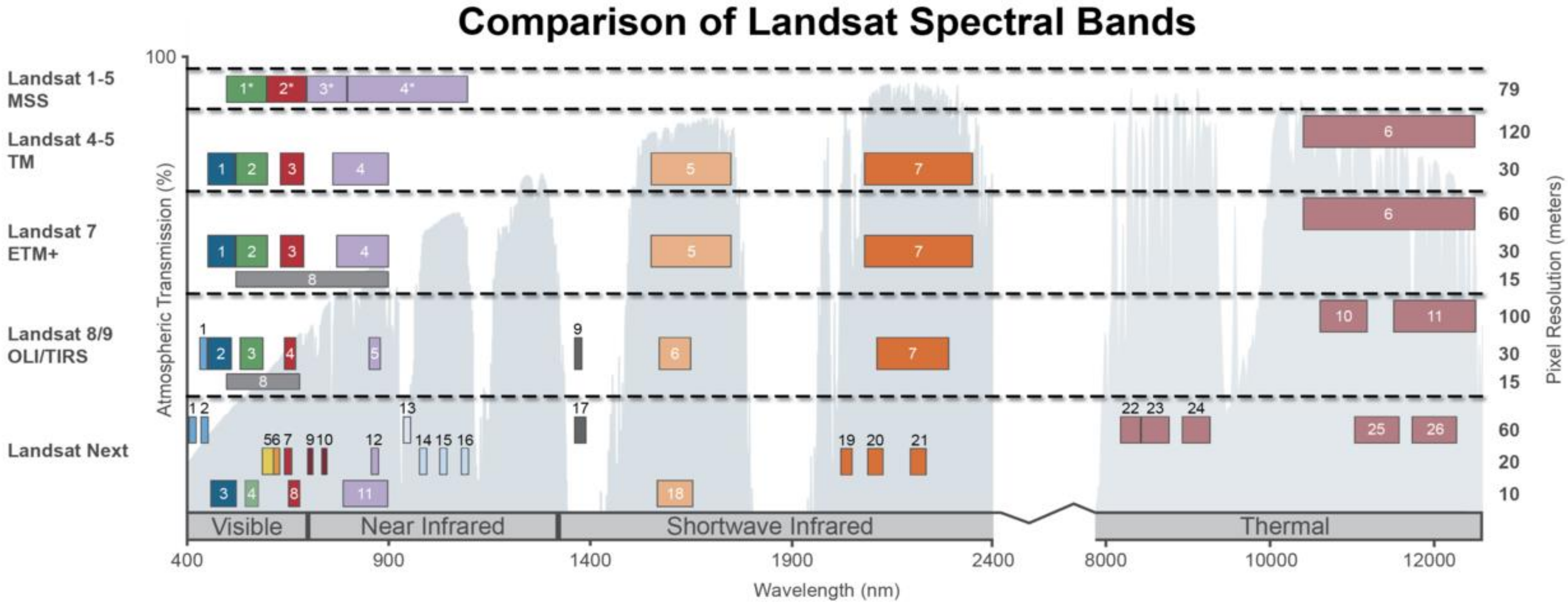
Google Earth Pro

Google Earth Pro is a free desktop application that lets users explore and visualize satellite imagery and other georeferenced data. It is ideal for basic mapping, viewing historical imagery, and exporting high-resolution map screenshots. While its analytical capabilities are limited, it remains a popular tool for quick visual exploration, presentations, and educational use.

Python libraries for processing and visualizing raster data

- rasterio - for reading and writing georeferenced raster data, built on GDAL;
- xarray - works well with multi-dimensional arrays (NetCDF, Zarr), good for climate and EO data;
- lexcube - A lesser-known tool, often used for 3D visualization of multi-dimensional arrays;
- numpy - core for numerical operations on raster arrays;
- matplotlib - standard plotting library for visualizing 2D raster data;
- dask - handles large raster datasets with parallel/distributed computing.

Example of visualization based on EO (Earth observation) data – Landsat 9 imagery



Comparison of Landsat Spectral Bands

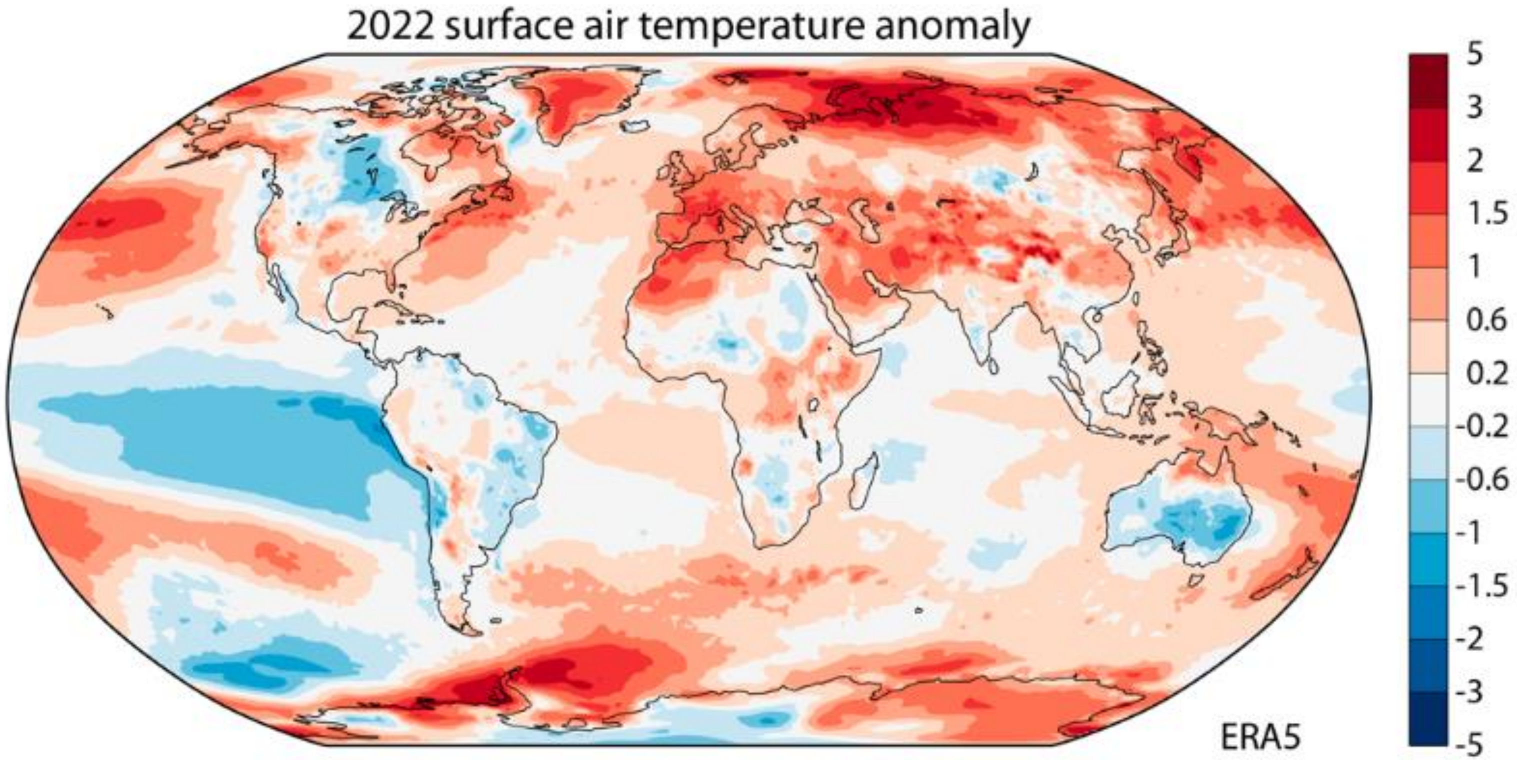
* MSS bands 1–4 were known as bands 4–7, respectively, on Landsats 1–3

USGS, March 2024

This image shows the bandpass wavelengths for the Landsat 1-9 sensors. *Landsat MSS = the numbers shown are for Landsat 4 and Landsat 5; Landsat 1-3 band numbers are 4, 5, 6 and 7

Source: https://www.usgs.gov/media/images/spectral-bandpasses-all-landsat-sensors

Example of visualization using Big Data – ERA5 weather data



Annual average surface air temperature anomaly (°C) for 2022,
relative to the average for the 1991–2020 reference period

Source: https://climate.copernicus.eu/how-c3s-era5-reanalysis-dataset-can-help-policymakers

Urban data science and quantitative methods

KSE Kyiv School of Economics

HAMBURG UNITAC

Resources and useful links

https://github.com/romanokhrimchuk/plot_geodata_kse

https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=download

https://earthexplorer.usgs.gov

Urban data science and quantitative methods

Дякую!
Thank you!

KSE Kyiv School of Economics

HAMBURG UNITAC

hcu HafenCity Universität Hamburg

UN-HABITAT

UNITED NATIONS OFFICE OF INFORMATION AND COMMUNICATIONS TECHNOLOGY