Spatial & Spatio-temporal Data Mining Challenges

By

Shashi Shekhar, University of Minnesota

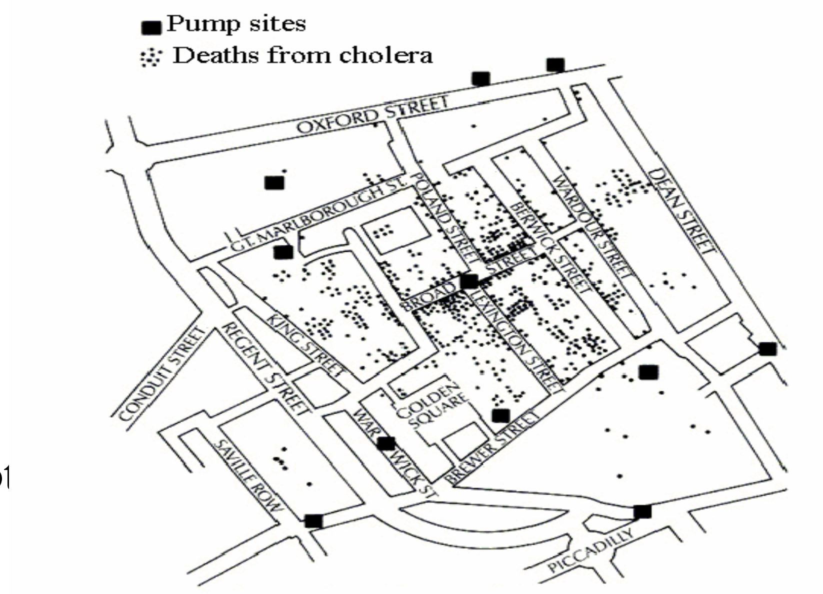
Bhavani Thuraisingham, Latifur Khan
U of.Texas, Dallas

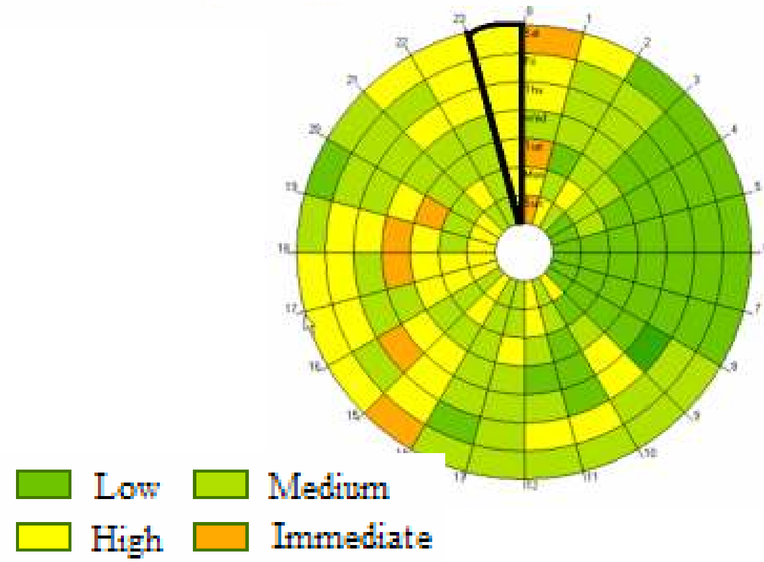
NSF Symposium on NGDM and CDI
Session of Security, Surveillance and Privacy

October 11th, 2007

Motivation

- Security: Geo-spatial Intelligence
- Surveillance:
  - Public Safety: Crime mapping & analysis
  - Public Health: (Emerging) Disease hotspo
- Privacy
  - Spatial location vs. HIPPA
  - Containing spread of infectious disease





Rings = weekdays; Slices = hour
(Source: US Army ERDC, TEC)

http://www.dublincrime.com/blog/wp-content/MappingOurMeanStreets.jpg

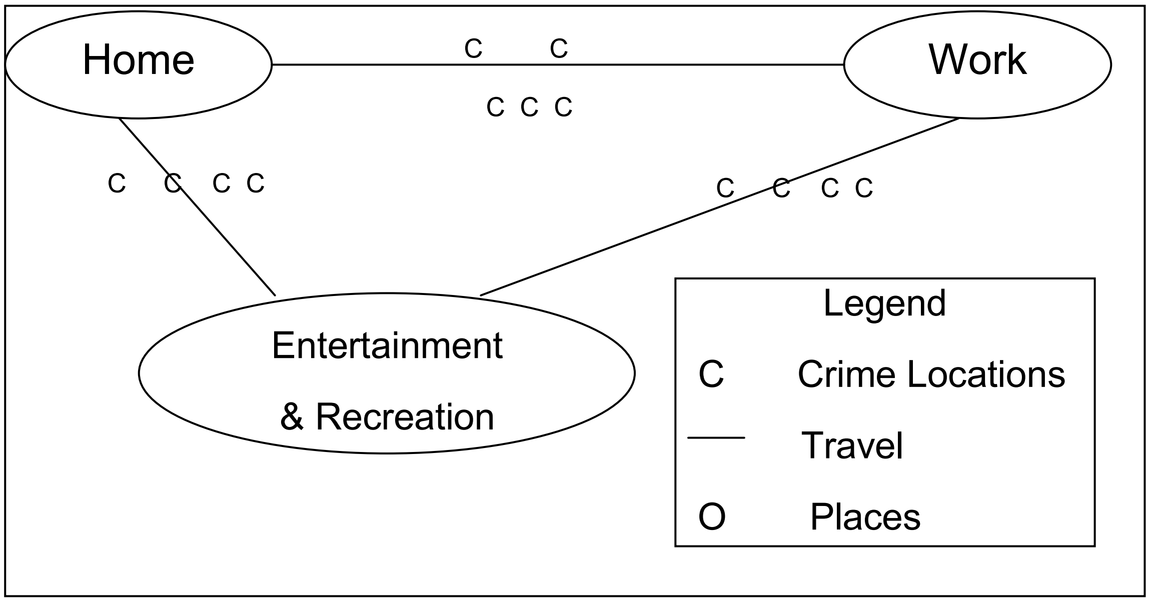
www.sentient.nl/crimeanabody.html http://www.esri.com/news/arcuser/0405/ss_crimestats2of2.html

Objectives, State of the Art

- Objectives:
  - to accurately track, monitor, and predict human activities

- State of the Art
  - Environmental Criminology
    - Routine Activity Theory (RAT), Crime Pattern Theory (CPT)
  - Spatial Data Analysis
    - Statistical, e.g. Knox test, Spatial Data Mining

Fig. 1: Crime Pattern Theory

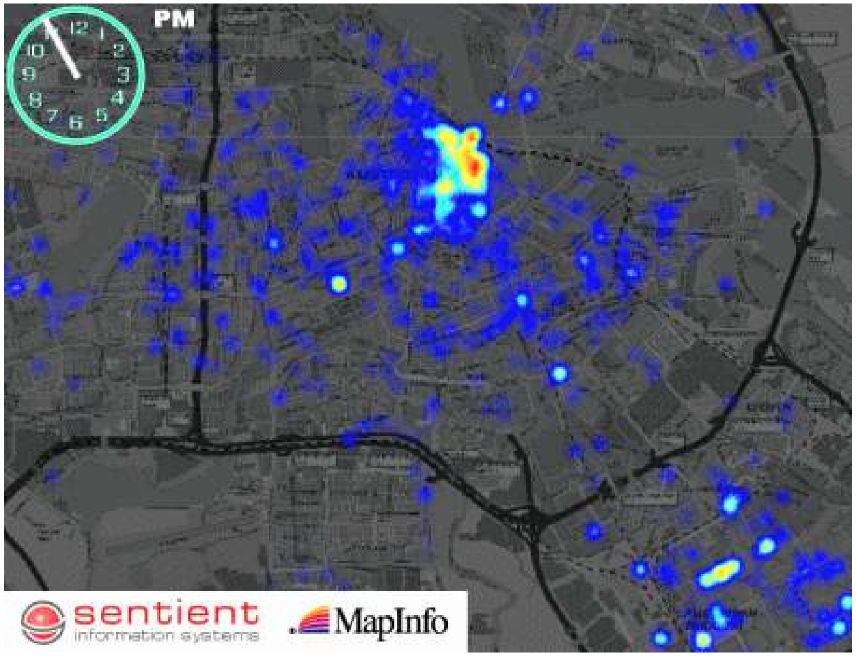


Limitations of State of the Art

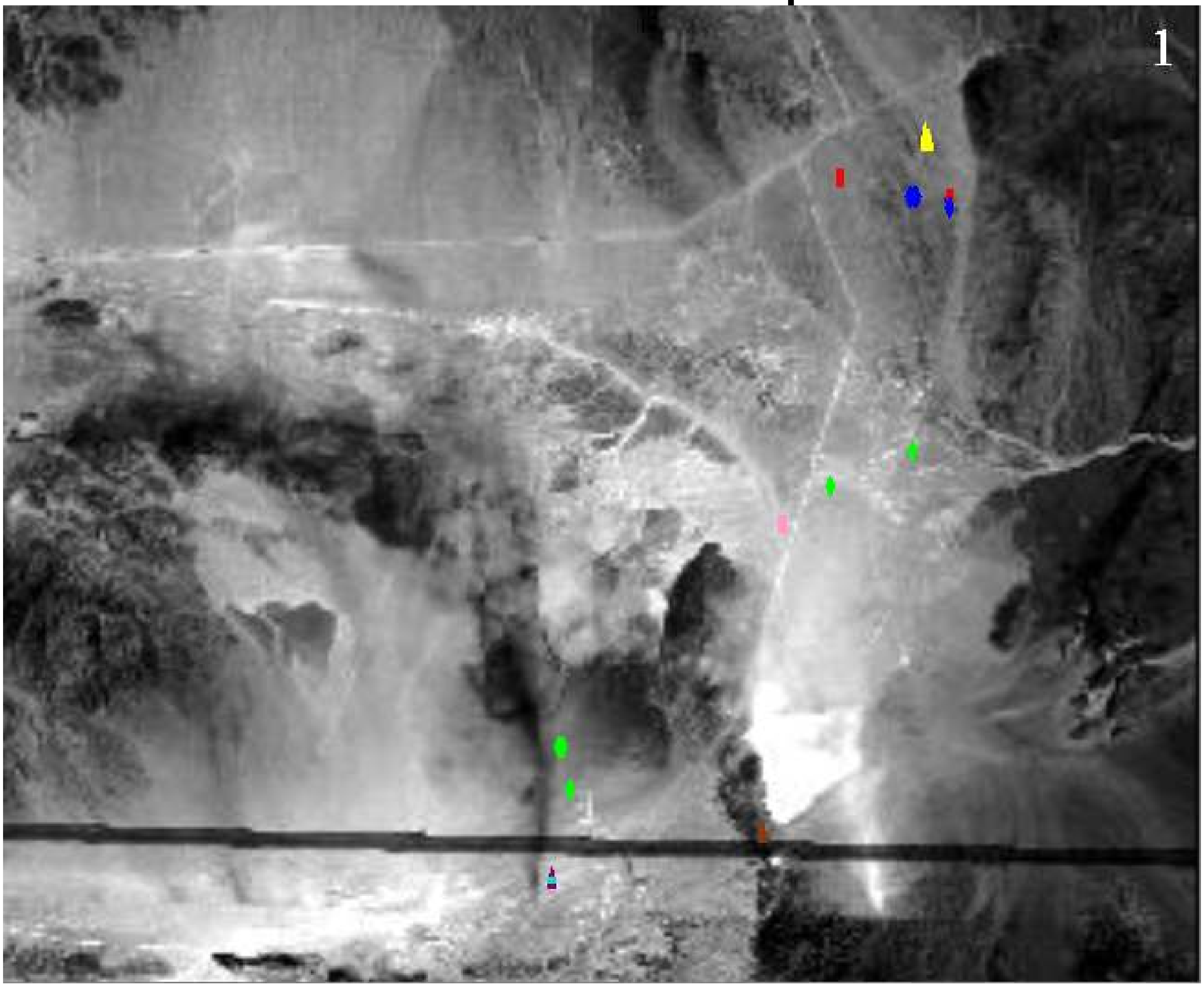
- do not adequately model richer temporal semantics
  - beyond space-time interaction (Knox test)
- do not satisfactorily explain the cause of detected hot spot locations on spatial networks,
  - such as roads, trains, …
- do not effectively model heterogeneities
  - across spatial networks
  - e.g. multi-modal urban transportation modes (such as light-rail subways and roads).

1: Spatio-Temporal (ST) Nature of Patterns

- State of the Art: Environmental Criminology
  - Spatial Methods: Hotspots, Spatial Regression
  - Space-time interaction (Knox test)
- Critical Barriers: richer ST semantics
  - Ex. Trends, periodicity, displacement
- Issues:
  - 1: Categorize pattern families
  - 2 : Quantify: interest measures
  - 3: Design scalable algorithms
  - 4: Evaluate with crime datasets
  - 5: Generalize beyond crimes
- Challenges: Trade-off b/w
  - Semantic richness and
  - Scalable algorithms

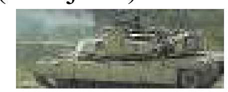


Co-occurrence in space and time!



Manpack stinger
(2 Objects)

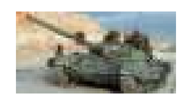
M1A1_tank
(3 Objects)



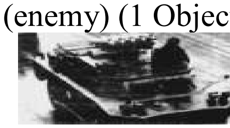
M2_IFV
(3 Objects)

Field_Marker
(6 Objects)

T80_tank
(2 Objects)

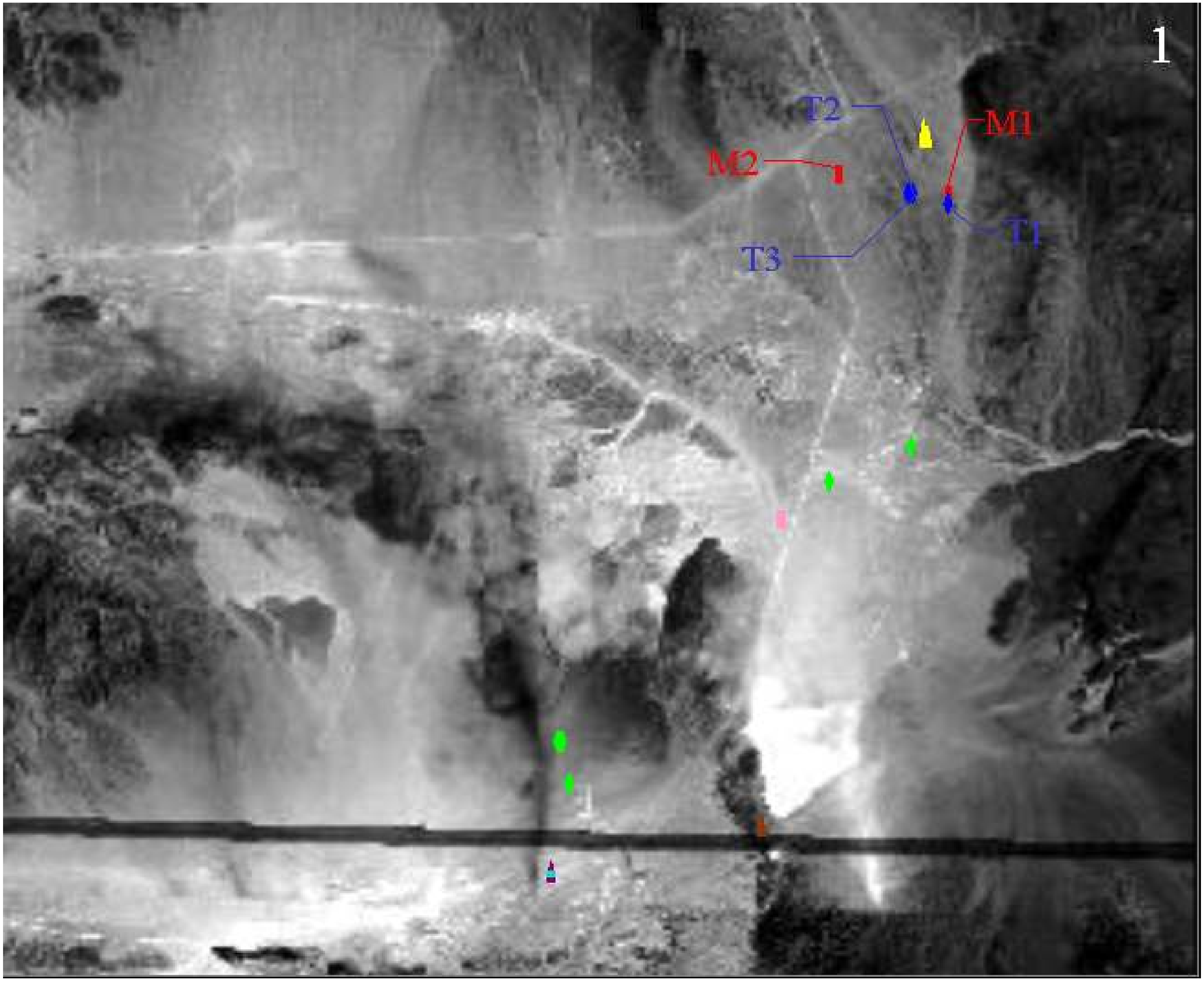


BRDM_AT5
(enemy) (1 Object)



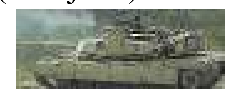
BMP1
(1 Object)

Co-occurring object-types



● Manpack stinger
(2 Objects)

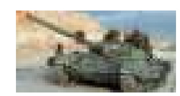
● M1A1_tank
(3 Objects)



● M2_IFV
(3 Objects)

● Field_Marker
(6 Objects)

● T80_tank
(2 Objects)



● BRDM_AT5
(enemy) (1 Object)

● BMP1
(1 Object)

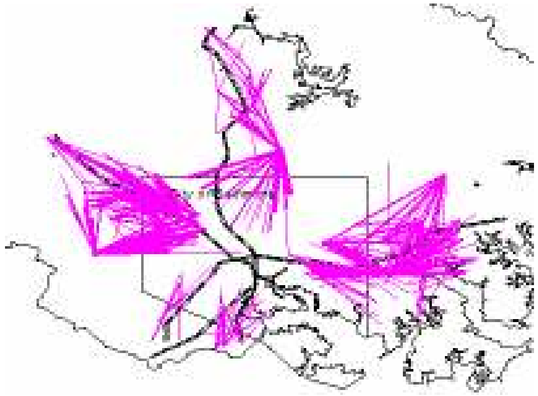
2: Activites on Urban Infrastructure ST Networks

- State of the Art: Environmental Criminology
  - Largely geometric Methods
  - Few Network Methods: Journey to Crime (J2C)

- Critical Barriers:
  - Scale: Houston – 100,000 crimes / year
  - Network based explanation
  - Spatio-temporal networks

- Issues:
  - 1: Network based explanatory models
  - 2: Scalable algorithms for J2C analysis
  - 3: ST Models for Networks
  - 4: ST Network Patterns
  - 5: Validation

- Challenges: Key assumptions violated!
  - Ex. Prefix optimality of shortest paths
  - Can't use Dijktra's, A*, etc.

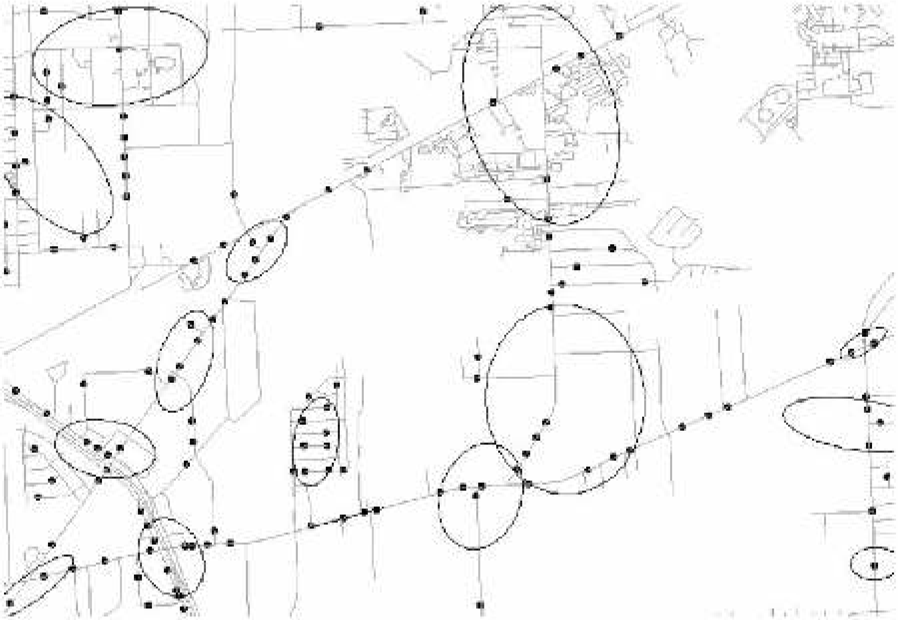


(a) Input: Pink lines connect crime location & criminal's residence

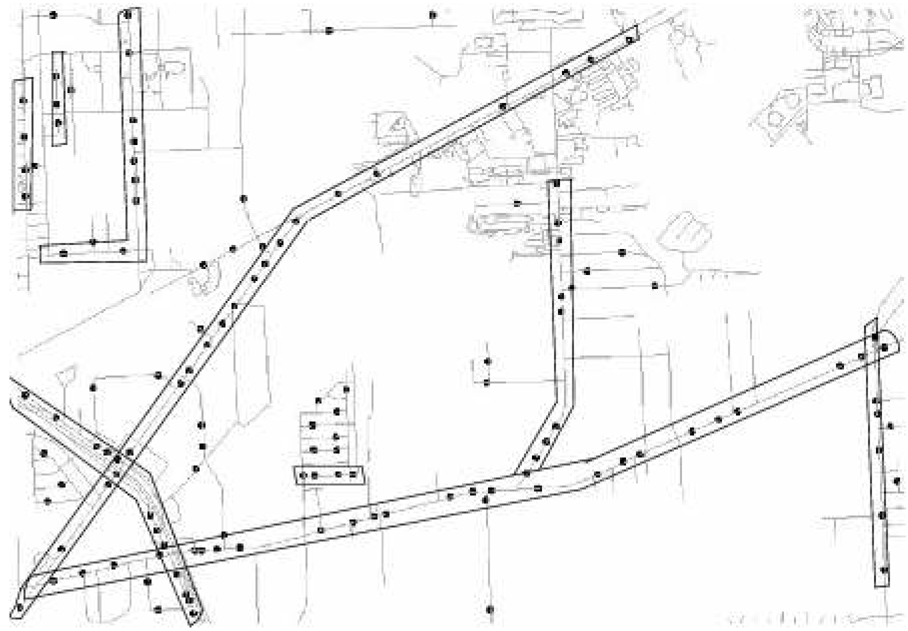
(b) Output: Journey- to-Crime (thickness = route popularity) Source: Crimestat

Hotspots: Euclidean vs. Streets

Houston Crime Dataset



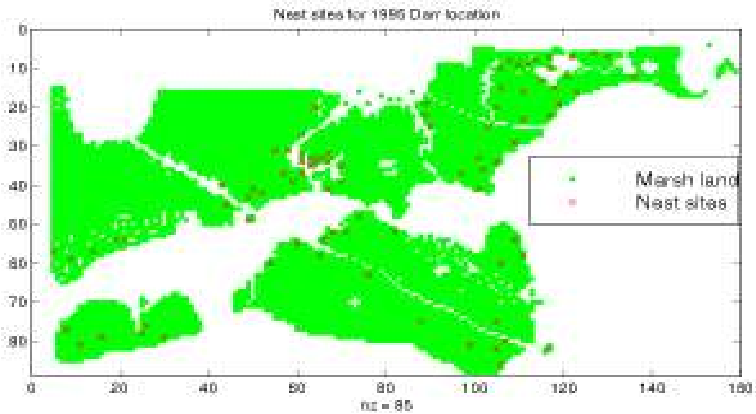
Hot Spots : CrimeStat using K Means clustering for 15 clusters



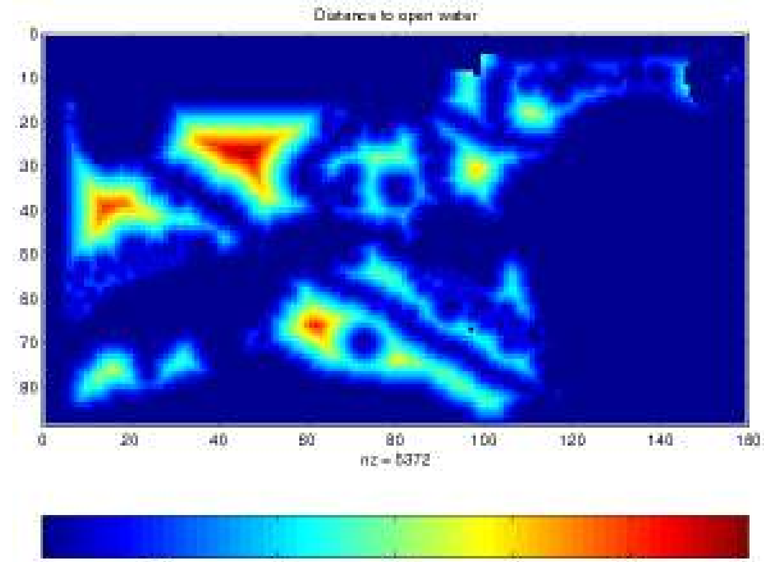
Mean Streets

- Traditional Hotspots:
  - Empty space
- Desirable:
  - Network based methods
  - Challenge: Statistics on networks

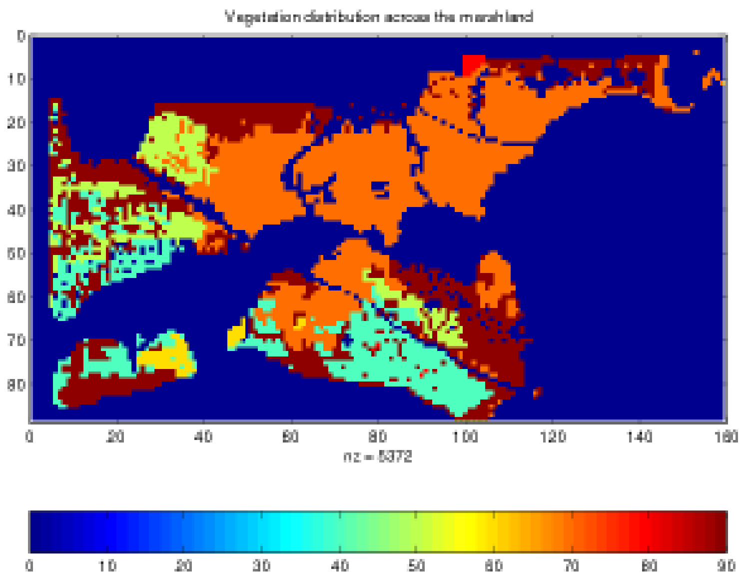
Challenge 1: Is I.I.D. assumption valid?



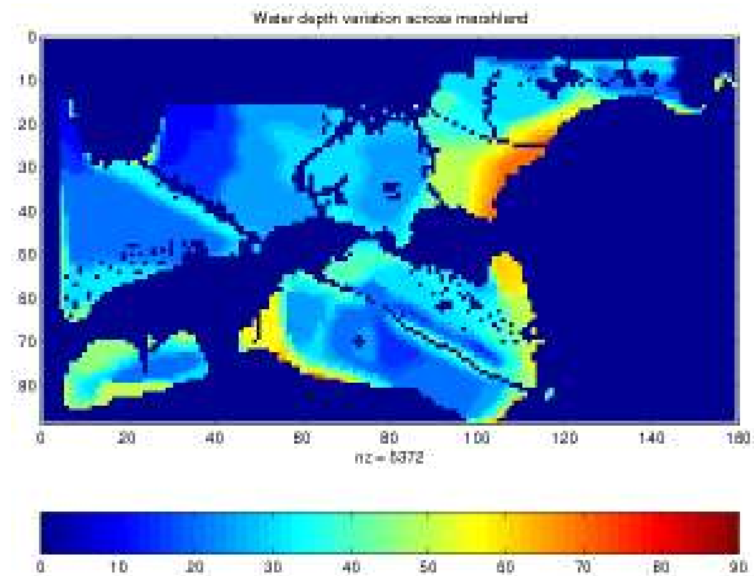
Nest locations



Distance to open water



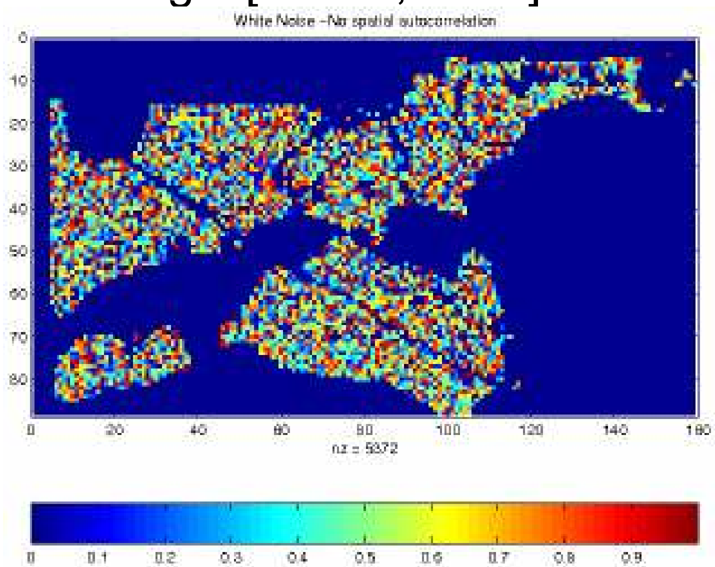
Vegetation durability



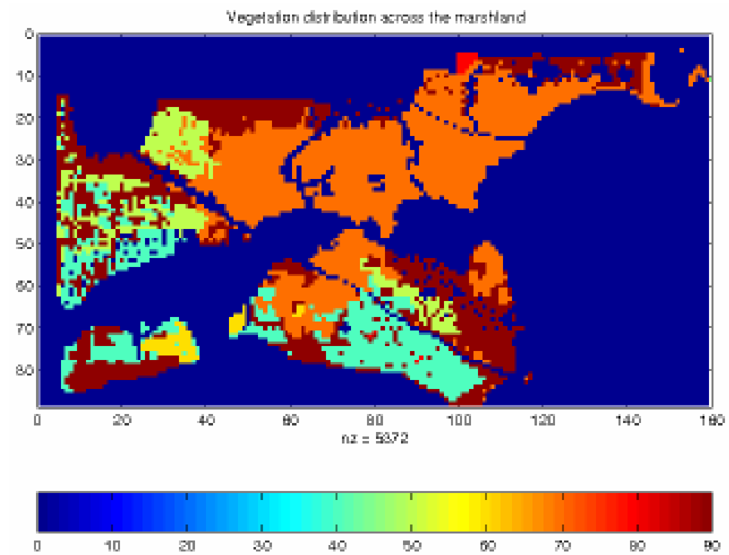
Water depth

Autocorrelation

- First Law of Geography
  - "All things are related, but nearby things are more related than distant things. [Tobler, 1970]"



Pixel property with independent identical distribution



Vegetation Durability with SA

- Autocorrelation
  - Traditional i.i.d. assumption is not valid
  - Measures: K-function, Moran's I, Variogram, …

Implication of Auto-correlation

Name	*Model*	*Classification Accuracy*
Classical Linear Regression	$\boxed{\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$	Low
Spatial Auto-Regression	$\boxed{\mathbf{y} = \boxed{\rho\mathbf{W}\mathbf{y}} + \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$	High

ρ : the spatial auto-regression (auto-correlation) parameter

$\mathbf{W}$: n-by-n neighborhood matrix over spatial framework

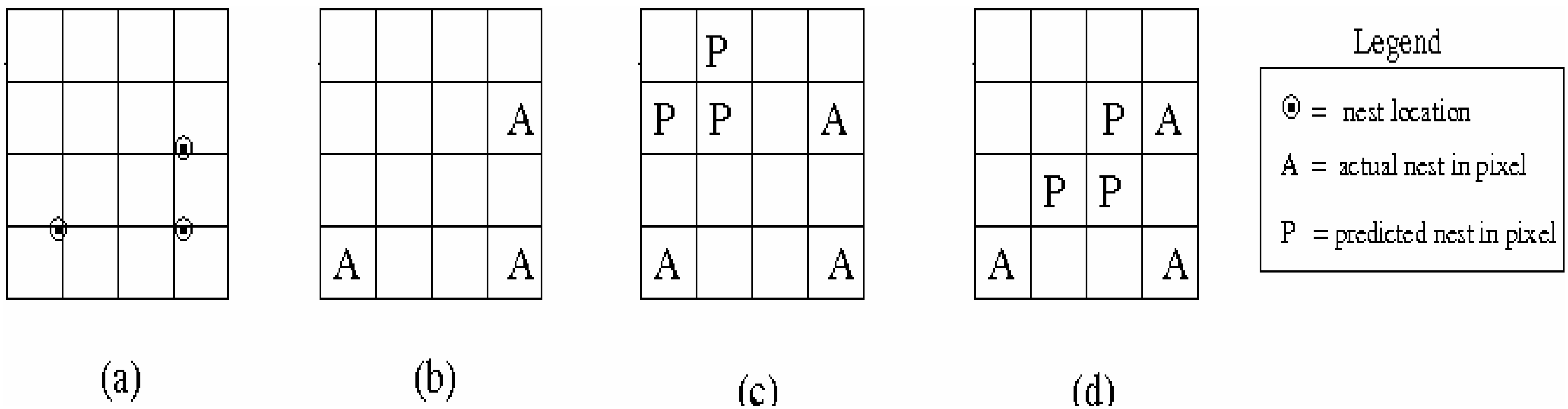
Computational Challenge:

Computing determinant of a very large matrix in the Maximum Likelihood Function:

$$\ln(L) = \boxed{\ln|\mathbf{I} - \rho\mathbf{W}|} - \frac{n\ln(2\pi)}{2} - \frac{n\ln(\sigma^2)}{2} - SSE$$

Research Needs in Location Prediction

- Additional Problems
  - Estimate W for SAR and MRF-BC
  - Scaling issue in SAR
    - Scale difference: $\rho W y$ vs. $X\beta$
  - Spatial error measure: e.g., avg, dist(actual, predicted)



Actual Sites

Pixels with actual sites

Prediction 1

Prediction 2. Spatially more accurate than Prediction 1

Challenge 2: Continuity

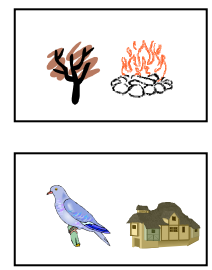
- Association rule e.g. (Diaper in T => Beer in T)

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
- Algorithm Apriori [Agarwal, Srikant, VLDB94]
  - Support based pruning using monotonicity
- Note: **Transaction is a core concept!**

Transactions → Neighborhoods

Q? Which Item-types co-occur in space (and time) ?



Co-location: A Neighborhood based Approach

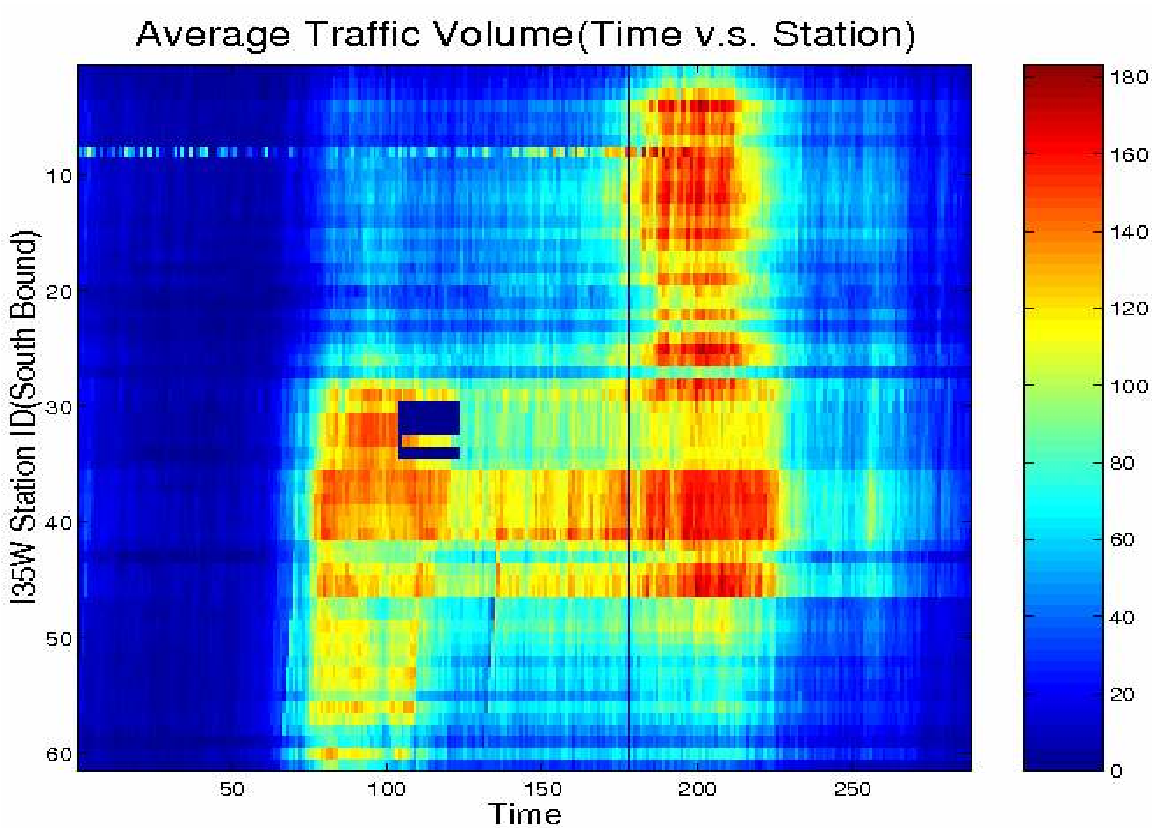
	Association rules	Colocation rules
underlying space	discrete sets	continuous space
item-types	item-types	events /Boolean spatial features
collections	Transactions	neighborhoods
prevalence measure	support	participation index
conditional probability measure	Pr.[A in T | B in T]	Pr.[A in N(L) | B at L]

Challenges:

1. Computational Scalability
   Needs a large number of spatial join, 1 per candidate colocation
2. Spatio-temporal Semantics
   Spatio-tempotal co-occurrences
   Emerging colocations
   …

Challenge 3: Spatial Anamolies

- Example – Sensor 9
  - Issue 1: Will sensor 9 be detected by traditional outlier detection ?
    - New tests: variograms, scatter plot, moran scatter plot,



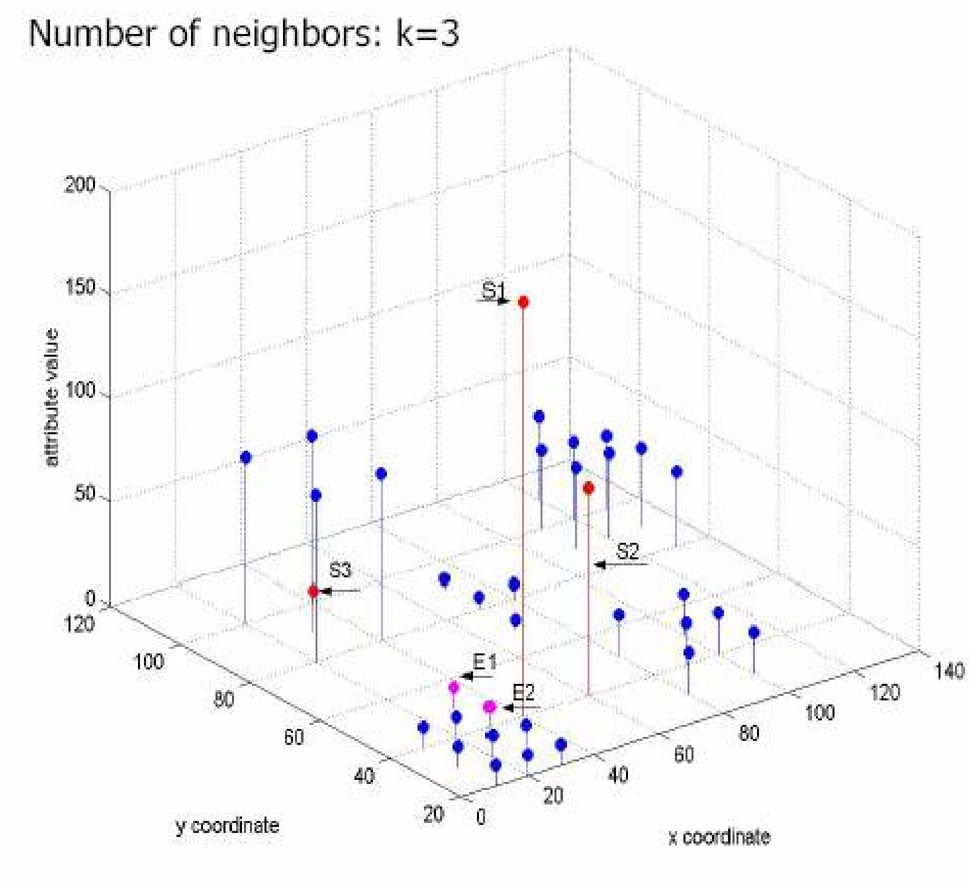
Challenge: Multiple Spatial Outlier Detection

Issue 2: A bad apple makes neighbors look anamolous

Expected Outliers: S1, S2, S3

Top 3 items flagged by traditional approaches: E1, E2, S1

Challenge:
Computational Scalability for detecting multiple spatial anamolies



Courtesy: C.T.Lu, Virginia Tech

3: Multi-Jurisdiction Multi-Temporal (MJMT) Data

- State of the Art:
  - Spatial, ST ontologies
  - Few network ontologies
- Critical Barriers:
  - Heterogeneity across networks
  - Uncertainty – map accuracy, gps, …
- Issues:
  - 1. Ontologies: Network activities
  - 2. Integration methods
  - 3. Location accuracy models
  - 4. Evaluation
- Challenges:
  - Test datasets
  - Evaluation methods