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 hotspots
- 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
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: Activities 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 Dijkstra’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

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

Hot Spots: CrimeStat using K Means clustering for 15 clusters

Mean Streets
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]”

• Autocorrelation
  – Traditional i.i.d. assumption is not valid
  – Measures: K-function, Moran’s I, Variogram, …
Implication of Auto-correlation

<table>
<thead>
<tr>
<th>Name</th>
<th>Model</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical Linear Regression</td>
<td>$y = x\beta + \varepsilon$</td>
<td>Low</td>
</tr>
<tr>
<td>Spatial Auto-Regression</td>
<td>$y = \rho W y + x\beta + \varepsilon$</td>
<td>High</td>
</tr>
</tbody>
</table>

$\rho$: the spatial auto-regression (auto-correlation) parameter

$W$: $n$-by-$n$ neighborhood matrix over spatial framework

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

$$\ln(L) = \ln \left| I - \rho W \right| - \frac{n \ln(2\pi)}{2} - \frac{n \ln(\sigma^2)}{2} - SSE$$
Research Needs in Location Prediction

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

Legend:
- $\oplus$ = nest location
- $A$ = actual nest in pixel
- $P$ = predicted nest in pixel

(a) Actual Sites
(b) Pixels with actual sites
(c) Prediction 1
(d) Prediction 2. Spatially more accurate than Prediction 1
Challenge 2: Continuity

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

<table>
<thead>
<tr>
<th>Transaction</th>
<th>Items Bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{socks, milk, beef, egg, …}</td>
</tr>
<tr>
<td>2</td>
<td>{pillow, toothbrush, ice-cream, muffin, …}</td>
</tr>
<tr>
<td>3</td>
<td>{ Pacifier, pacifier, formula, blanket, …}</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>n</td>
<td>{battery, juice, beef, egg, chicken, …}</td>
</tr>
</tbody>
</table>

- 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 $\rightarrow$ Neighborhoods

Q? Which Item-types co-occur in space (and time)?
**Co-location: A Neighborhood based Approach**

<table>
<thead>
<tr>
<th></th>
<th>Association rules</th>
<th>Colocation rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>underlying space</td>
<td>discrete sets</td>
<td>continuous space</td>
</tr>
<tr>
<td>item-types</td>
<td>item-types</td>
<td>events /Boolean spatial features</td>
</tr>
<tr>
<td>collections</td>
<td>Transactions</td>
<td>neighborhoods</td>
</tr>
<tr>
<td>prevalence measure</td>
<td>support</td>
<td>participation index</td>
</tr>
<tr>
<td>conditional probability measure</td>
<td>Pr.[ A in T</td>
<td>B in T ]</td>
</tr>
</tbody>
</table>

**Challenges:**

1. **Computational Scalability**
   - Needs a large number of spatial join, 1 per candidate colocation

2. **Spatio-temporal Semantics**
   - Spatio-temporal co-occurrences
   - Emerging colocations

...
Challenge 3: Spatial Anomalies

- 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 anomalous

Expected Outliers: S1, S2, S3

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

Challenge:
  Computational Scalability for detecting multiple spatial anomalies

Number of neighbors: k=3

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