Lecture 1. Overview of Spatial Data Problems

Spatial Big Data Analysis with GIS

Korean Statistical Society, Winter School, February 24, 2023

Goals of spatial statistics (or spatio-temporal statistics)

- In the physical world, phenomena evolve in space and time following deterministic, perhaps "chaotic," physical rules, so we need to consider **randomness** and **uncertainty**.
- Statistical models give us the ability to model components in a physical system that appear to be random.
- Main goals with a spatial statistical model
  1. prediction in space and time (filtering and smoothing)
  2. inference on parameters
  3. forecasting in time

Introduction to spatial data and models

Researchers in diverse areas such as ecology, epidemiology, climatology, hydrology, and real estate marketing are faced with the task of analyzing data that are[1]

1. highly multivariate, with many important explanatory and response variables,
2. geographically referenced, and often presented as maps, and
3. temporally correlated, as in longitudinal or other time series structures.

[1] Banergee, Carlin, and Gelfand (2015), Hierarchical Modeling and Analysis for Spatial Data.

Types of spatial data

1. **Geostatistical data (Point referenced data)**
   - Regularly spaced data vs irregularly spaced data
2. **Lattice data (Areal data)**
   - Point measurement vs block averages
3. **Point pattern data**

Other types: directional data, data from moving stations, etc.

General description

- Temporal: $\{Z(t), t \geq 0\}$
- Spatial: $\{Z(\mathbf{s}), \mathbf{s} \in D\}$
- Spatio-temporal: $\{Z(\mathbf{s}, t), \mathbf{s} \in D, t \geq 0\}$
- Multivariate: $\{\mathbf{Z}(\mathbf{s}), \mathbf{s} \in D\}$, $\mathbf{Z} \in \mathbb{R}^p$
- Use latitude/longitude for $\mathbf{s}$ on (the surface of) the sphere

Law of Geography: Nearby things tend to be more alike than those far apart.

Geostatistical data

- When a spatial process that varies continuously is observed only at points.
- $Z(\mathbf{s})$ is a random vector at a location $\mathbf{s} \in D$
- $\mathbf{s}$ varies continuously over $D \in \mathbb{R}^d, d = 1, 2, 3$, D is a continuous, fixed set.
- Examples: Mining (coal ash), Pollution (soil, PM2.5), Rainfall, Temperature, Pressure, Wind speed and direction
- Remote sensing (satellite), climate model output

Lattice data

- When a spatial process is observed at countably many (often finitely many) locations. Usually this arises due to aggregation of some sort, e.g., total over counties, average over a pixel etc.
- D is a fixed collection of data (of regular or irregular shape), "discrete" spatial index.
- Partitioned into a finite number of areal units with well-defined boundaries
- Examples: Crime rates, Census data (the poverty level in some counties, the number of children in the area's zip codes), Agriculture

Point pattern data

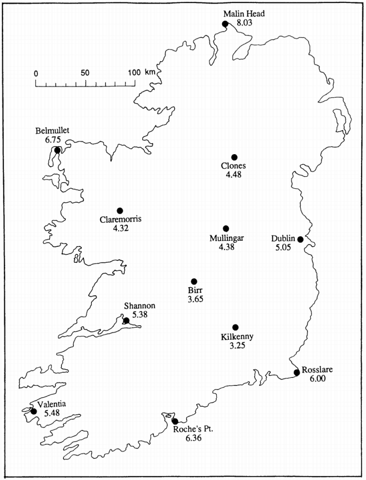
- When a spatial process is observed at points and the locations themselves are of interest. Example research questions are: **Is the pattern random or does it exhibit clustering?**

- D is a random set, i.e., random locations. Its index set gives the locations of random events that are spatial point pattern.

- $Z(\mathbf{s})$ can equal 1 for all $\mathbf{s} \in D$ (indicating occurrence of the event)

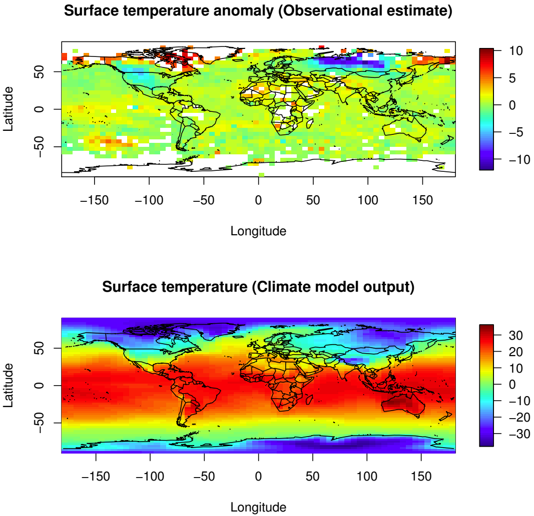
- Examples: Earth quake locations, Spread of certain disease, Wild fires, Mine fields, Lansing wood trees in Michigan (hickory, maple)

Irish mean wind speeds data for 1961-78



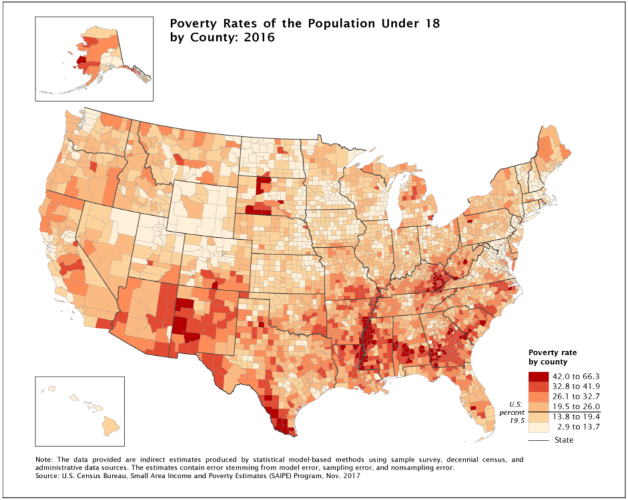
Source: Haslett and Raftery (1989, Applied Statistics)

Surface temperature anomaly and temperature, Dec 2009



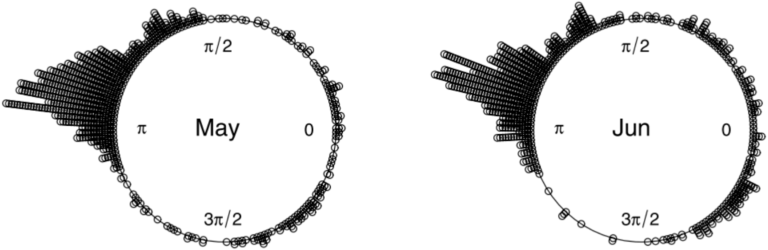
Source: Jeong, Jun, and Genton (2017, Statistical Science)

Poverty rates of the population under 18 by county



Source: www.census.gov

Wind speed and direction data

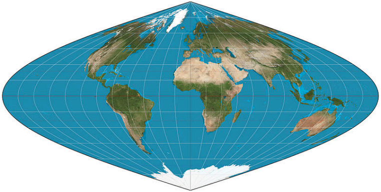


Circular histograms of wind directions at Goodnoe Hilles for May and June in 2003.
Source: Hering and Genton (2010, JASA)

Fundamentals of cartography

- For global data, a popular approach among geographical information system (GIS) users is map projection.
- From (latitude,longitude) $= (L, l)$, construct an appropriate rectangular coordinate system.
- Examples
  - Conformal (preserving angles) projection - the Mercator projection
  - Equal-area (preserving areas) projection - the sinusoidal projection
  - Neither conformal or equal-area projection - the Kavrayskiy VII, Robinson, and Winkel-Tripel
  - ... Jenney, Patterson, and Hurni (2008) - Natural Earth Projection (blends characteristics of the Kavrayskiy VII and Robinson)

Map projections



The Mercator and sinusoidal projections (Source: Wikipedia)

Calculating distances on the surface of a sphere

- The Gauss's Theorema Egregium in differential geometry - the planar map preserving all inter-site distances does not exist.
- For two points, (L_1, l_1) and (L_2, l_2), on $\mathcal{S}^2$, the central angle between them is

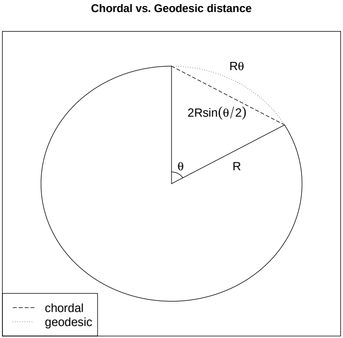
$$\theta = \arccos\{\sin L_1 \sin L_2 + \cos L_1 \cos L_2 \cos(l_1 - l_2)\},\ \theta \in [0, \pi]$$

- The geodesic (arc or great circle) distance : $d_G = R\theta$.

Calculating distances on the surface of a sphere (conti-)

- Note that most statistical theories developed for the Euclidean space. Geodesic distance may cause issues (Earth surface is NOT Euclidean).
- The chordal distance : Euclidean distance in $\mathbb{R}^3$. All theories in Euclidean space applies. $d_C = 2R\sin(\theta/2)$.
- The chordal distance provides accurate approximations for short-mid distances, e.g., Chicago-Minneapolis 562 km (geodesic), 561.8 km (chordal). New York-New Orleans 1897.2 km (geodesic), 1890.2km (chordal)

Calculating distances on the surface of a sphere



- Gneiting (2013, Bernoulli) - "The chordal distance is counter to spherical geometry for larger values of the geodesic distance, and thus may result in physically unrealistic distortion".

Need for spatial statistics and features of spatial analysis

- Roots: geology (mining), geography, meteorology, environmetrics
- Classical statistics: $X_1, \ldots, X_n \overset{iid}{\sim} F$, e.g., F is a normal (Gaussian) distribution
- Spatial data: measurements/observations taken at specific locations or within specific regions
- Key features of spatial data: autocorrelation of observations in space, i.e., observations spatially close tend to be more similar.

Non-Spatial Analysis

- Spatial (geographical) data are analyzed using conventional statistical methods.
- The geographical coordinates are excluded from the computational procedures.
- The results are independent of the spatial arrangement of the geographical entities.
- Observations or entities are assumed to be independent and identically distributed, or in some occasions temporal dependence are also explored.

Non-Spatial Analysis (conti-)

	ATTRIBUTE			
	Variable 1	Variable 2	$\cdots$	Variable n
Entity 1	$attribute_{11}$	$attribute_{12}$	$\cdots$	$attribute_{1n}$
Entity 2	$attribute_{21}$	$attribute_{22}$	$\cdots$	$attribute_{2n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Entity m	$attribute_{m1}$	$attribute_{m2}$	$\cdots$	$attribute_{mn}$

Spatial Analysis

- Spatial (geographical) data are analyzed using **spatial statistical methods**.
- The geographical coordinates are **included** into the computational procedures.
- The results **depend on** the spatial arrangement of the geographical entities.
- It can also include temporal dependence.

Spatial Analysis

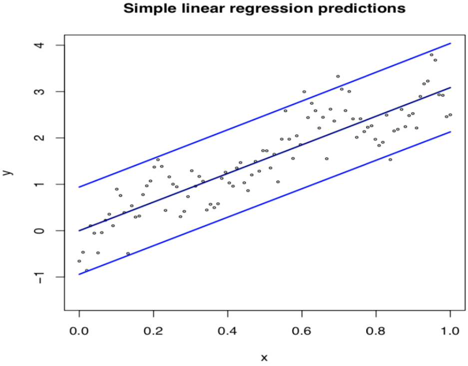
	Geo-Coord		ATTRIBUTE		
	(X, Y)	Variable 1	Variable 2	$\cdots$	Variable n
Entity 1	(X_1, Y_1)	*attribute*$_{11}$	*attribute*$_{12}$	$\cdots$	*attribute*$_{1n}$
Entity 2	(X_2, Y_2)	*attribute*$_{21}$	*attribute*$_{22}$	$\cdots$	*attribute*$_{2n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Entity m	(X_m, Y_m)	*attribute*$_{m1}$	*attribute*$_{m2}$	$\cdots$	*attribute*$_{mn}$

The importance of dependence

- Model will be a poor fit to the data, hence ignoring dependence can lead to poor estimates and poor prediction based on the estimated model.
- Not only do we have poor estimates and predictions, we will underestimate the variability of our estimates (variability of estimates is higher due to dependence).
- Toy example: consider the following simulated realization from a dependent process. For easy visualization, we consider a simple 1-D scenario:
  - Simulate $Z(s_i) = \beta s_i + \epsilon_i$ where $s_i \in (0, 1)$ and $i = 1, \ldots, N$.
  - $(\epsilon_1, \ldots, \epsilon_N)^\top \sim$ zero mean dependent (Gaussian) process.

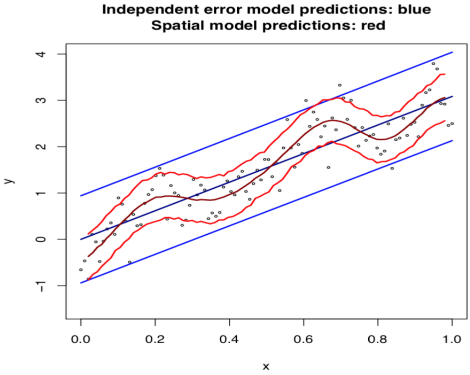
The importance of dependence (conti-)

- Model: simple linear regression with the correct mean but assuming iid error structure. $Z(s_i) = \beta s_i + \epsilon_i$ where ϵ_is are iid.
- Does not capture the data/data generating process well even though trend (β) is estimated correctly.



The importance of dependence (conti-)

- Model: linear regression with correct mean, now assuming dependent error structure. This picks up the ‘wiggles’.
- Independent error model (blue) and dependent error model (red).



Dependence and understanding variability

- Simple example (Cressie, 1993): $Z(1), \dots, Z(n) \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with σ^2 known.
- Estimator of μ, $\hat{\mu} = \bar{Z} = \sum_{i=1}^{n} Z(i)/n$.
- 95% confidence interval for μ: $(\bar{Z} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{Z} + 1.96\frac{\sigma}{\sqrt{n}})$.
- Let spatial data in $\mathbb{R}^1$ (on a line) be dependent:

  $\mathrm{Cov}(Z(i), Z(j)) = \sigma^2 \rho^{|i-j|}, i, j = 1, \dots, n, \rho \in (0, 1)$.

  $\mathrm{Var}(\bar{Z}) = \sum_i \sum_j \mathrm{Cov}(Z(i), Z(j))/n^2$

  $$= \frac{\sigma^2}{n}\Big(1 + 2\Big(\frac{\rho}{1-\rho}\Big)\Big(1 - \frac{1}{n}\Big) - 2\Big(\frac{\rho}{1-\rho}\Big)^2 (1 - \rho^{n-1})/n\Big).$$

Nearby things tend to be more alike

- Spatial (and temporal) dependence is the rule.

  1. Nearby (in space and time) observations tend to be more alike than those far apart, e.g., spatial interaction, contagion, spill-overs, copycatting.
  2. Competition: opposite may happen.
  3, Physical barriers can affect what is meant by 'nearby' or 'neighboring', e.g., rivers, mountains.

- **Spatio-temporal data should not be modeled as being statistically independent**.

- Tobler (1970) called this **the first law of geography**: *everything depends on everything else, but closer things more*.

Differences between spatial and time series problems

- Seems reasonable to think of spatial modeling as "2-D/3-D time series modeling."
- One-dimensional time domain is **fully ordered** while we can only **partially order** the spatial domain.
- Time series: dependence is from past to present to future while spatial dependence is in all directions.
- With time series, we are most often interested in **extrapolation**, i.e., predicting what happens in the future, while with spatial data, we are most often interested in **interpolation**, i.e., what happens at unobserved locations between sites (extrapolation is usually inappropriate).
- **Space is different from time**: Modeling spatio-temporal phenomena needs to respect these differences.

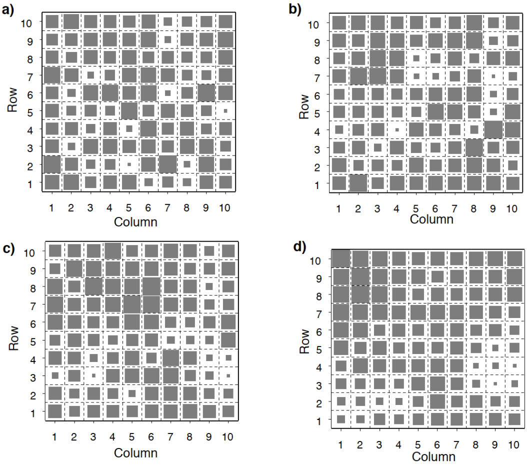
Large sample theory: spatial and time series problems

- **Large sample theory** for time series involves imagining increasing number of data points in time ($t \to \infty$).
- Large sample analysis for spatial process data domain involves **infill asymptotics**, i.e., envisioning an increasing amount of information available for the same region. Makes sense since interest is in interpolation (Some authors study **increasing domain asymptotics**).
- Often only have a single realization whether spatial or time series process: i) Makes us worry about inference based on a single realization and ii) Usual large sample theory seems awkward/inappropriate (Aside: with longitudinal data, usually have lots of replication, so this is not an issue).

Example: simulated data on 10×10 lattice $\overset{iid}{\sim} N(5, 1)$

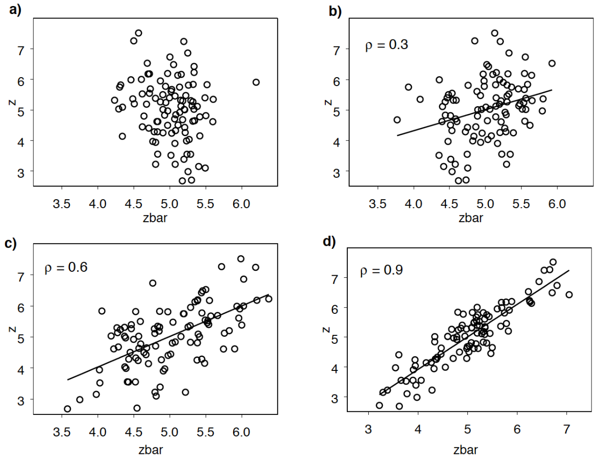
- a) observations assigned randomly to lattice coordinates
- b)-d) data rearranged: each value surrounded by more similar values (by simulated annealing algorithm)
- Define nearest neighbors: move queen piece on chess board
- $(\mathbf{s}_i, \bar{Z}_i), i = 1, \dots, 100, \bar{Z}_i$=average of neighboring sites of $\mathbf{s}_i$
- Plot: $(\bar{Z}_i, Z(\mathbf{s}_i))$.

Different autocorrelations



Source: Schabenberger and Gotway (2005)

Different autocorrelations



Source: Schabenberger and Gotway (2005)

General reasons to use spatial models

- Utilizing spatial dependence leads to superior estimators.
- Ignoring dependence may underestimate variability.
- Learning about spatial dependence may be of interest in its own right.
- Spatial dependence can be surrogate for unknown and important covariates: can 'adjust' for these covariates. Spatial dependence component can project against misspecification on mean structure (by accounting for a variable that is spatially varying). "What is one person's (spatial) covariance structure may be another person's mean structure." (Cressie, 1993).
- Dependent models: useful for modeling complicated functional forms (even when there is no dependence!).

Spatial modeling

- Scientists are often interested in one or more of the following:
  - Modeling of trends and correlation structures
  - Estimation of the model parameters
  - Hypothesis Testing (or comparison of competing models)
  - Prediction of observations at unobserved times/locations
  - Experimental design: location of experimental units for optimal inference
- When spatial *dynamics* (mechanism of spatial spread) are of interest, need different tools:
  - Spatio-temporal models
  - Spatial point process
  - Emulation of complex models

Spatial statistics using R

- R is a free statistical package (`http://r-project.org`)
- There are many resources for you to get started:
  - `https://www.statmethods.net/index.html`
  - `https://r4ds.had.co.nz`
- R packages for spatial statistics and point patterns:
  - STARbook (`https://github.com/andrewzm/STRbook`)
  - fields (`https://www.image.ucar.edu/Software`)
  - geoR (`http://www.leg.ufpr.br/geoR`)
  - RandomFields (`https://cran.r-project.org/web/packages/RandomFields/index.html`)
  - spatstat (`https://cran.r-project.org/web/packages/spatstat/index.html`)

Reference

- Cressie, N. Statistics for Spatial Data. Wiley.
- Banerjee, S., Carlin, B., and Gelfand, A. Hierarchical Modeling and Analysis for Spatial Data (2nd). CRC Press.
- Jun, M., Genton, M. G., and Jeong, J. Lecture Notes for Spatial Statistics. UH, KAUST, and HYU.