Learning Objectives

- After this segment, students will be able to
  - List limitations of traditional statistics for spatial data
  - Describe simple concepts in spatial statistics
    - Spatial auto-correlation
    - Spatial heterogeneity
  - Describe first law of Geography

Limitations of Traditional Statistics

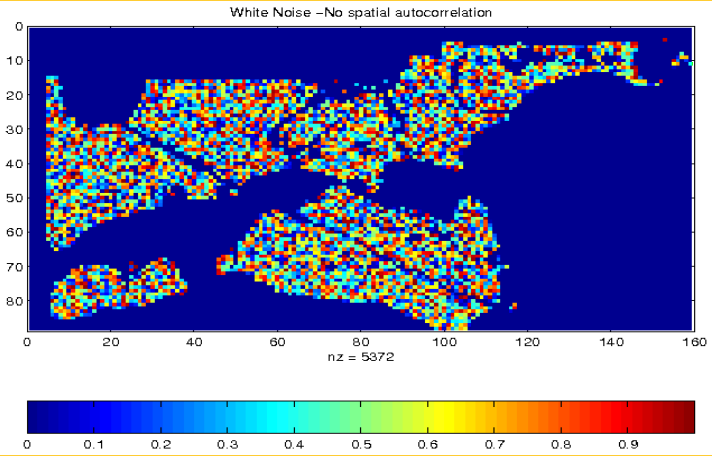
- Classical Statistics
  - Data samples: independent and identically distributed (i.i.d.)
  - Simplifies mathematics underlying statistical methods, e.g., Linear Regression
- Spatial data samples are not independent
  - Spatial Autocorrelation metrics
    - distance-based (e.g., K-function), neighbor-based (e.g., Moran's I)
  - Spatial Cross-Correlation metrics
- Spatial Heterogeneity
  - Spatial data samples may not be identically distributed!
  - No two places on Earth are exactly alike!
- …

Spatial Statistics: An Overview

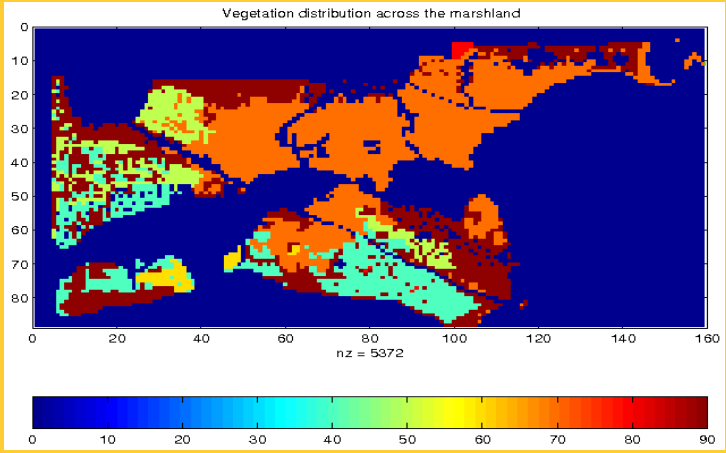
- Point process
  - Discrete points, e.g., locations of trees, accidents, crimes, …
  - Complete spatial randomness (CSR): Poisson process in space
  - K-function: test of CSR
- Geostatistics
  - Continuous phenomena, e.g., rainfall, snow depth, …
  - Methods: Variogram measure how similarity decreases with distance
  - Spatial interpolation, e.g., Kriging
- Lattice-based statistics
  - Polygonal aggregate data, e.g., census, disease rates, pixels in a raster
  - Spatial Gaussian models, Markov Random Fields, Spatial Autoregressive Model

Spatial Autocorrelation (SA)

- First Law of Geography
  - All things are related, but nearby things are more related than distant things. [Tobler70]
- Spatial autocorrelation
  - Traditional i.i.d. assumption is not valid
  - Measures: K-function, Moran's I, Variogram, …



Independent, Identically Distributed pixel property



Vegetation Durability with SA

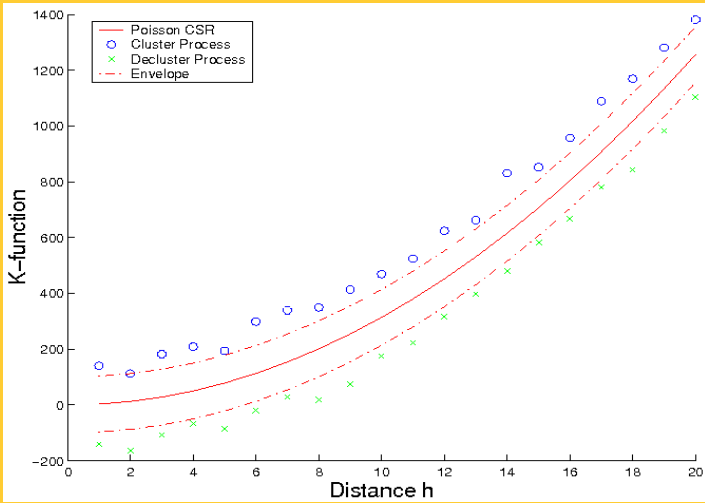
Spatial Autocorrelation: K-Function

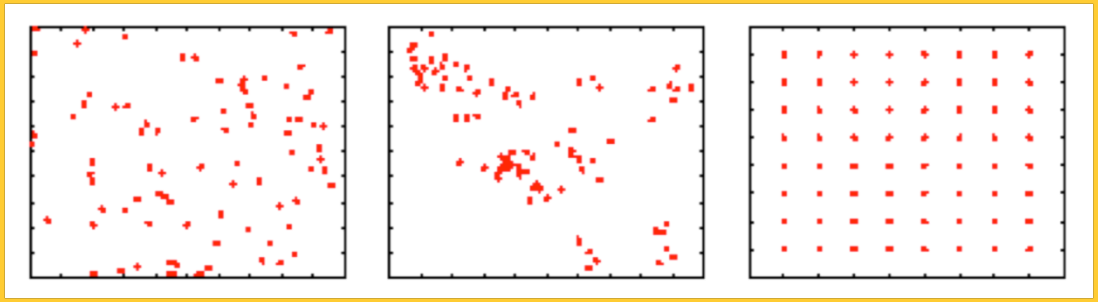
- **Purpose:** Compare a point dataset with a complete spatial random (CSR) data
- **Input:** A set of points

  $K(h, data) = \lambda^{-1} E$ [number of events within distance *h* of an arbitrary event]
  - where λ is intensity of event
- **Interpretation:** Compare k(h, data) with *K(h,* CSR)
  - *K(h, data) = k(h, CSR):* Points are CSR

    > means Points are clustered

    < means Points are de-clustered





CSR Clustered De-clustered

Cross-Correlation

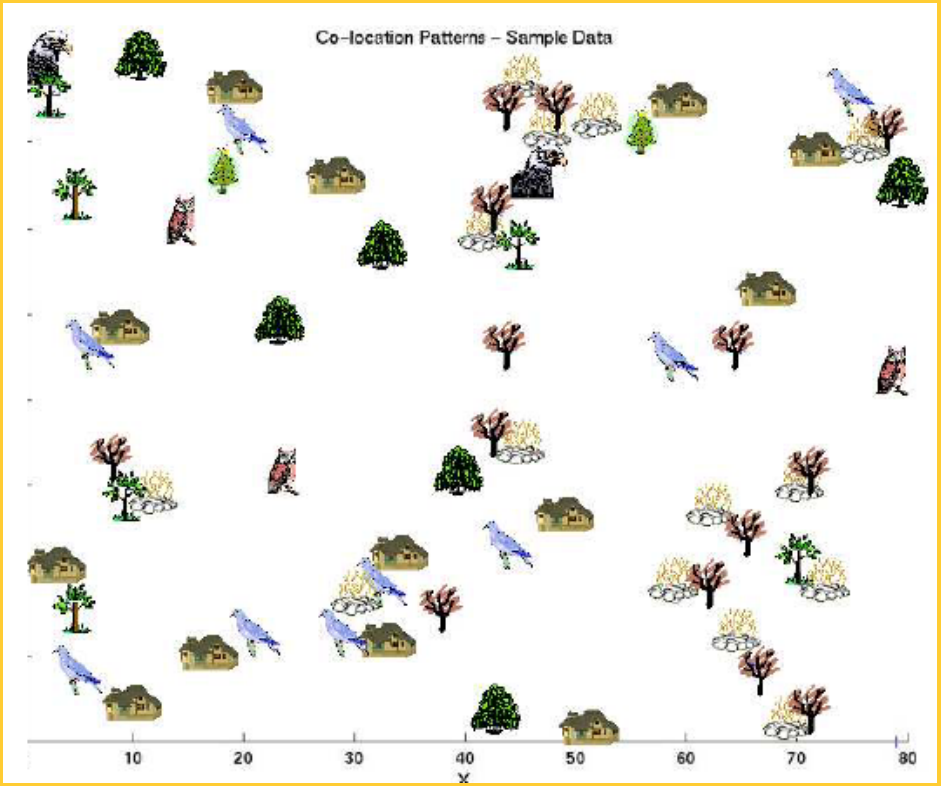
- Cross K-Function Definition

$$K_{ij}(h) = \lambda_j^{-1} E \left[\text{number of type } \boldsymbol{j} \text{ event within distance } h \text{ of a randomly chosen type } \boldsymbol{i} \text{ event}\right]$$

  - Cross K-function of some pair of spatial feature types
  - Example
    - Which pairs are frequently co-located
    - Statistical significance

Recall Pattern Family 4: Co-locations

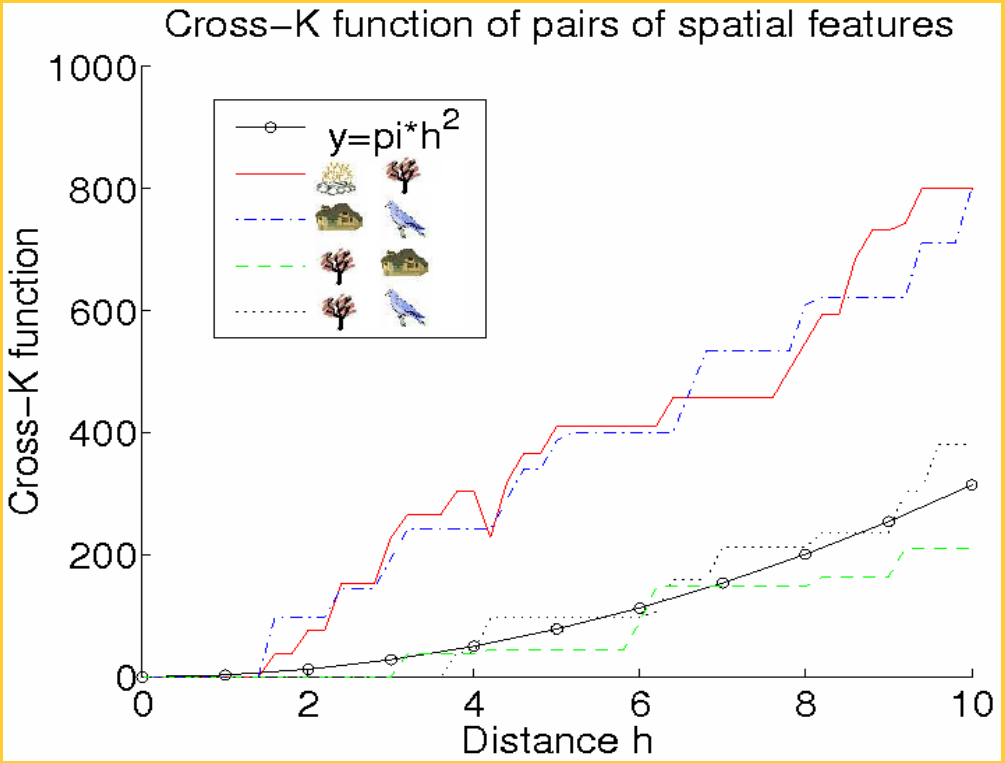
- Given: A collection of different types of spatial events
- Find: Co-located subsets of event types



Source: Discovering Spatial Co-location Patterns: A General Approach, IEEE Transactions on Knowledge and Data Eng., 16(12), December 2004 (w/ H.Yan, H.Xiong).

Illustration of Cross-Correlation

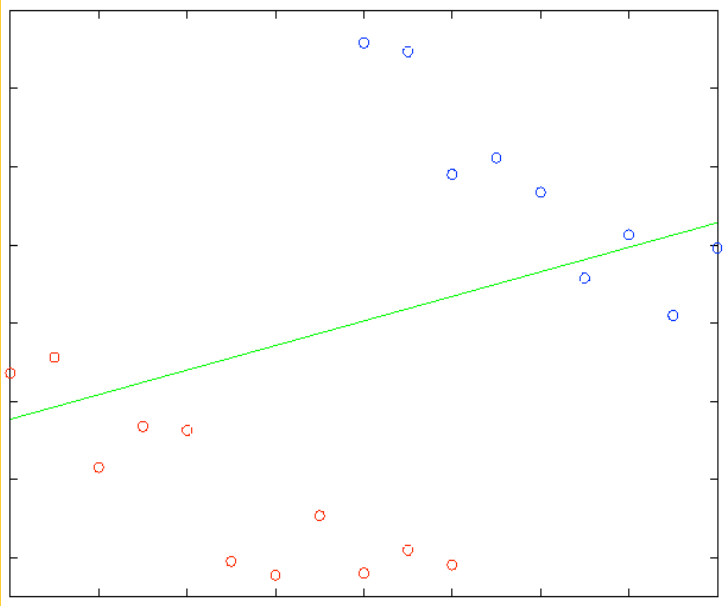
- Illustration of Cross K-function for Example Data

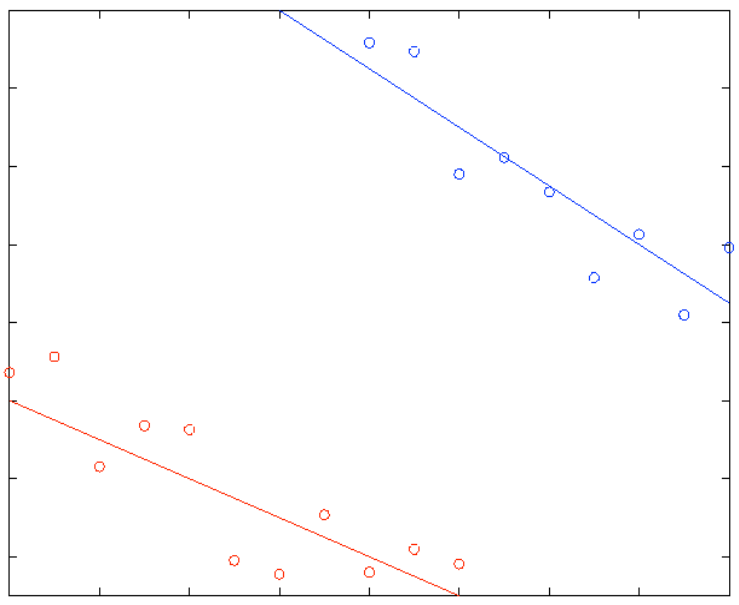


Cross-K Function for Example Data

Spatial Heterogeneity

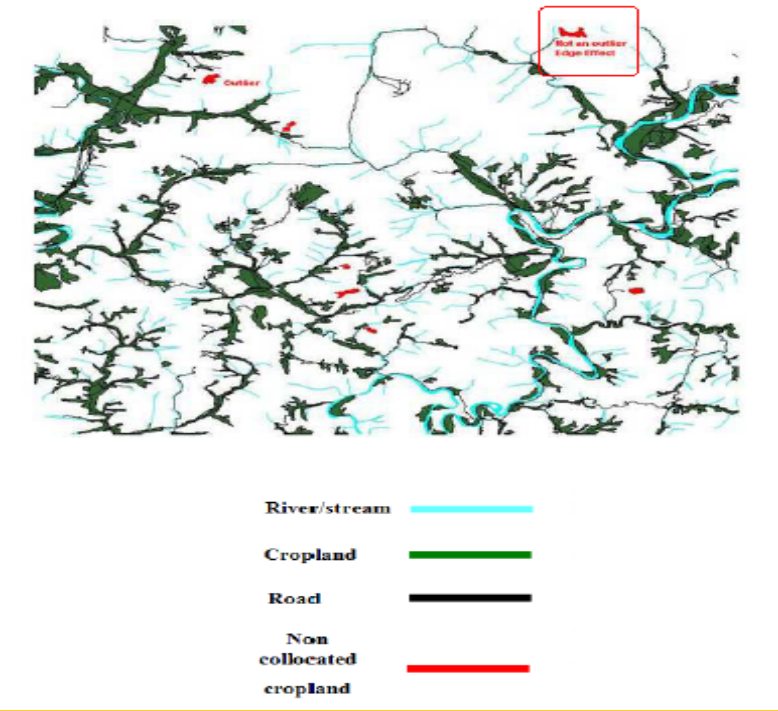
- "Second law of geography" [M. Goodchild, UCGIS 2003]
- Global model might be inconsistent with regional models
  - Spatial Simpson's Paradox
- May improve the effectiveness of SDM, show support regions of a pattern





Edge Effect

- Cropland on edges may not be classified as outliers
- No concept of spatial edges in classical data mining



Korea Dataset, Courtesy: Architecture Technology Corporation

Research Challenges of Spatial Statistics

- State-of-the-art of Spatial Statistics

		Point Process	Lattice	Geostatistics
raster			√	√
Vector	Point	√	√	√
	Line			√
	Polygon		√	√
graph				

Data Types and Statistical Models

- Research Needs
  - Correlating extended features, road, rivers, cropland
  - Spatio-temporal statistics
  - Spatial graphs, e.g., reports with street address