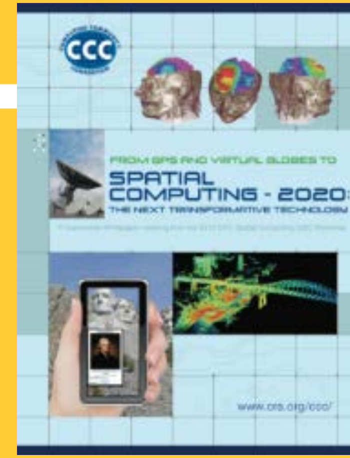
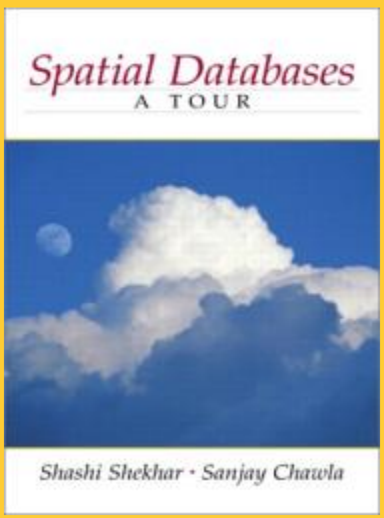
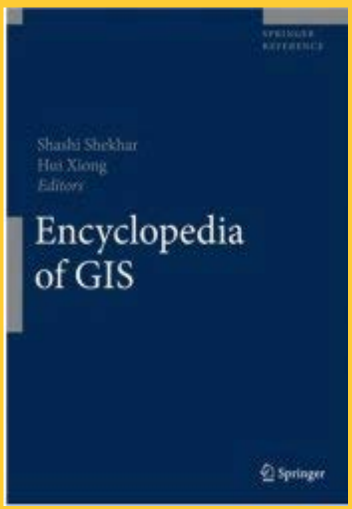
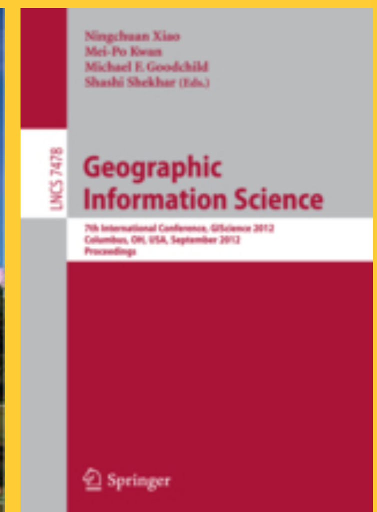
What is special about mining spatial data?



Shashi Shekhar
McKnight Distinguished University Professor
Dept. of Computer Sc. and Eng.
University of Minnesota
www.cs.umn.edu/~shekhar

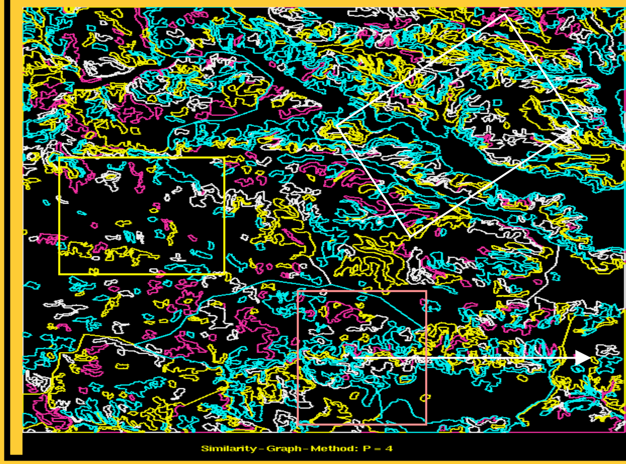






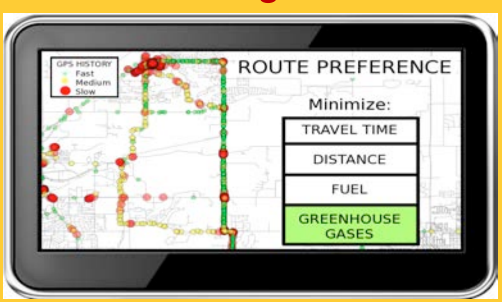
Spatial Databases: Representative Projects

Details: Spatial Databases: Accomplishments and Research Needs, IEEE Transactions on Knowledge and Data Engineering, 11(1), 1999. (and recent update via a technical report)



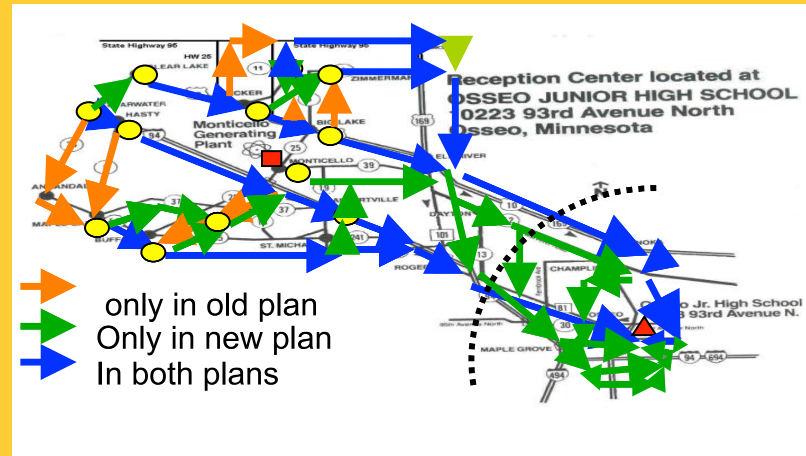
Parallelize Range Queries

Eco-Routing



Storing graphs on disk

Evacutation Route Planning

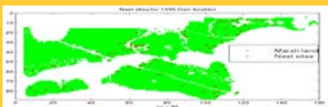


Spatial Data Mining: Example Projects

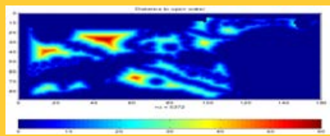
Details: Identifying patterns in spatial information: a survey of methods, Wiley Interdisc. Reviews: Data Mining and Know. Discovery , 1(3):193-214, May/June 2011

Location prediction: nesting sites

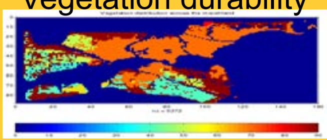
Nest locations



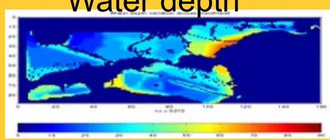
Distance to open water



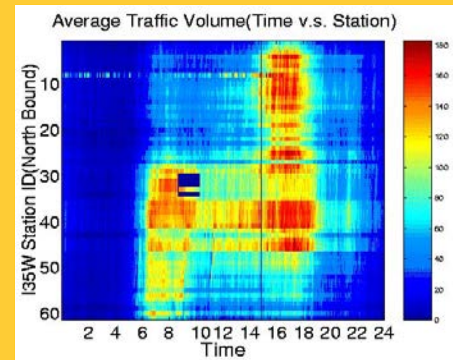
Vegetation durability



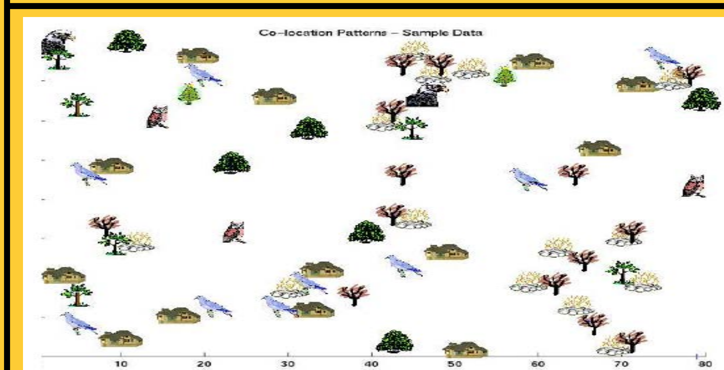
Water depth



Spatial outliers: sensor (#9) on I-35



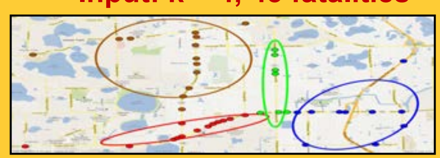
Co-location Patterns



Spatial Network Activity Summarization

Input: k = 4, 43 fatalities

Euclidean Distance



Network Distance

KMR

Outline

- Motivation
  - Use cases
  - Pattern families
- Spatial Data Types
- Spatial Statistical Foundations
- Spatial Data Mining
- Conclusions

Why Data Mining?

- <u>Holy Grail</u> - <u>Informed</u> Decision Making
- Sensors & Databases increased rate of Data Collection
  - Transactions, Web logs, GPS-track, Remote sensing, …
- Challenges:
  - Volume (data) >> number of human analysts
  - Some automation needed
- Approaches
  - Database Querying, e.g., SQL3/OGIS
  - Data Mining for Patterns
  - …

Data Mining vs. Database Querying

- Database Querying (e.g., SQL3/OGIS)
  - Does not answer questions about items not in the database!
    - Ex. Predict tomorrow's weather or credit-worthiness of a new customer
  - Does not efficiently answer complex questions beyond joins
    - Ex. What are natural groups of customers?
    - Ex. Which subsets of items are bought together?
- Data Mining may help with above questions!
  - Prediction Models
  - Clustering, Associations, …

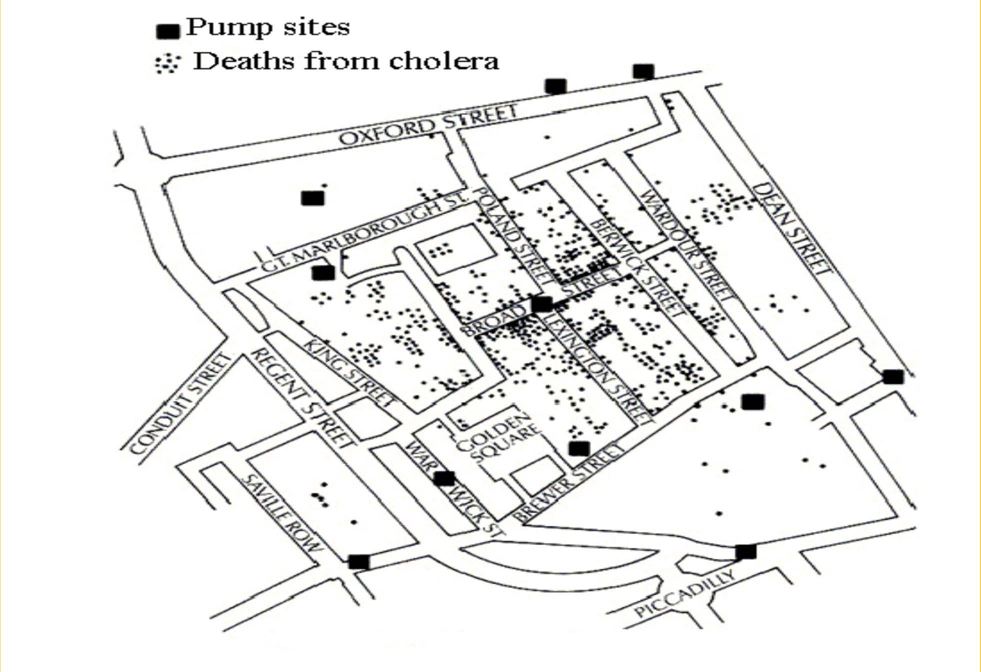
Spatial Data Mining (SDM)

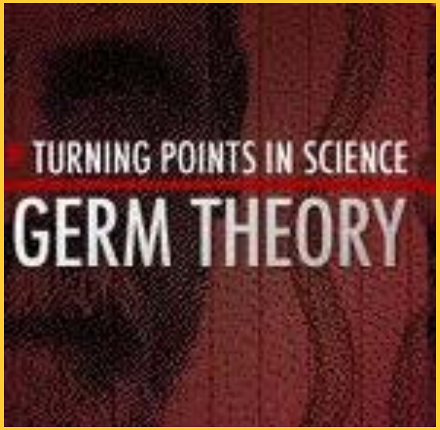
- The process of discovering
  - interesting, useful, non-trivial patterns
    - patterns: non-specialist
    - exception to patterns: specialist
  - from large spatial datasets

- Spatial pattern families
  - Hotspots, Spatial clusters
  - Spatial outlier, discontinuities
  - Co-locations, co-occurrences
  - Location prediction models
  - …

Pattern Family 1: Hotspots, Spatial Cluster

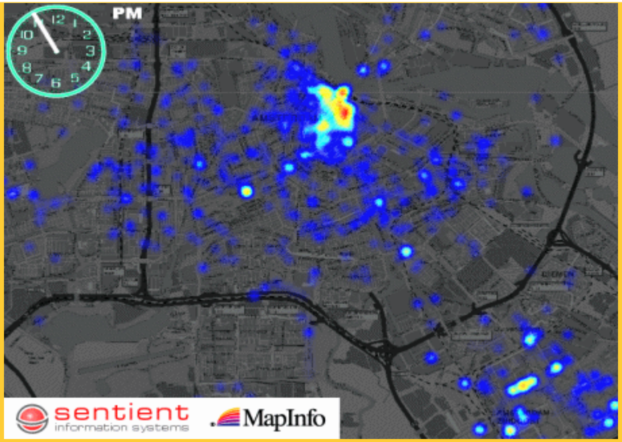
- The 1854 Asiatic Cholera in London
  - Near Broad St. water pump except a brewery

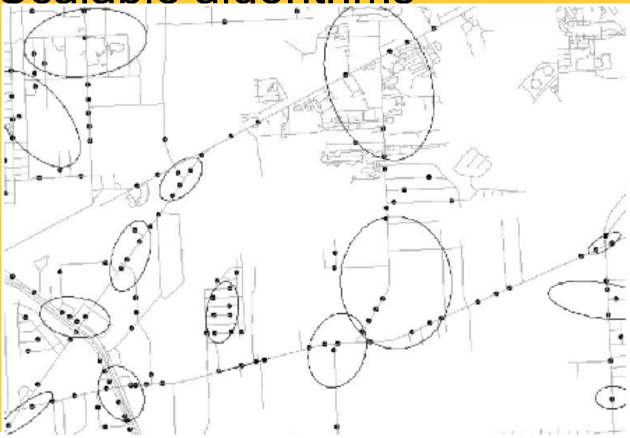


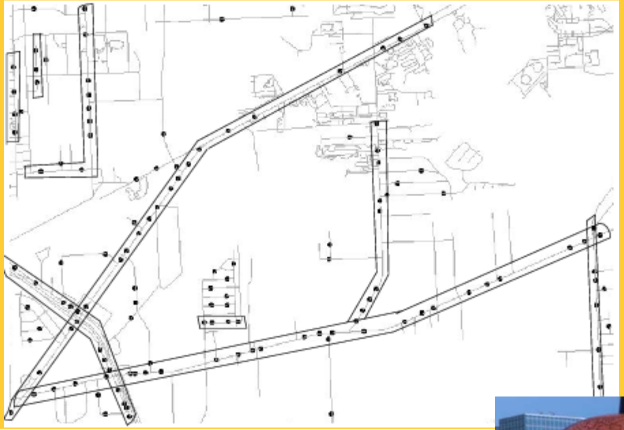


Complicated Hotspots

- Complication Dimensions
  - Time
  - Spatial Networks
- Challenges: Trade-off b/w
  - Semantic richness and
  - Scalable algorithms

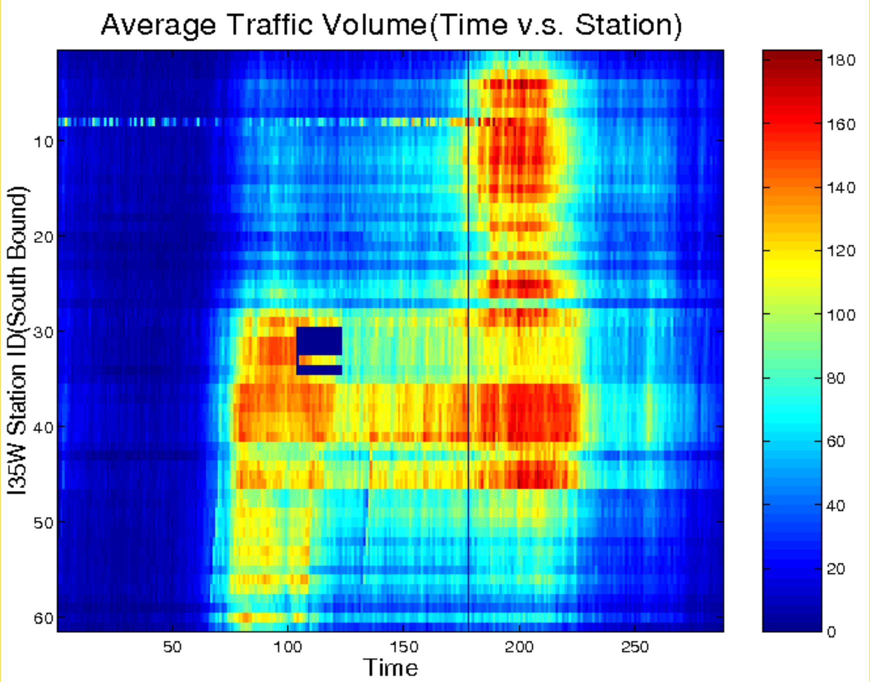






Pattern Family 2: Spatial Outliers

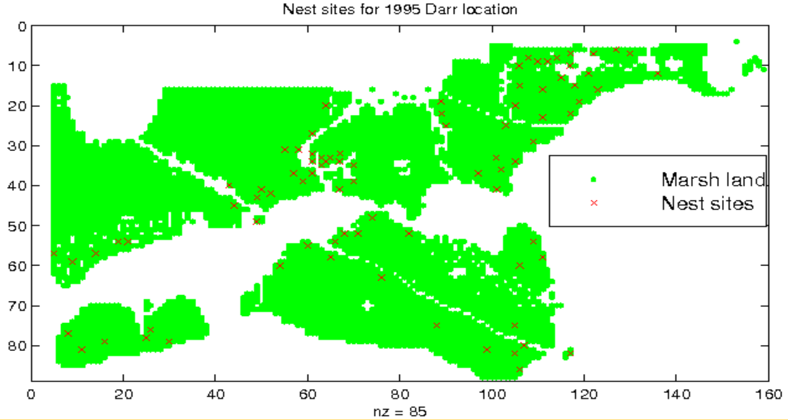
- Spatial Outliers, Anomalies, Discontinuities
  - Traffic Data in Twin Cities
  - Abnormal Sensor Detections
  - Spatial and Temporal Outliers

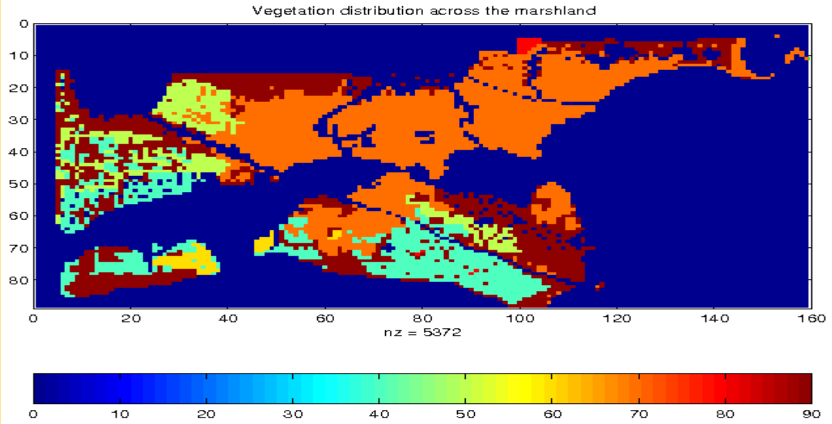


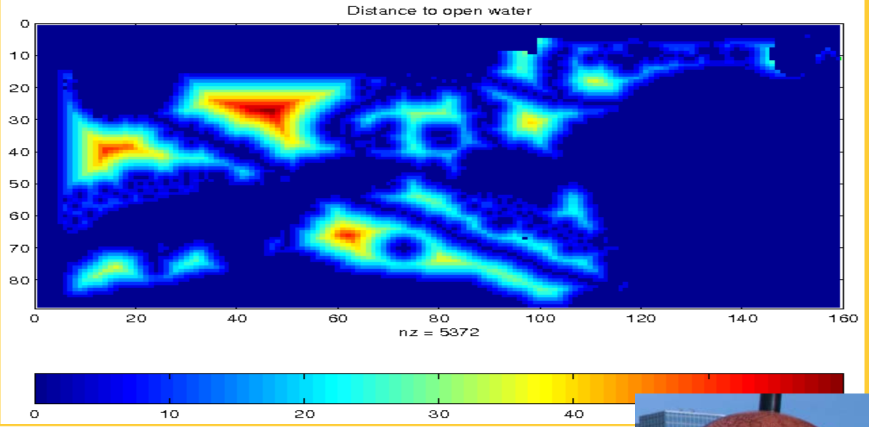
<u>Source: </u>A Unified Approach to Detecting Spatial Outliers, GeoInformatica, 7(2), Springer, June 2003. (A Summary in Proc. ACM SIGKDD 2001) with C.-T. Lu, P. Zhang.

Pattern Family 3: Predictive Models

- Location Prediction:
  - Predict Bird Habitat Prediction
  - Using environmental variables

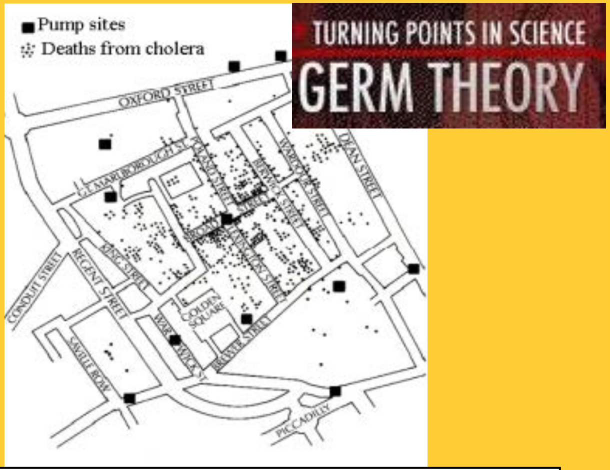


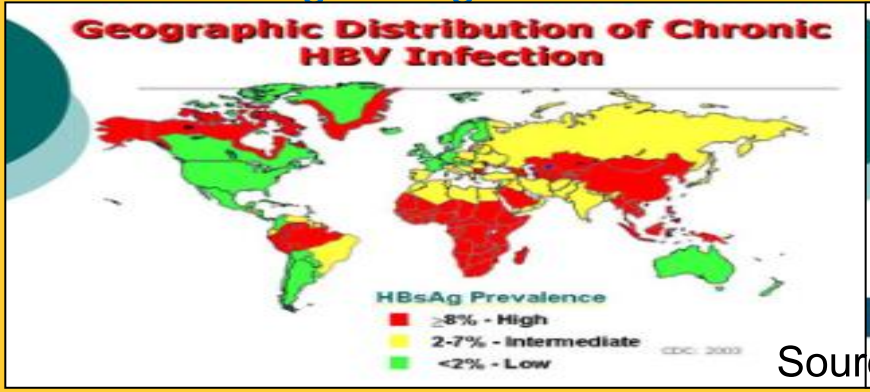


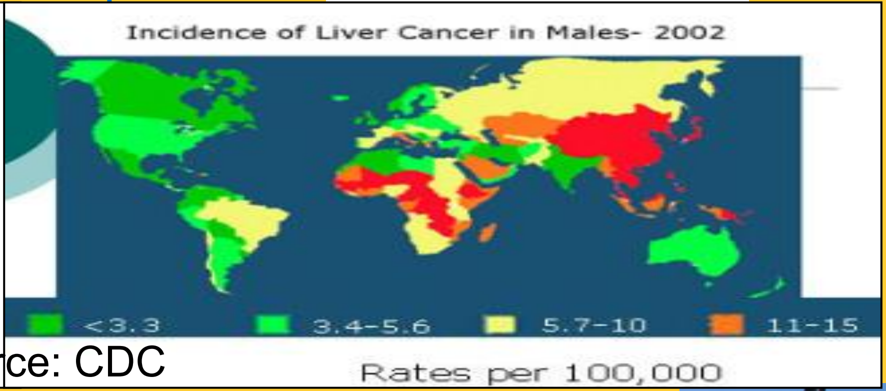


Family 4: Co-location, Co-occurrence

- Co-location (Cholera Deaths, Water Pump)
  - Hypothesis: Cholera is water-borne (1854)
  - Miasama theory => Germ Theory
- Co-location (Liver Cancer, HBV infection)
- Which exposures and cancers are co-located?
  - Challenge: Large number of candidate pairs!



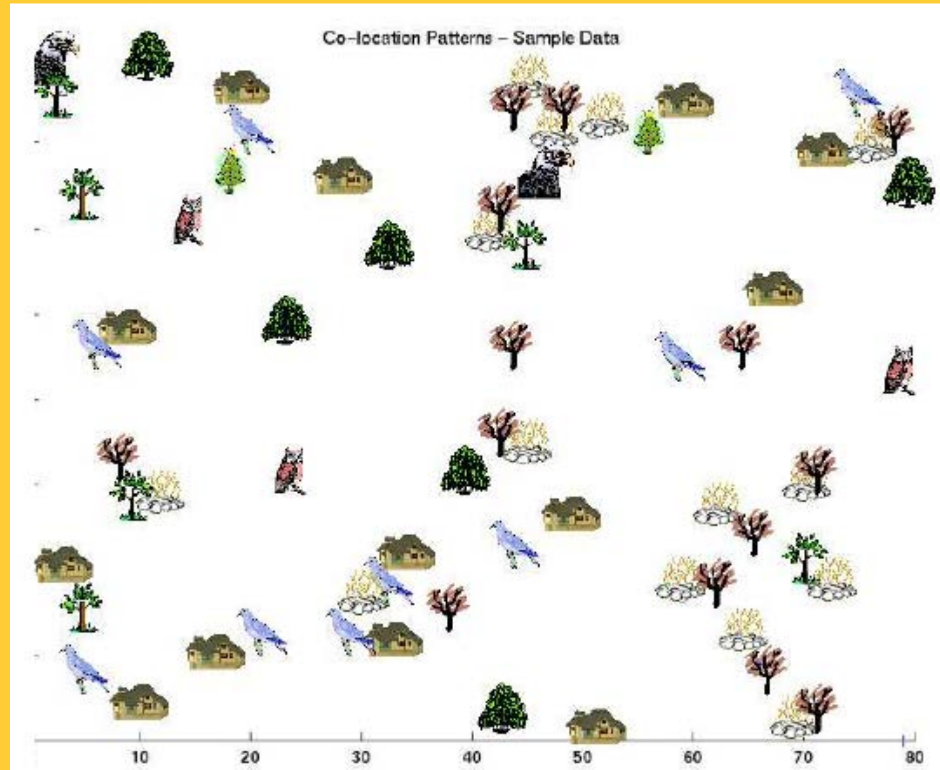




Source: CDC

Family 4: Co-locations/Co-occurrence

- 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).

What's NOT Spatial Data Mining (SDM)

- Simple Querying of Spatial Data
  - Find neighbors of Canada, or shortest path from Boston to Houston
- Testing a hypothesis via a primary data analysis
  - Ex. Is cancer rate inside Hinkley, CA higher than outside ?
  - SDM: Which places have significantly higher cancer rates?
- Uninteresting, obvious or well-known patterns
  - Ex. (Warmer winter in St. Paul, MN) => (warmer winter in Minneapolis, MN)
  - SDM: (Pacific warming, e.g. El Nino) => (warmer winter in Minneapolis, MN)
- Non-spatial data or pattern
  - Ex. Diaper and beer sales are correlated
  - SDM: Diaper and beer sales are correlated in blue-collar areas (weekday evening)

Review Quiz: Spatial Data Mining

- Categorize following into queries, hotspots, spatial outlier, colocation, location prediction:
  - (a) Which countries are very different from their neighbors?
  - (b) Which highway-stretches have abnormally high accident rates ?
  - (c) Forecast landfall location for a Hurricane brewing over an ocean?
  - (d) Which retail-store-types often co-locate in shopping malls?
  - (e) What is the distance between Beijing and Chicago?

Outline

- Motivation
- Spatial Data
  - Spatial Data Types & Relationships
  - OGIS Simple Feature Types
- Spatial Statistical Foundations
- Spatial Data Mining
- Conclusions

Data-Types: Non-Spatial vs. Spatial

- Non-spatial
  - Numbers, text-string, …
  - e.g., city name, population
- Spatial (Geographically referenced)
  - Location, e.g., longitude, latitude, elevation
  - Neighborhood and extent
- Spatial Data-types
  - Raster: gridded space
  - Vector: point, line, polygon, …
  - Graph: node, edge, path

Raster (Courtesy: UMN)

Vector (Courtesy: MapQuest)

Relationships: Non-spatial vs. Spatial

- Non-spatial Relationships
  - Explicitly stored in a database
  - Ex. New Delhi is the capital of India
- Spatial Relationships
  - Implicit, computed on demand
  - Topological: meet, within, overlap, …
  - Directional: North, NE, left, above, behind, …
  - Metric: distance, area, perimeter
  - Focal: slope
  - Zonal: highest point in a country
  - …

OGC Simple Features

- Open GIS Consortium: Simple Feature Types
  - Vector data types: e.g. point, line, polygons
  - Spatial operations :

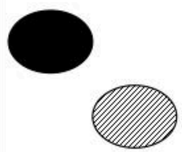
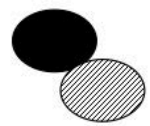
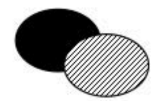
Operator Type	Operator Name
Basic Function	SpatialReference, Envelope, Boundary, Export, IsEmpty, IsSimple
Topological/Set Operations	Equal, Disjoint, Intersect, Touch, Cross, Within, Contains, Overlap
Spatial Analysis	Distance, Buffer, ConvexHull, Intersection, Union, Difference, SymmDiff

Examples of Operations in OGC Model

OGIS - Topological Operations

- Topology: 9-intersections
  - interior
  - boundary
  - exterior

$$
\begin{array}{cccc}
\text{Interior(B)} & \text{Boundary(B)} & \text{Exterior(B)} & \\
\left(\begin{array}{ccc}
(A^{\circ} \cap B^{\circ}) & (A^{\circ} \cap \partial B) & (A^{\circ} \cap B^{-}) \\
(\partial A \cap B^{\circ}) & (\partial A \cap \partial B) & (\partial A \cap B^{-}) \\
(A^{-} \cap B^{\circ}) & (A^{-} \cap \partial B) & (A^{-} \cap B^{-})
\end{array}\right) & & & \begin{array}{l}\text{Interior(A)}\\ \text{Boundary(A)}\\ \text{Exterior(A)}\end{array}
\end{array}
$$

Topological Relationship	 disjoint	 meet	 overlap	 equal
9-intersection model	$\begin{pmatrix} 0&0&1\\0&0&1\\1&1&1 \end{pmatrix}$	$\begin{pmatrix} 0&0&1\\0&1&1\\1&1&1 \end{pmatrix}$	$\begin{pmatrix} 1&1&1\\1&1&1\\1&1&1 \end{pmatrix}$	$\begin{pmatrix} 1&0&0\\0&1&0\\0&0&1 \end{pmatrix}$

Research Needs for Data

- Limitations of OGC Model
  - Direction predicates - e.g. absolute, ego-centric
  - 3D and visibility, Network analysis, Raster operations
  - Spatio-temporal
- Needs for New Standards & Research
  - Modeling richer spatial properties listed above
  - Spatio-temporal data, e.g., moving objects

Outline

- Motivation
- Spatial Data Types
- Spatial Statistical Foundations
  - Spatial Auto-correlation
  - Heterogeneity
  - Edge Effect
- Spatial Data Mining
- Conclusions

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

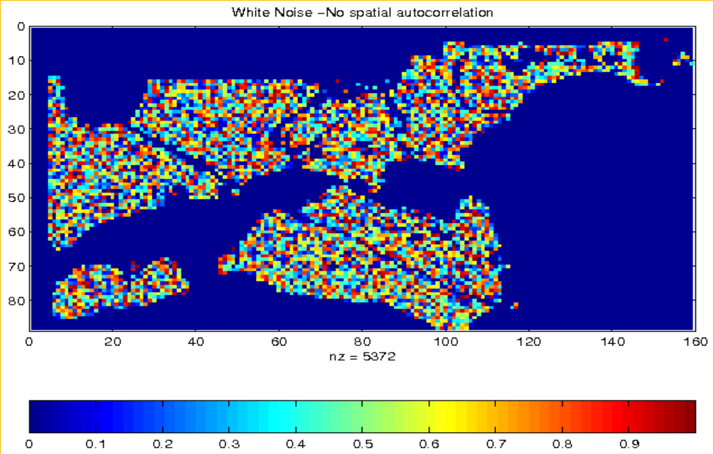
- Point process
  - Discrete points, e.g., locations of trees, accidents, crimes, …
  - Complete spatial randomness (CSR): Poisson process in space
  - K-function: test of CSR

- Geostatistics
  - Continuous phenomena, e.g., rainfall, snow depth, …
  - Methods: Variogram measure how similarity decreases with distance
  - Spatial interpolation, e.g., Kriging

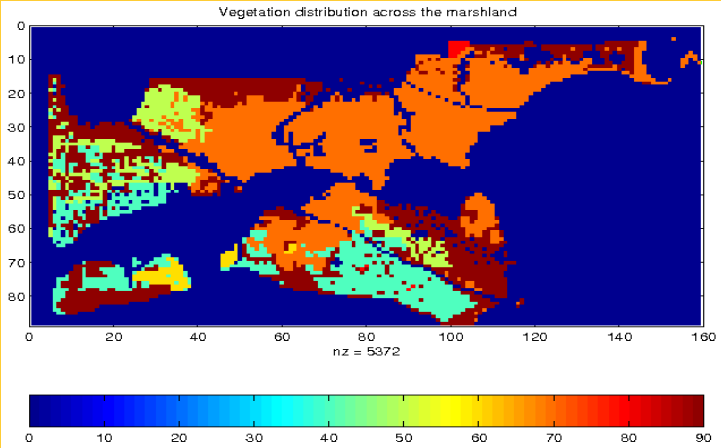
- Lattice-based statistics
  - Polygonal aggregate data, e.g., census, disease rates, pixels in a raster
  - Spatial Gaussian models, Markov Random Fields, Spatial Autoregressive Model

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

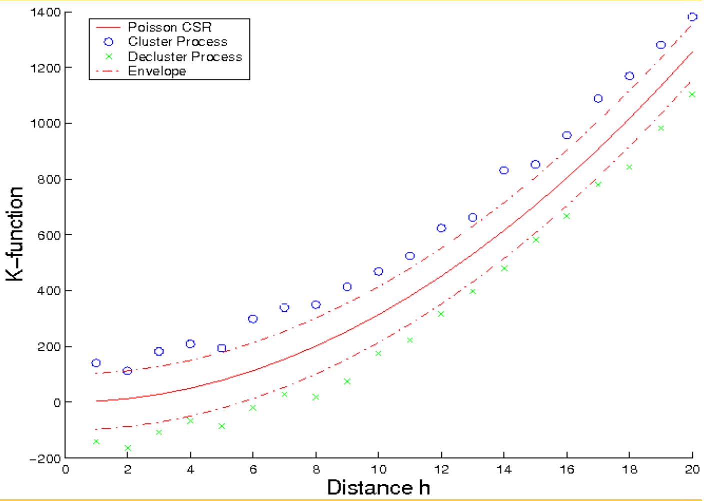
- **Purpose:** Compare a point dataset with a complete spatial random (CSR) data
- **Input:** A set of points

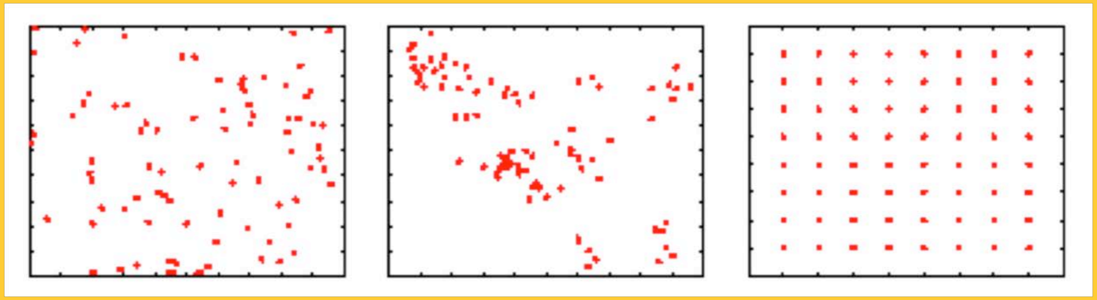
$K(h, data) = \lambda^{-1} E$ [number of events within distance h of an arbitrary event]

  - 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

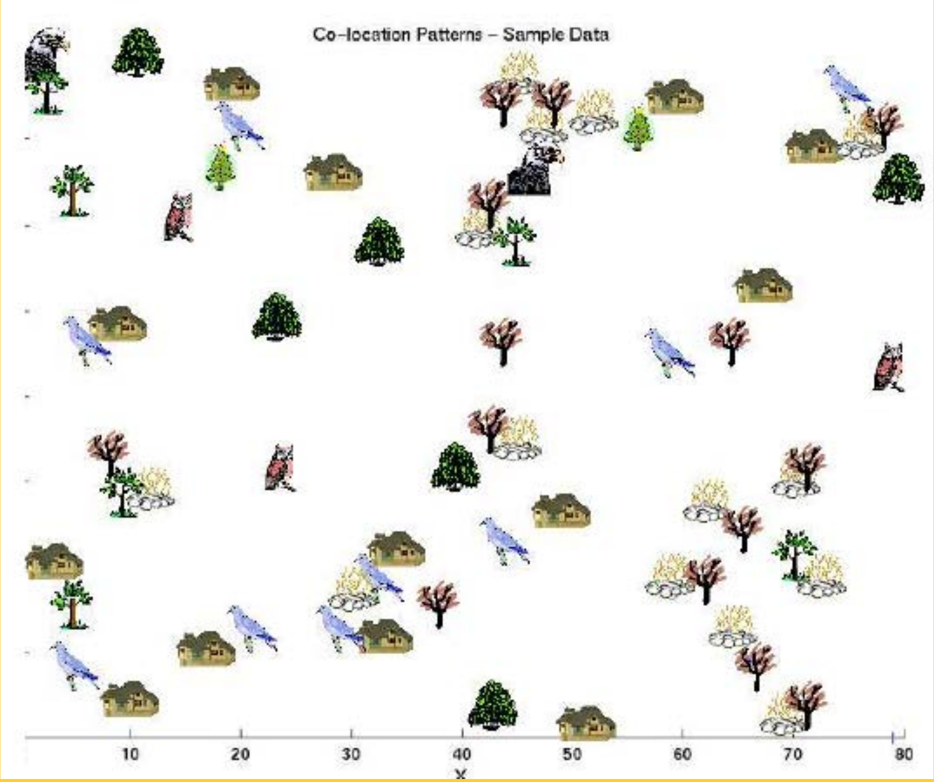
- Cross K-Function Definition

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

  - 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

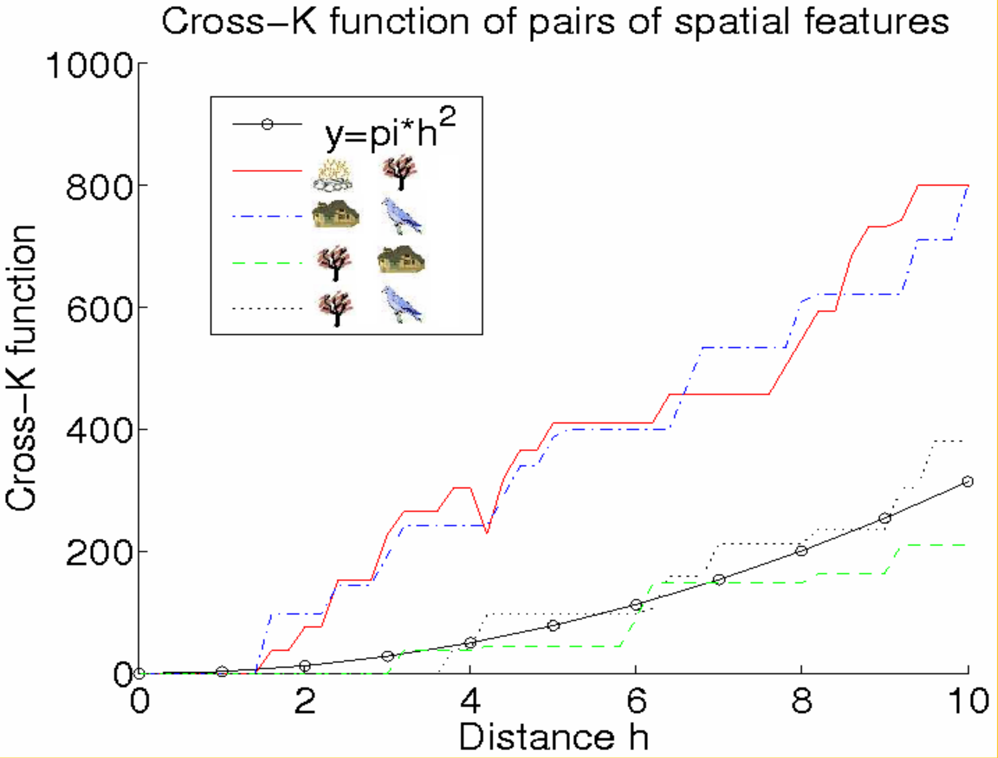
- Given: A collection of different types of spatial events
- Find: Co-located subsets of event types



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

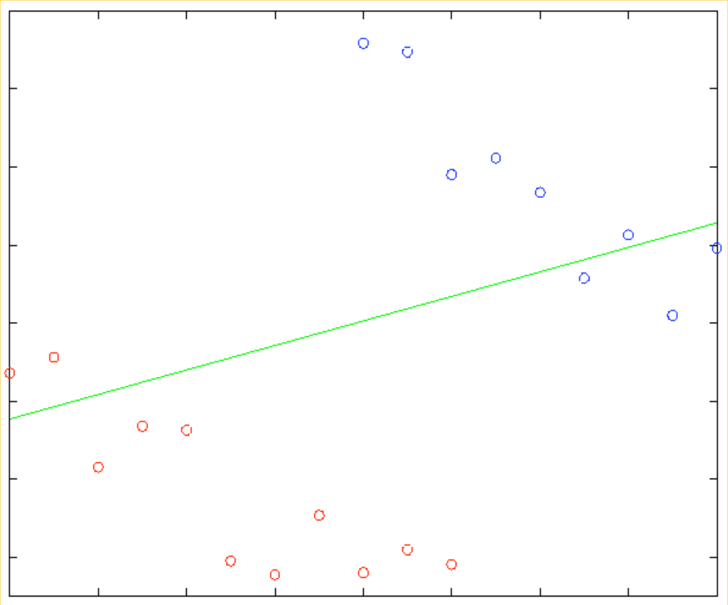
- Illustration of Cross K-function for Example Data

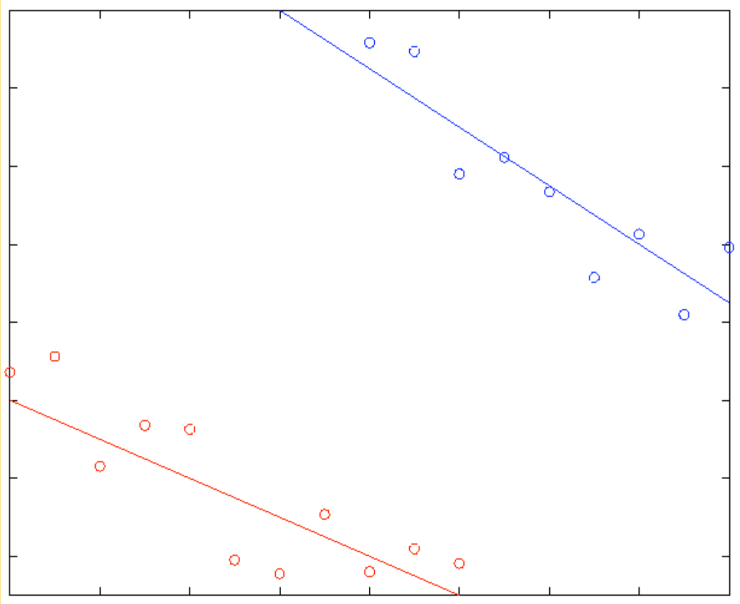


Cross-K Function for Example Data

Spatial Heterogeneity

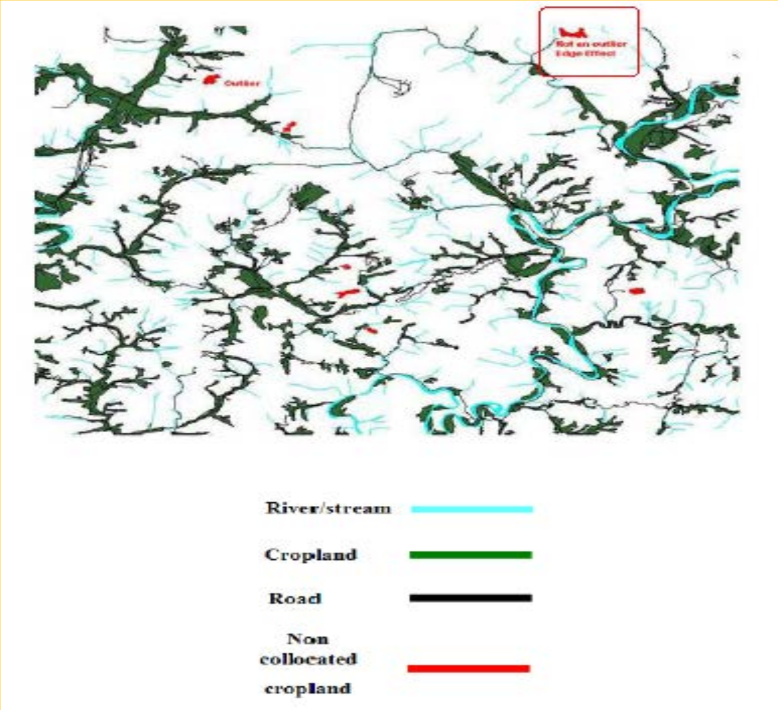
- "Second law of geography" [M. Goodchild, UCGIS 2003]
- Global model might be inconsistent with regional models
  - Spatial Simpson's Paradox
- May improve the effectiveness of SDM, show support regions of a pattern





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

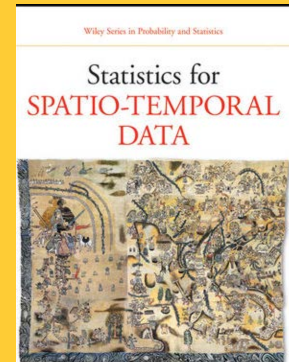
Research Challenges of Spatial Statistics

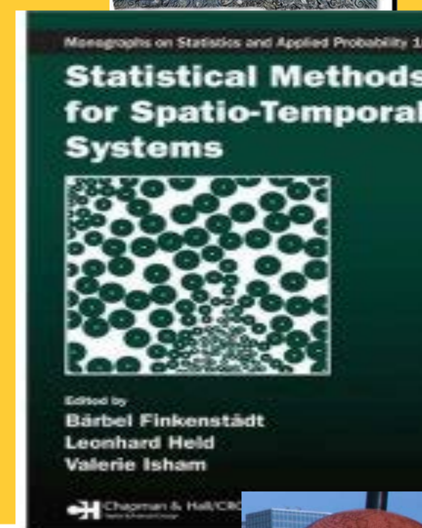
- State-of-the-art of Spatial Statistics

		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

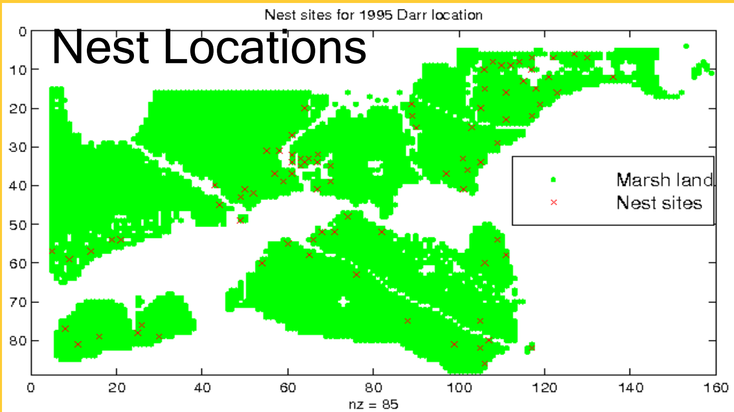


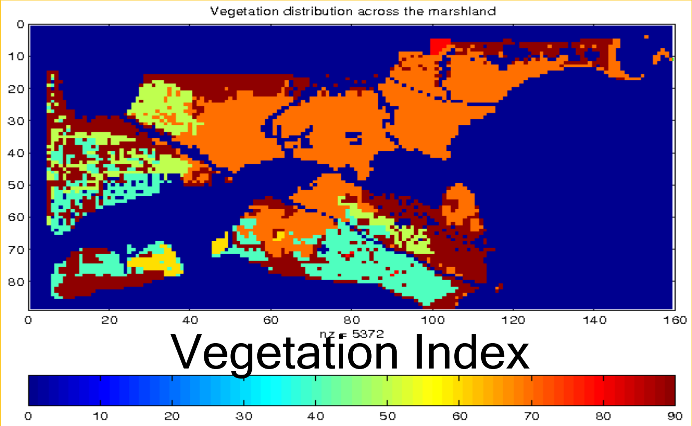


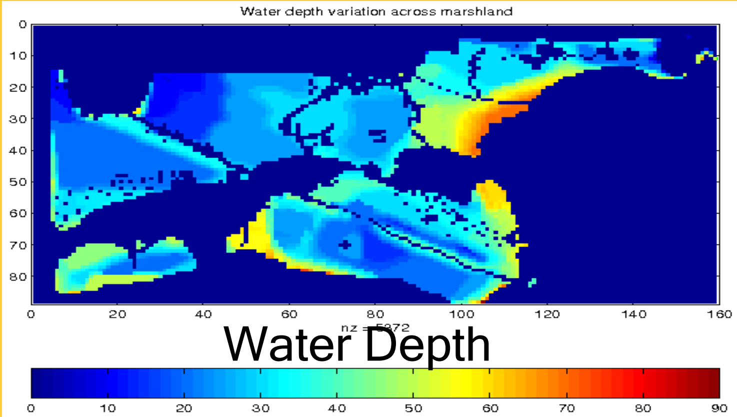
Outline

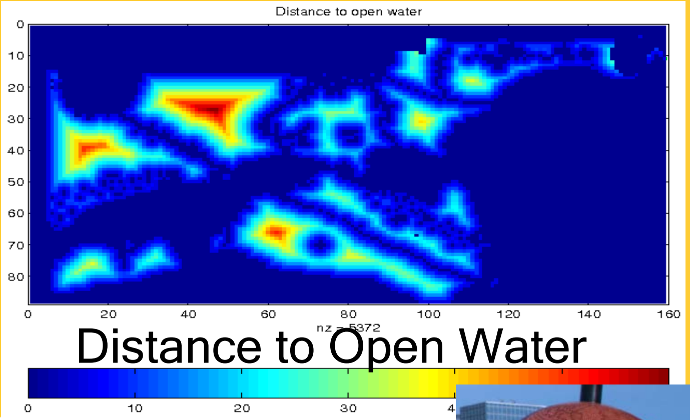
- Motivation
- Spatial Data Types
- Spatial Statistical Foundations
- Spatial Data Mining
  - Location Prediction
  - Hotspots
  - Spatial Outliers
  - Colocations
- Conclusions

Illustration of Location Prediction Problem





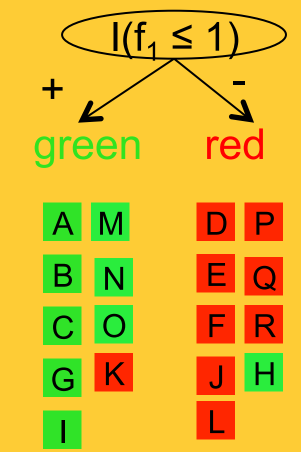




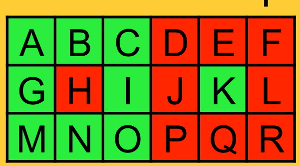
Decision Tree vs. Spatial Decision Tree

Inputs: table of records
Output: Decision Tree

ID	f_1	f_2	class
A	1	1	green
B	1	1	green
C	1	3	green
G	1	1	green
I	1	3	green
K	1	2	red
M	1	1	green
N	1	1	green
O	1	3	green
D	3	2	red
E	3	2	red
F	3	2	red
H	3	1	green
J	3	2	red
L	3	2	red
P	3	2	red
Q	3	2	red
R	3	2	red

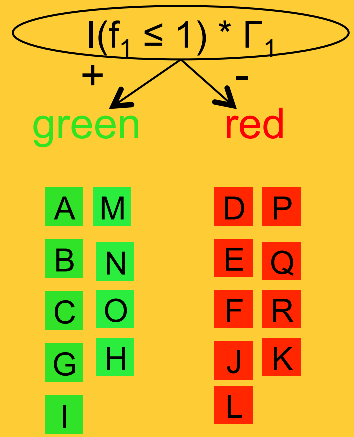


Predicted map



feature test	information gain
$f_1 \le 1$	0.50
$f_2 \le 1$	0.46
$f_2 \le 2$	0.19

Inputs: feature n class maps, (rook) neighborhood
Output: Spatial Decision Tree



Predicted map

Focal function Γ_1

1	.3	.3	.3	.3	1
.3	-1	0	0	-1	.3
1	.3	.3	.3	.3	1

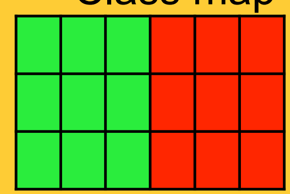
Feature f_1

1	1	1	3	3	3
1	3	1	3	1	3
1	1	1	3	3	3

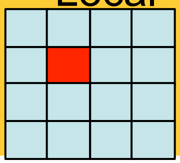
Feature f_2

1	1	3	2	2	2
1	1	3	2	2	2
1	1	3	2	2	2

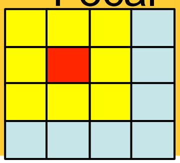
Class map



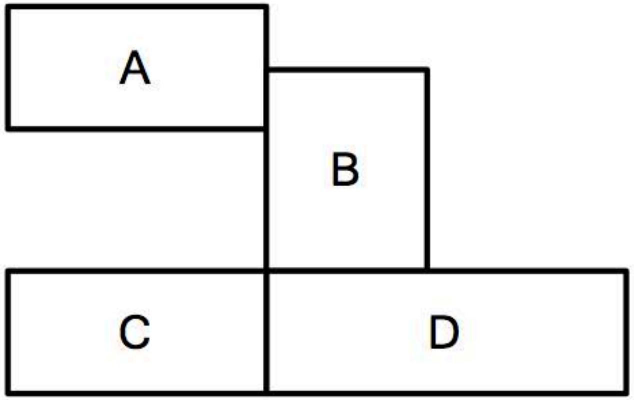
Local



Focal



Neighbor Relationship: W Matrix



(a) Map

$$
\begin{array}{c|cccc|}
 & A & B & C & D \\
\hline
A & 0 & 1 & 0 & 0 \\
B & 1 & 0 & 1 & 1 \\
C & 0 & 1 & 0 & 1 \\
D & 0 & 1 & 1 & 0 \\
\hline
\end{array}
$$

(b) Boolean W

$$
\begin{array}{c|cccc|}
 & A & B & C & D \\
\hline
A & 0 & 1 & 0 & 0 \\
B & 0.3 & 0 & 0.3 & 0.3 \\
C & 0 & 0.5 & 0 & 0.5 \\
D & 0 & 0.5 & 0.5 & 0 \\
\hline
\end{array}
$$

(c) Row-normalized W

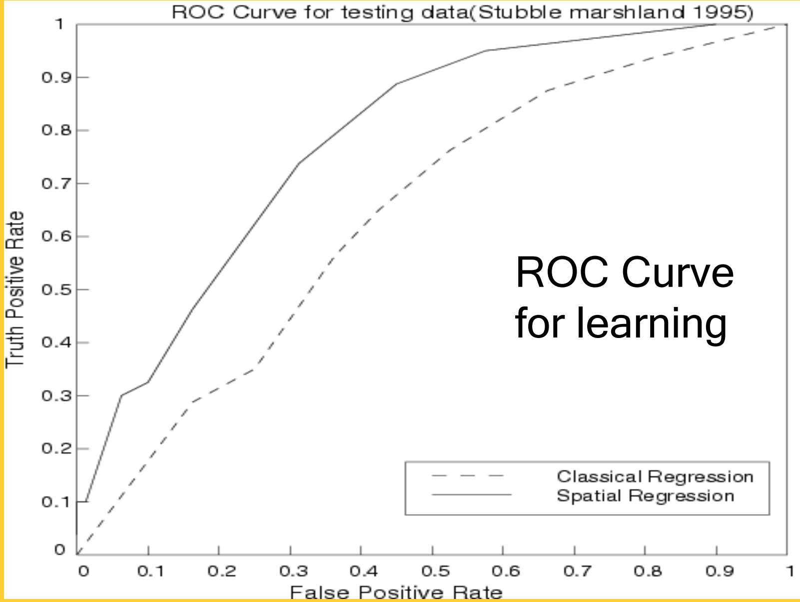
Location Prediction Models

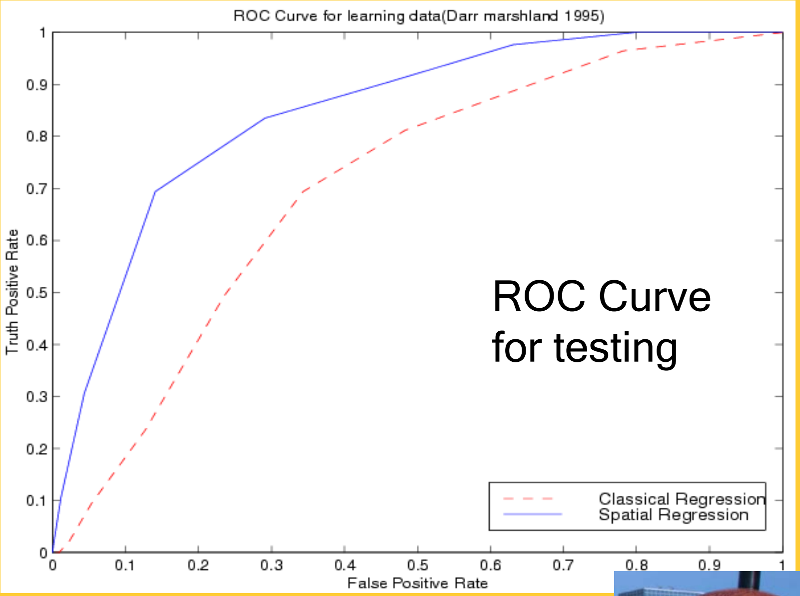
- Traditional Models, e.g., Regression (with Logit or Probit),
  - Bayes Classifier, …
- Spatial Models
  - Spatial autoregressive model (SAR)
  - Markov random field (MRF) based Bayesian Classifier

Classical	Spatial
$y = X\beta + \varepsilon$	$y = \rho W y + X\beta + \varepsilon$
$\Pr(C_i \mid X) = \dfrac{\Pr(X \mid C_i)\Pr(C_i)}{\Pr(X)}$	$\Pr(c_i \mid X, C_N) = \dfrac{\Pr(C_i)\Pr(X, C_N \mid c_i)}{\Pr(X, C_N)}$

Comparing Traditional and Spatial Models

- Dataset: Bird Nest prediction
- Linear Regression
  - Lower prediction accuracy, coefficient of determination,
  - Residual error with spatial auto-correlation
- Spatial Auto-regression outperformed linear regression

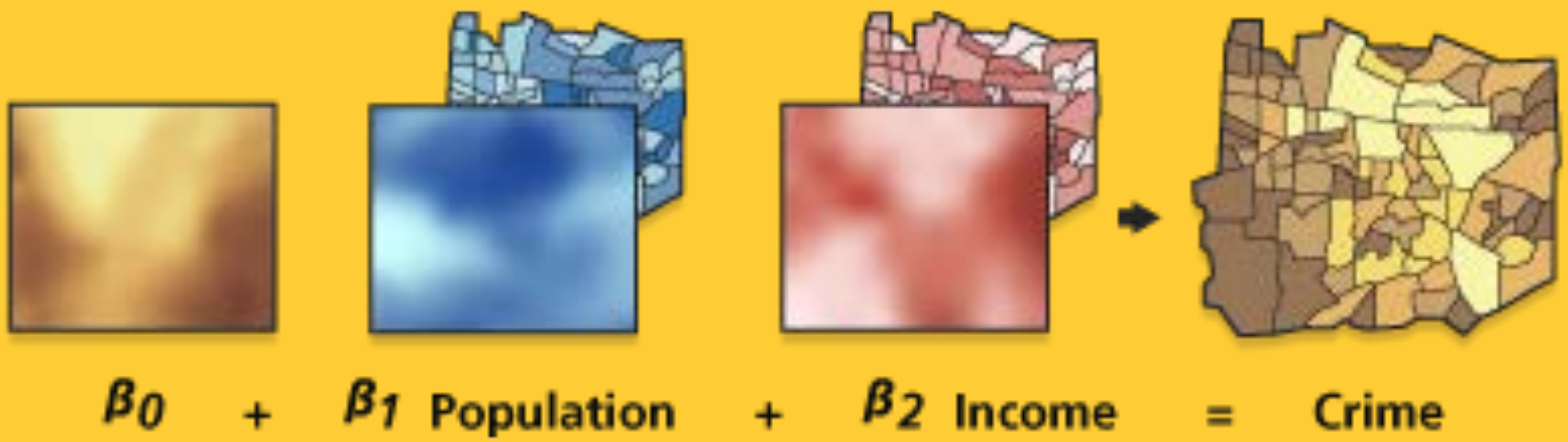




Modeling Spatial Heterogeneity: GWR

- Geographically Weighted Regression (GWR)
  - Goal: Model spatially varying relationships
  - Example: $y = X\beta^{'} + \varepsilon^{'}$

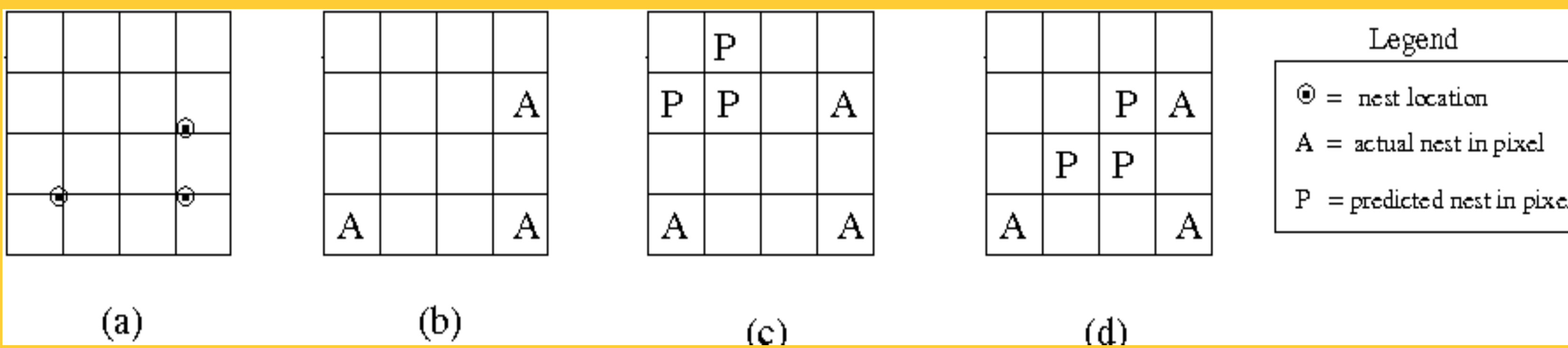
    Where $\beta^{'}$ and $\varepsilon^{'}$ are location dependent



Source: resources.arcgis.com

Research Needs for Location Prediction

- Spatial Auto-Regression
  - Estimate W
  - Scaling issue $\rho \mathrm{W} y$ vs. $\mathrm{X}\beta$
- Spatial interest measure
  - e.g., distance(actual, predicted)



(a) Actual Sites

(b) Pixels with actual sites

(c) Prediction 1

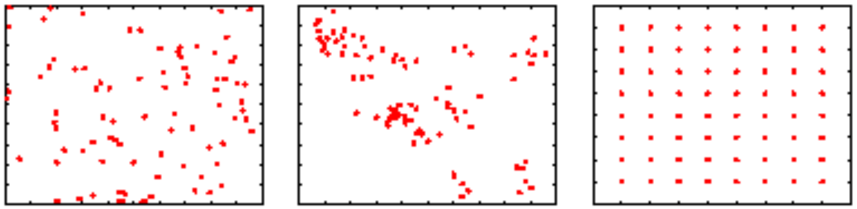
(d) Prediction 2. Spatially more interesting than Prediction 1

Outline

- Motivation
- Spatial Data Types
- Spatial Statistical Foundations
- Spatial Data Mining
  - Location Prediction
  - Hotspots
  - Spatial Outliers
  - Colocations
- Conclusions

Limitations of K-Means

- K-Means does test Statistical Significance
  - Finds chance clusters in complete spatial randomness (CSR)



Classical Clustering

Spatial Clustering

SaTScan™

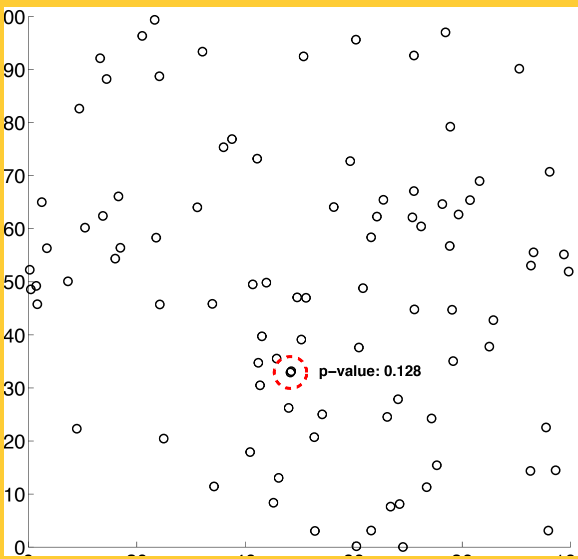
Software for the spatial, temporal, and space-time scan statistics

Spatial Scan Statistics (SatScan)

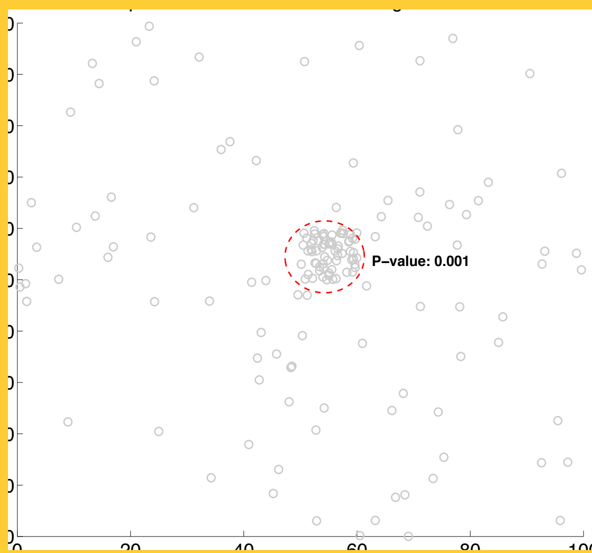
- Goal: Omit chance clusters
- Ideas: Likelihood Ratio, Statistical Significance
- Steps
  - Enumerate candidate zones & choose zone X with highest likelihood ratio (LR)
    - LR(X) = p(H1|data) / p(H0|data)
    - H0: points in zone X show complete spatial randomness (CSR)
    - H1: points in zone X are clustered
  - If LR(Z) >> 1 then test statistical significance
    - Check how often is LR(CSR) > LR(Z) using 1000 Monte Carlo simulations

SatScan Examples

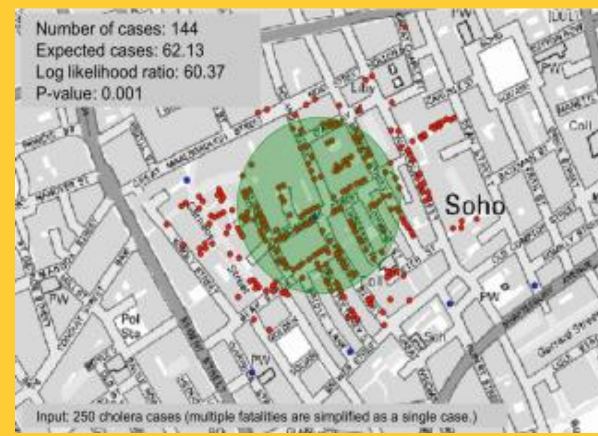
Complete Spatial Randomness
Output: No hotspots !
Highest LR circle p-value = 0.128



Data with a hotspot
Output: A hotspot!
p-value = 0.001



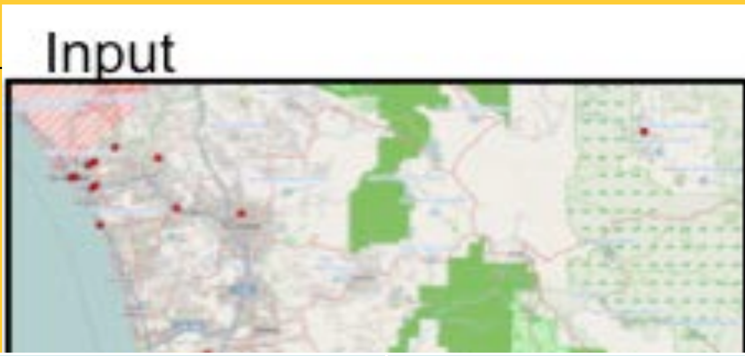
1854 London Cholera
Output: A hotspot!
p-value = 0.001

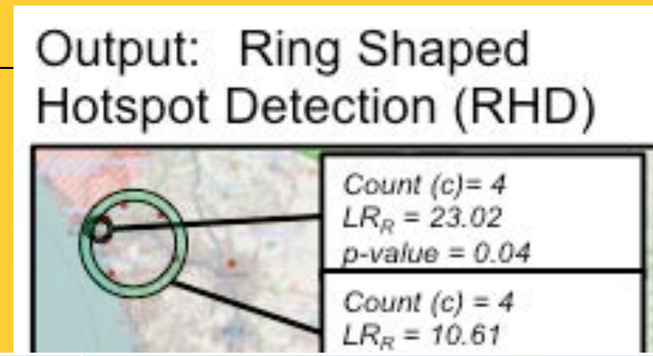


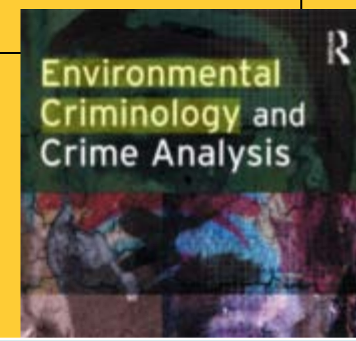
Complex Hotspots

Semantic Gap between Spatial and Machine Learning

- Environmental Criminology
  - Routine Activities Theory, Crime Pattern Theory, Doughnut Hole pattern
- Formulation: rings, where inside density is significantly higher than outside …





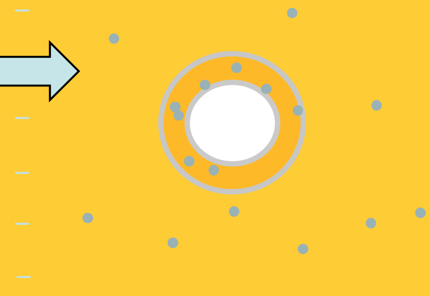


Mathematics	Concepts	Relationships
Sets	Set Theory	Member, set-union, set-difference, …
Vector Space	Linear Algebra	Matrix & vector operations
Euclidean Spaces	Geometry	Circle, Ring, Polygon, Line_String, Convex hull, …
Boundaries, Graphs, Spatial Graphs	Topology, Graph Theory, Spatial graphs, …	Interior, boundary, Neighbor, inside, surrounds, …, Nodes, edges, paths, trees, … Path with turns, dynamic segmentation, …

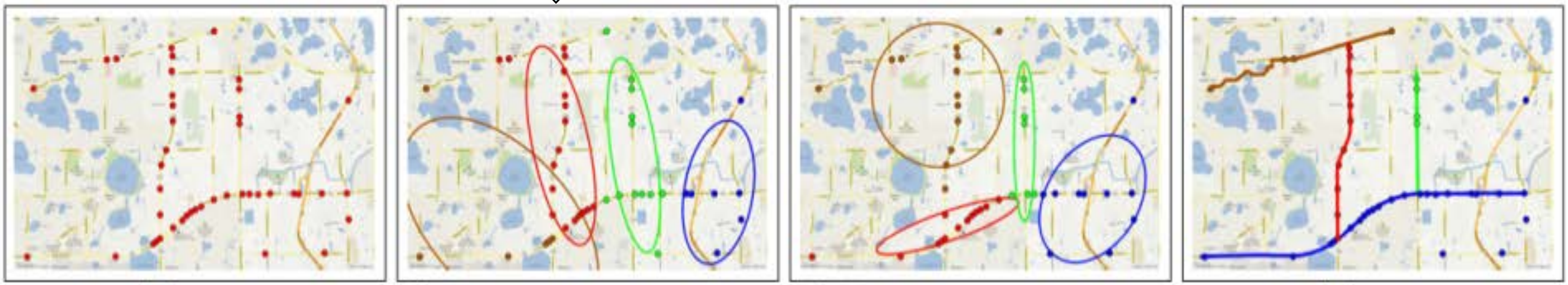
Source: Ring-Shaped Hotspot Detection: A Summary of Results, IEEE ICDM 2014 (w/ E. Eftelioglu et al.)

Spatial-Concept/Theory-Aware Clusters

- Spatial Theories, e.g,, environmental criminology
  - Circles ➔ Doughnut holes



- Geographic features, e.g., rivers, streams, roads, …
  - Hot-spots => Hot Geographic-features



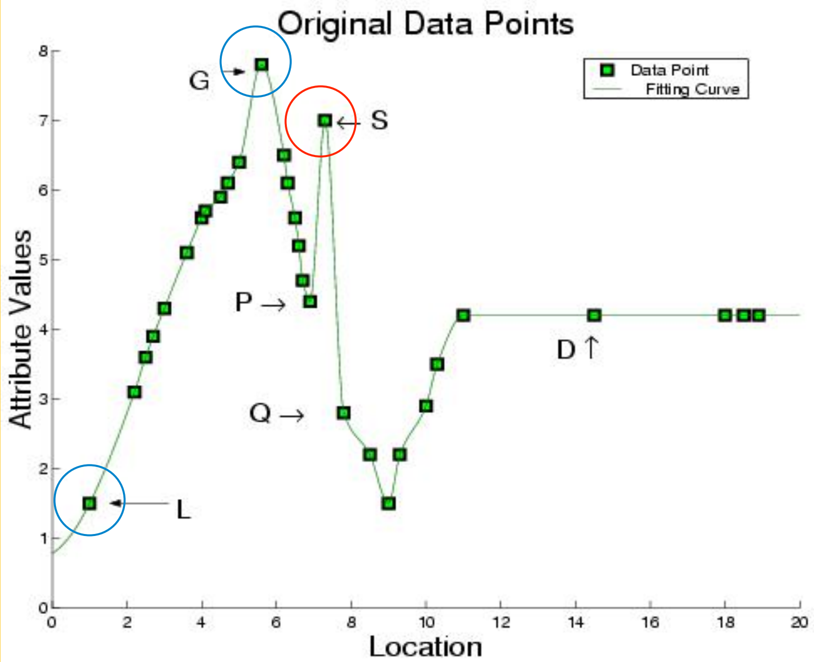
(a) Input (b) Crimestat K-means with Euclidean Distance (c) Crimestat K-means with Network Distance (d) KMR

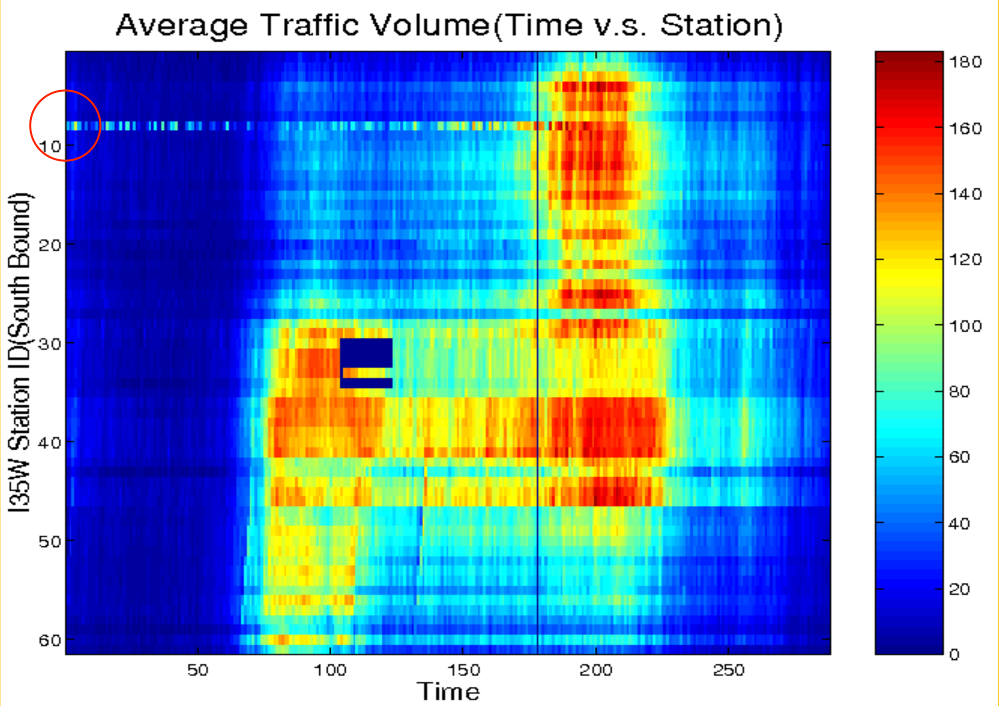
Source: A K-Main Routes Approach to Spatial Network Activity Summarization, IEEE Transactions on Knowledge and Data Eng., 26(6), 2014.)

Outline

- Motivation
- Spatial Data Types
- Spatial Statistical Foundations
- Spatial Data Mining
  - Location Prediction
  - Hotspots
  - Spatial Outliers
  - Colocations
- Conclusions

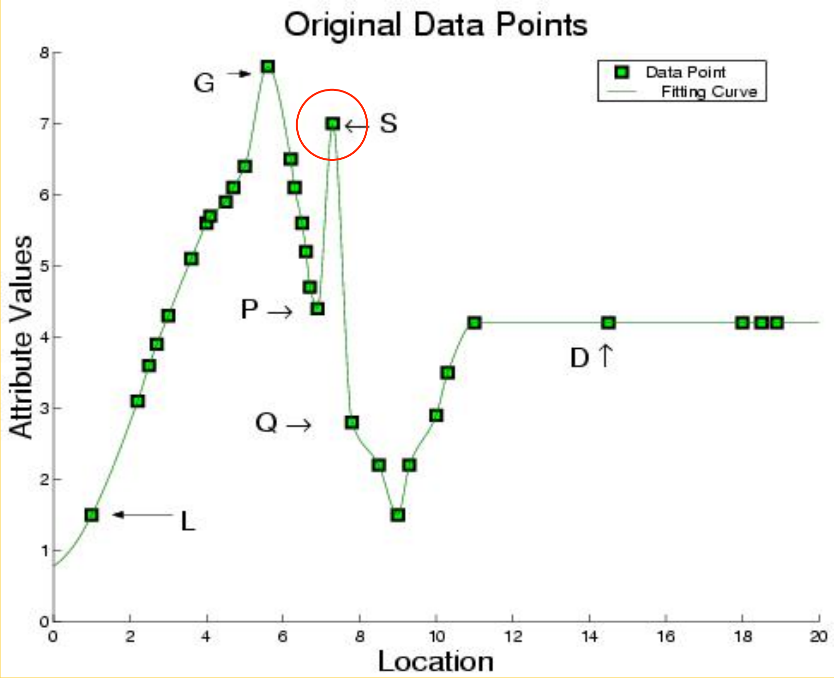
Outliers: Global (G) vs. Spatial (S)

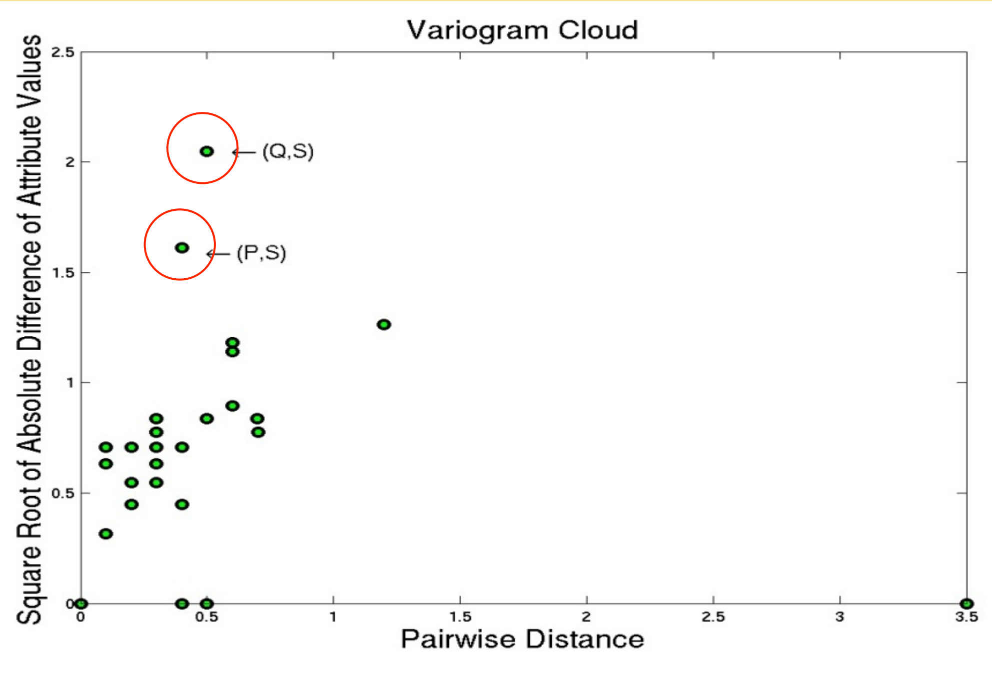




Outlier Detection Tests: Variogram Cloud

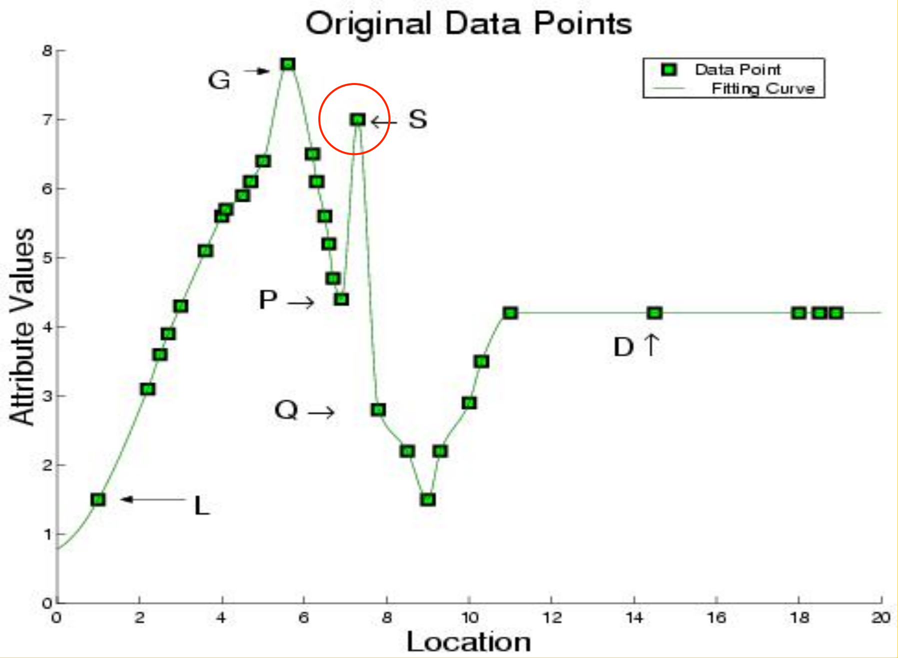
- Graphical Test: Variogram Cloud

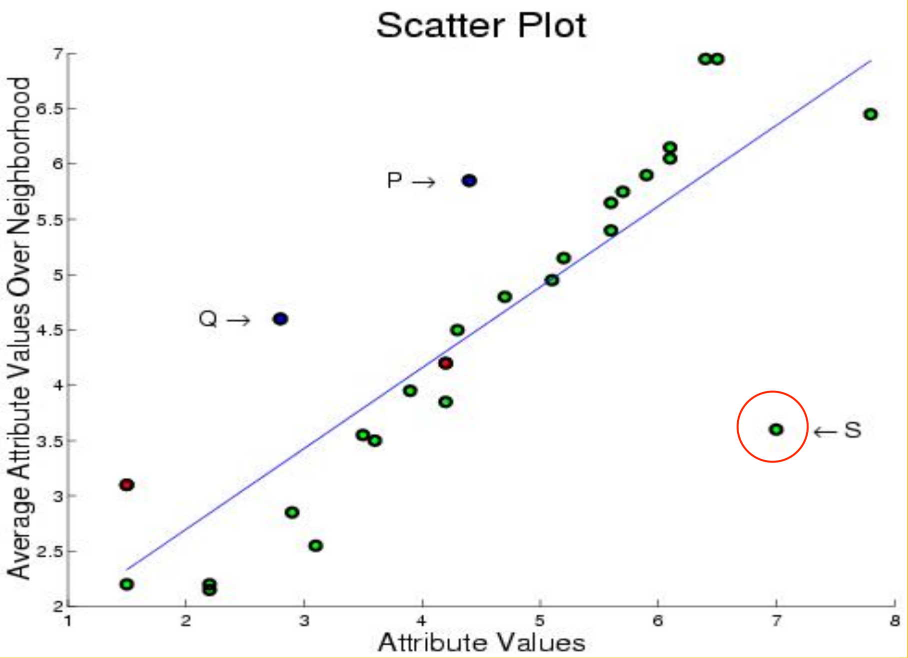




Outlier Detection - Scatterplot

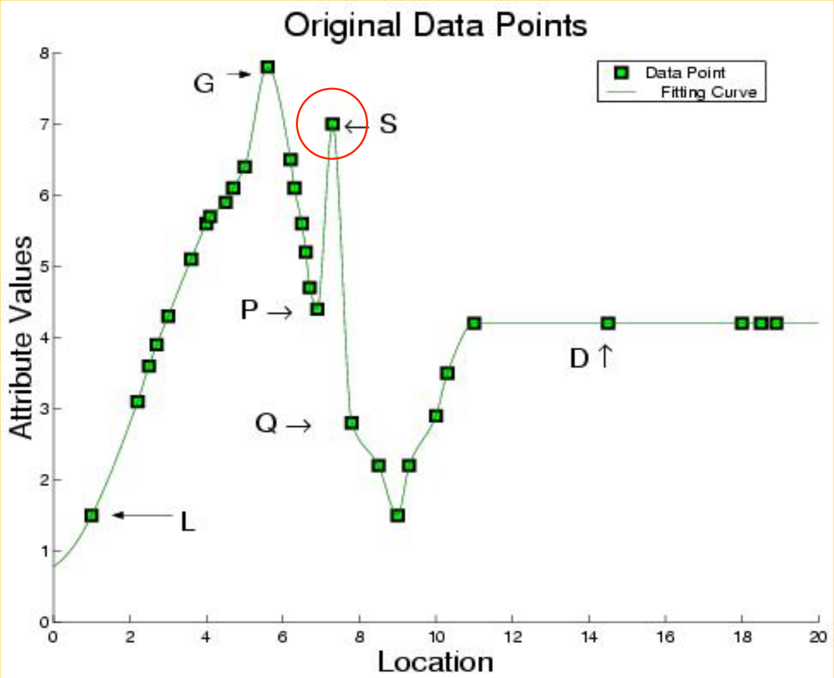
- Quantitative Tests: Scatter Plot

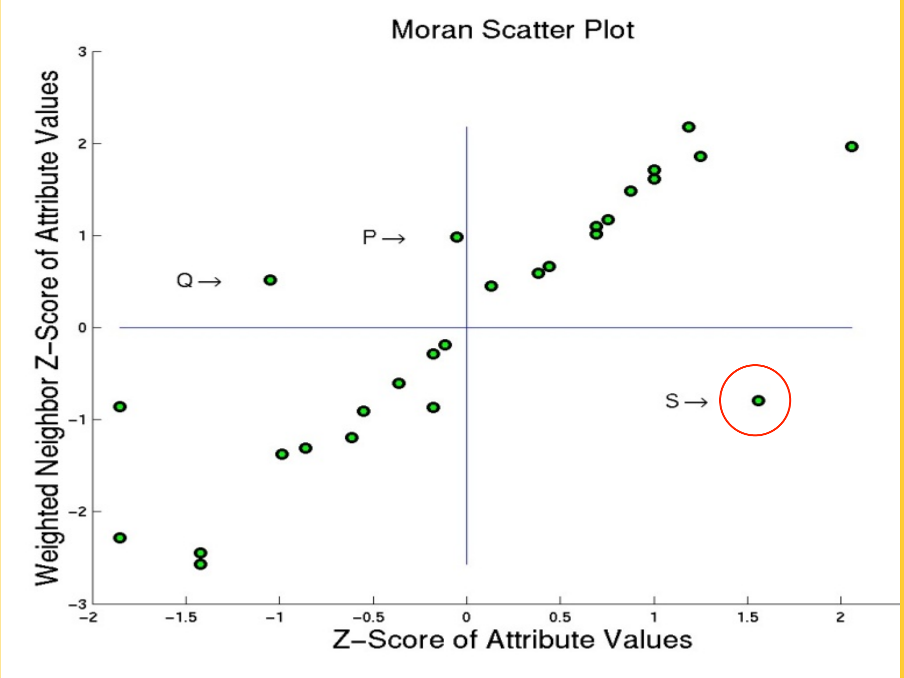




Outlier Detection Test: Moran Scatterplot

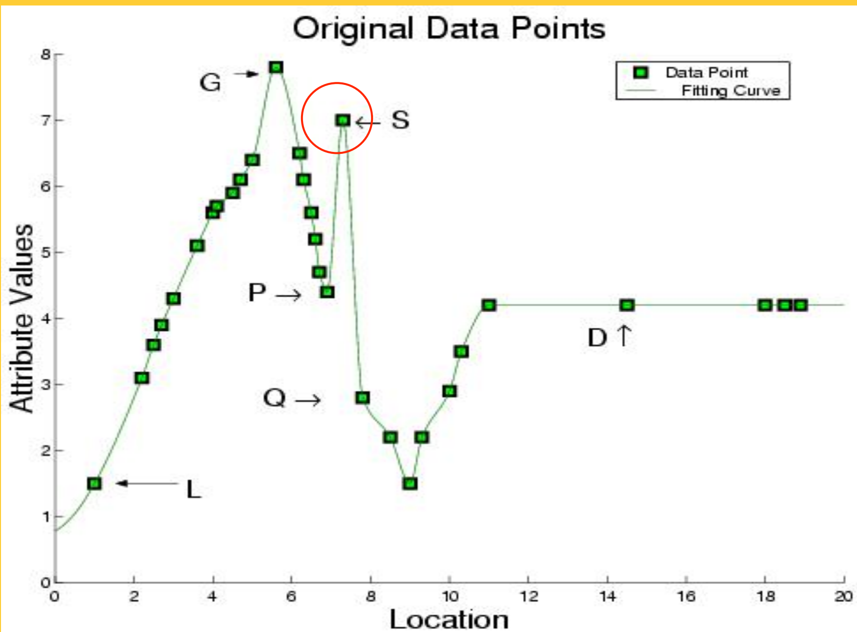
- Graphical Test: Moran Scatter Plot

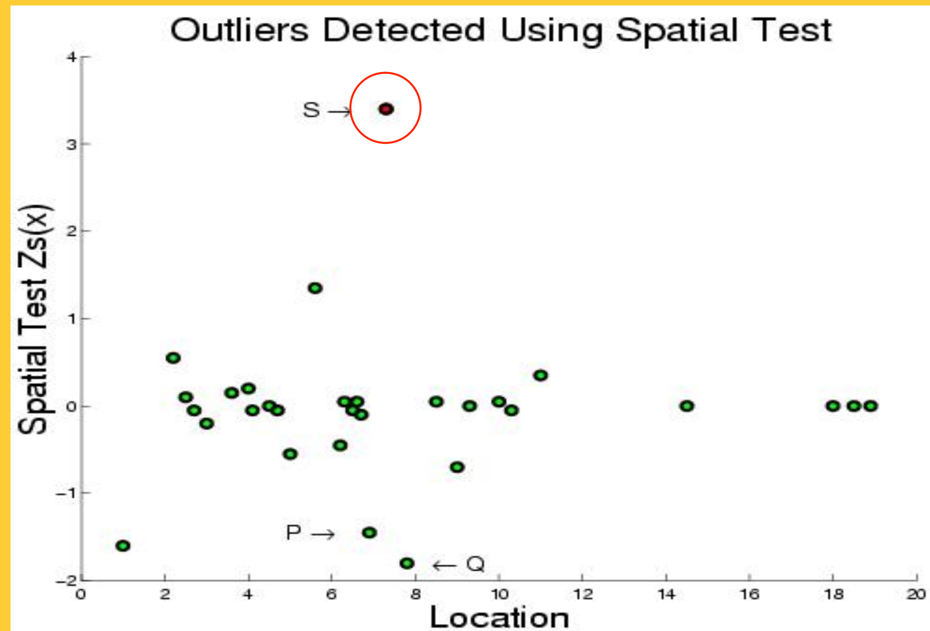




Outlier Detection Tests: Spatial Z-test

- Quantitative Tests: Spatial Z-test
  - Algorithmic Structure: Spatial Join on neighbor relation





Spatial Outlier Detection: Computation

- Separate two phases
  - Model Building
  - Testing: test a node (or a set of nodes)

- Computation Structure of Model Building
  - Key insights:
    - Spatial self join using N(x) relationship
    - Algebraic aggregate function computed in one scan of spatial join

Trends in Spatial Outlier Detection

- Multiple spatial outlier detection
  - Eliminating the influence of neighboring outliers

- Multi-attribute spatial outlier detection
  - Use multiple attributes as features

- Scale up for large data

Outline

- Motivation
- Spatial Data Types
- Spatial Statistical Foundations
- Spatial Data Mining
  - Location Prediction
  - Hotspots
  - Spatial Outliers
  - Colocations
- Conclusions

Learning Objectives

- After this segment, students will be able to
  - Contrast colocations and associations
  - Describe colocation interest measures

Background: Association Rules

- Association rule e.g. (Diaper in T => Beer in T)

Transaction	Items Bought
1	{socks, , milk, , beef, egg, …}
2	{pillow, , toothbrush, ice-cream, muffin, …}
3	{ , , pacifier, formula, blanket, …}
…	…
n	{battery, juice, beef, egg, chicken, …}

  - Support: probability (Diaper and Beer in T) = 2/5
  - Confidence: probability (Beer in T \| Diaper in T) = 2/2

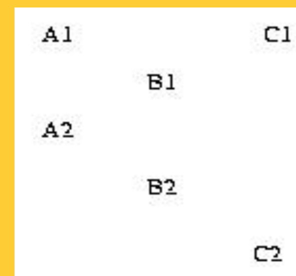
- Apriori Algorithm
  - Support based pruning using monotonicity

Association Rules Limitations

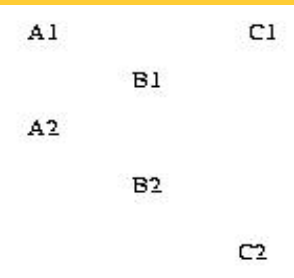
- **Transaction is a core concept!**
  - Support is defined using transactions
  - Apriori algorithm uses transaction based Support for pruning

Transaction	Items Bought
1	{socks, , milk, , beef, egg, …}
2	{pillow, , toothbrush, ice-cream, muffin, …}
3	{ , , pacifier, formula, blanket, …}
…	…

- However, spatial data is embedded in continuous space
  - Transactionizing continuous space is non-trivial !



Spatial Association Rule vs. Colocation



Input = Spatial feature A,B, C, & their instances

- Spatial Association Rule (Han 95)
- Output = (B,C)
- Transactions by Reference feature C

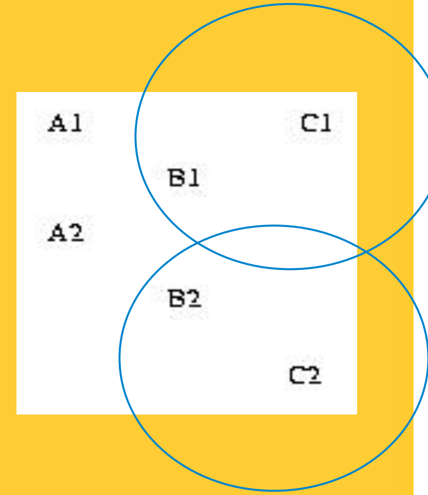
  Transactions: (C1, B1), (C2, B2)
  Support (A,B) = Ǿ, Support(B,C)=2 / 2 = 1

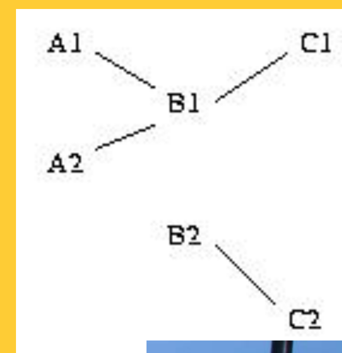
- Cross-K Function

  Cross-K (A, B) = 2/4 = 0.5
  Cross-K(B, C) = 2/4 = 0.5
  Output = (A,B), (B, C)

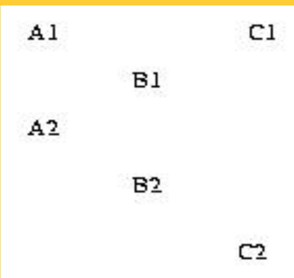
- Colocation - Neighborhood graph

  Output = (A,B), (B, C)
  PI(A,B) = min(2/2,1/2) = 0.5
  PI(B,C) = min(2/2,2/2) = 1





Spatial Association vs. Cross-K Function



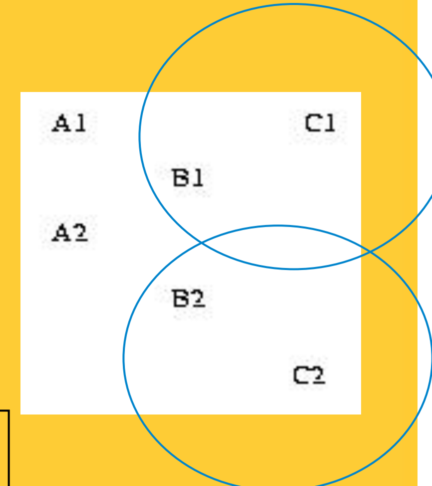
Input = Feature A,B, and, C, & instances A1, A2, B1, B2, C1, C2

- Spatial Association Rule (Han 95)
- Output = (B,C) with threshold 0.5
- Transactions by Reference feature, e.g. C

  Transactions: (C1, B1), (C2, B2)

  Support (A,B) = Ǿ

  Support(B,C)=2 / 2 = 1



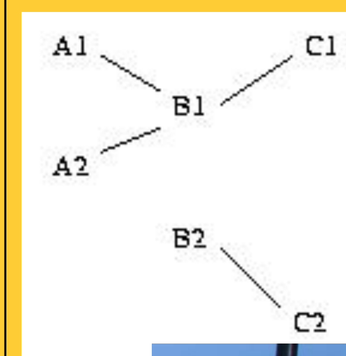
- Cross-K Function

  Cross-K (A, B) = 2/4 = 0.5

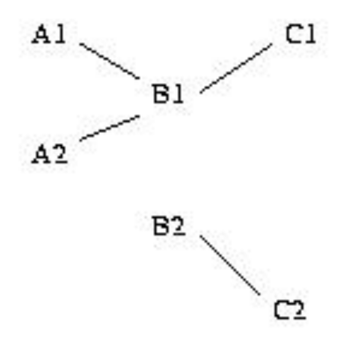
  Cross-K(B, C) = 2/4 = 0.5

  Cross-K(A, C) = 0

  Output = (A,B), (B, C) with threshold 0.5



Spatial Colocation



Features: A. B. C

Feature Instances: A1, A2, B1, B2, C1, C2

Feature Subsets: (A,B), (A,C), (B,C), (A,B,C)

Participation ratio (pr):

pr(A, (A,B)) = fraction of A instances neighboring feature {B} = 2/2 = 1

pr(B, (A,B)) = ½ = 0.5

Participation index (A,B) = **pi**(A,B) = min{ **pr**(A, (A,B)), **pr**(B, (A,B)) } = min (1, ½) = 0.5

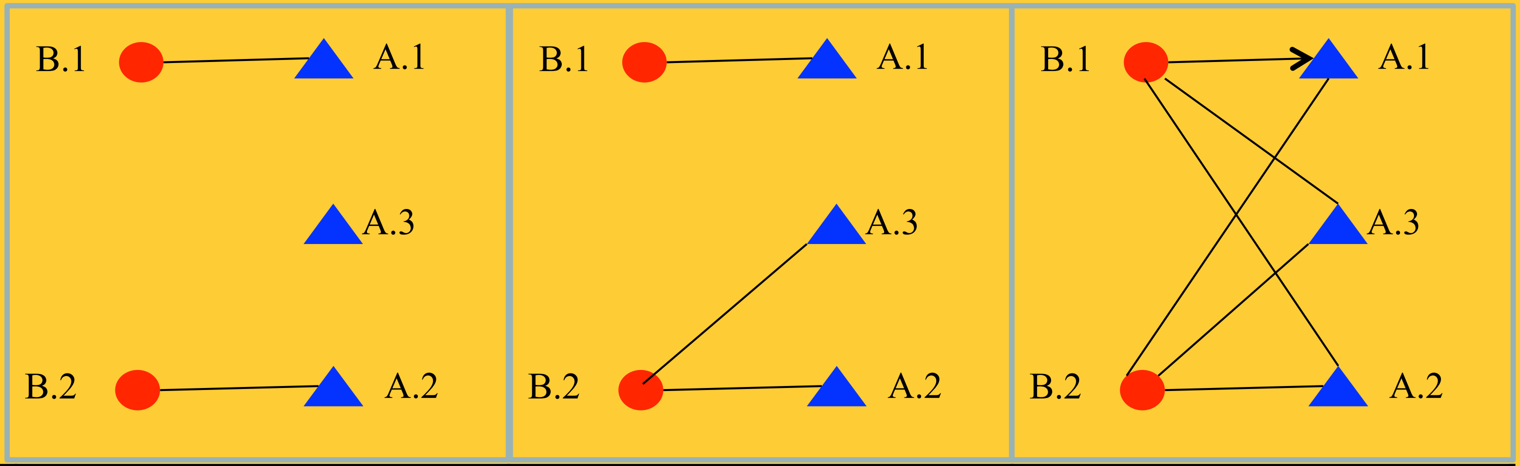
pi(B, C) = min{ **pr**(B, (B,C)), **pr**(C, (B,C)) } = min (1,1) = 1

Participation Index Properties:

(1) Computational: Non-monotonically decreasing like support measure

(2) Statistical: Upper bound on Ripley's Cross-K function

Participation Index >= Cross-K Function



Cross-K (A,B)			
PI (A,B)			

Association Vs. Colocation

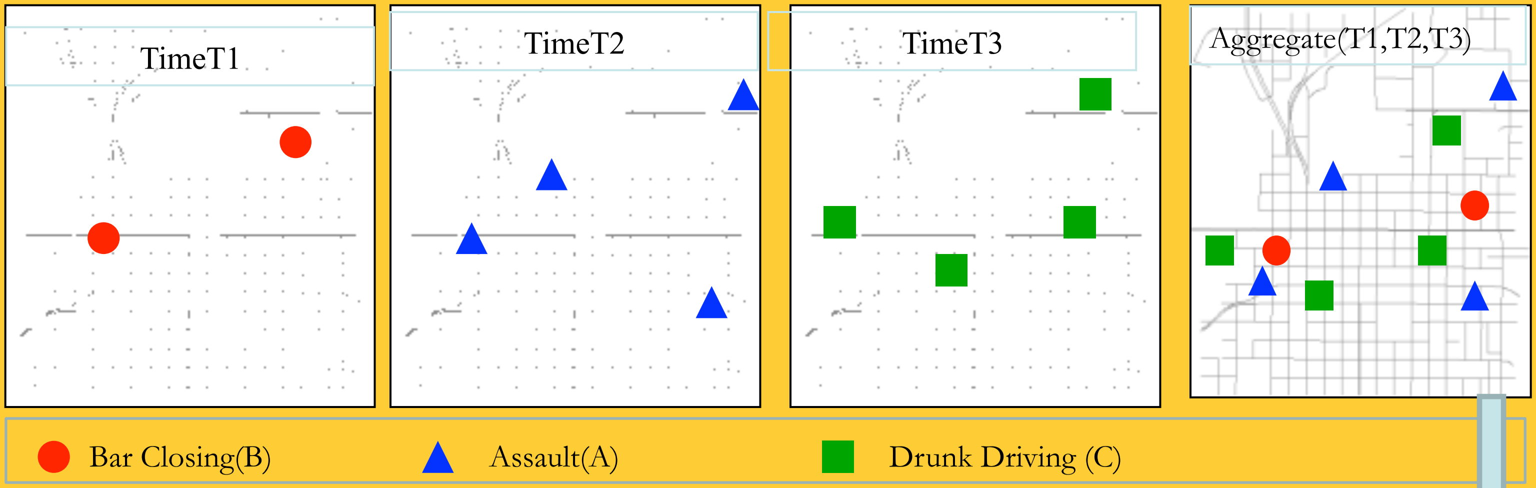
	Associations	Colocations
underlying space	Discrete market baskets	
event-types	item-types, e.g., Beer	
collections	Transaction (T)	
prevalence measure	Support, e.g., Pr.[Beer in T]	
conditional probability measure	Pr.[Beer in T | Diaper in T]	

Spatial Colocation: Trends

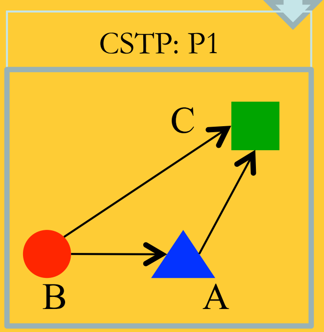
- Algorithms
  - Join-based algorithms
    - One spatial join per candidate colocation
  - Join-less algorithms

- Spatio-temporal
  - Which events co-occur in space and time?
    - (bar-closing, minor offenses, drunk-driving citations)
  - Which types of objects move together?

Cascading spatio-temporal pattern (CSTP)

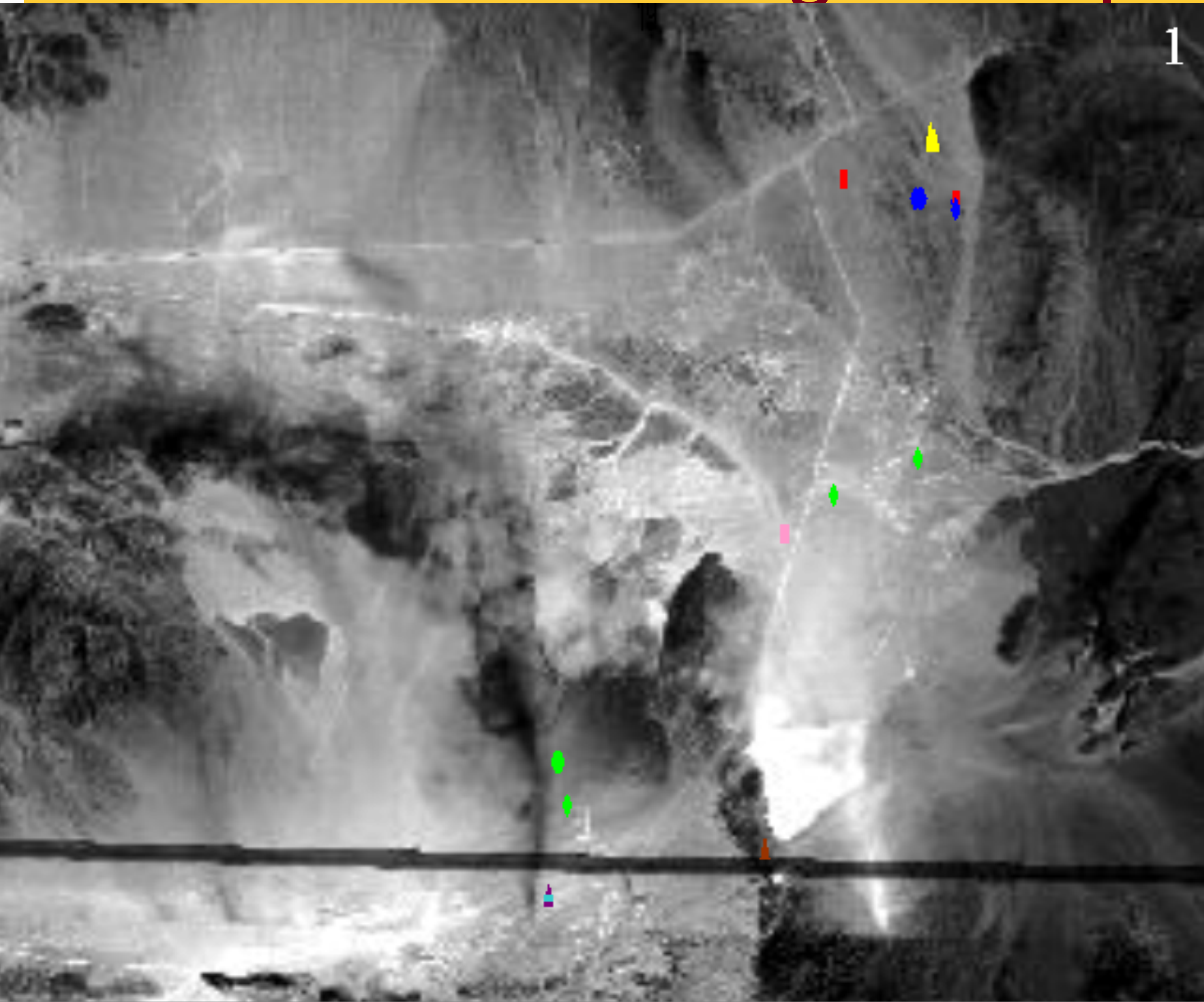


- *Input:* Urban Activity Reports
- *Output: CSTP*
  - *Partially ordered* subsets of ST event types.
  - Located together in space.
  - Occur in *stages* over time.
- Applications: Public Health, Public Safety, …



Details: Cascading Spatio-Temporal Pattern Discovery, IEEE Trans. on Know. & Data Eng, 24(11), 2012.

MDCOP Motivating Example :



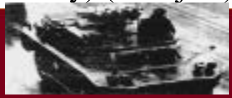
- Manpack stinger (2 Objects)

- M1A1_tank (3 Objects)

- M2_IFV (3 Objects)

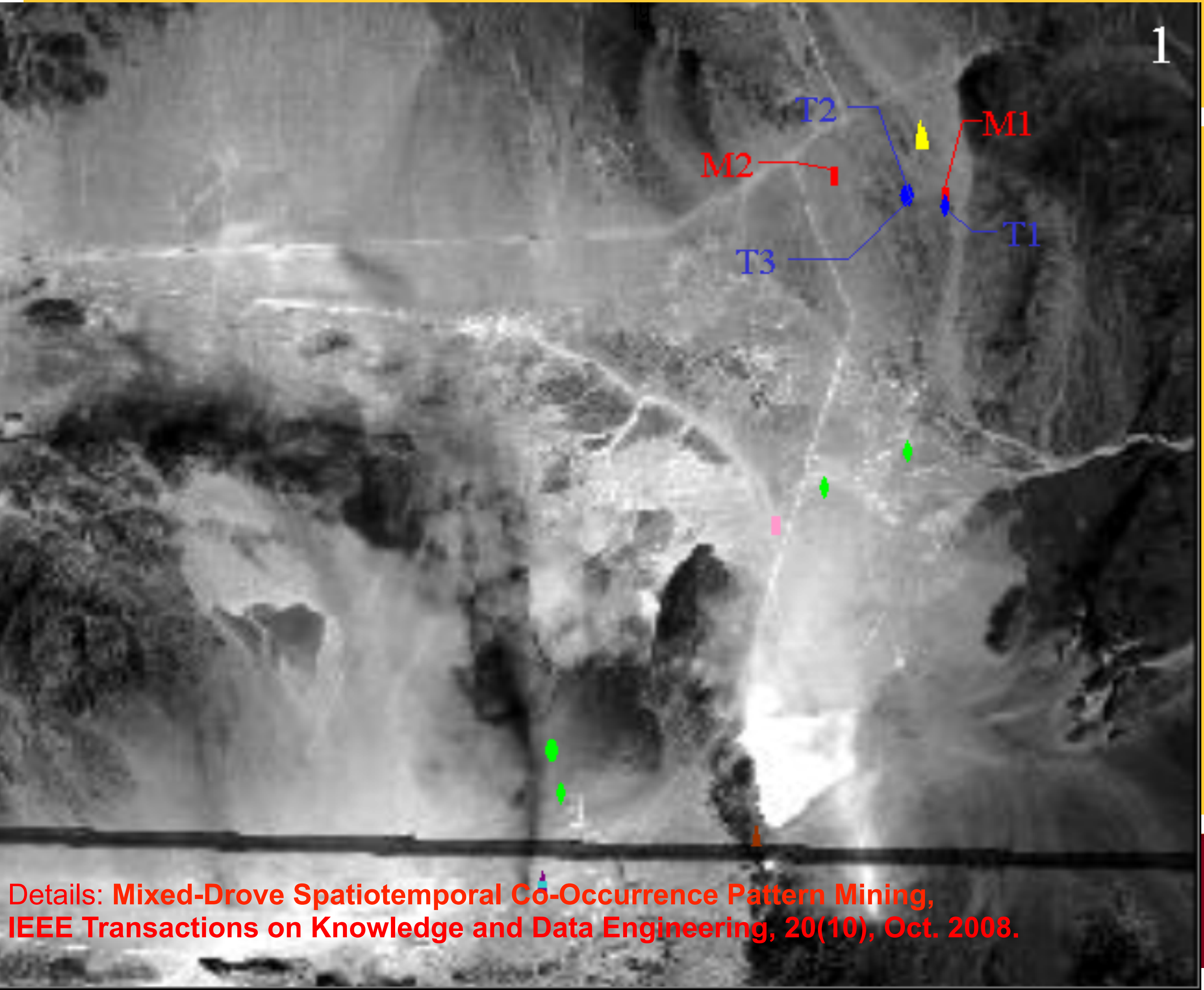
- Field_Marker (6 Objects)
- T80_tank (2 Objects)

- BRDM_AT5 (enemy) (1 Object)



- BMP1 (1 Object)

MDCOP Motivating Example : Output



Details: **Mixed-Drove Spatiotemporal Co-Occurrence Pattern Mining, IEEE Transactions on Knowledge and Data Engineering, 20(10), Oct. 2008.**

- Manpack stinger (2 Objects)

- M1A1_tank (3 Objects)

- M2_IFV (3 Objects)

- Field_Marker (6 Objects)
- T80_tank (2 Objects)

- BRDM_AT5 (enemy) (1 Object)

- BMP1 (1 Object)

Outline

- Motivation
  - Use cases
  - Pattern families
- Spatial Data Types
- Spatial Statistical Foundations
- Spatial Data Mining
- Conclusions

Summary

What's Special About Mining Spatial Data ?

		Spatial DM	
Input Data		Often implicit relationships, complex types	
Statistical Foundation			
Output	Association		
	Clusters		
	Outlier		
	Prediction		

Acknowledgements

National Science Foundation (Current Grants)

- 1320580 : III:Investigating Spatial Big Data for Next Generation Routing Services
- 0940818: Expedition: Understanding Climate Change: A Data Driven Approach
- IIS-1218168 : III:Towards Spatial Database Management Systems for Flash Memory Storage
- 1029711 :: Datanet: Terra Populus: A Global Population / Environment Data Network

USDOD (Current Grants)

- HM0210-13-1-0005: Identifying and Analyzing Patterns of Evasion
- SBIR Phase II: Spatio-Temporal Analysis in GIS Environments (STAGE) (with Architecture Technology Corporation)

University of Minnesota (Current Grants)

- Infrastructure Initiative: U-Spatial - Support for Spatial Research
- MOOC Initiative: From GPS and Google Earth to Spatial Computing

- Past Sponsors, e.g., NASA, ARL, AGC/TEC, Mn/DOT, …

References

Colocations

- Discovering colocation patterns from spatial data sets: a general approach, *IEEE Transactions on Knowledge and Data Eng*ineering, 16(12), 2004 (with Y. Huang et al.).
- A join-less approach for mining spatial colocation patterns, IEEE Transactions on Knowledge and Data Engineering,18 (10), 2006. (with J. Yoo).

Spatial Outliers

- Detecting graph-based spatial outliers: algorithms and applications (a summary of results), Proc.: ACM International Conference on Knoweldge Discovery & Data Mining, 2001 (with Q. Lu et al.)
- A unified approach to detecting spatial outliers, Springer GeoInformatica, 7 (2),), 2003. (w/ C. T. Lu, et al.)

Hot-Spots

- Discovering personally meaningful places: An interactive clustering approach, ACM Transactions on Information Systems (TOIS) 25 (3), 2007. (with C. Zhou et al.)
- A K-Main Routes Approach to Spatial Network Activity Summarization, IEEE Transactions on Knowledge & Data Engineering, 26(6), 2014. (with D. Oliver et al.)

Location Prediction

- Spatial contextual classification and prediction models for mining geospatial data, IEEE Transactions on Multimedia, 4 (2), 2002. (with P. Schrater et al.)
- Focal-Test-Based Spatial Decision Tree Learning, to appear in IEEE Transactions on Knowledge and Data Eng. (a summary in Proc. IEEE Intl. Conference on Data Mining, 2013).

Change Detection

- Spatiotemporal change footprint pattern discovery: an inter-disciplinary survey. Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery 4(1), 2014. (with X. Zhou et al.)