Spatial Analysis and Modeling — Intro

Why spatial is important and different?

Corso di formazione su ML e DL
Fondazione LINKS
25/09/2025

ISI Foundation

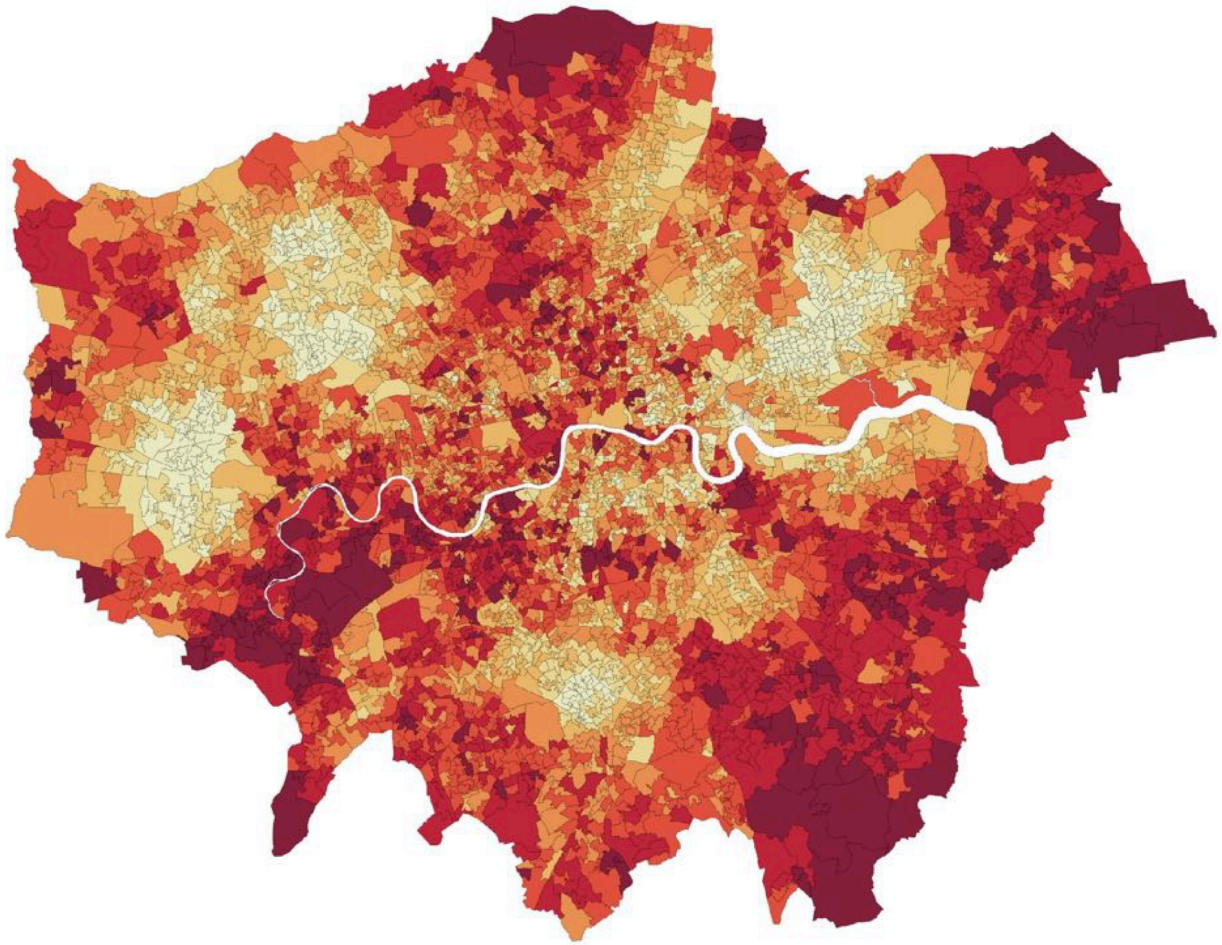

@Nokia Bell Labs
CAMBRIDGE
@IU
LONDON
@UNITO + @ISI Foundation
BLOOMINGTON
@Yahoo
TURIN
NEW YORK
BARCELONA
@UDD
SANTIAGO

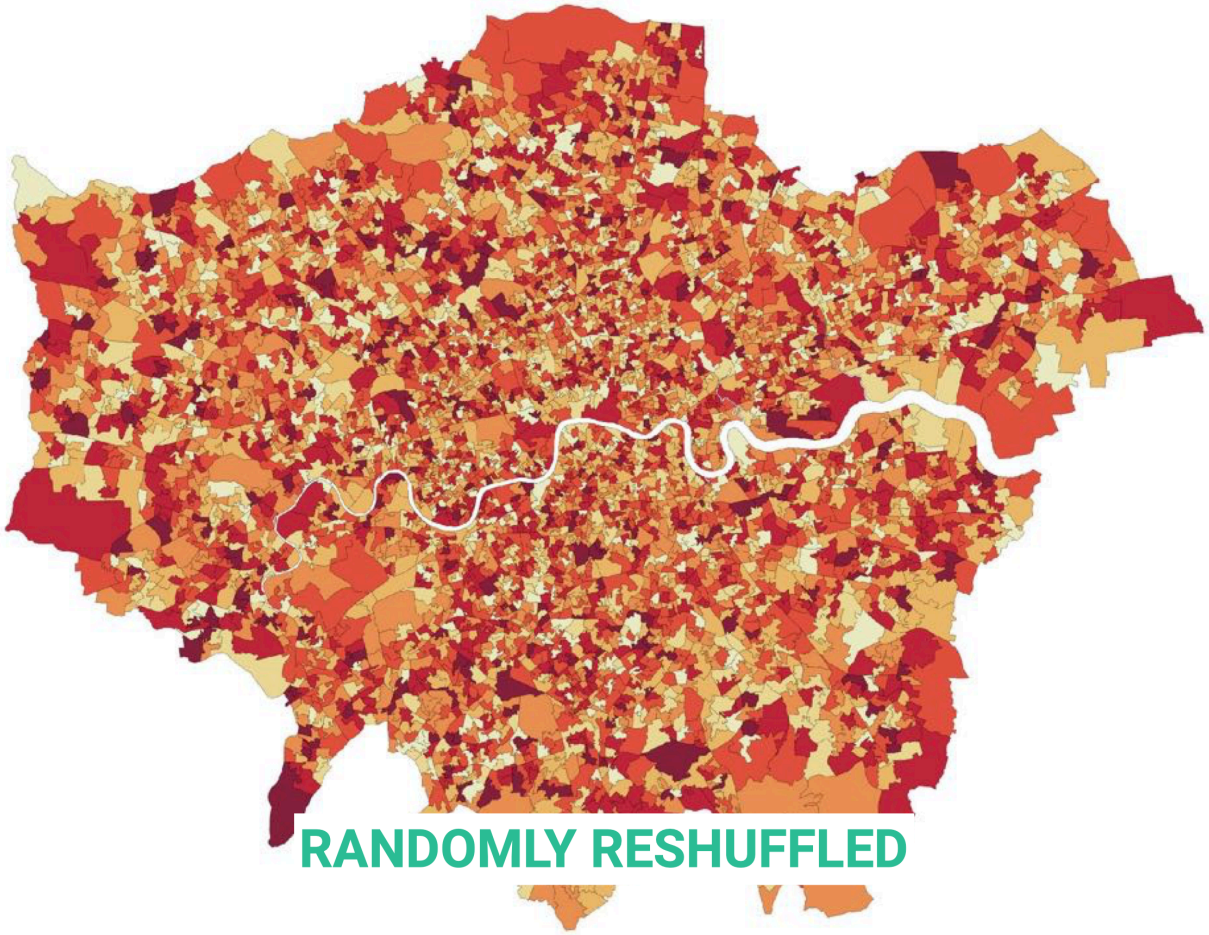
Why spatial thinking is important?


DUFOURS PLACE
MARSHALL STREET
SOUTH ROW
MARLBOROUGH ROW
BROAD STREET
PUMP
CAMBRIDGE ST
BREWERY
NEW STREET
HOPKINS STREET
COCK CT.
PULTENEY
EDWARD
DUCK LANE
HUSBAND ST
LITTLE

How does summary statistics appear?

Percentage of white people in London (LSOA, census 2011)

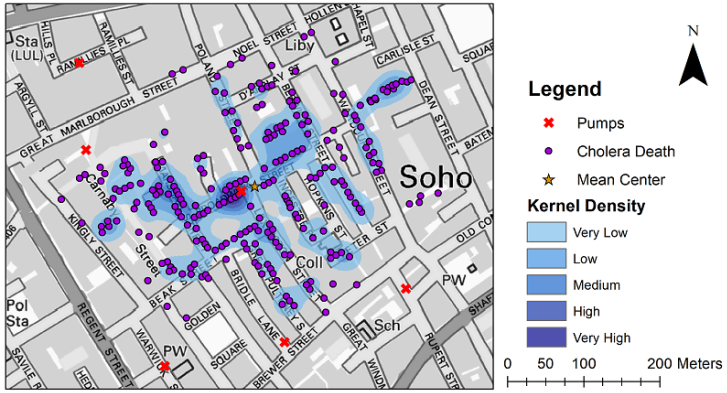




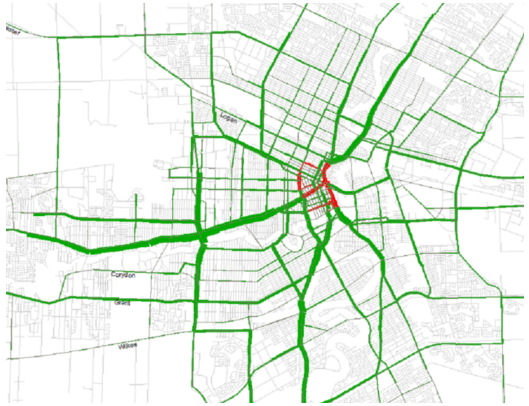
Spatial Data is Everywhere

Object Data (discrete features)

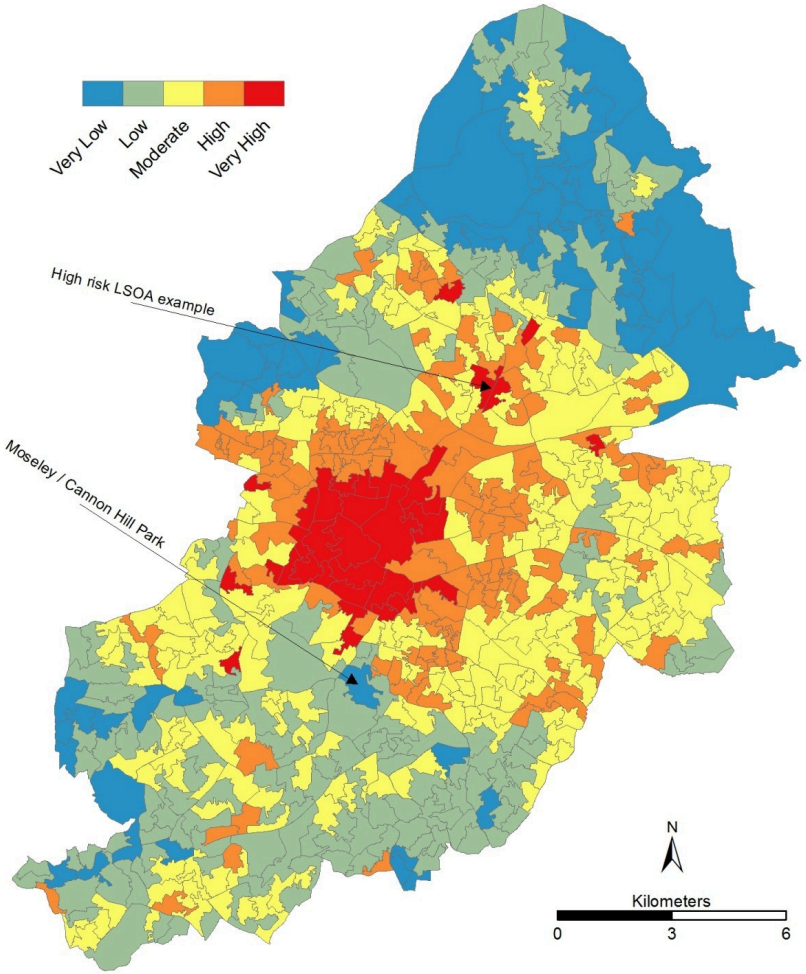
- **Points**: Locations (e.g., trees, accidents)



- **Lines**: Paths (e.g., roads, rivers)



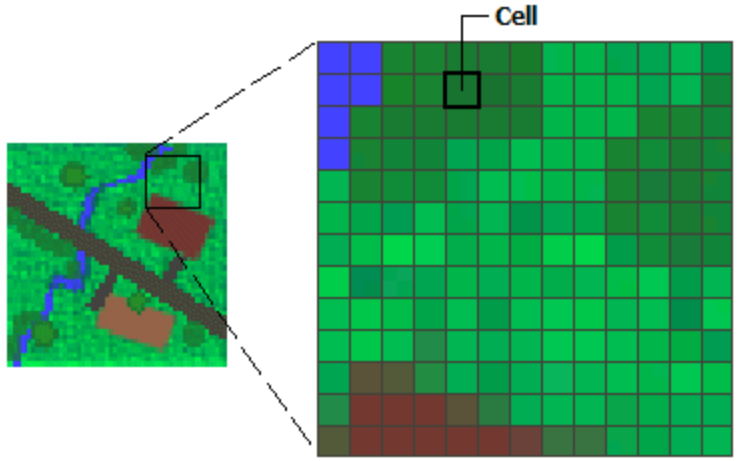
- **Polygons**: Areas (e.g., city boundaries, lakes)



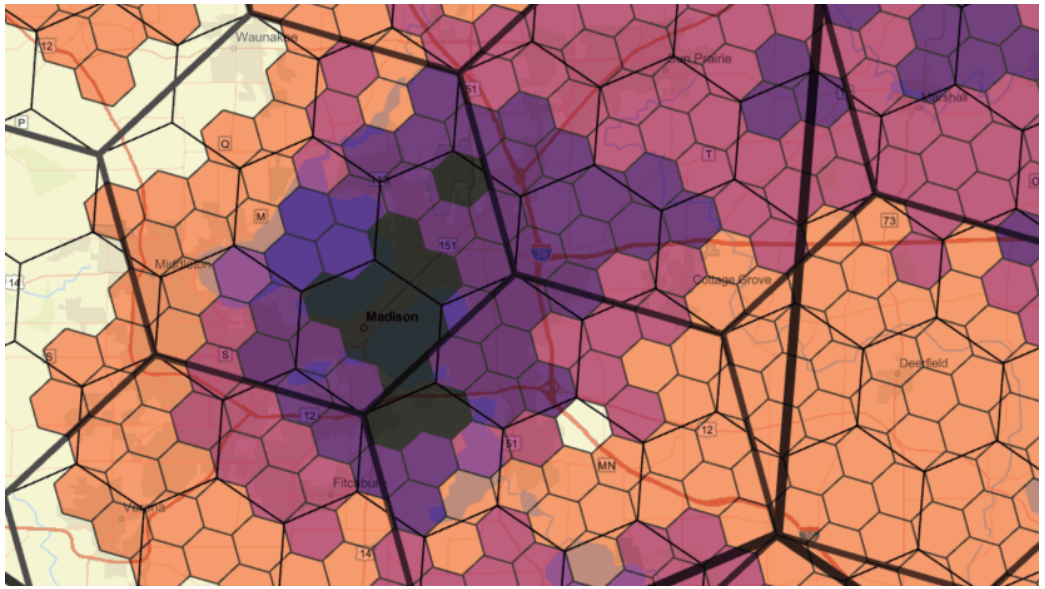
Spatial Data

Field Data (continuous surfaces)

- **Raster/Grid**: Continuous surfaces (e.g., elevation, temperature)



- **Regular/Irregular Cells**: Grids or tessellations



Wide Range of Applications

Earth Science

- Land use/cover mapping using satellite imagery
- Deforestation monitoring
- Disaster mapping (floods, earthquakes, hurricanes)
- Species distribution modeling
- Soil property mapping

Urban Informatics

- Traffic prediction and routing
- Crime hotspot prediction
- Air quality mapping
- Ride-sharing demand forecasting

Public Health

- Disease risk mapping
- Outbreak detection and monitoring
- Environmental health factor analysis
- Drug epidemic understanding

Geosocial Media Analytics

- Real-time event detection from tweets
- Political unrest forecasting
- Travel recommendation systems

Problem formulation

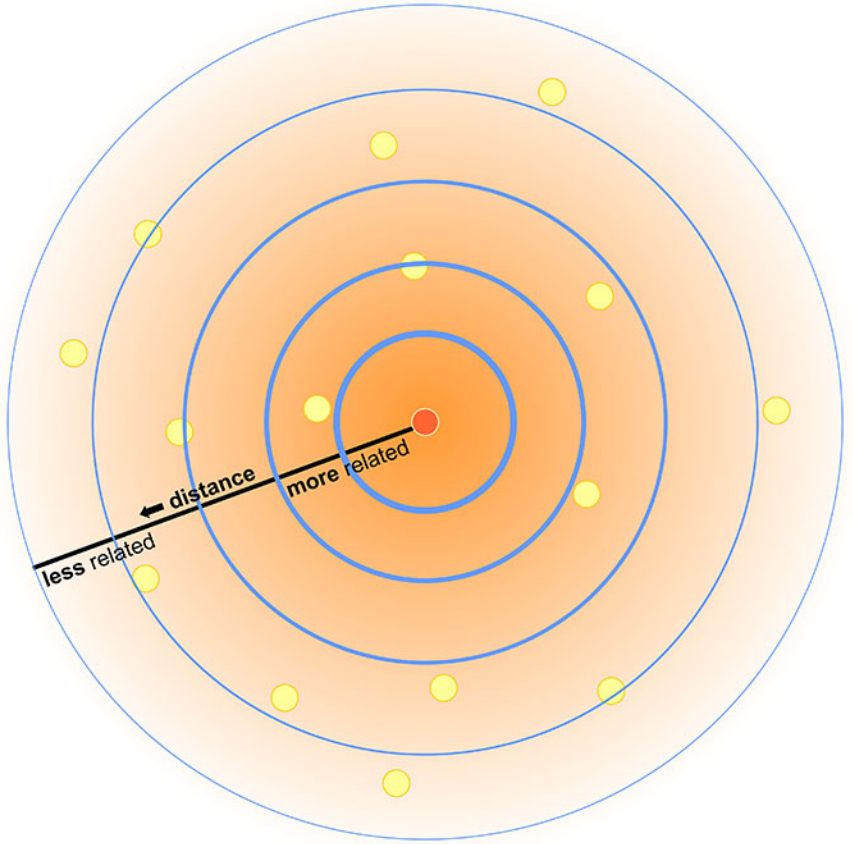
Given samples $\{(x(s_i), y(s_i))\}_{i=1}^{n}$, learn f such that

- Regression: $y(s) \approx f(x(s))$
- Classification: $y(s) \approx f(x(s)) \in \mathcal{C}$
    - where $\mathcal{C}$ is the set of possible classes (e.g. land-cover types)

Spatial twist: **observations are not i.i.d. (independent & identically distributed)**; dependence & heterogeneity matter

What is special about spatial data

- **Spatial dependence**
  - "Near things are more related than distant things" - Tobler's First Law
- **Spatial heterogeneity**
  - Sample distributions vary across regions (**non-stationarity**, **anisotropy**)
- **Scale and aggregation**
  - MAUP and resolution/extent effects
- **Limited/biased ground truth**
  - Expensive field collection, travel costs between locations
- **Constraints** and **networks**: roads, rivers, admin borders guide interactions
- **Spatio-temporal dynamics**



Challenge 1: Spatial Autocorrelation

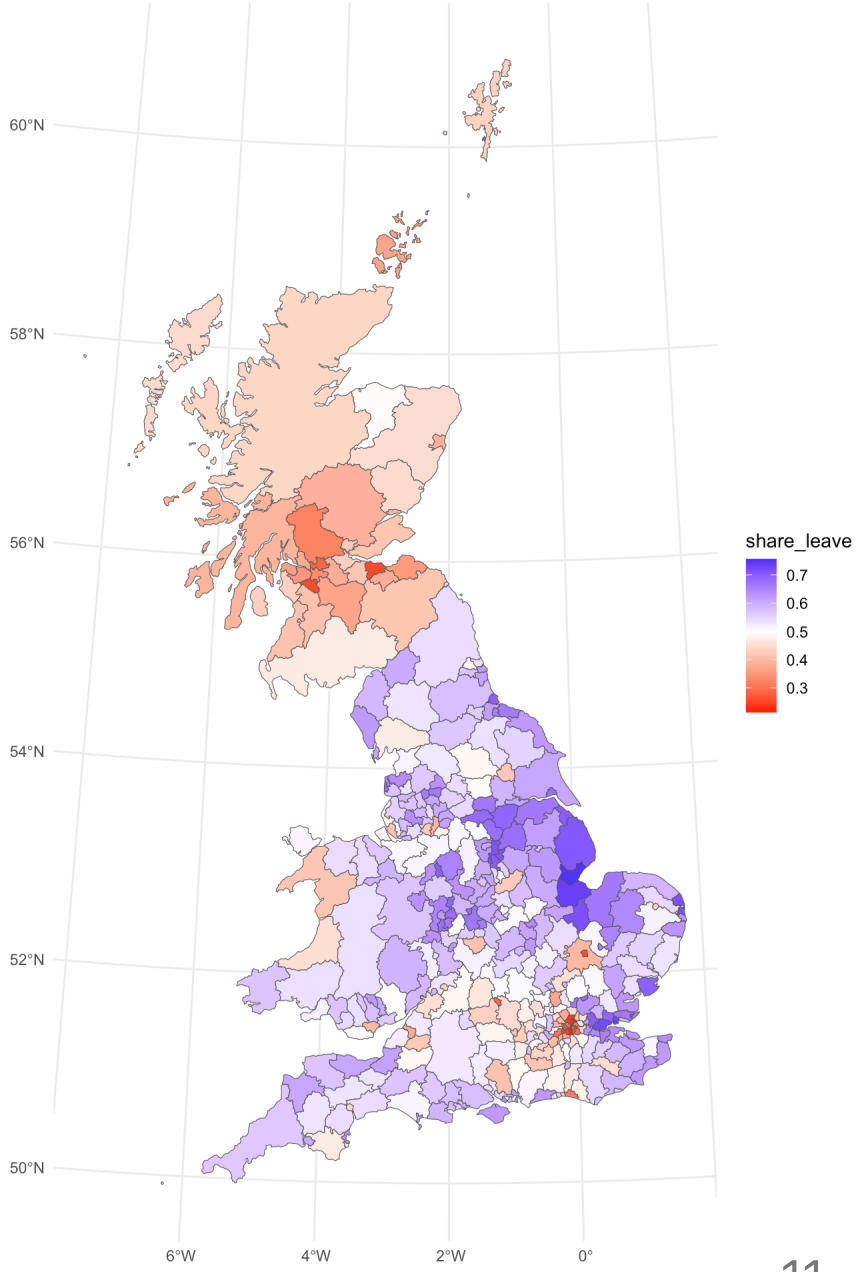
What It Means

- Nearby locations tend to have similar values
- Violates the **i.i.d. assumption** of traditional ML
- Temperature example: two nearby cities have similar temperatures

Problems It Causes

- **Linear regression:** Residual errors are correlated, not i.i.d.
- **Classification:** Salt-and-pepper noise artifacts
- **Model assumptions:** Statistical independence violated

Example: Decision trees on satellite imagery produce noisy, unrealistic boundaries



Challenge 2: Spatial Heterogeneity

Non-stationarity

- **Sample distribution varies across sub-regions**
  - Same spectral signature $\rightarrow$ different land cover (tropical vs. temperate)
- **Ecological fallacy:** Global models fail locally

Anisotropy

- Spatial dependency varies by **direction**
  - Climate influenced by mountain ranges
  - Water flow follows terrain and gravity

Implication: One-size-fits-all models often fail!

Challenge 3: Multiple Scales & Resolutions

The Problem

- **Earth imagery:** Sub-meter (aerial) to 100m+ (MODIS satellite)
- **Agriculture:** Point sensors + plot-level yields + aerial imagery
- **MAUP:** Modifiable Areal Unit Problem - relationships change with scale

Implications

- Models assume same scale/resolution (violated)
- Simple aggregation loses critical information
- Need multi-scale fusion methods

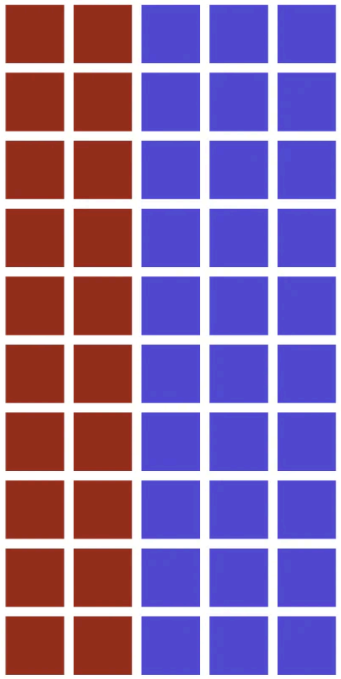
MAUP

- **The same basic data yield different results when aggregated in different ways**
  - Refer to: **A million or so correlation coefficient: three experiments on the modifiable area unit problem** (Openshaw and Taylor, 1979)
- **Zonal effect**
  - Similar size and number of units, but different boundaries
    - Zip codes versus census tracts, postal zones versus city neighborhoods
- **Scale effect**
  - Increases size and decreases number of units
    - US counties versus states
  - Global model might be inconsistent with local models
- The take home message is that **how we aggregate the input units will impact the values of the output units**

Gerrymandering, explained

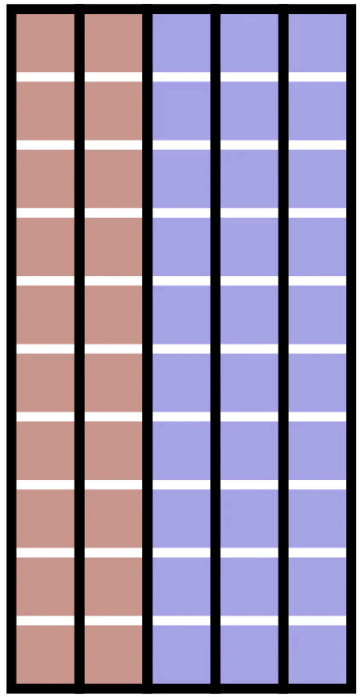
Three different ways to divide 50 people into five districts

50 people



60% blue, 40% red

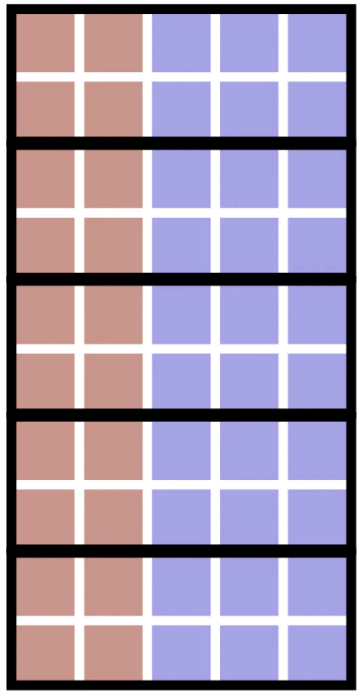
1. Perfect representation



3 blue districts, 2 red districts

BLUE WINS

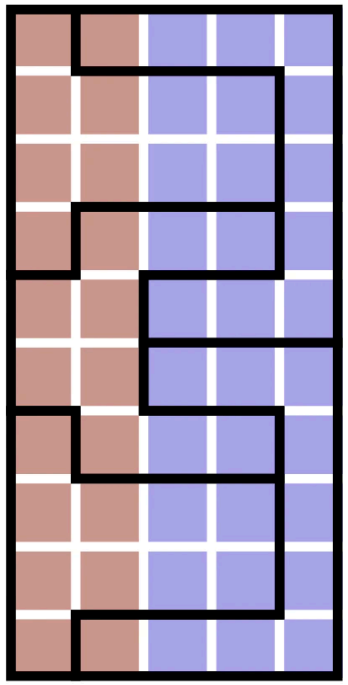
2. Compact, but unfair



5 blue districts, 0 red districts

BLUE WINS

3. Neither compact nor fair



2 blue districts, 3 red districts

RED WINS

Ecological Fallacy

- Occurs when **relationships measured between variables at an aggregate level** (e.g., areas, counties) are assumed to hold for individuals.
  - The group-level association can differ in sign or magnitude from the individual-level association.

- Why :
  - **Aggregation hides within-group variability** and **individual heterogeneity**.
  - Confounding and composition effects within units can produce spurious group-level correlations.
  - Simpson's paradox — **aggregated and disaggregated trends can point in opposite directions**.

- **Do not make individual-level causal claims from area-level coefficients**.

- Can lead to **misleading policy decisions**.

- **Over-** or **under-estimation of risk factors** when within-area heterogeneity is large.

- **Good practice:**
  - Prefer individual-level data for individual-level inferences. If unavailable, restrict claims to the aggregate level.
  - Use **multilevel** / **hierarchical models** (partial pooling) to model both individual and area effects when mixed data exist.
  - Apply **small-area estimation** or **downscaling** methods with explicit uncertainty quantification for finer inference.
  - Include **contextual covariates** and measures of **within-area variance** to reduce ecological bias.
  - Perform **sensitivity analyses** across scales and aggregation schemes; be explicit about the inference level in reporting.

- Rule of thumb: **match the inference target to the data level**
  - aggregate data $\longrightarrow$ aggregate conclusions;
  - individual decisions $\longrightarrow$ individual data or appropriate hierarchical modeling.

Challenge 4: Limited Ground Truth

Why Unique

- **High collection costs:** Field crews, travel time between locations
- **Labeling costs:** Visual interpretation by experts
- **Geographical representativeness:** Sample selection must be spatially balanced

Impact on Modeling

- Traditional active learning doesn't account for travel costs
- Semi-supervised methods become critical
- Need spatial considerations in sample selection

Addressing the challenges

- Methods grouped by major challenge addressed
  - **Spatial autocorrelation**: contextual features, dependency in model, regularization
  - **Spatial heterogeneity**: location-dependent models, ensembles, multi-task learning
  - **Limited ground truth**: semi-supervised, active learning
  - **Multiple scales**: spatial hierarchical models

(Today we focus on the first two challenges)

Roadmap

- **Why spatial is special** → Special nature of spatial data
- **Key challenges** → Dependence, heterogeneity, scale, limited labels, anisotropy
- **Spatial statistics** → Indices, SAR/SEM/SDM, CAR/MRF, Kriging/GP
- **Machine learning approaches** → Contextual features, Regularization, Ensembles, (M)GWR, Clustering, Regionalization
- **Deep learning approaches** → CNN/UNet, GNNs (GCN/GAT), spatio-temporal
- **Evaluation and practice** → Spatial CV, diagnostics, uncertainty
- **Bibliography** → References and resources