Transportation Data Mining: Vision & Challenges

Shashi Shekhar
McKnight Distinguished University Professor
University of Minnesota
www.cs.umn.edu/~shekhar

Pervasive Data for Transportation:
Innovations in Distributed and Mobile Information Discovery in ITS & LBS
Transportation Research Board Meeting 182
January 23rd, 2011.
Spatial Databases: Representative Projects

Parallelize Range Queries

Evacuation Route Planning

- only in old plan
- Only in new plan
- In both plans

Shortest Paths

Storing graphs in disk blocks
Spatial Data Mining: Representative Projects

**Location prediction: nesting sites**
- Nest locations
- Distance to open water
- Vegetation durability
- Water depth

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

**Co-location Patterns**

**Tele connections**
Outline

• Motivation
  – Transportation Questions
  – Transportation Theories
  – Limitations of theories

• Data mining

• Conclusions
Questions in Transportation Domain

• Traveler, Commuter
  – What will be the travel time on a route?
  – Will I make to destination in time for a meeting?
  – Where are the incident and events?

• Transportation Manager
  – How the freeway system performed yesterday?
  – Which locations are worst performers?

• Traffic Engineering
  – Which loop detection are not working properly?
  – Where are the congestion (in time and space)?
  – How congestion start and spread?

• Planner and Researchers
  – What will be travel demand in future?
  – What will be the effect of hybrid cars?
  – What are future bottlenecks? Where should capacity be added?

• Policy
  – What is an appropriate congestion-pricing function?
  – Road user charges: How much more should trucks pay relative to cars?
Theories in Transportation Domain

• Physics
  • Traffic: Fluid flow models (e.g. reduce turbulence), control theory
  • How to reduce icing on pavements?
• Chemistry
  • Environmental impact (e.g. salt, incomplete combustion)
• Biology
  • How to reduce crash-injury severity?
  • Effect of age, sleep deprivation, toxins, ...
• Psychology
  • Human factors: design of highway signage, vehicle dashboard
  • Activity and agent based models
• Sociology
  • Household decisions, Homophily and social networks
  • Lack of trust => aggressive driving
• Economics, Game Theory
  • Incentive mechanisms
  • Wardrop equilibrium among commuters
    • Ex. All comparable paths have same travel time!
Limitations of Theories

• Multi-disciplinary questions:
  • Will hybrid cars reduce environmental impact of transportation?
  • Extreme events – evacuation, conventions, …
  • Impact of context – weather, climate, economy, politics, crime, police cars, …

• Mono-disciplinary questions
  • Non-equilibrium phenomena, e.g. location, time and path
  • Critical places & moments: Accident hotspots (hot-moments)? Why?
  • Normality & anomalies: e.g. traffic flow discontinuities – location, cause
  • Regional difference: effectiveness of Ramp meters across places & time-periods

• What are the **options** to complement theory based approaches?
Data-Intensive Scientific Discovery

• Classical Approach
  • Travel diaries, NHTS survey (OD matrix), Lab. (mpg rating)
  • Hypothesis driven data collection, Statistical hypothesis testing

• Emerging Data-Intensive Approach
  • Secondary Data: Engine computer, gps, cell-phones, face-book,VGI,
  • Exploratory data analysis for hypothesis generation
  • Ex. Data Mining and Knowledge Discovery
Outline

• Motivation
• Data mining
  – Case Studies
  – Definition
• Pattern Families
• Conclusions
Adoption of Data Mining

- Example: IBM Smarter Planet Initiative, SAS, …
  - Large Organizations: Walmart, USDOD, …

- 1990s: Data Mining
  - Scale up to traditional models to large relational databases
    - Linear regression, Decision Trees, …
  - New pattern families: Association rules
    - Which items are bought together? E.g. (Diaper, beer)

- Spatial customers
  - Walmart
    - Which items are bought just before/after events, e.g. hurricanes?
    - How to send these items to appropriate stores?
    - Where is (diaper-beer) pattern prevalent?
  - Center for Disease Control: cancer clusters
  - Police: crime hotspots
  - USDOD, intelligence: anomaly detection, link analysis
Serious Scientists are also using Data Mining!

Example: NASA IVMS Data Mining Laboratory
The lab enables the dissemination of Integrated Vehicle Health Management data, algorithms, and results to the public. It will serve as a national asset for research and development of discovery algorithms for detection, diagnosis, prognosis, and prediction for NASA missions.
Data Mining

• What is it?
  – Identifying interesting, useful, non-trivial patterns
    • Hot-spots, anomalies, associations, precursors
  – in large datasets
    • Infrastructure:
      – Aerial surveillance (e.g. ARGUS-IS)
      – Geo-sensor network (loop detector, cameras), …
    • Volunteered: cell-phone, gps, social network

• Importance
  – Potential of discoveries and insights to improve lives
    • Traffic Management: Where and when are traffic flow anomalies? Why?
    • Safety: Where are accident hotspots? Why?
    • (Tele)-connection: traffic-congestion & events (e.g. weather, conventions)
    • Transportation Planning: How is demand changing? Consequences?

• Challenge:
  • \((d/dt)\) (Data Volume) \(\gg\) \((d/dt)\) (Number of Human Analysts)
  • Need automated methods to mine patterns from data
  • Need tools to amplify human capabilities to analyze data
Outline

• Motivation
• Data mining
• Pattern Families
  – Spatial outliers
  – Hotspots
  – Co-occurrences
  – Prediction
• Conclusions
Example 1: Anomalies

- Example – Sensor 9
  - Will sensor 9 be detected by traditional outlier detection?
  - Is it a global outlier?
Example 2: HotSpots

- What is it?
  - Unusually high spatial concentration of a phenomena
    - Accident hotspots
    - Used in epidemiology, crime analysis

- Solved
  - Spatial statistics based ellipsoids

- Almost solved
  - Transportation network based hotspots

- Next
  - Emerging hot-spots
Example 3: Associations, Co-locations, Co-occurrences

• Road user-charges:
  – Is technology available for road-type based policy?
  – Which road segments are vulnerable for mis-classification?
• Issue: accuracy or GPS & digital roadmaps
Example 3b: Associations

- Which following transportation networks co-occur? Where? Why?
  - e.g. roads, river, railroads, air, etc.. in North Korea

Road-River/Stream Colocation
Example 4: Prediction

• Impact
  – Deforestation – Brazil lost 150,000 sq. km. of forest between 2000 and 2006
  – Urban Sprawl

• Environmental Aspects
  – Deforestation
  – Habitat loss, endangered species
  – Water and air quality
  – Climate change (?)
  – …

• Urgent issues => Policy changes
  – Brazil: real-time monitoring of forests
  – USA: from VMT to access
  – …
Example 4: Prediction

• Transportation Planning
  – What will be the impact of a new office building?
  – What will be travel demand? future bottlenecks?
  – What will be the effect of hybrid cars on traffic?
  – How will better bicycle facility impact vehicle traffic?

• Q? Are classical techniques (e.g. Decision trees, SVM, …) adequate?

• Challenges
  – Spatio-temporal auto-correlation – violates independence assumption
  – Network : routes, edge capacities, …
  – Individual behavior: urban sprawl?
  – Group dynamics: game theory, Wardrop equilibrium, …
Outline

• Motivation
• Data mining
• Conclusions
  – Summary
  – Research Challenges
Summary

• It’s time for transportation community to give serious consideration to data mining and knowledge discovery!

• Transportation is facing new challenges
  – Climate change driven policy changes

• Classical approaches are limited
  – Multi-disciplinary problems, non-equilibrium scenarios,
  – Extreme events

• Data-Intensive Scientific Discovery
  – Complements classical approaches: Hypothesis generation
  – Secondary datasets are growing
  – Data mining technology is maturing
Datasets in Transportation Domain

• Datasets
  – Reports on accidents, traffic law violation
  – Travel diaries and surveys
  – Traffic simulator (e.g. DYNASMART) outputs
  – Loop-detector: traffic volume, density, occupancy, …
  – Traffic camera - videos
  – Automatic vehicle location and identification
    • from GPS, cell-phone, automatic tolling transponder, etc.
  – Other sensors: bridge strain, visibility (in fog), ice, …
  – Yellow Pages, street addresses

• Characteristics
  – Spatio-temporal networks
New Datasets: Speed Profiles

- Transportation
  - Road networks: Nodes = road intersections, Edge = road segments
  - Edge attribute: travel time; Navteq reports it a function of time!
- Operations:
  - Hot moments (i.e. rush hours), Hotspots (i.e. congestion)
  - Fastest Path, Evacuation capacities of routes
Transportation Data Mining: Computational Challenges

• Violates assumptions of classical data mining
  – Lack of independence among samples - ? Decision trees, …
  – No natural transactions -? Association rule, …

• Two kinds of spaces
  – Embedding space, e.g. Geography, Network, Time
  – Feature space, e.g. Traffic volume, accidents, …

• Lessons from Spatial thinking
  – 1st Law: Auto-correlation: Nearby things are related
  – Heterogeneity
  – Edge effect
  – …
Spatial/Spatio-temporal Outliers Challenges

- **What is it?**
  - Location different from their neighbors
    - Discontinuities, flow anomalies

- **Solved**
  - Transient spatial outliers

- **Almost solved**
  - Anomalous trajectories

- **Failed**

- **Missing**
  - Persistent anomalies
  - Multiple object types, Scale

- **Next**
  - Dominant Persistent Anomalies
HotSpots

What is it?
- Unusually high spatial concentration of a phenomena
  - Accident hotspots
  - Used in epidemiology, crime analysis

Solved
- Spatial statistics based ellipsoids

Almost solved
- Transportation network based hotspots

Failed
- Classical clustering methods, e.g. K-means

Missing
- Spatio-temporal

Next
- Emerging hot-spots
Colocation, Co-occurrence, Interaction

- **What is it?**
  - Subset of event types, whose instances occur together
  - Ex. Symbiosis, (bar, misdemeanors), …

- **Solved**
  - Colocation of point event-types

- **Almost solved**
  - Co-location of extended (e.g. linear) objects
  - Object-types that move together

- **Failed**
  - Neighbor-unaware Transaction based approaches

- **Missing**
  - Consideration of flow, richer interactions

- **Next**
  - Spatio-temporal interactions, e.g. item-types that sell well before or after a hurricane
  - Tele-connections
Space/Time Prediction

What is it?
- Models to predict location, time, path, …
  - Nest sites, minerals, earthquakes, tornadoes, …

Solved
- Interpolation, e.g. Krigging
- Heterogeneity, e.g. geo. weighted regression

Almost solved
- Auto-correlation, e.g. spatial auto-regression

Failed: Independence assumption
- Models, e.g. Decision trees, linear regression, …
- Measures, e.g. total square error, precision, recall

Missing
- Spatio-temporal vector fields (e.g. flows, motion), physics

Next
- Scalable algorithms for parameter estimation
- Distance based errors

\[ y = \rho Wy + x\beta + \varepsilon \]
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>$y = x\beta + \epsilon$</td>
<td>Low</td>
</tr>
<tr>
<td>Spatial Auto-Regression</td>
<td>$y = \rho Wy + x\beta + \epsilon$</td>
<td>High</td>
</tr>
</tbody>
</table>

$\rho$: the spatial auto-regression (auto-correlation) parameter  
$W$: $n$ - by - $n$ neighbor-od matrix over spatial framework

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

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