Data Science & AI for Economists

Lecture 3: Intro GIS in R

Zhaopeng Qu
Business School,Nanjing University
September 15 2025

Roadmap

Today and next Thursday's Agenda

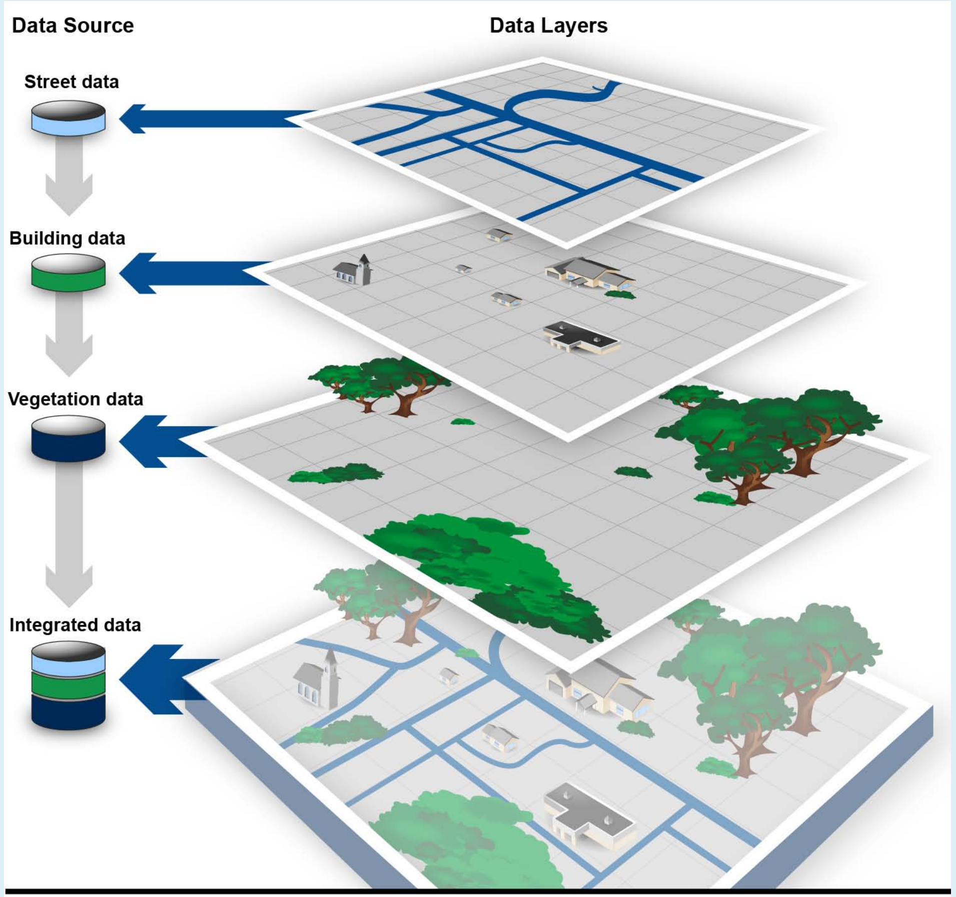
- Intro to GIS
  - What is GIS?
  - Why is GIS useful for Economists?
  - GIS Data Types and Coordinate Systems
- R in Practice:
  - Intro to R packages: sf, raster, tmap, leaflet
  - Using R to open map data
- Nightlights Data in Economics
  - What and Why NTL?
  - Some Applications of NTL in Economics
  - How to get NTL data?
  - How to process NTL data?
- Using NTL data to practice GIS Analysis in R
  - Using package blackmarbleR to download NTL data
  - Using package raster and sf to process NTL data

Introduction to GIS

Introduction to GIS

What is GIS?

- GIS, thus **Geographic Information System**, is a system that allows you to create, manage, and analyze spatial data.
- The main conceptual feature of GIS: **layer data based on their co-location** such that seemingly unrelated data can be combined based on the fact they are located in the same area.
- It helps users understand patterns, relationships, and geographic context. The benefits include improved communication, efficiency, management, and decision-making.



Source: GAO. | GAO-15-193

GIS for Economists

- GIS is a powerful tool for economists to visualize and analyze data in spatial dimension.
- The use of geographic information systems (GIS) has become increasingly popular.
- 1.Visualizing important variables in maps which is very intuitive for variation across space.
  - GDP per capita, Population Density, etc. across regions or countries.
  - Environmental variables, such as PM2.5, air quality, etc. across regions.
  - Trade flows, migration patterns, infrastructure development.
- 2.Measuring the previously unobserved and makes research feasible:
  - Satellite images provide comprehensive views of economic activities.
  - Historical maps for long-term economic development studies.
  - Real-time monitoring of economic shocks (e.g., COVID-19 lockdowns via mobility data).

GIS for Economists

- 3.Identifying causal relationships using spatial variation:
  - Distance to markets, ports, or borders as instrumental variables.
  - Elevation, slope, and geological features as exogenous variation.
  - Spatial Regression Discontinuity Design (RDD) at administrative boundaries.

Landmark Studies Using Spatial Data

- Henderson, Storeygard, and Weil (2012) - "Measuring Economic Growth from Outer Space"
  - Innovation: Use nighttime lights as proxy for economic activity.
  - Method: Statistical framework linking nighttime luminosity to GDP growth.
  - Impact: Particularly valuable for countries with poor statistical capacity.
- Notable Applications by Chinese Economists:
- Chen, Kung and Ma (2020): Use GIS to calculate the river distance to the nearest pine and bamboo location as IV of jinshi density in Ming and Qing Dynasties to estimate the long term effect of Keju on human capital accumulation today.
- Jia, Liang and Ma(2021): Use nightlights data and GIS to estimate the effect political hierarchy on economic development in China. More specifically, they use township level nightlights data and GIS info to compare the economic development along with the new border between Chongqing and Sichuan before and after Chongqing became a municipality.

GIS Softwares

- Traditional GIS softwares:
  - ArcGIS(commercial)
  - QGIS(open source)
- Modern GIS softwares:
  - R
  - Python
  - Stata
- R or Python are the most popular ones and can deal with 99% of GIS tasks.

GIS Data

- Spatial data usually focus on locations of objects.
- There are two types of GIS data:

- Vector Data（矢量数据）
  - Spatial objects: points, lines, polygons, etc.
  - Examples: rivers, roads, buildings, etc.
  - It has boundaries and shapes.

- Raster Data（栅格数据）
  - Gridded data
  - Examples: nighttime light, elevation, temperature, etc.
  - It has no natural boundaries and shapes.

Vector Data

- It usually includes two kind of information:

- Non-spatial Attributes:
  - Descriptive information about the spatial objects, like country name, population size, GDP per capita, etc.

- Spatial Attributes:
  - The coordinates that specify the spatial location of the objects.
  - The coordinates represent the shape of the objects.

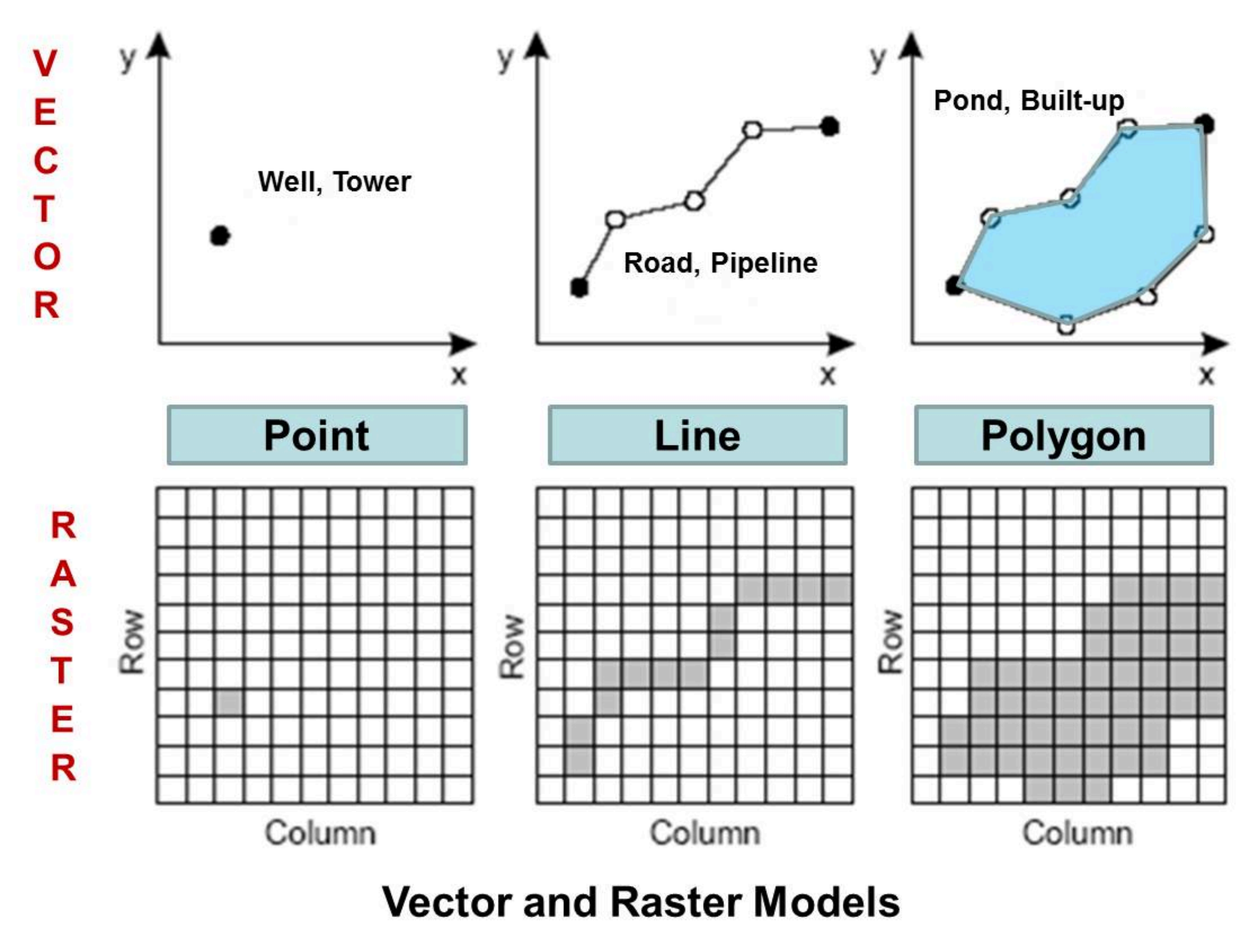
Vector Data

- According to their spatial attributes, spatial objects can be classified into several types:
  - **Point**: A single point in space.
  - **Line**: A line in space.
  - **Polygon**: A polygon in space.
- In all cases, the geometry of these data structures consists of sets of coordinate pairs `(x, y)`.
- Normally, it is stored in a folder containing several files with the same base name but different extensions: `.shp`, `.shx`, `.prj`, `.dbf`, etc.

Raster Data

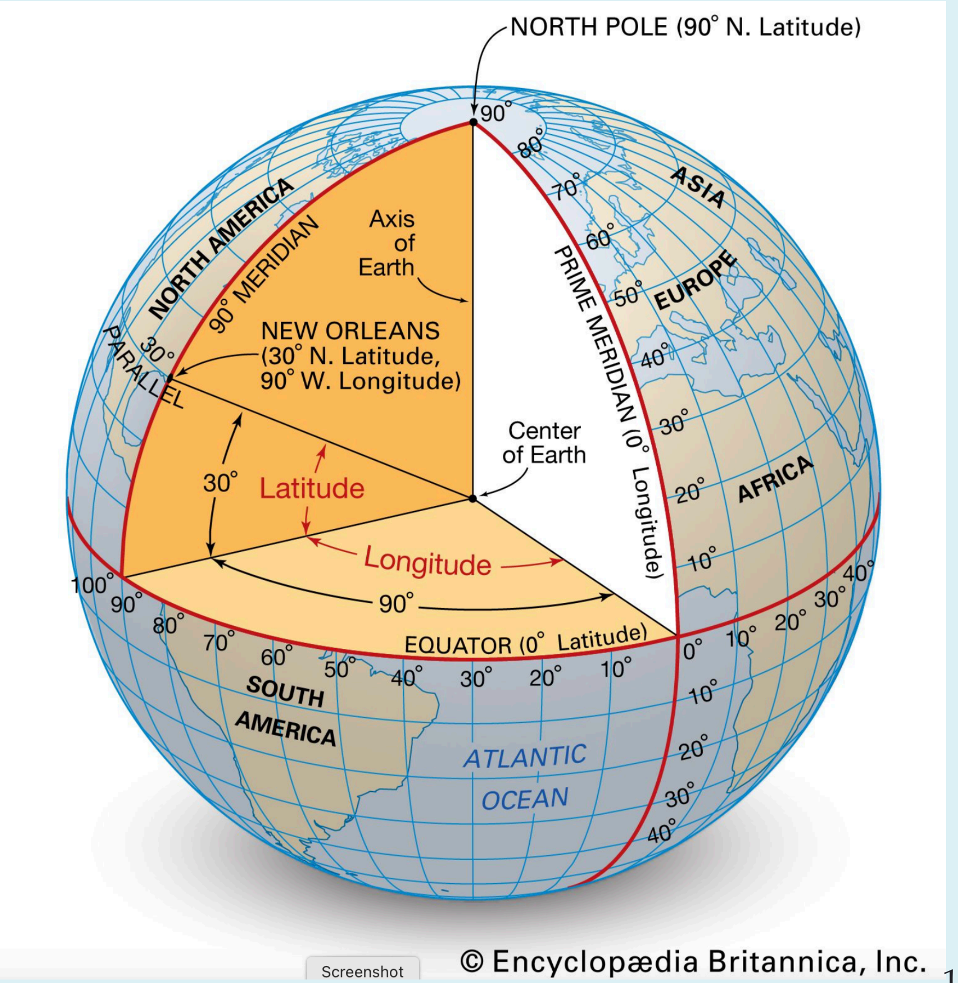
- Spatial fields are usually represented by raster data too.
    - It is commonly used to represent continuous variables which can naturally cross the boundaries of the spatial objects.
    - It is stored in a grid of cells, each of which has a value.
- The value of each cell is determined by the location of the cell in the grid and the value of the variable being represented.
    - A raster cell value should normally represent the average value of the variable over the area of the cell.
- Normally, it is stored in a file with the extension `.tif` or `.img`, which are apparently figure for raster data.

Vector and Raster Data



Coordinate Systems

- The distinguishing feature of spatial data is that each data point has a geographic identifier attached to it.
- Geographic Coordinate System(GCS) a system for locating the position of a given location on the earth surface using a 3D model of the globe (spheroid) and angular measurements called latitude and longitude.
  - latitude: north-south(90°N to 90°S) or (+90° to -90°)
  - longitude: east-west(180°E to 180°W) or (+180° to -180°)



Geographic Coordinate System(GCS)

- **Question**: Since we have the latitude and longitude, why don't we use them directly?
- **NO**. There are many reasons:

  1. The earth is not a perfect sphere, it is an irregular ellipsoid. There are many different ellipsoids.
  2. The latitude and longitude are good, but the origin of the coordinate system is not acceptable for different countries or regions.
  3. The angular unit(degree) is not the best unit for different countries or regions.

- Therefore, there are many different coordinate systems for different countries or regions. A full geographic coordinate system consists of:
  - **Datum**: The reference ellipsoid that is used to define the shape of the earth.
  - **Prime meridian**: The meridian of the origin of the coordinate system.
  - **Angular unit**: The unit of the angle.
  - Others: Like **WKID** which is the unique identifier of the coordinate system.

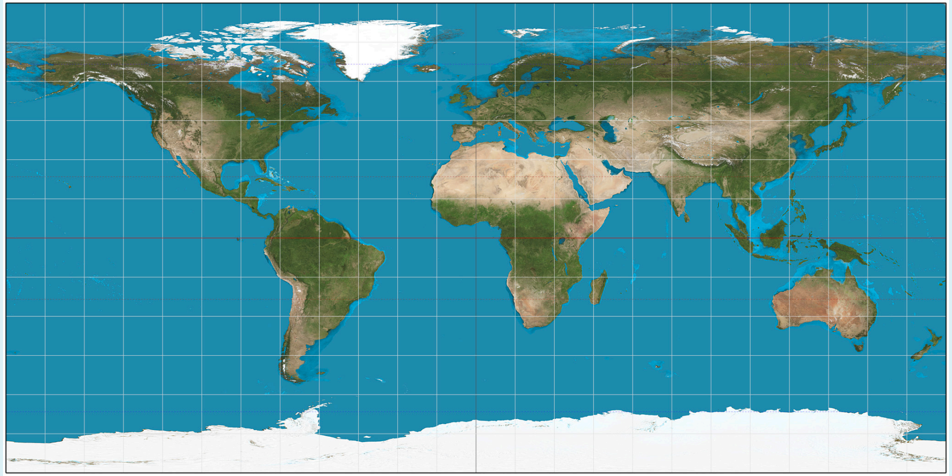
Geographic Coordinate System Components

- The most widely used geographic coordinate system globally is WGS84, which is the basis for the GPS system.
  - the datum is WGS84,
  - the prime meridian is Greenwich
  - the angular unit is degree.
  - the WKID is 4326.
- One of popular coordinate systems in China is *CGCS2000* coordinate system,which is based on WGS84. The differences are negligible.
  - the datum is CGCS2000,
  - the prime meridian is Greenwich
  - the angular unit is degree.
  - the WKID is 4490.
- Another popular coordinate system in China is *CGJ02* coordinate system,which is based on WGS84 with obfustication.("火星坐标系")

Map Projections(Projection Coordinate System)

- A key question: *How transform the 3D earth to 2D map?*
- Map Projection It is the mathematical algorithm that defines how to present the round earth(3D) on a flat map(2D).
- Projection Coordinate System
    - Geographic Coordinate System
    - Projection Method(Mercator, Equirectangular, etc.)
    - Unit(Meter, Kilometer, etc.)
    - Others
- There are many different map projections,each with its own strengths and weaknesses.
    - The most widely used map projection is Mercator Projection and Equirectangular Projection.

Map Projections



Equirectangular Projection

- 保持经纬度的简单线性关系
- 适用于全球尺度的初步分析
- 高纬度地区变形严重

Mercator Projection

- 保持角度和方向
- 适用于航海、导航分析
- 面积变形随纬度增加

Wrap Up

Key Points for GIS Data

1. Choose appropriate projection based on study area scale and location.
   - Global/Continental/National Scale:
   - Regional/Provincial Scale:
   - Local/City Scale:
2. Ensure consistency - all datasets must use **the exact same GCS and map projection**.
   - Otherwise, the results will be misleading.

GIS Analysis Toolkit

Understanding Spatial Metrics in GIS

1. Topological Metrics

- Adjacent(相邻)
- Overlaps(重叠)
- Contains(包含)
- Disjoint(分离)
- Intersects(相交)

2. Metric Relationships

- Distance or Length (Euclidean, Manhattan, Network, Terrain)
- Area (Road Area, River Area, etc.)
- Shape (Aspect, Slope, Curvature, etc.)
- Density (Population Density, Land Use Density, etc.)
- Centrality (Center of Gravity, Centroid, Median, Mean)

3. Directional Metrics

- Relative Direction (North, South, East, West)
- Absolute Direction (Based on Coordinate System)

Understanding Spatial Relationships in GIS

- Based on the types of spatial objects, we can classify the spatial relationships into the following types:

	点(Point)	**线(Line)**	**面(Polygon)**
点(Point)	point-point	point-line	point-polygon
线(Line)	point-line	line-line	line-polygon
面(Polygon)	point-polygon	line-polygon	polygon-polygon

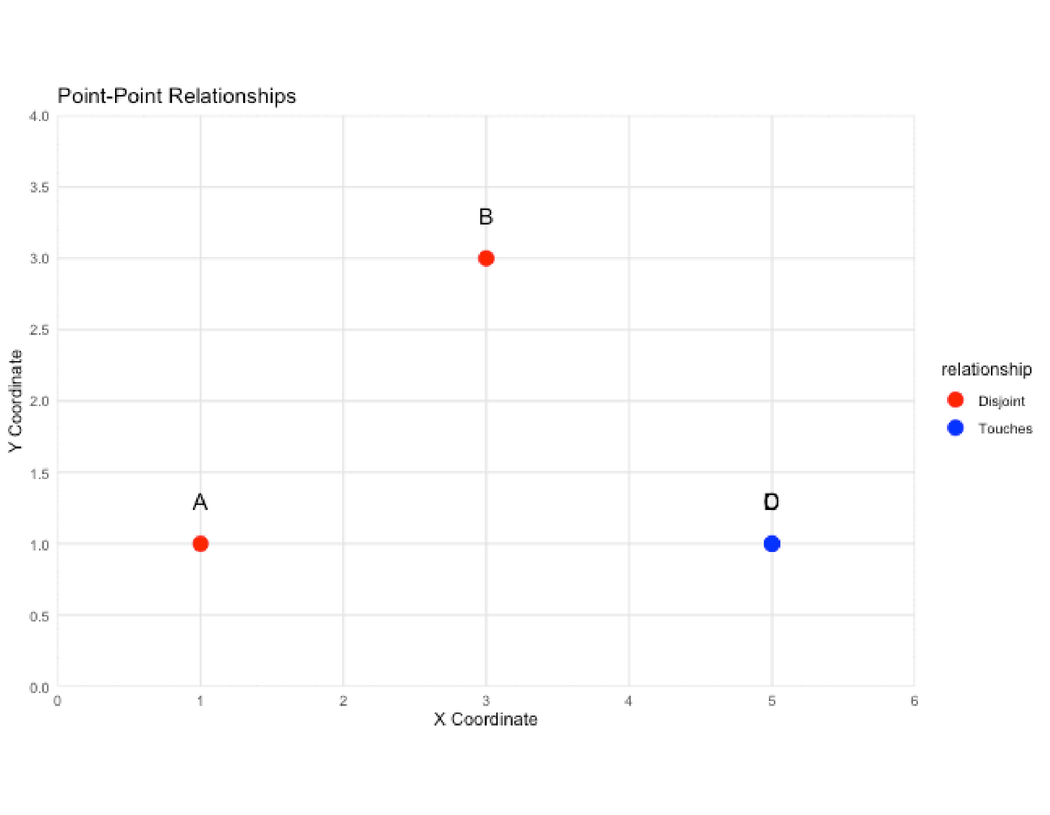
Actually, 1:1 relationships can be more simple, only 6 types:

- Point-Point(点对点)
- Point-Line(点对线)
- Point-Polygon(点对面)
- Line-Line(线对线)
- Line-Polygon(线对面)
- Polygon-Polygon(面对面)

Topological Relationships: Point-Point

1. Point-Point

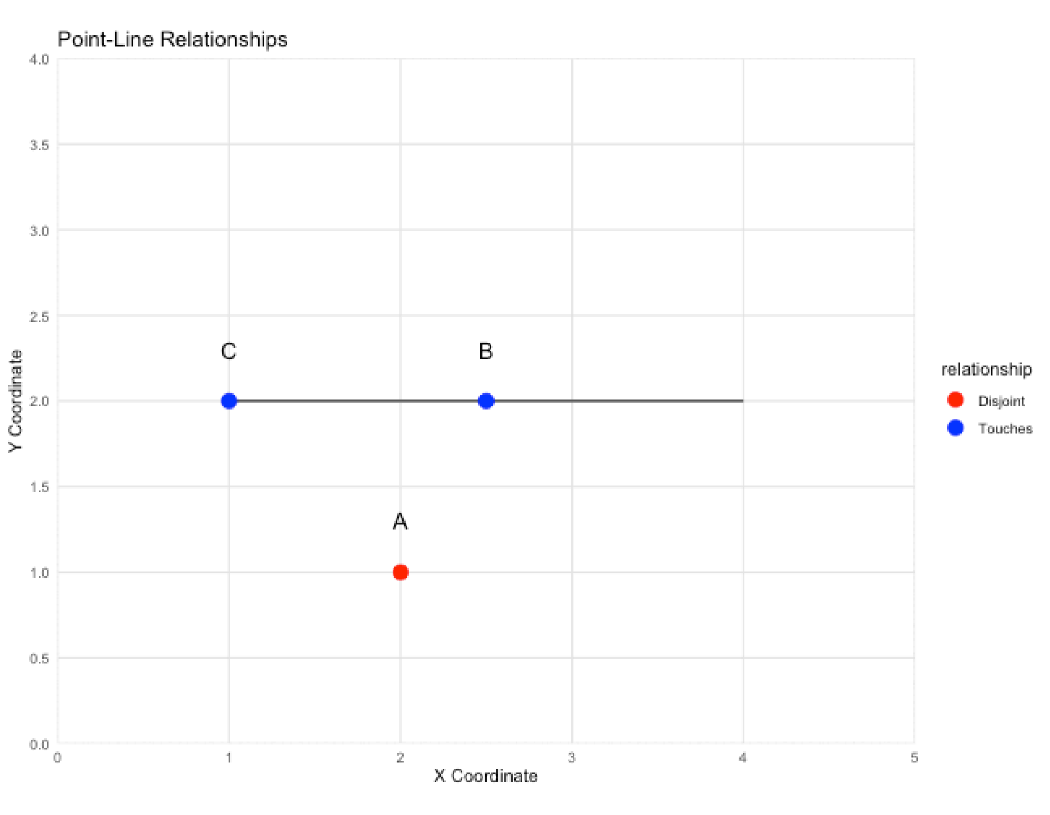
- Disjoint（分离）- 两个点不重合(A and B are disjoint)
- Touches or Adjacent (相邻) - 两个点重合（相同坐标）(C and D are touching or adjacent)



Topological Relationships: Point-Line

2. Point-Line

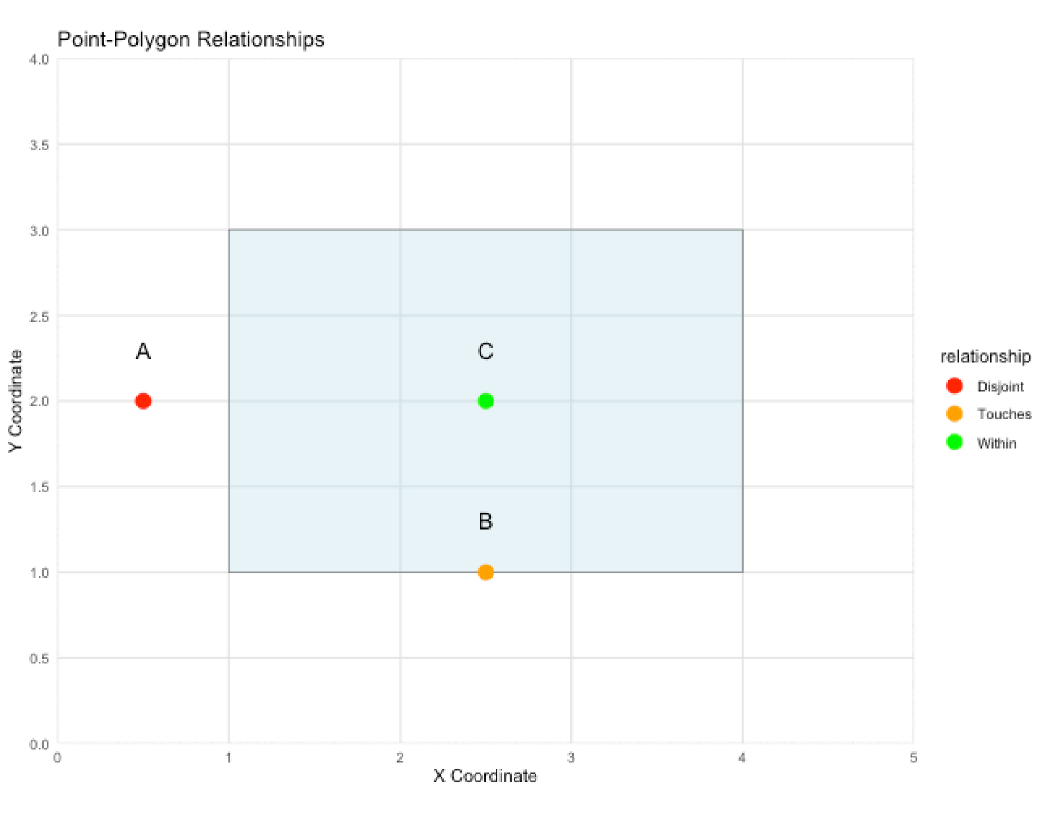
- Disjoint (分离) - 点不在线上(A and the line are disjoint)
- Touches (相邻) - 点在线的端点或线上(B and C and the line are touching)



Topological Relationships: Point-Polygon

3. Point-Polygon

- Disjoint (分离) - 点在面外(A and the polygon are disjoint)
- Touches (相邻) - 点在面的边界上(B and the polygon are touching)
- Contains (包含) - 点在面内（内部）(C and the polygon are containing)



Topological Relationships: Line-Line

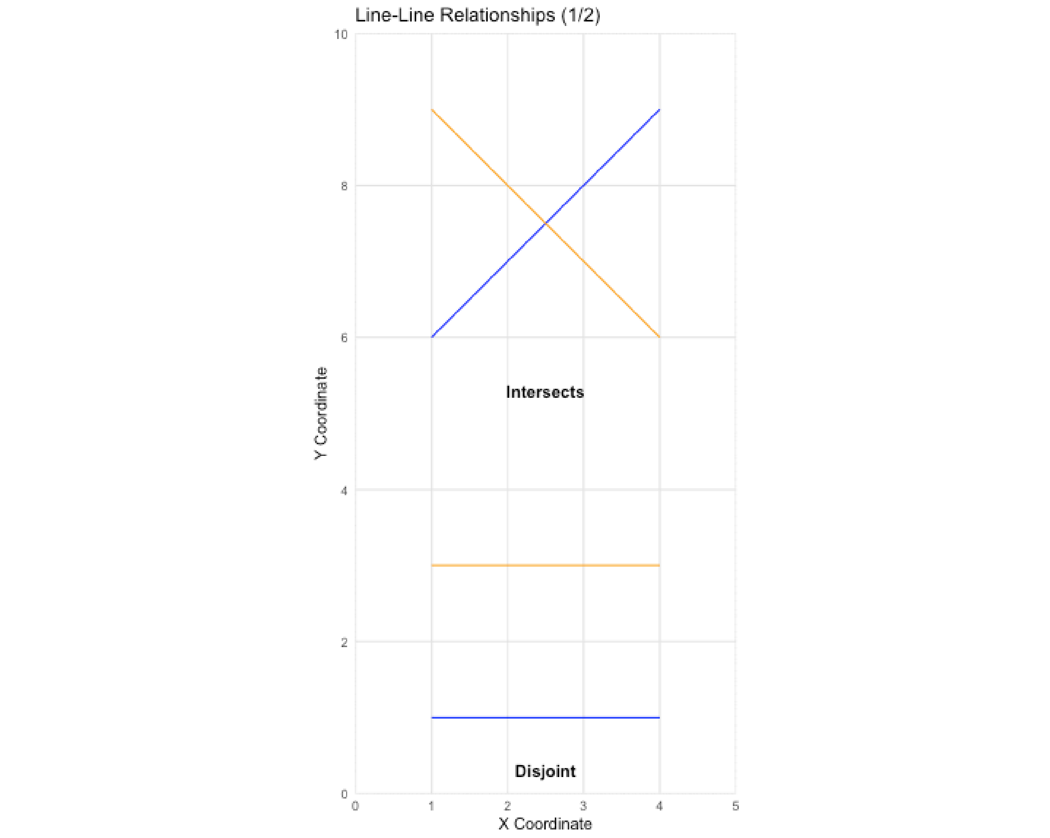
4. Line-Line

Disjoint (分离)

- Lines do not touch or intersect
- Example: Parallel roads

Cross (穿越)

- Lines cross at one or more points
- Example: Road intersections



Topological Relationships: Line-Line

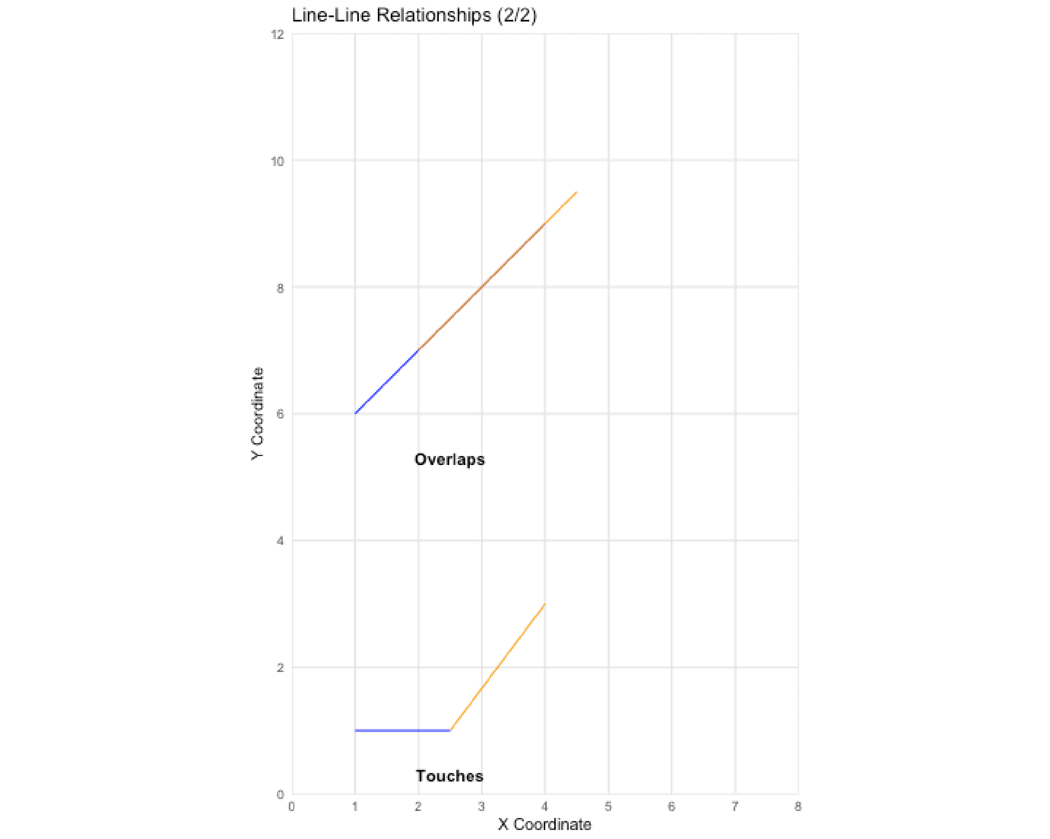
4. Line-Line

Overlaps (重叠)

- Lines share common segments
- Example: Overlapping transit routes

Touches (相接)

- Lines share endpoint(s) but don't cross
- Example: Road segments meeting



Topological Relationships: Line-Polygon

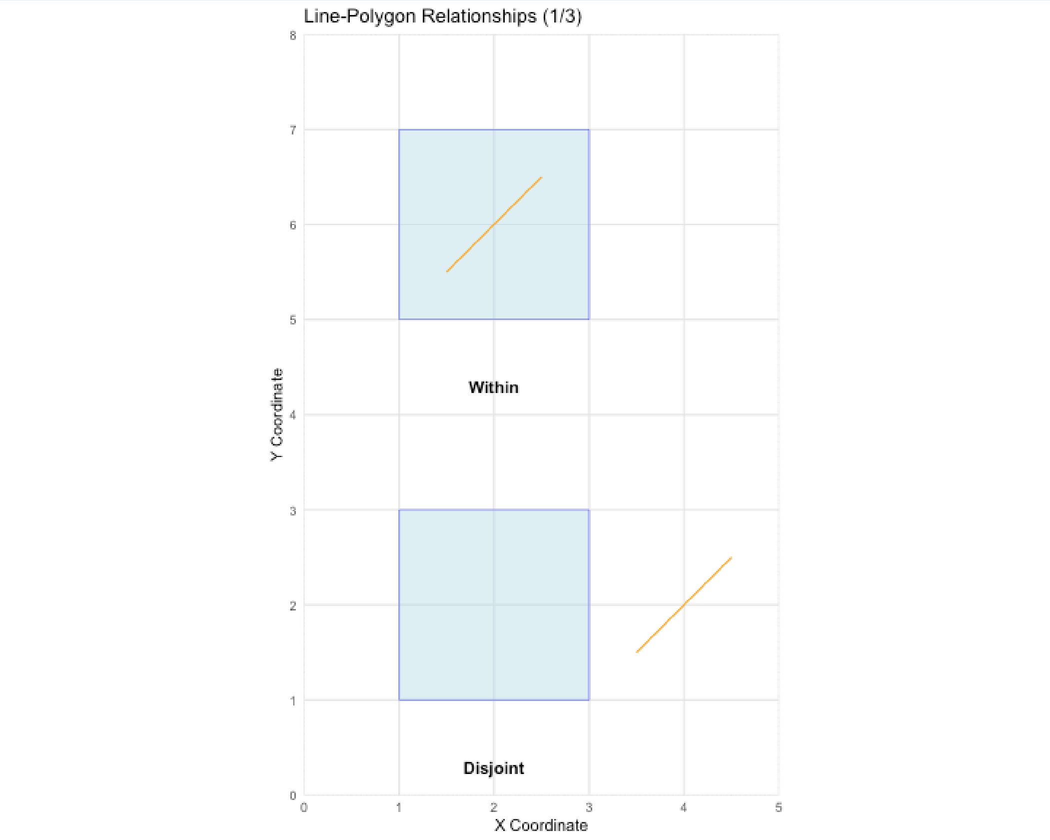
5. Line-Polygon

Within (包含于)

- Line completely inside polygon
- Example: Street within city block

Disjoint (分离)

- Line completely outside polygon
- Example: Highway outside city boundary



Topological Relationships: Line-Polygon

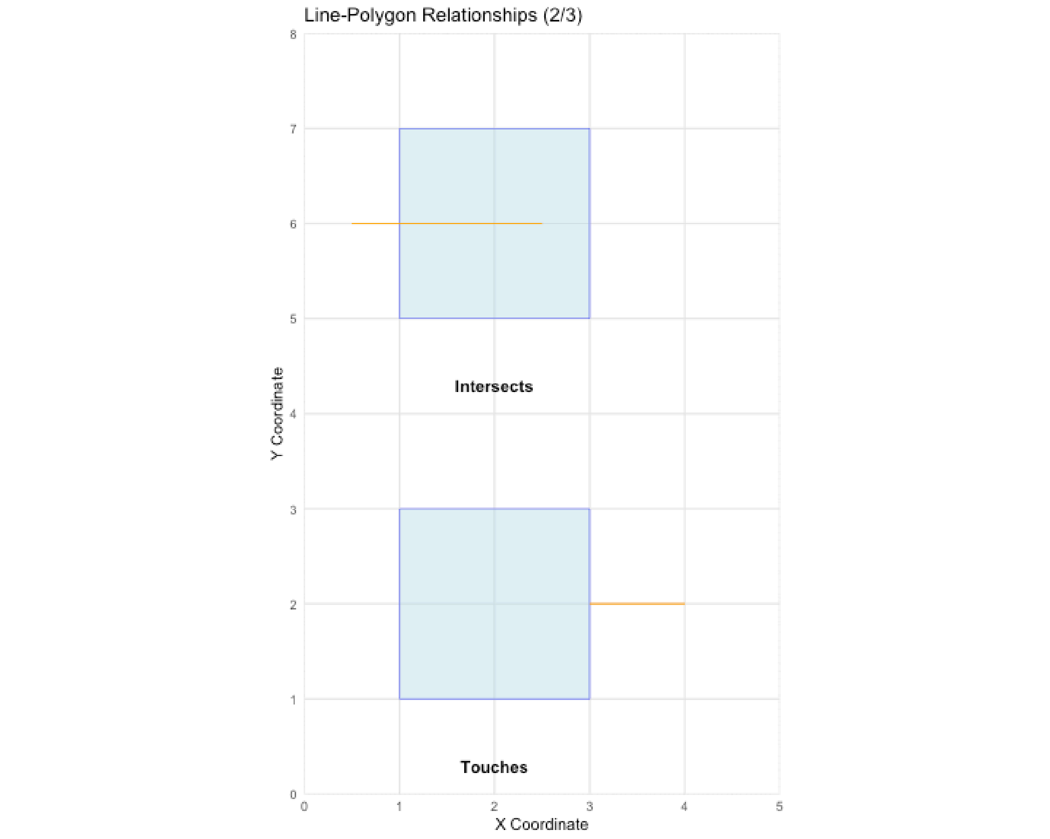
5. Line-Polygon

Intersects (相交)

- Line may only partially overlap with polygon
- Example: River intersects with administrative boundary

Touches (相接)

- Line touches polygon boundary at specific points
- Example: Road tangent to park boundary

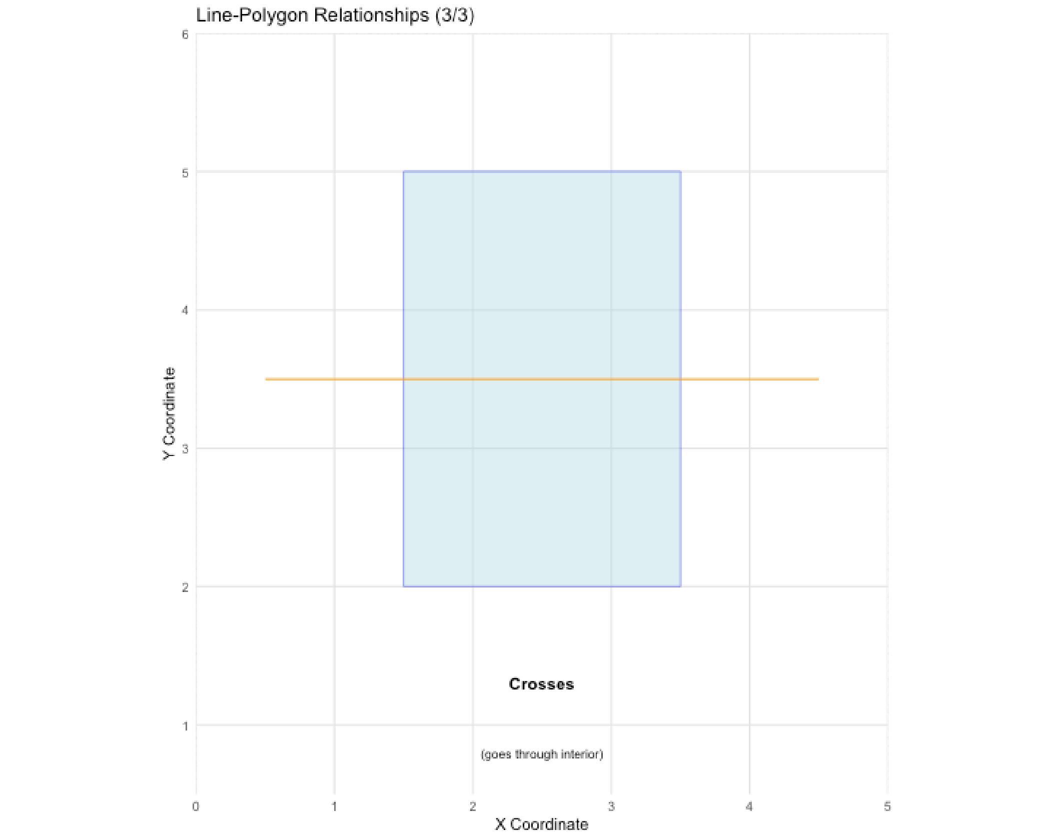


Topological Relationships: Line-Polygon

5. Line-Polygon

Crosses (穿越)

- Line completely passes through polygon interior
- Line enters polygon, goes through interior, exits on other side
- Example: Highway crossing county boundary
- Key difference from Intersects: Line must go through interior



Topological Relationships: Polygon-Polygon

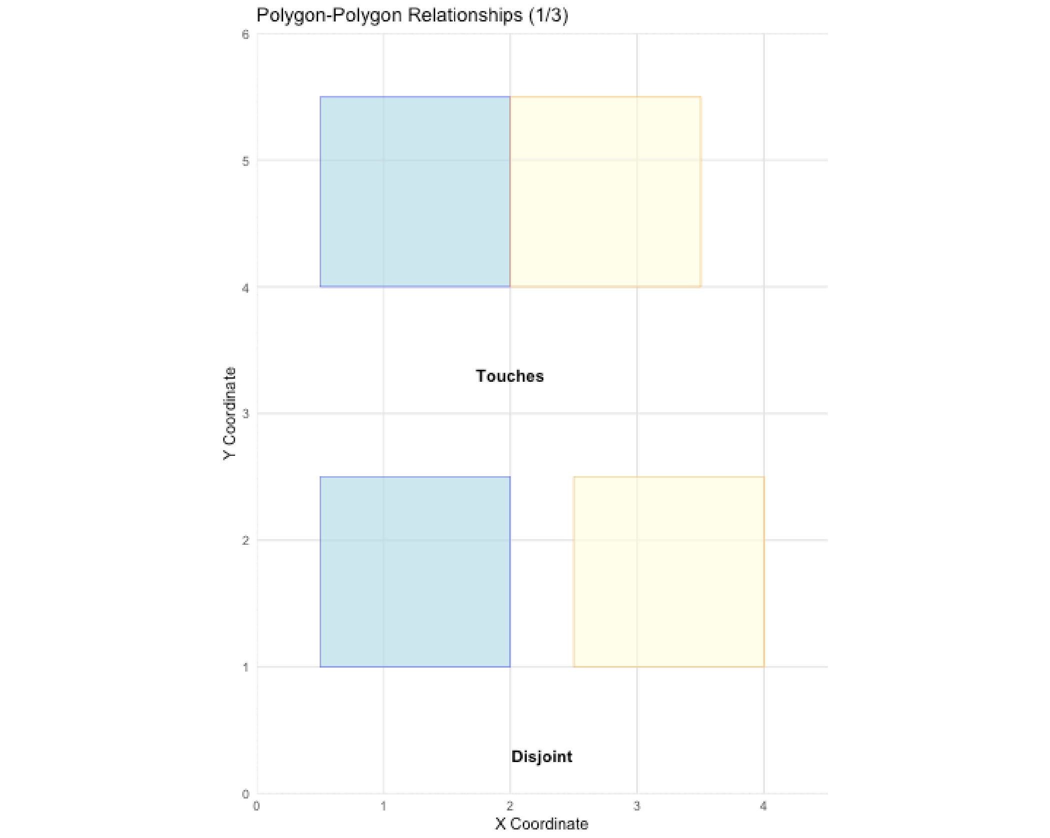
3. Polygon-Polygon

1)Disjoint (分离)

- Polygons don't touch
- Example: Separate administrative districts

2)Touches (相接)

- Polygons share boundary points
- Example: Adjacent counties



Topological Relationships: Polygon-Polygon

3. Polygon-Polygon

3)Overlaps (重叠)

- Polygons partially overlap
- Example: Overlapping economic zones

4)Within/Contains (包含)

- One polygon inside another
- Example: City within province

Topological Relationships: Polygon-Polygon

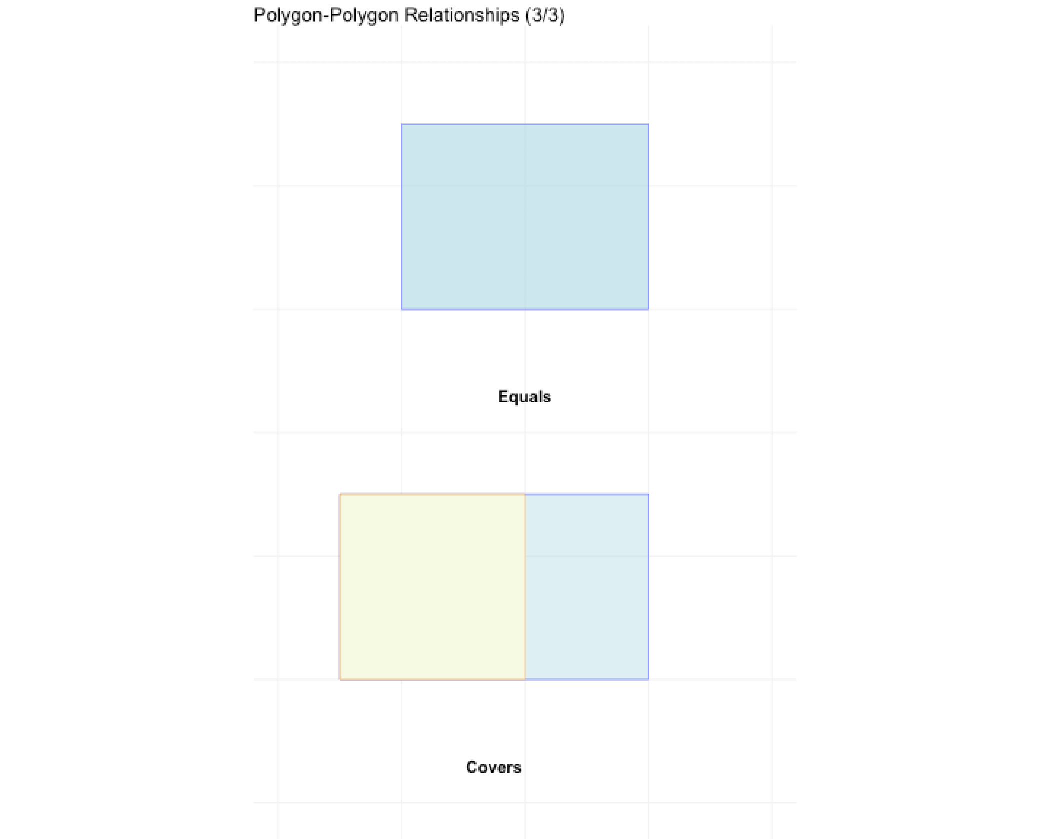
3. Polygon-Polygon

5)Covers/CoveredBy (覆盖)

- Complete coverage relationship
- Example: Electoral district coverage

6)Equals (相等)

- Identical polygons
- Example: Same boundary definitions



Distance Relationships

1. Euclidean Distance (欧几里得距离)

Definition: Straight-line distance between two points - "as the crow flies"

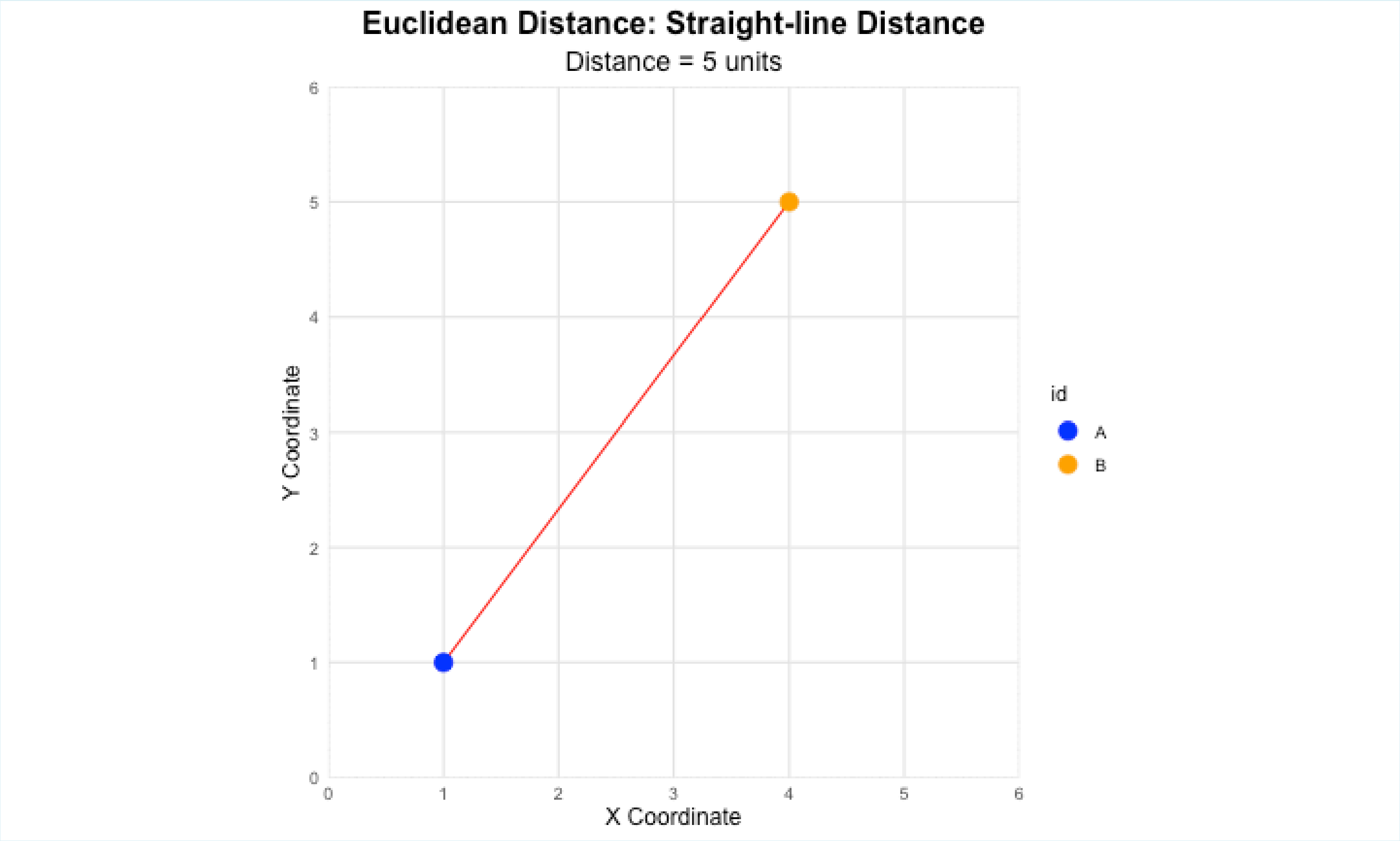
$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Use Cases:

- Air travel distances
- Radio signal propagation
- Emergency response planning

Distance Relationships

1. Euclidean Distance (欧几里得距离)



- $$distance = \sqrt{(5-1)^2 + (4-1)^2} = \sqrt{16+9} = \sqrt{25} = 5$$

Distance Relationships

2. Manhattan Distance (曼哈顿距离)

Definition: Distance measured along grid lines - "city block" distance

Formula:

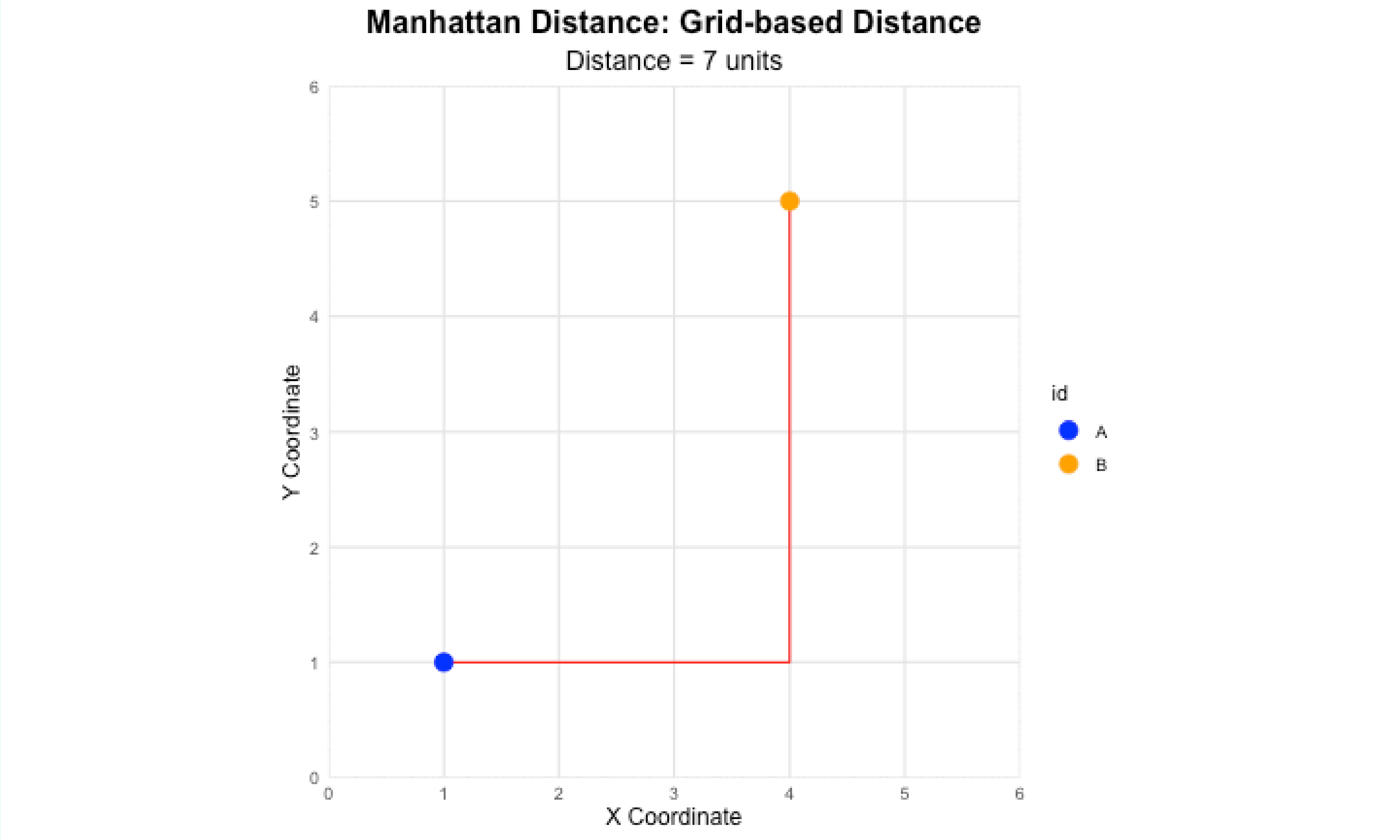
$$d = |x_2 - x_1| + |y_2 - y_1|$$

Use Cases:

- Urban planning in grid-based cities
- Taxi navigation
- Network routing

Distance Relationships

2. Manhattan Distance (曼哈顿距离)



- $$distance = |4 - 1| + |5 - 1| = 3 + 4 = 7$$

Distance Relationships

3. Network Distance (网络距离)

Definition: Distance along actual transportation networks (roads, rivers, etc.)

Characteristics:

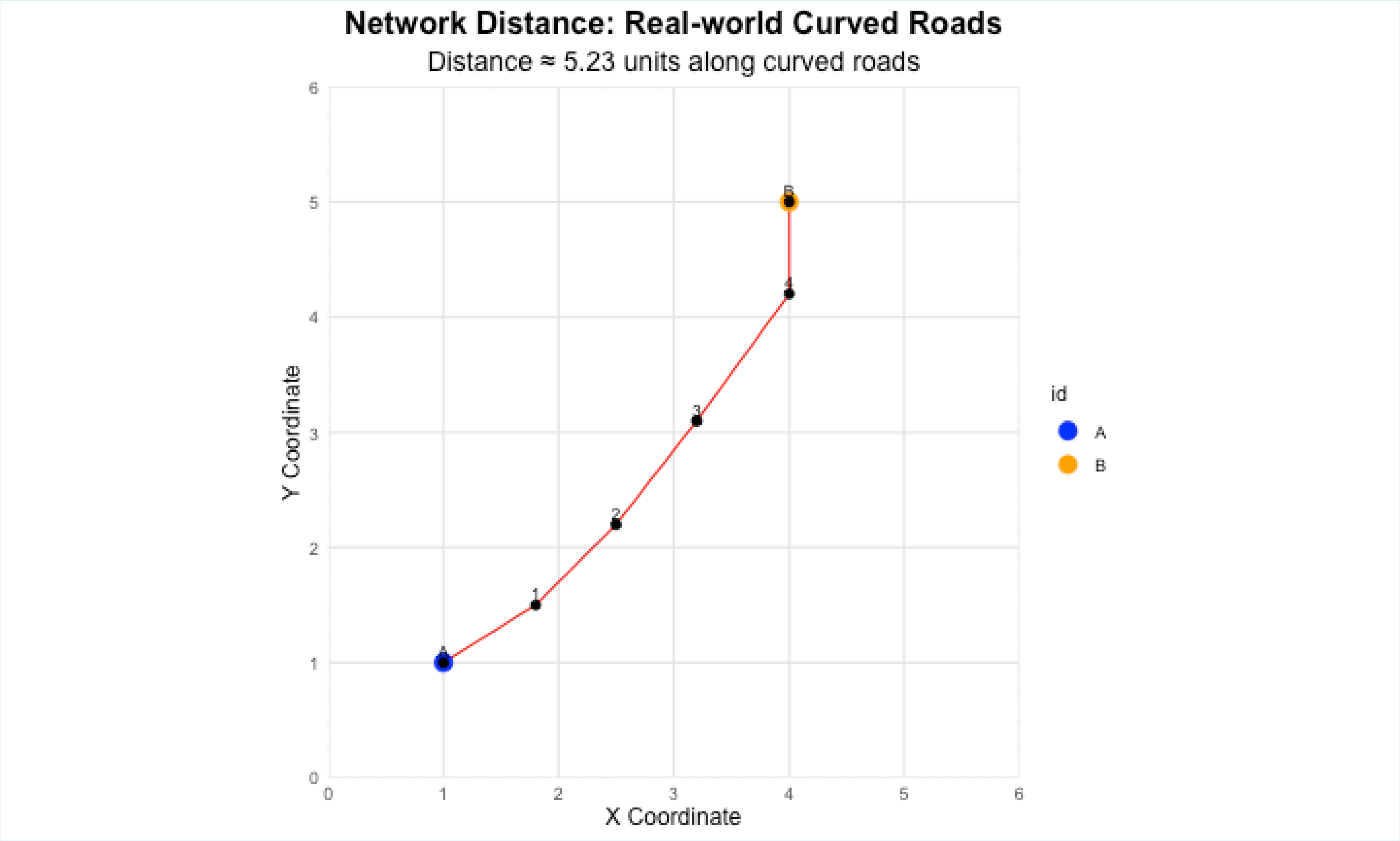
- Follows real-world infrastructure
- Accounts for barriers and constraints
- Most realistic for practical applications

Use Cases:

- Logistics and delivery planning
- Emergency services routing
- Accessibility analysis

Distance Relationships

3. Network Distance (网络距离)



- $$distance = \sum^{n} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$$

4. Terrain Distance (地形距离)

Definition: Distance accounting for terrain elevation and slope - "as the hiker walks"

Formula:

$$d = \sum_{i=1}^{n} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2}$$

Characteristics:

- Includes elevation changes (z-coordinate)
- Accounts for slope difficulty
- Most realistic for hiking, cycling, or vehicle travel

Use Cases:

- Hiking trail planning
- Cycling route optimization

4. Terrain Distance (地形距离)

Terrain Distance: 3D Visualization with Elevation

3D Distance ≈ 5.79 units

Distance Relationships

5. Great Circle Distance

Definition and Characteristics

Great Circle Distance is the shortest distance between two points on a sphere, measured along the great circle arc on the Earth's surface.More specifically,

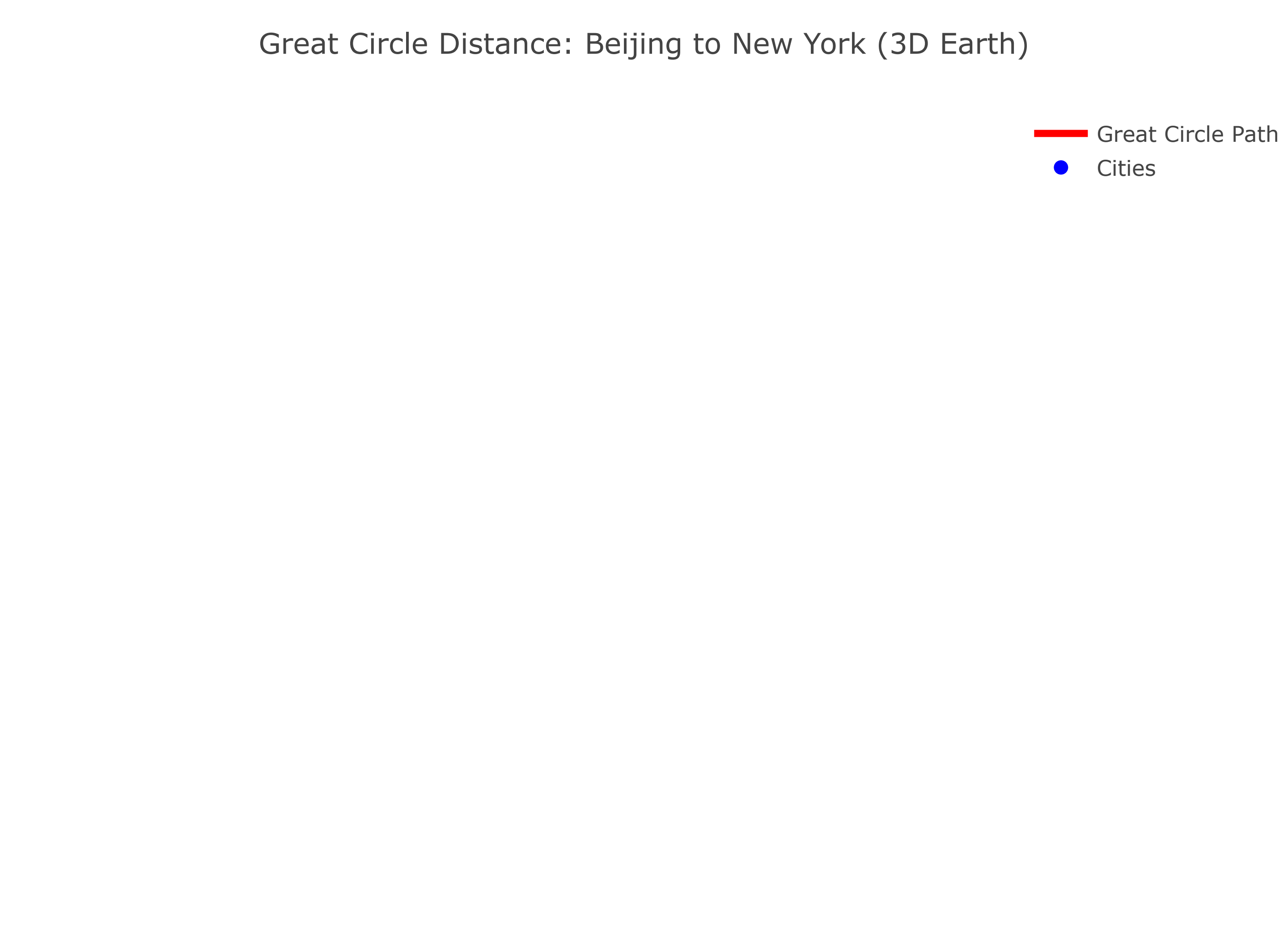
$$d = R \cdot \arccos(\sin(\phi_1)\sin(\phi_2) + \cos(\phi_1)\cos(\phi_2)\cos(\Delta\lambda))$$

- Where R = Earth radius (6,371 km),ϕ_1, ϕ_2 = Latitude of the two points,$\Delta\lambda$ = Longitude difference

Applications:

- Aviation route planning
- Global navigation systems
- International logistics

5. Great Circle Distance (大圆距离)



Distance Relationships

- As to the types of spatial objects, we can classify the distance relationships into the following types:
    - Point-Point(点对点)
    - Point-Line(点对线)
    - Point-Polygon(点对面)
    - Line-Line(线对线)
    - Line-Polygon(线对面)
    - Polygon-Polygon(面对面)

Other Metrics

- Other Metrics
    - Area(面积) only for Polygon
    - Shape(形状)
    - Density(密度)
    - Centrality(中心性)
- Other Analysis methods
    - Buffer Analysis
    - Overlay Analysis
    - Network Analysis
    - Spatial Statistical and Econometrical Analysis
- We have no enough time to cover these topics in this course, but you can learn more by yourself or in other courses if you are interested.

Next Lecture and Homework

Next Lecture

- Introduction to nighttime light data and its economic applications
- Introduction to NASA Black Marble data

Homework

- Register an account in NASA Earthdata here
- Read the guide of how to download NTL Data from NASA Earthdata
- Try to download NTL Data of China in 2020 from NASA Earthdata