Transportation Data Mining Challenges

Shashi Shekhar

McKnight Distinguished University Professor

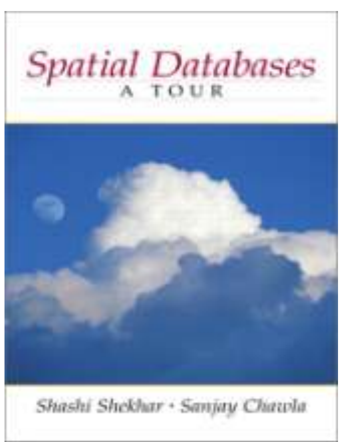
University of Minnesota

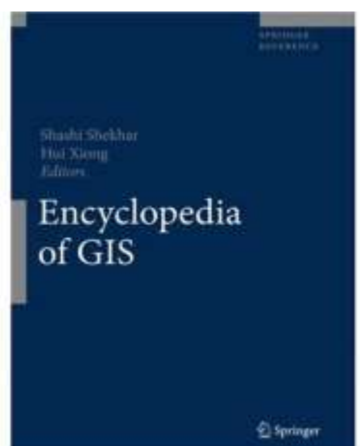
www.cs.umn.edu/~shekhar

Next Generation Data Mining

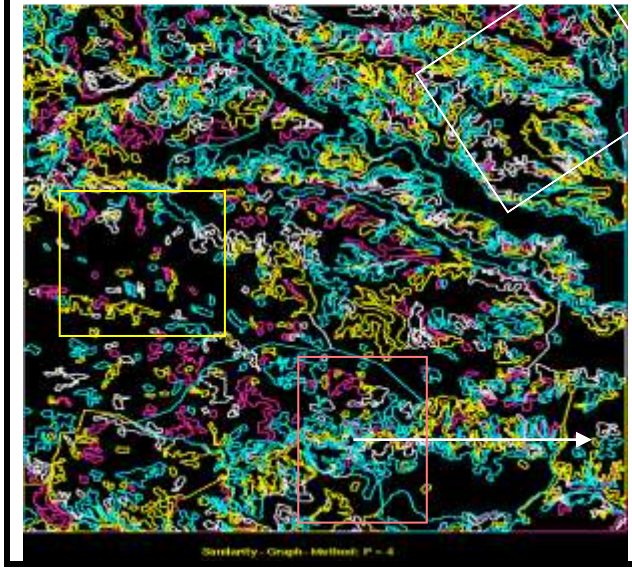
Session of Transportation

October 2nd, 2009



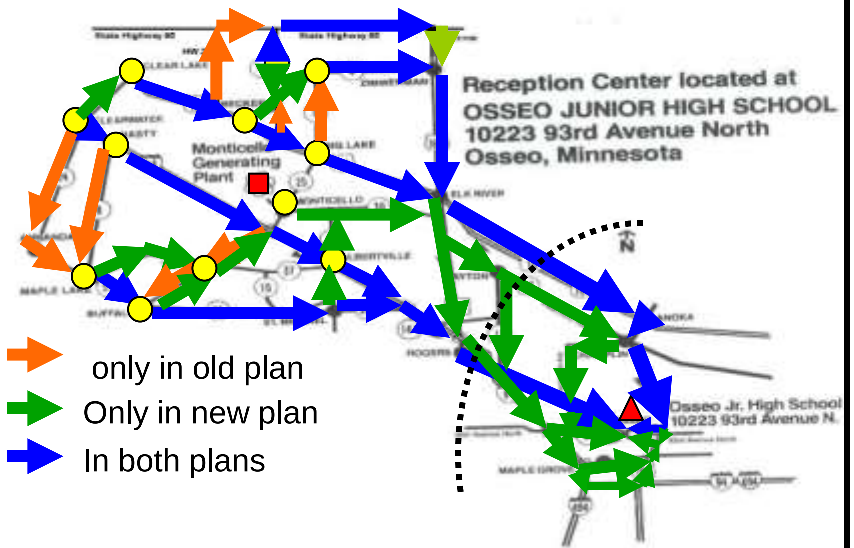


Spatial Databases: Representative Projects



Parallelize
Range Queries

Evacutation Route Planning

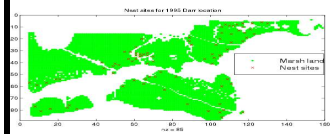


Shortest Paths Storing graphs in disk blocks

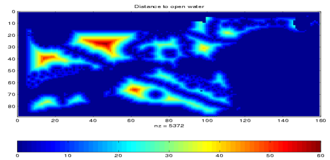
Spatial Data Mining : Representative Projects

Location prediction: nesting sites

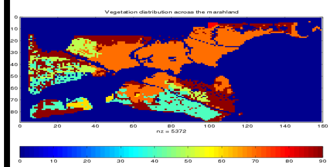
Nest locations



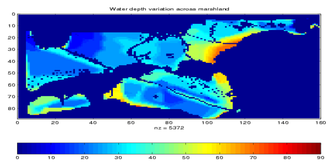
Distance to open water



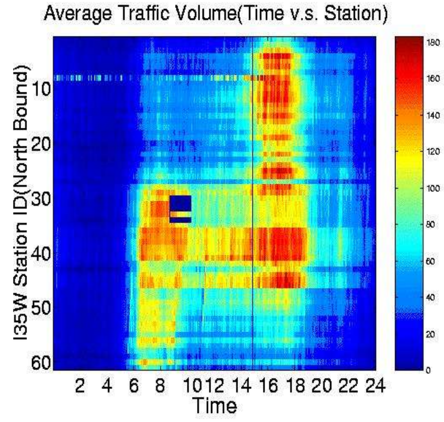
Vegetation durability



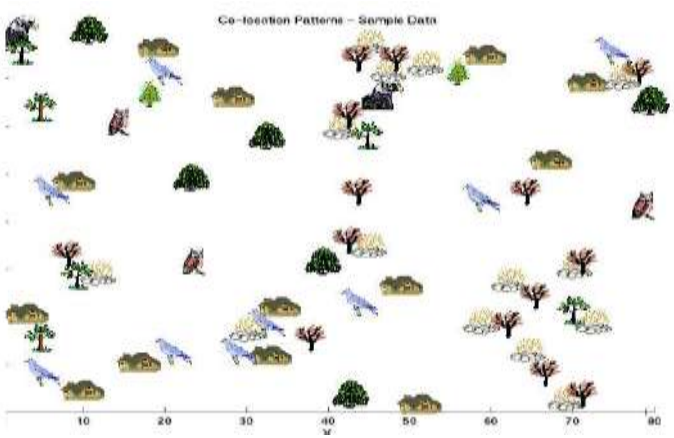
Water depth



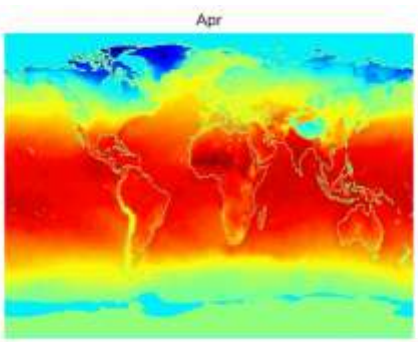
Spatial outliers: sensor (#9) on I-35



Co-location Patterns



Tele connections



Outline

- Transportation domain
  - Questions
  - Stakeholders
  - Datasets
- A transportation dataset
- Data Mining Challenges
- Summary

Transportation Questions

- 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?

Transportation Knowledge

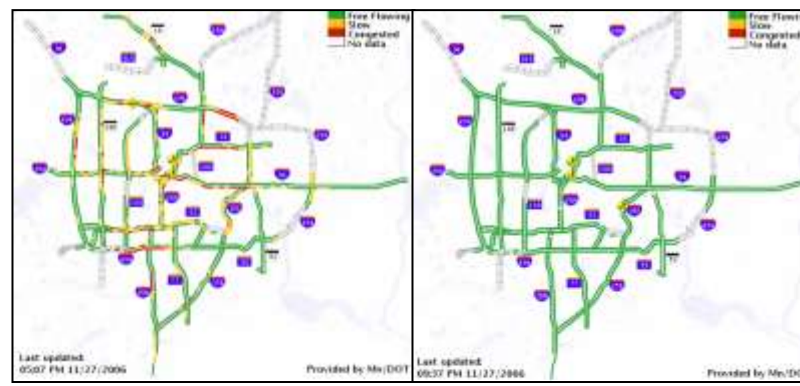
- Classical data:
  - travel diaries, NHTS survey (e..g. OD matrix), Lab. (mpg rating)
- Physics
  - Fluid flow models for traffic
  - Reduce turbulence (i.e. lane weaving) to improve flow
- Chemistry, Biology
  - Environmental impact analysis (e.g. salt)
- Psychology: Individual Behavior
  - Lack of trust => aggressive driving,
  - Activity leads to travel, agent based model
- Socio-Economics: Group Dynamics
  - Social interaction: Household
  - Game thoery: Wardrop equilibrium in commuter traffic
    - All comparable paths have same travel time!
  - Incentive mechanism
- Why data mining?
  - New datasets – engine computers, traffic sensors, gps-tracks,
  - Finer resolution – non-equilibrium phenomena, …
  - Extreme events – evacuation, conventions, …
  - Causal insights ?

New Datasets Datasets

- 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

I94 @ Hamline Ave at 8AM & 10AM



Traffic sensors on Twin-Cities, MN Road Network monitor traffic levels/travel time on the road network.

(Courtesy: MN-DoT (www.dot.state.mn.us))

bing
55455 pizza
Businesses | People | Collections | Locations
business results 1-10 of 250 for pizza near 55455, MN Modify search
Sort by: Relevance Distance | Map all
SPONSORED LISTINGS FROM YELLOWPAGES.COM
Broadway Pizza/Eagles Nest Bar · Website
4106 Lakeland Ave N, Minneapolis, MN
(763) 533-0800
1-click directions
2D 3D Road Aerial
SE 5th St
SE 12th Ave
Elliot Hall
Pleasant St SE
River Rd E
SE Washington Ave
Huron Bl
Harvard St SE
9:42 AM
0.7 miles 3 minutes
Metropolitan Museum of Art
Search
Directions
15° NW
NW
NE
SE
9'55"N, 122°1'46"W
The Metropolitan Museum of Art
70
Keep left on High Street
E15
Arrival
4:48 pm
Menu
Turn In
200 m
GARMIN nüvi

Transportation Domain

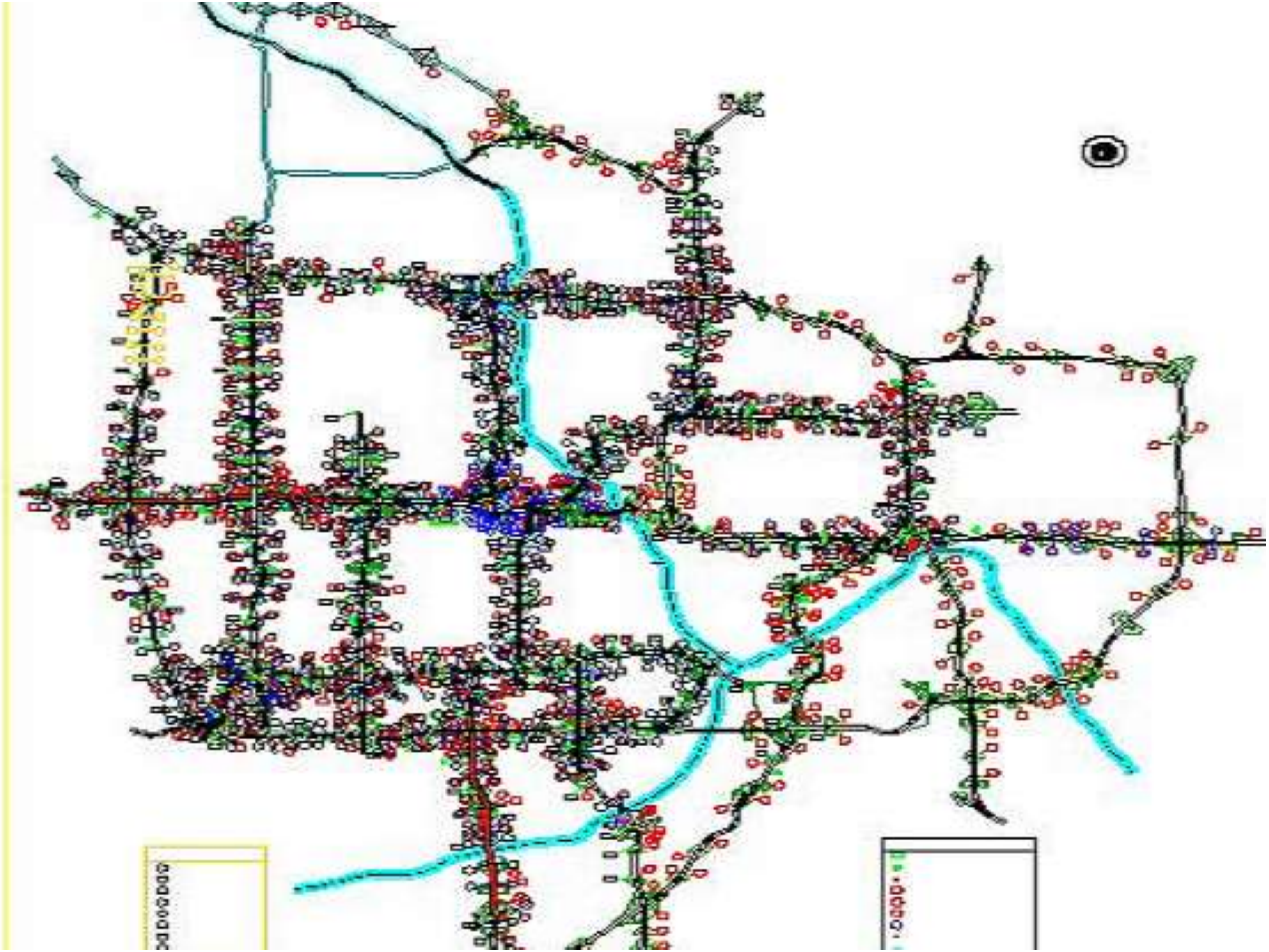
- Datasets
  - Travel diaries and surveys
  - Traffic simulator outputs
  - Accident reports, traffic law violation reports
  - Loop-detector measurement of traffic volume, density, occupancy, etc.
  - Traffic camera - videos
  - Automatic vehicle location and identification
    - from automatic tolling transponder, gps, etc.
  - Other sensors: bridge strain, visibility (in fog), ice, …
  - Yellow Pages, street addresses

- Characteristics
  - Spatio-temporal networks

Outline

- Transportation domain
- A transportation dataset
  - Map of sensor network
  - Spatio-temporal dimensions
  - Summary visualizations
- Data Mining Challenges
- Summary

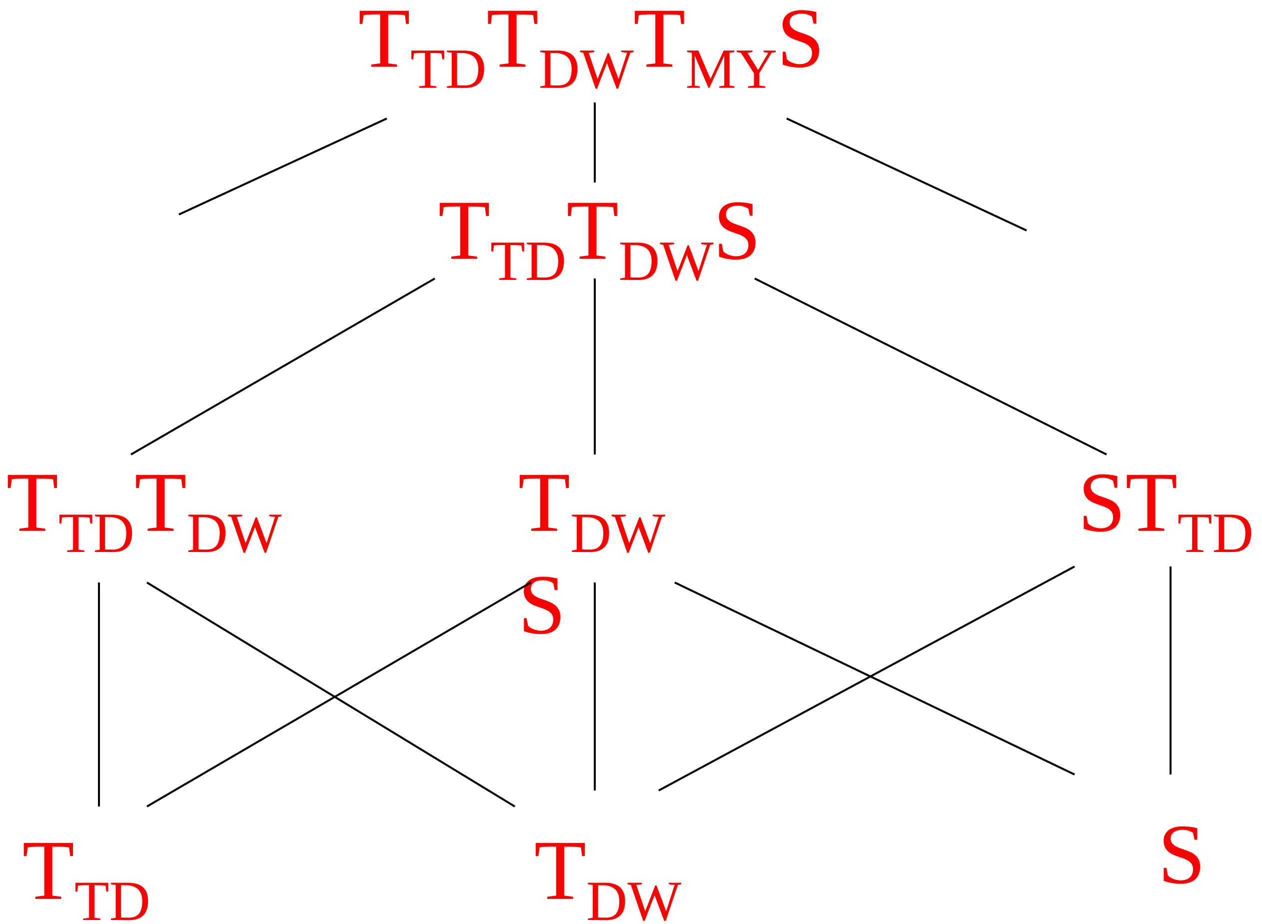
Loop-detector on Twincities Highways



Dimensions

- Available
  - T_{TD} : Time of Day
  - T_{DW} : Day of Week
  - T_{MY} : Month of Year
  - S : Station, Highway, All Stations
- Others
  - Scale, Weather, Seasons, Event types, …

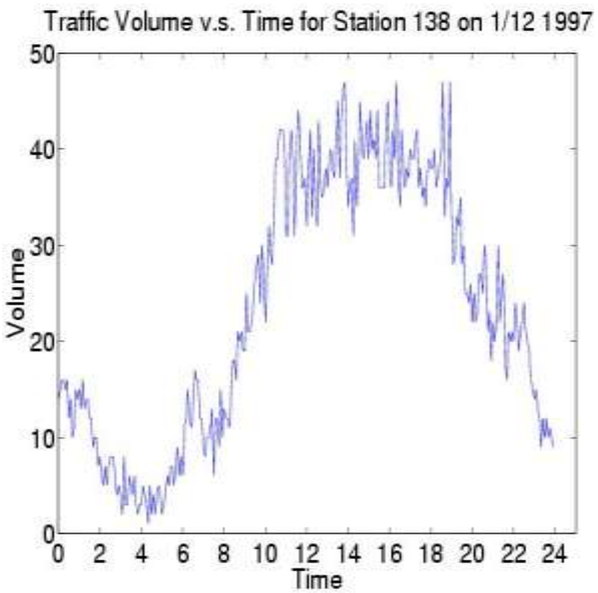
Mapcube :
Which Subset of Dimensions ?

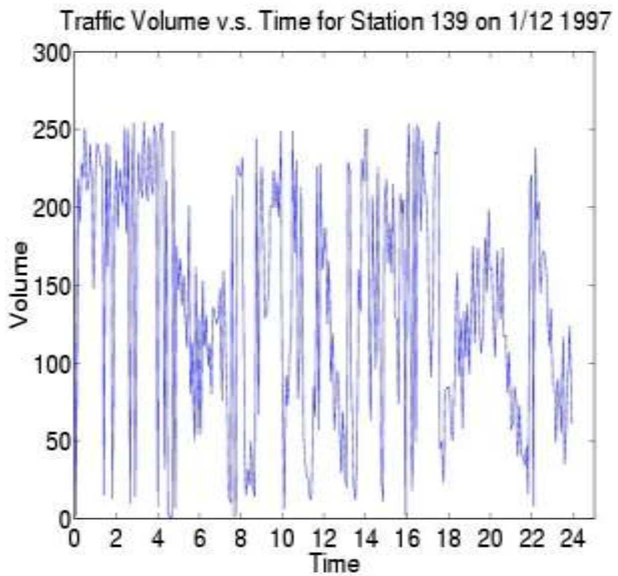


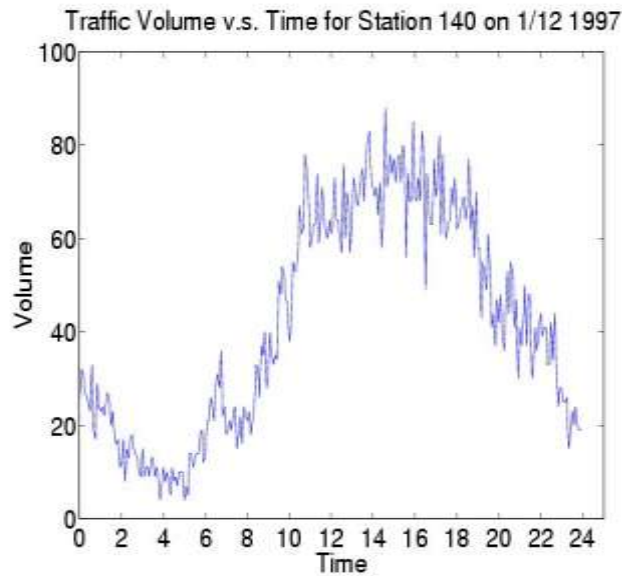
[Next Project]

Singleton Subset : T_{TD}

Configuration:
- X-axis: time of day; Y-axis: Volume
- For station sid 138, sid 139, sid 140, on 1/12/1997



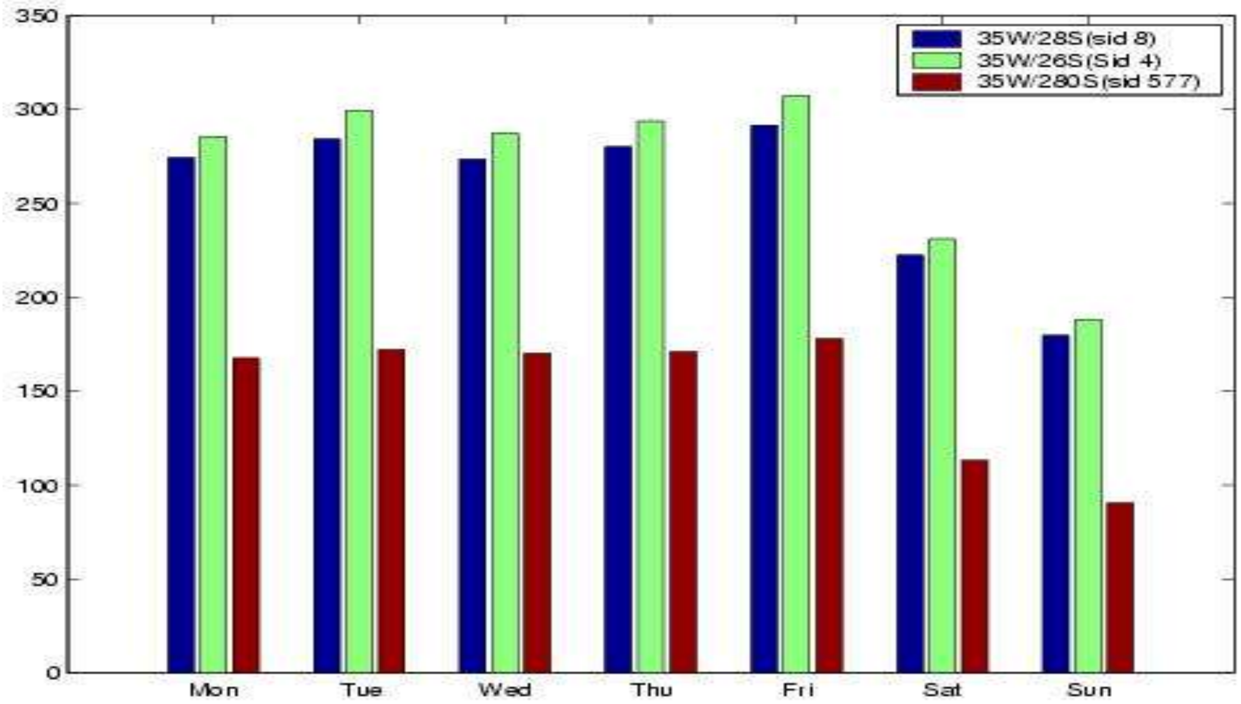




Trends:
- Station sid 139: rush hour all day long
- Station sid 139 is an S-outlier

Singleton Subset: T_{DW}

- Configuration:
  - X axis: Day of week; Y axis: Avg. volume.
  - For stations 4, 8, 577
  - Avg. volume for Jan 1997

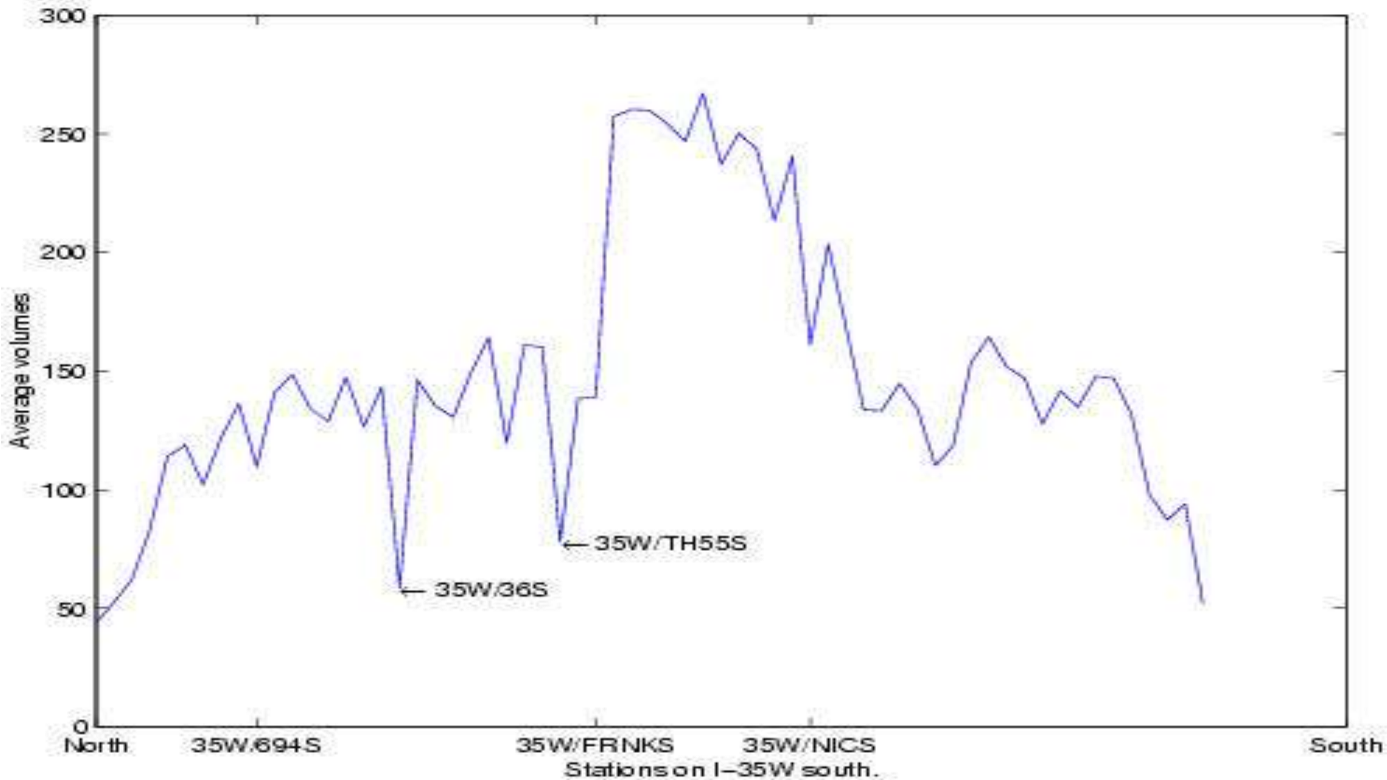


Trends:
- Friday is the busiest day of week
- Tuesday is the second busiest day of week

Singleton Subset: S

Configuration:

- X-axis: I-35W South; Y-axis: Avg. traffic volume
- Avg. traffic volume for January 1997

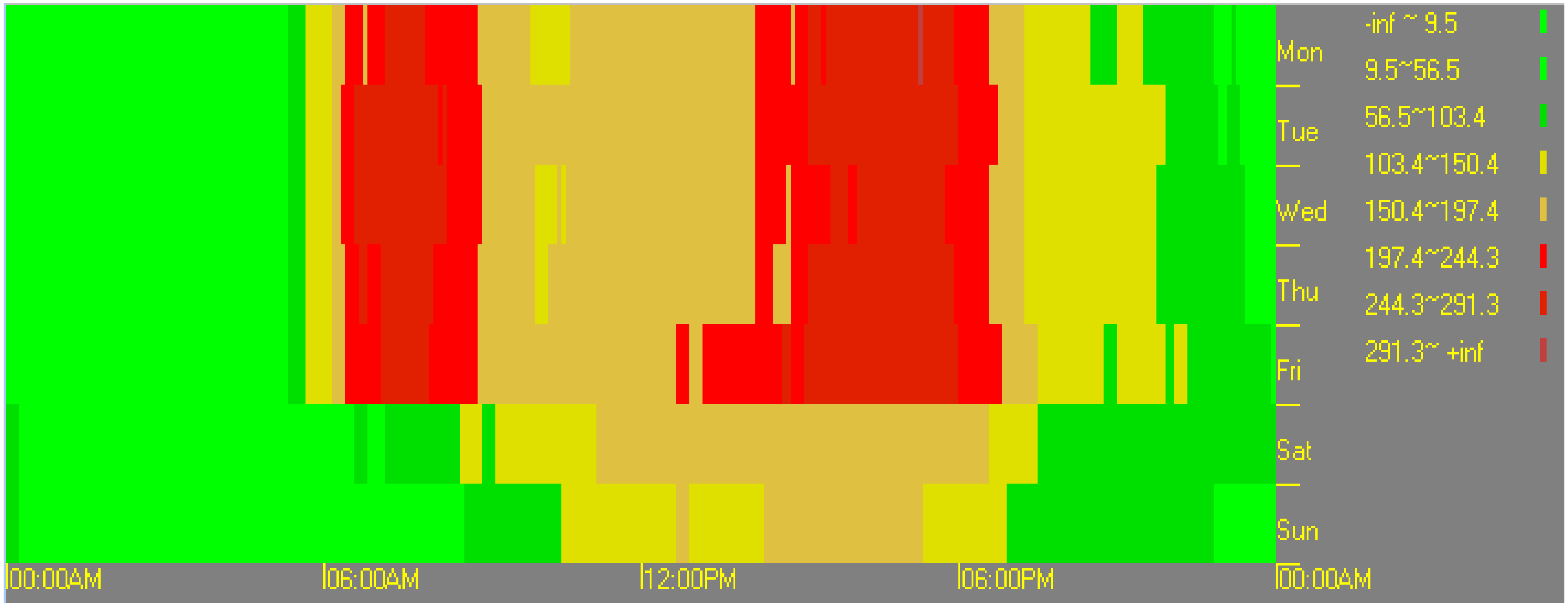


Trends?:

- High avg. traffic volume from Franklin Ave to Nicollet Ave
- Two outliers: 35W/26S(sid 576) and 35W/TH55S(sid 585)

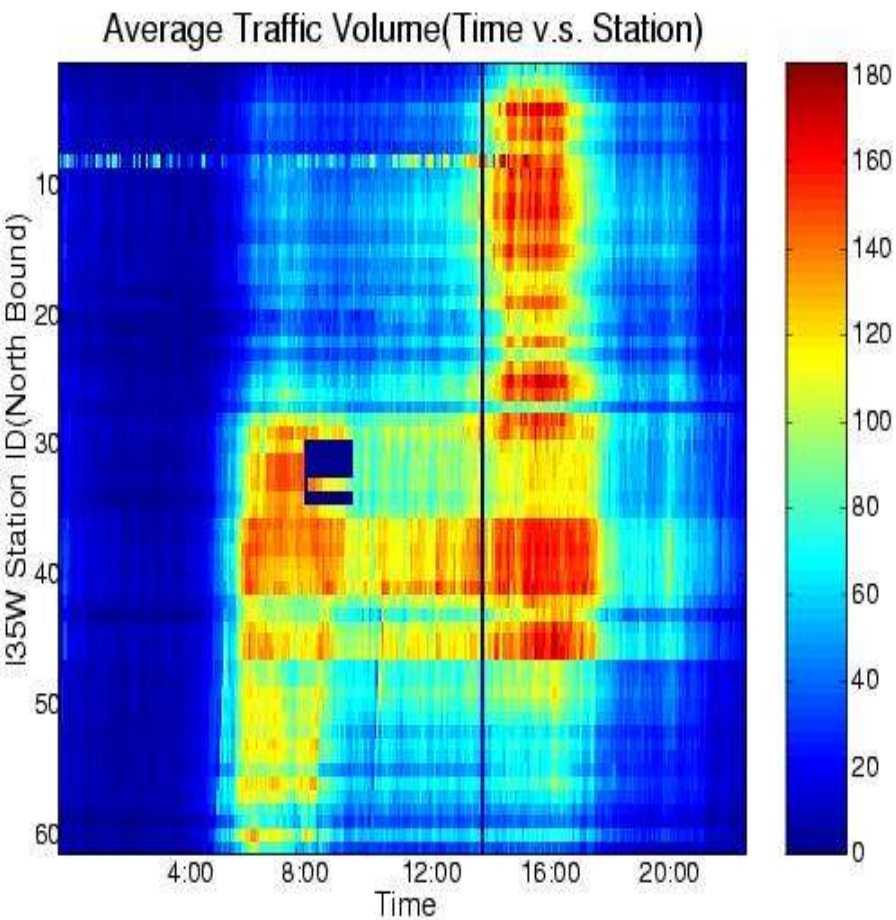
Dimension Pair: T_{TD}-T_{DW}

Configuration:
- X-axis: time of date; Y-axis: day of Week
- f(x,y): Avg. volume over all stations for Jan 1997, except Jan 1, 1997



Trends:
- Evening rush hour broader than morning rush hour
- Rush hour starts early on Friday.
- Wednesday - narrower evening rush hour

Dimension Pair: S-T_{TD}



Configuration:

- X-axis: Time of Day
- Y-axis: Highway
- f(x,y): Avg. volume over all stations for 1/15, 1997

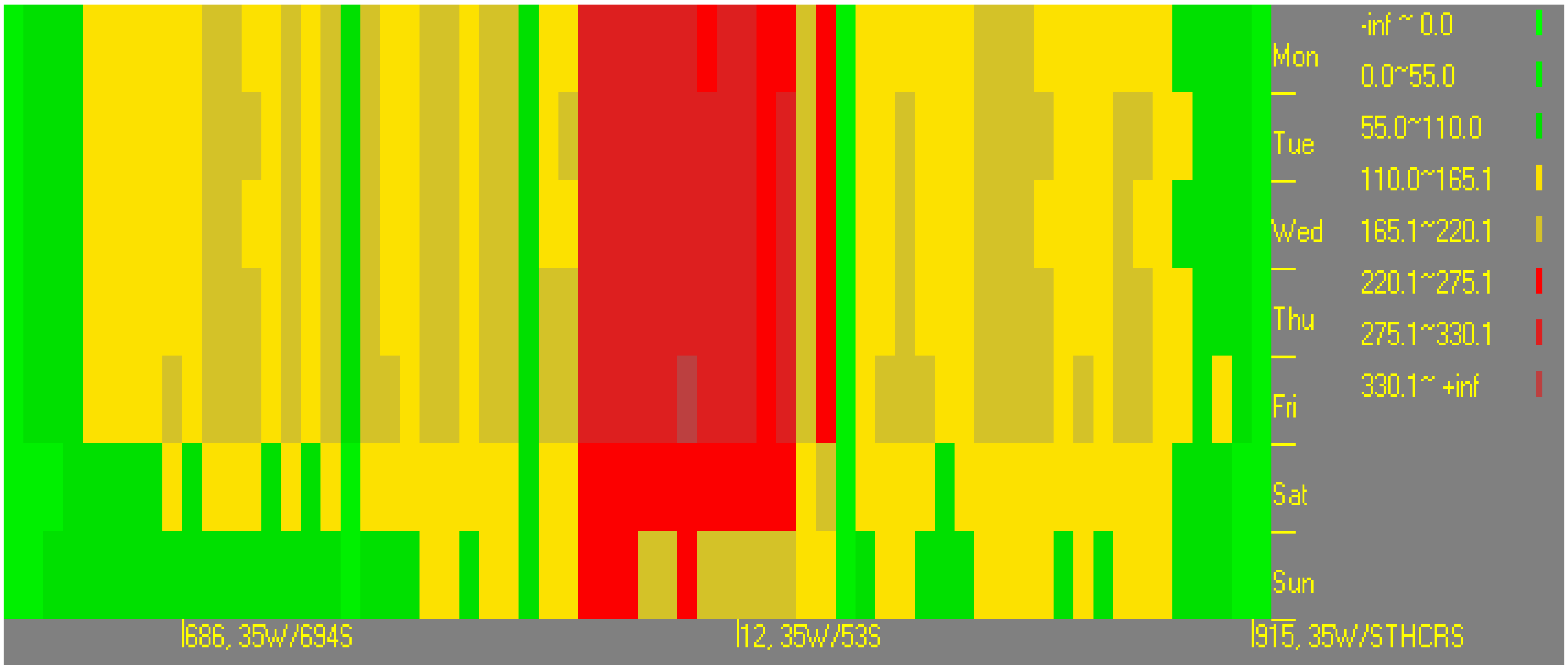
Trends:

- 3-Cluster
  - North section:Evening rush hour
  - Downtown area: All day rush hour
  - South section:Morning rush hour
- S-Outliers
  - station ranked 9th
  - Time: 2:35pm
- Missing Data

Dimension Pair: T_{DW}-S

Configuration:

- X-axis: stations; Y-axis: day of week
- f(x,y): Avg. volume over all stations for Jan-Mar 1997

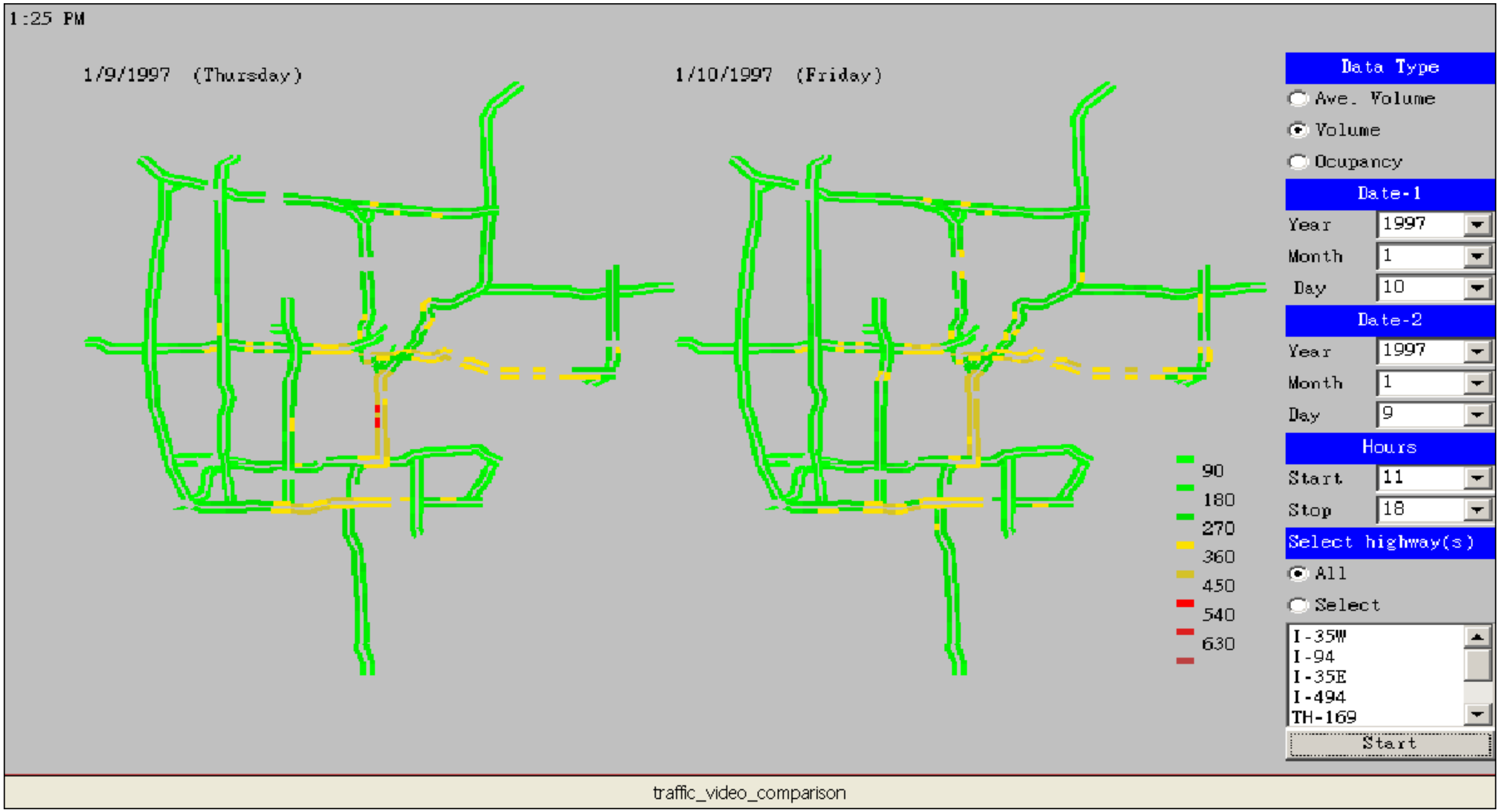


Trends:

- Busiest segment of I-35 SW is b/w Downtown MPLS & I-62
- Saturday has more traffic than Sunday
- Outliers – highway branch

Triplet: $T_{TD}T_{DW}S$: Compare Traffic Videos

Configuration: Traffic volume on Jan 9 (Th) and 10 (F), 1997



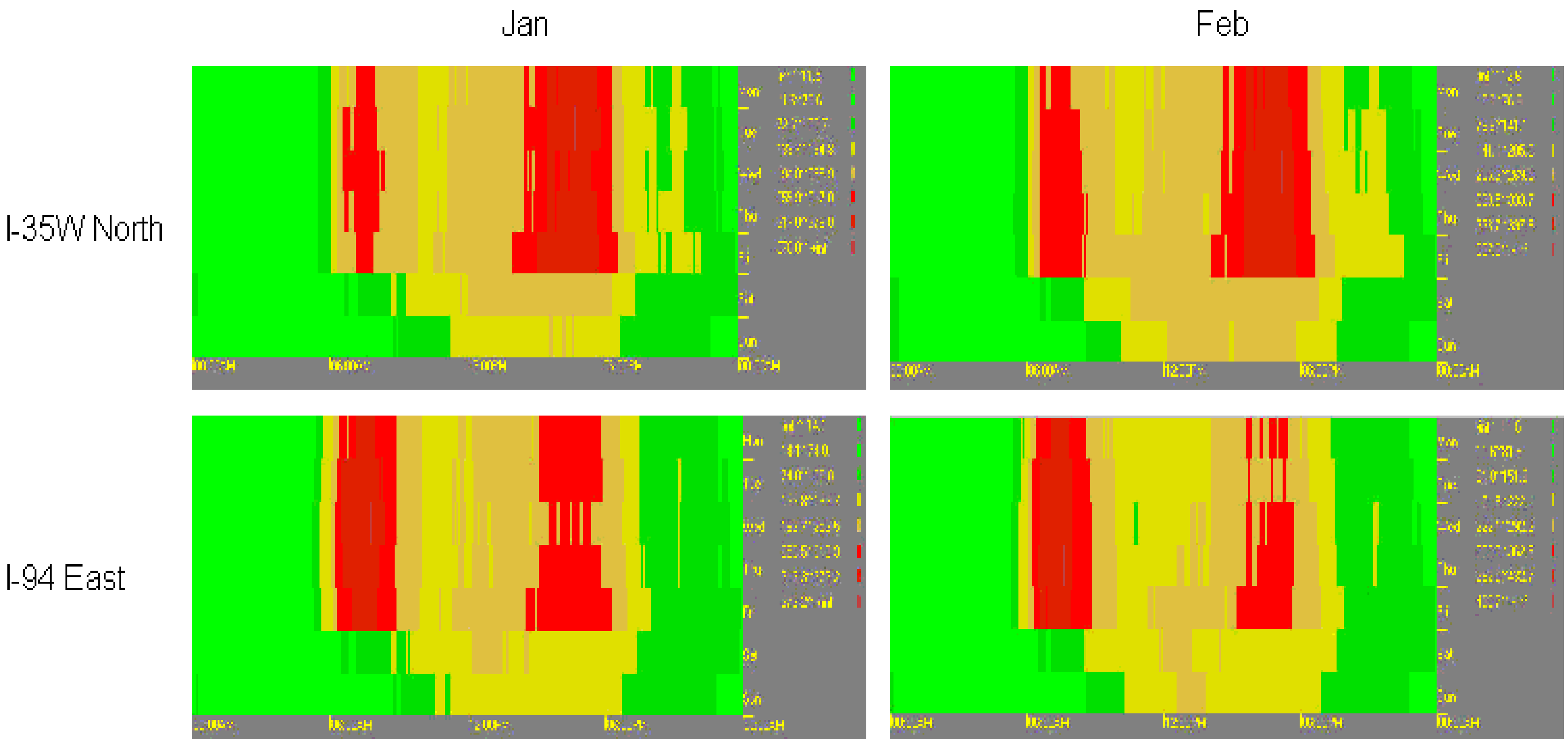
Trends:

- Evening rush hour starts earlier on Friday
- Congested segments: I-35W (downtown Mpls – I-62); I-94 (Mpls – St. Paul); I-494 (intersection I-35W)

Size 4 Subset: $T_{TD}T_{DW}T_{MY}S$(Album)

Configuration:

- Outer: X-axis (month of year); Y-axis (highway)
- Inner: X-axis (time of day); Y-axis (day of week)



Trends:

- Morning rush hour: I-94 East longer than I-35 W North
- Evening rush hour: I-35W North longer than I-94 East
- Evening rush hour on I-94 East: Jan longer than Feb

Outline

- Transportation domain
- A transportation dataset
- Data mining issues
  - Spatio-temporal networks
  - Spatial outliers
  - Hotspots
  - Co-occurrences
  - Location prediction
- Summary

Data Mining

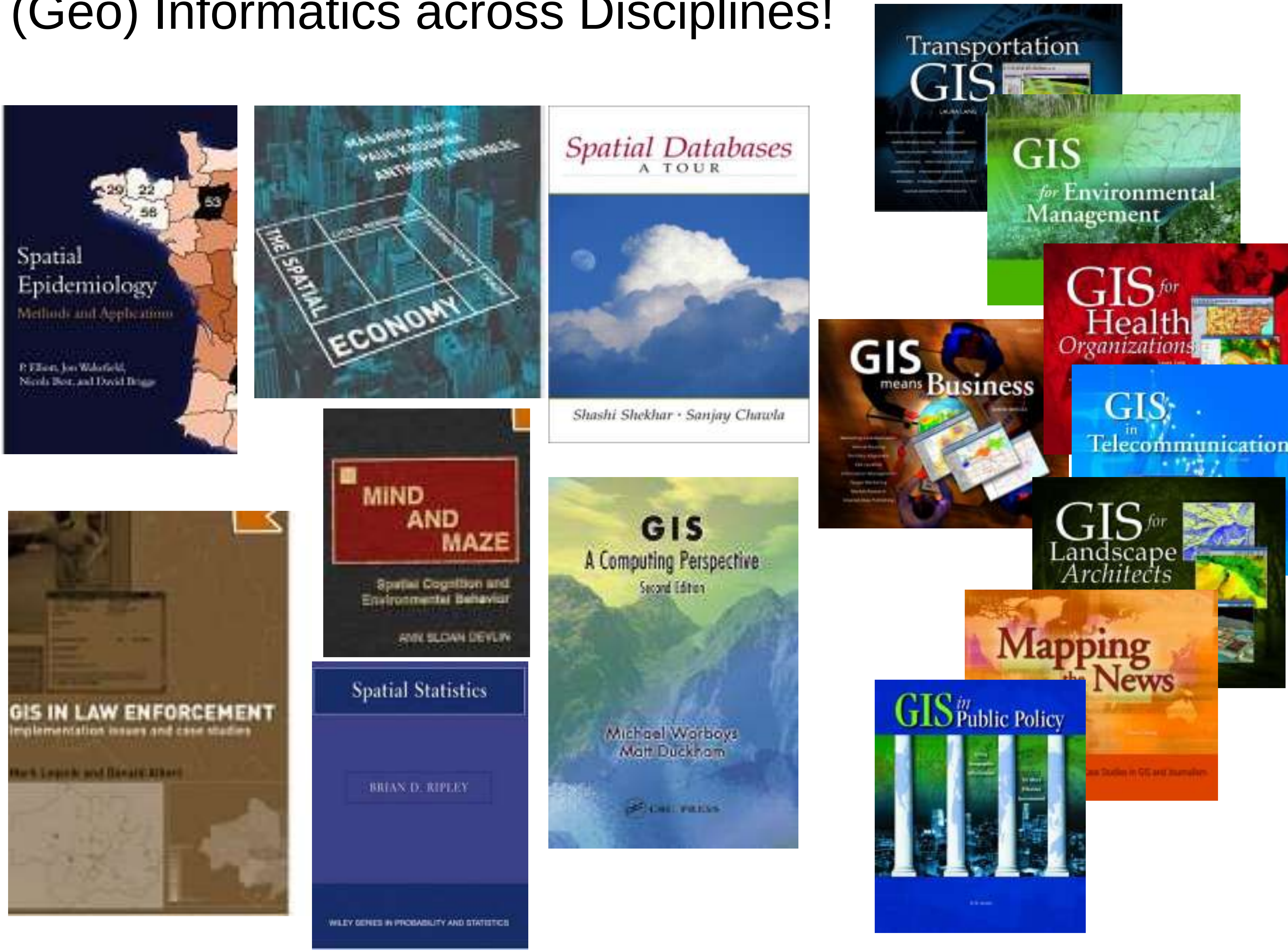
- What is it?
  - Identifying interesting, useful, non-trivial patterns
    - Hot-spots,
  - in large spatial or spatio-temporal datasets
    - Satellite imagery, geo-referenced data, e.g. census
    - gps-tracks, geo-sensor network, …

- Why is it important ?
  - Potential of discoveries and insights to improve human lives
    - Environment: How is Earth system changing? Consequences for humans?
    - Public safety: Where are hotspots of crime? Why?
    - Public health: Where are cancer clusters? Environmental reasons?
    - Transportation, National Security, …
  - However, (d/dt) (Spatial Data Volume) >> (d/dt) (Number of Human Analysts)
    - Need automated methods to mine patterns from spatial data
    - Need tools to amplify human capabilities to analyze spatial data

Transportation Data Mining: Some Challenges

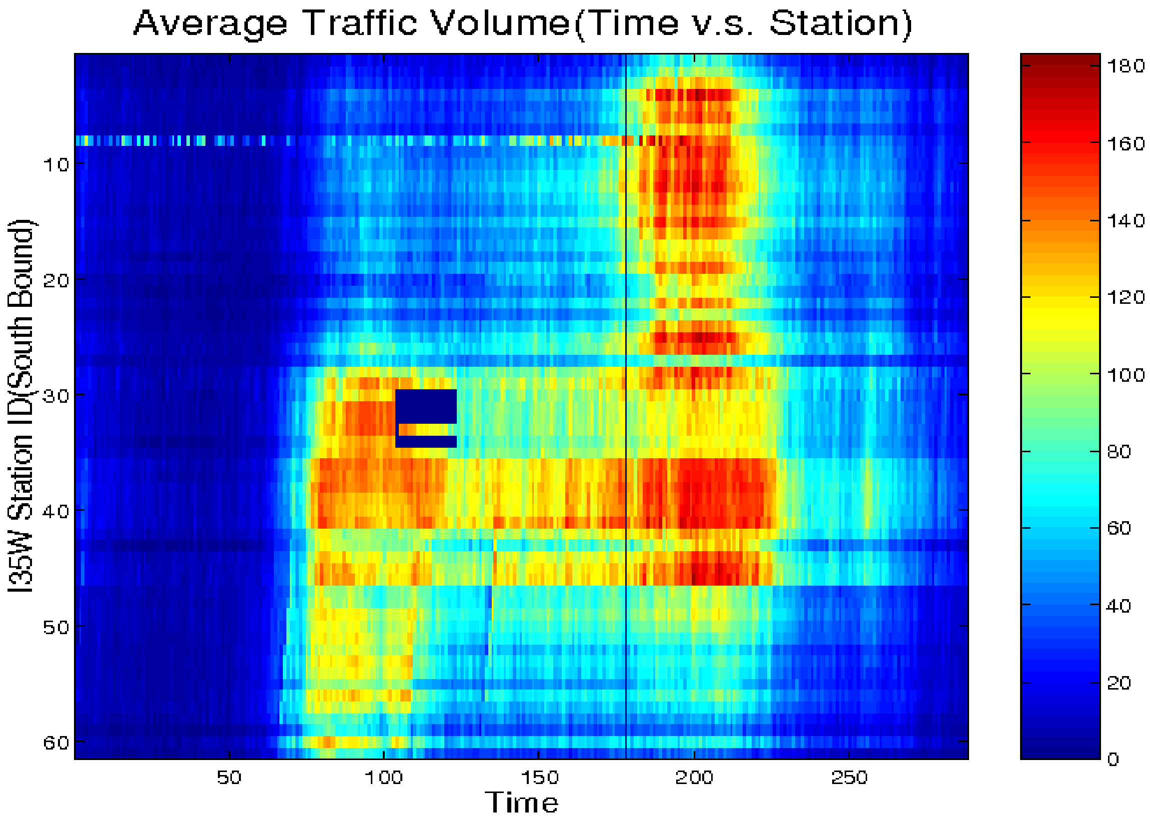
- Violates assumptions of classical data mining
  - Lack of independence among samples - ? Decision trees, …
  - No natural transactions -? Association rule, …
- Two kinds of spaces
  - Embedding space, e.g. Geography, Network, Time
  - Feature space, e.g. Traffic volume, accidents, …
- Lessons from Spatial thinking
  - 1^{st} Law: Auto-correlation: Nearby things are related
  - Heterogeneity
  - Edge effect
  - …

(Geo) Informatics across Disciplines!



Example 1: Spatial Anomalies

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



Global vs. Spatial outliers (SIGKDD 2001)

Spatial outlier

A data point that is extreme relative to it neighbors

Given

A spatial graph G={V,E}

A neighbor relationship (K neighbors)

An attribute function f: V -> R

Test T for spatial outliers

Find

$O = \{v_i \mid v_i \in V, v_i \text{ is a spatial outlier}\}$

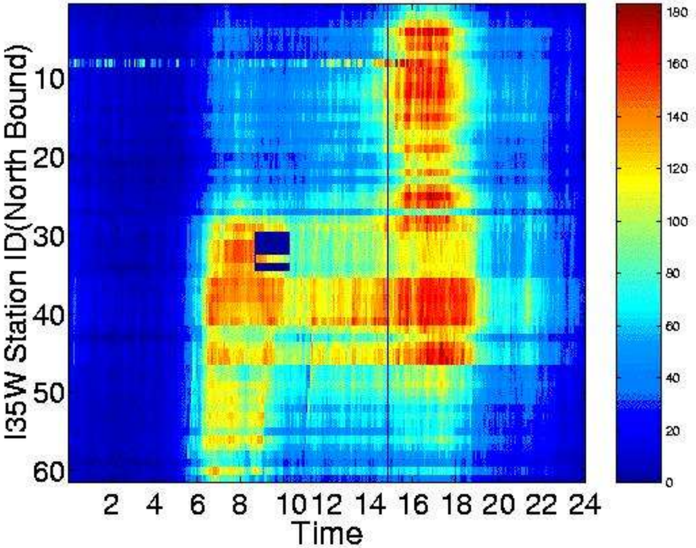
Objective

Correctness, Computational efficiency

Constraints

Test T is an algebraic aggregate function

Average Traffic Volume(Time v.s. Station)



Spatial outlier detection

Spatial outlier and its neighbors

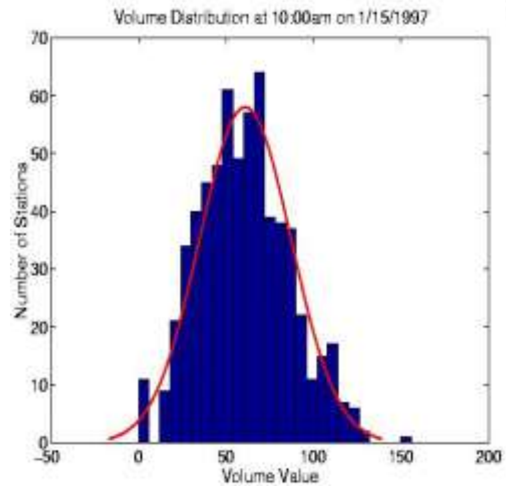
1. Choice of Spatial Statistic

   $S(x) = [f(x) – E_{y \in N(x)}(f(y))]$

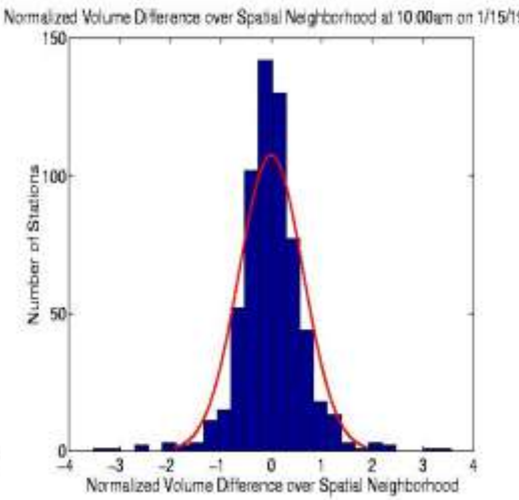
   Theorem: S(x) is normally distributed if f(x) is normally distributed

2. Test for Outlier Detection

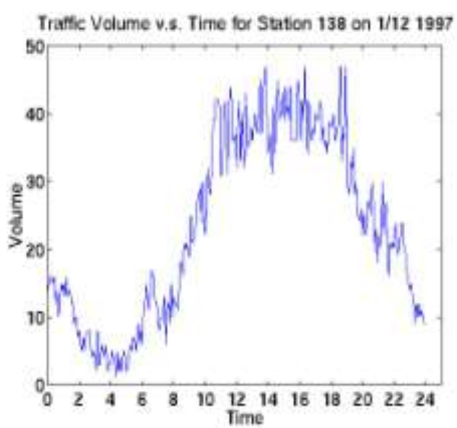
   $| (S(x) - \mu_s) / \sigma_s | > \theta$

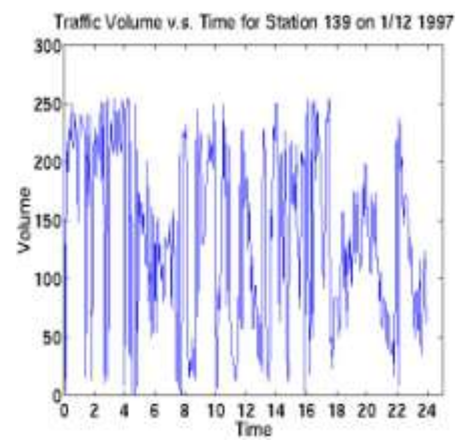


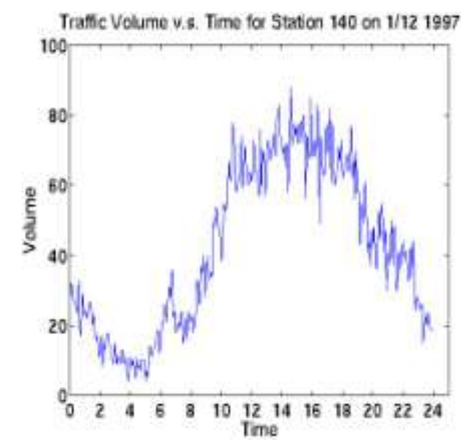
f(x)



S(x)

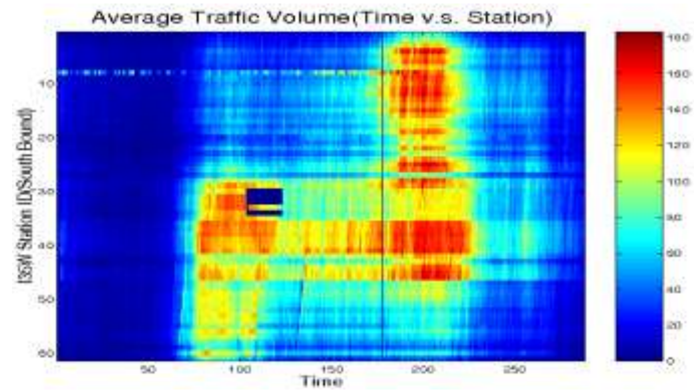


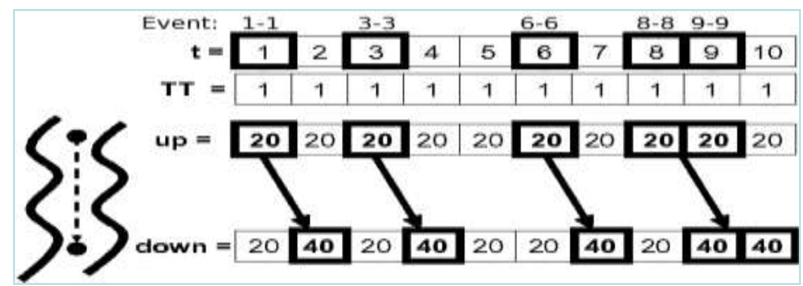


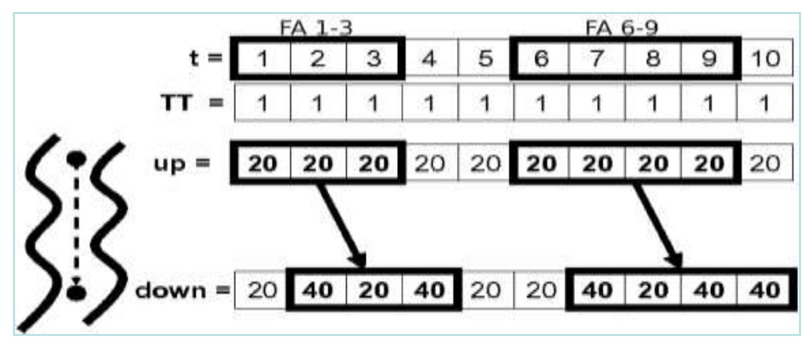


Spatial/Spatio-temporal Outliers Challenges

- 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

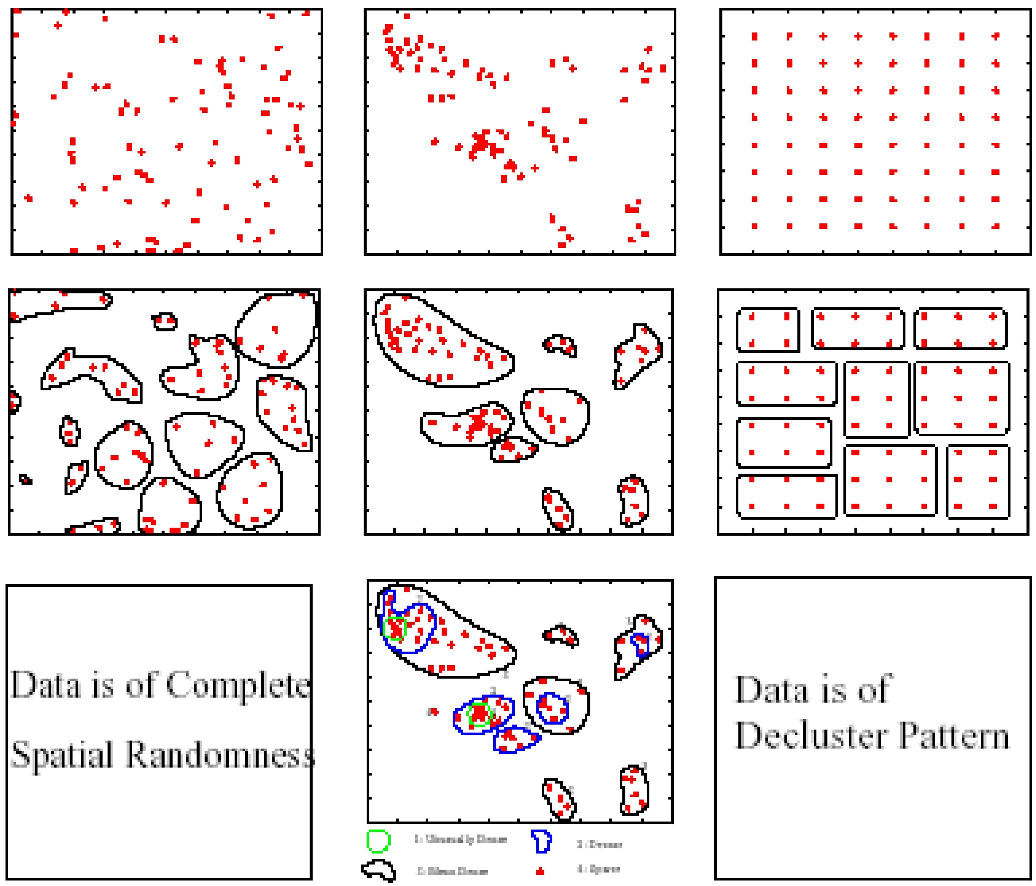






Example 2: Hotspots

- Is classical clustering (e.g. K-mean) effective?



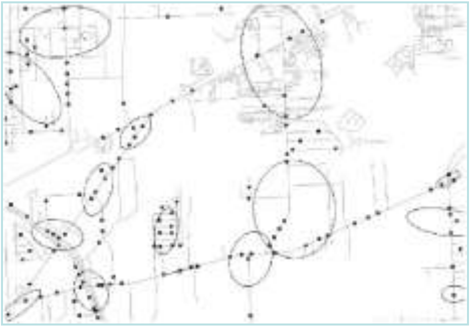
Inputs: locations of potholes, accidents, sensors

Outputs of K-mean Clustering

Spatial Statistical view

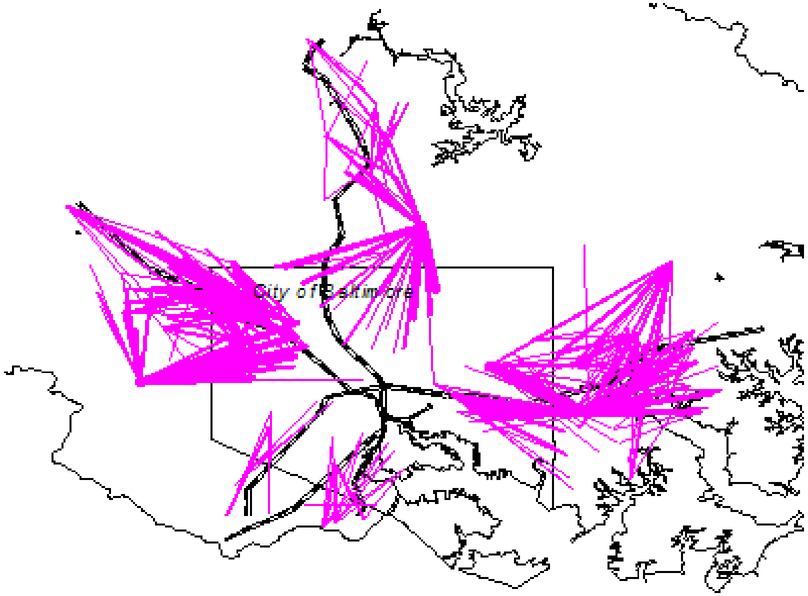
HotSpots

- What is it?
  - Unusally 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

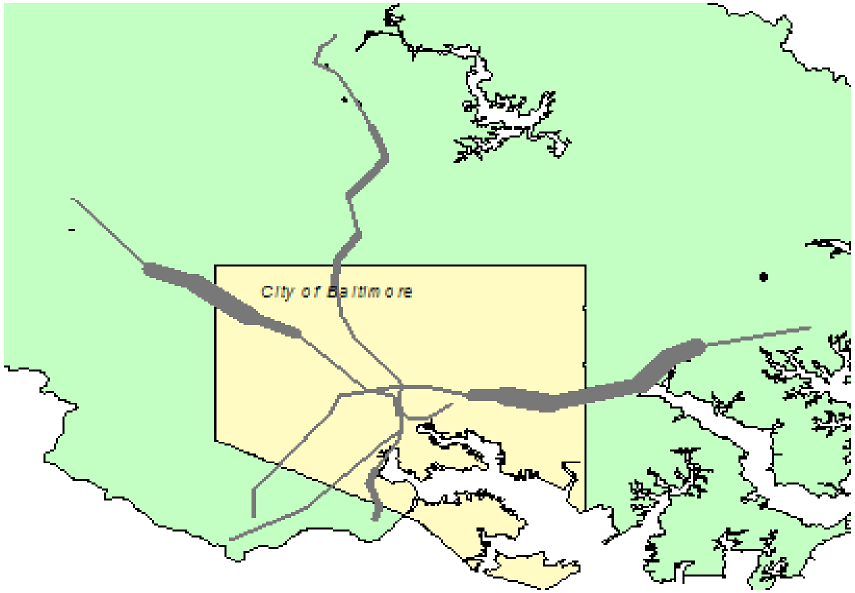


Network Semantics: Implicit Routes

- Complicated Feature
  - Urban environment
  - Transportation Networks
- Patterns
  - Journey to crime
  - Network based explanation



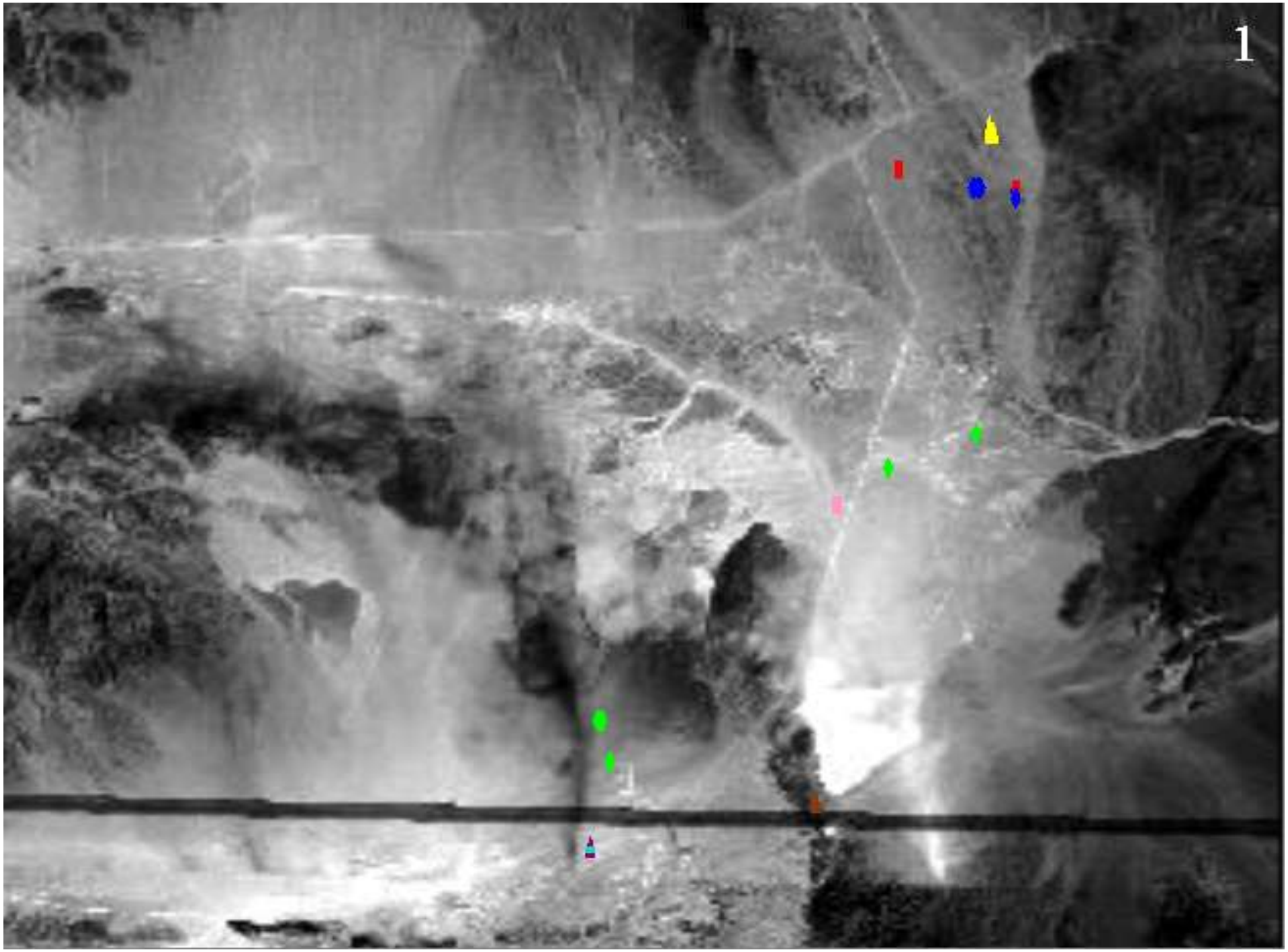
(a) Input: Pink lines connect crime location & criminal's residence



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

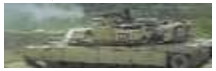
Example 3b: Associations

- Given a set of tracks of different types, can association mining find subset of types that often move together?



Manpack stinger
(2 Objects)

M1A1_tank
(3 Objects)



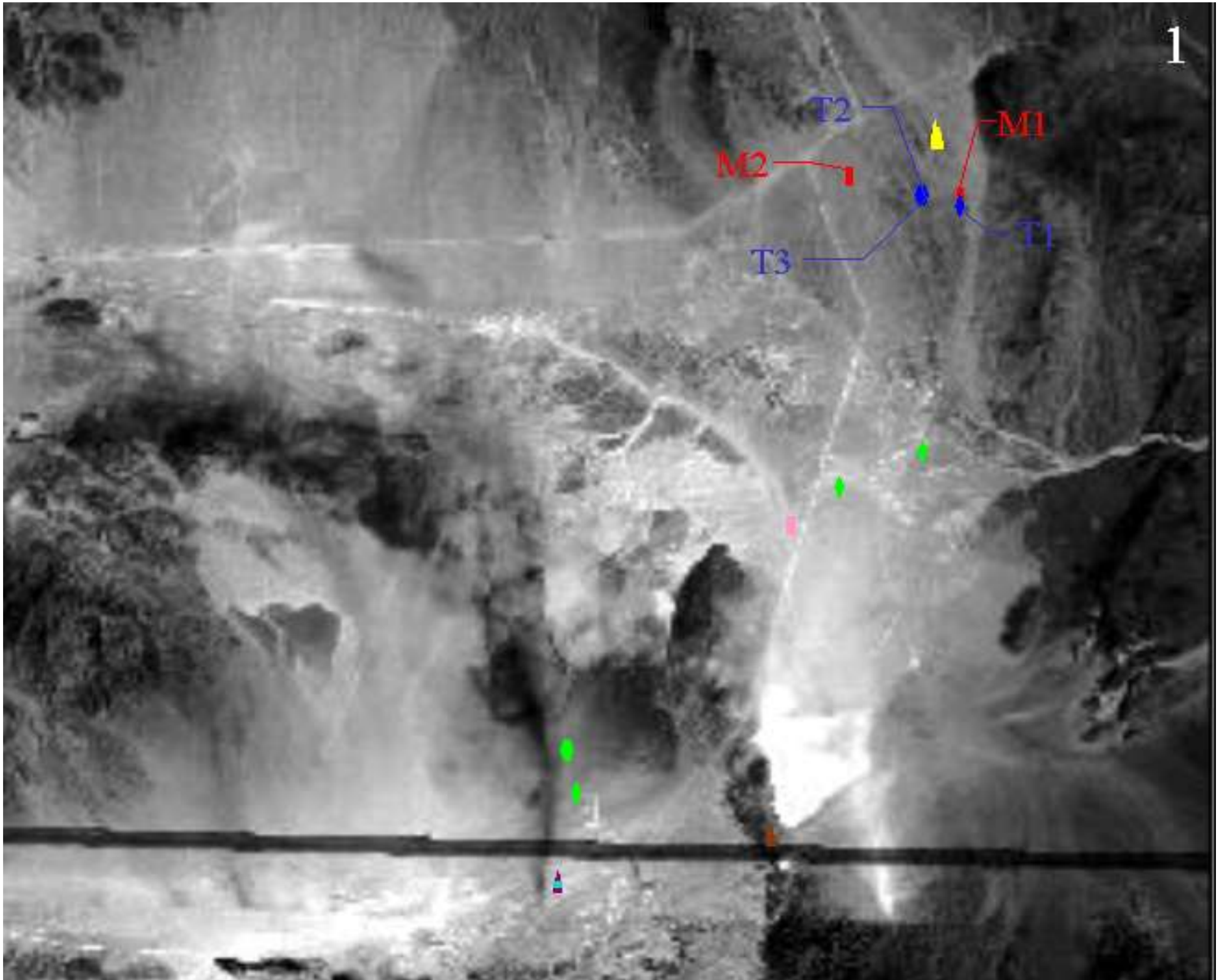
M2_IFV
(3 Objects)

Field_Marker
(6 Objects)

T80_tank
(2 Objects)

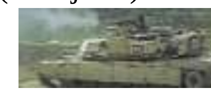
BRDM_AT5
(enemy) (1 Object)

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)

Challenge: 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!**

Co-location Patterns (SSTD 2001, TKDE 2004)

	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. Spatial Statistical Interpretation

   Related to Ripley's K-function in Spatial Statistics

…

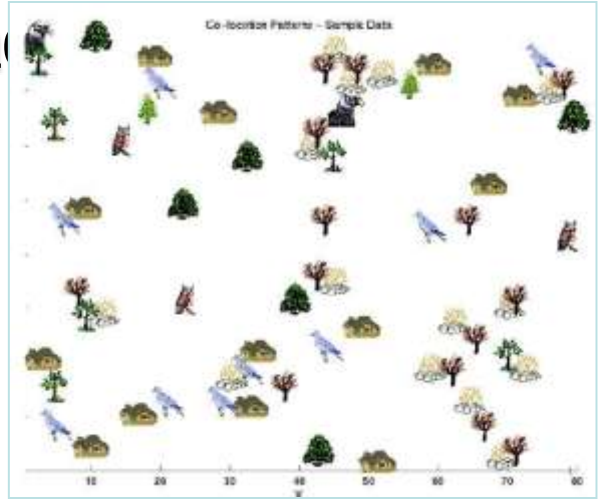
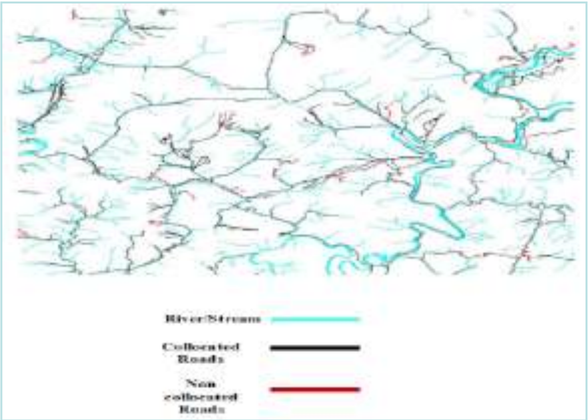
Spatio-temporal Association: Cascade Patterns

- Time Geography theory
  - Processes = a collection of events
  - Events
    - Have specific endpoint
    - (Partially) ordered by time-footprints
- Instance level model
  - Nodes = instances of events
  - Edges = spatio-temporal neighbors
    - Direction defined by time-footprints
- Cascade Patterns = Schema-level summary
  - Nodes = Event-types (ET)
  - Edge(ET1, ET2, N) =>N compatible edges at instance level
  - Cycles are possible, e.g. ST overlapping processes

- Similar to Graphical Models, Bayesian Networks, Graph mining…
  - Simpler interest measure, e.g. Pr(Pattern P | an event instance)
  - Cheaper than joint probability distribution, max. independent set
  - Computationally more scalable

Colocation, Co-occurrence, Intera

- 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

Example 4: Spatio-temporal Prediction

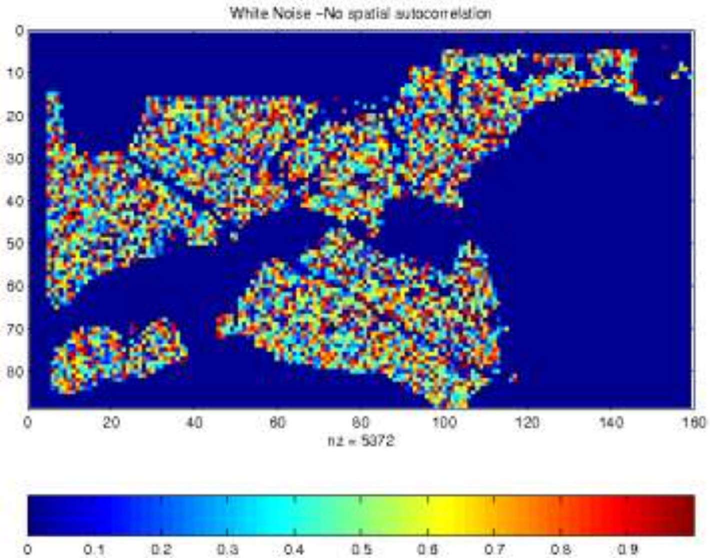
- Transportation Planning
  - What will be the impact of a new office building?
  - What will be travel demand? future bottlenecks?
  - What will be the effect of hybrid cars on traffic?
  - How will better bicycle facility impact vehicle traffic?

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

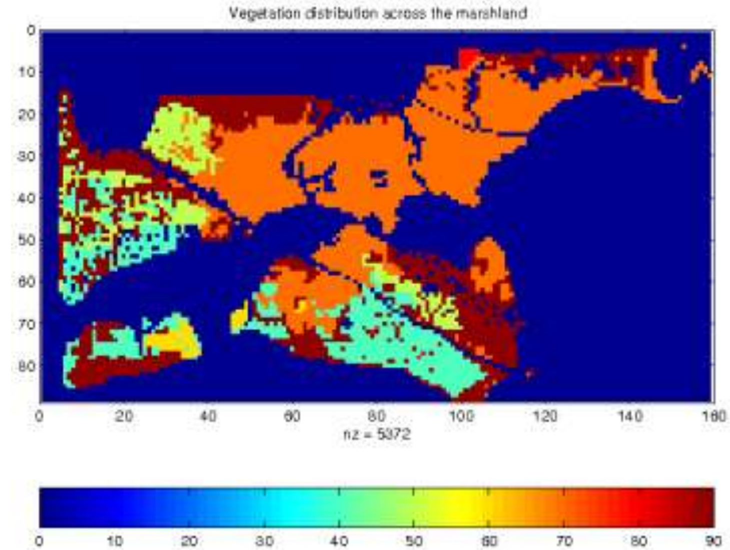
- Challenges
  - Spatio-temporal auto-correlation – violates independence assumption
  - Network : routes, edge capacities, …
  - Individual behavior: urban sprawl?
  - Group dynamics: game theory, Wardrop equilibrium, …

Autocorrelation

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



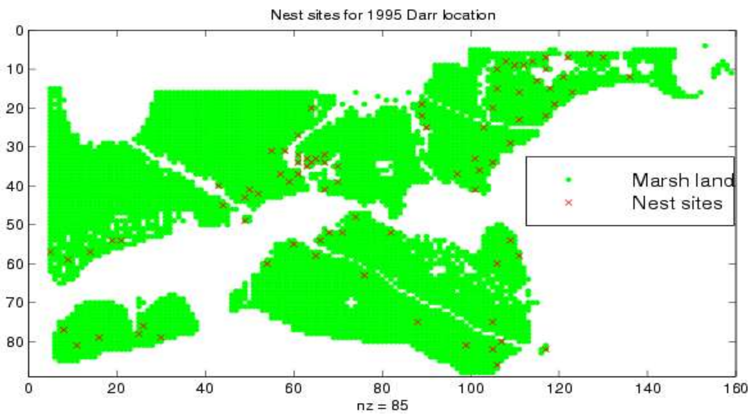
Pixel property with independent identical distribution



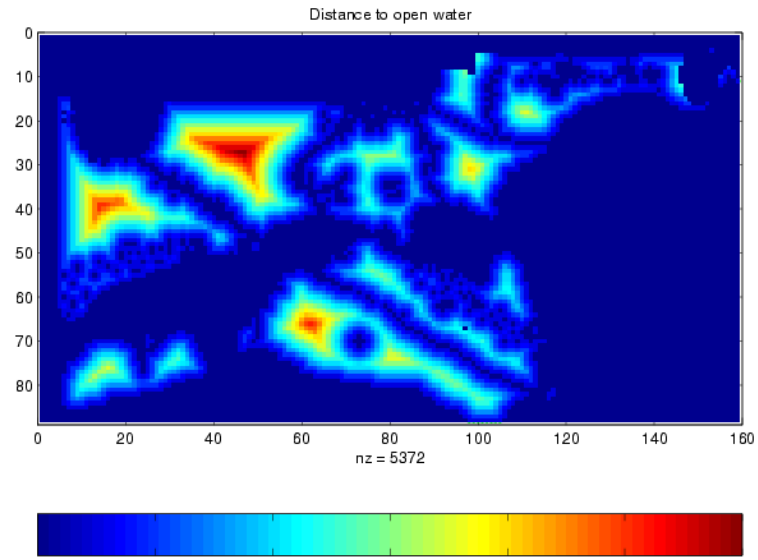
Vegetation Durability with SA

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

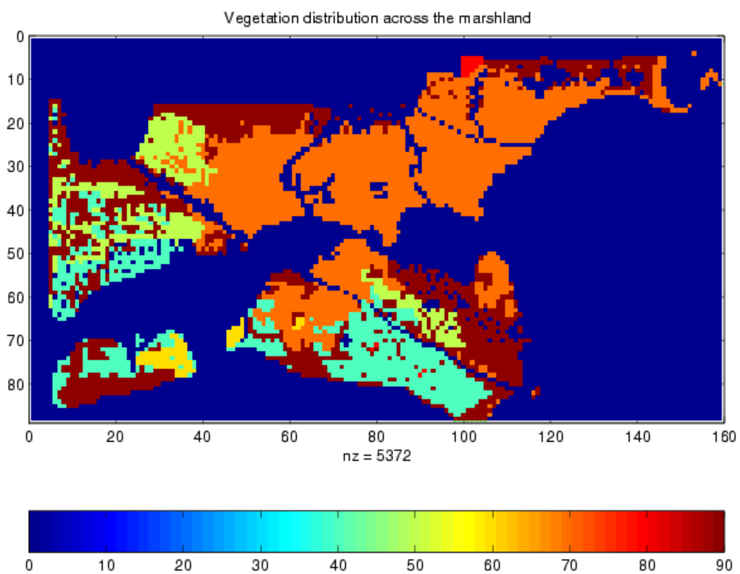
Challenge 1: Is I.I.D. assumption valid?



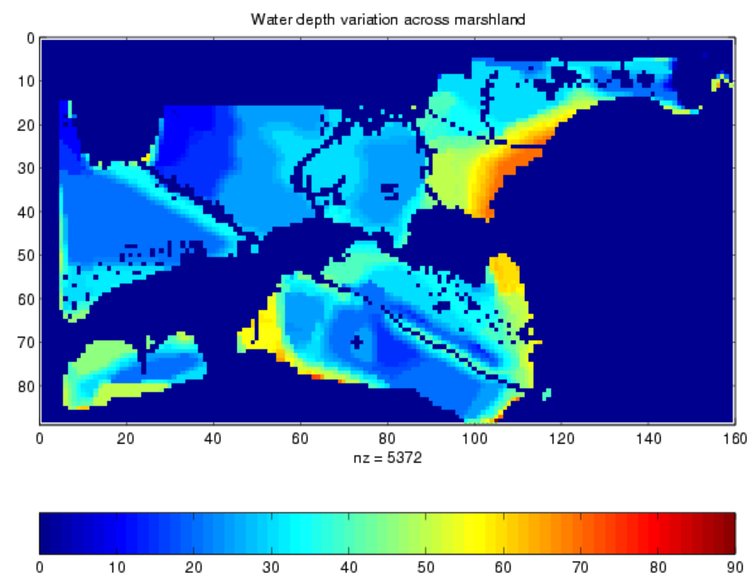
Nest locations



Distance to open water



Vegetation durability



Water depth

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

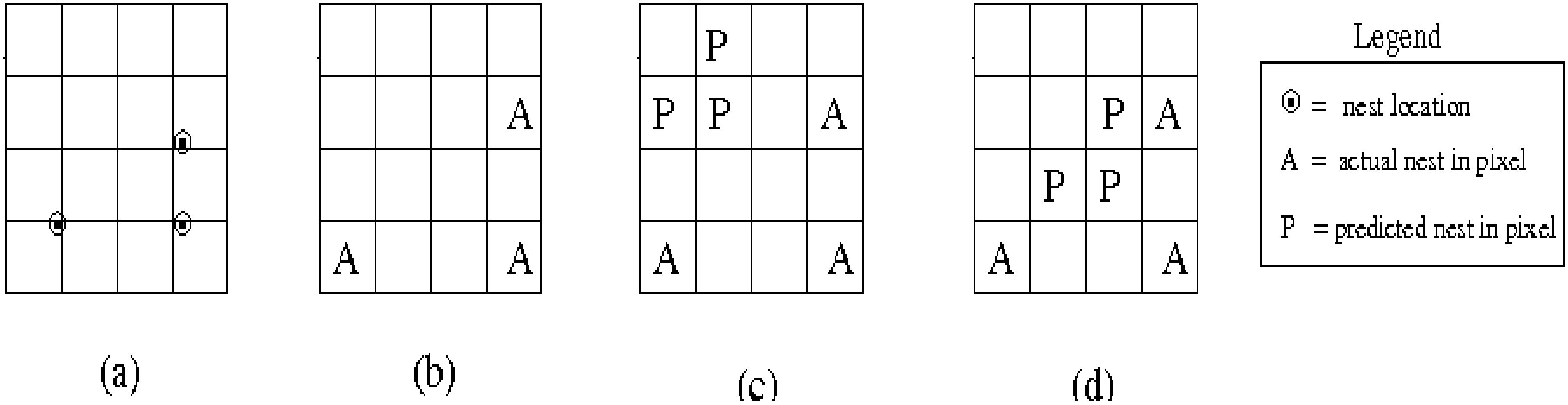
<u>**Computational Challenge**</u>:

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

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

Research Needs in Location Prediction

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



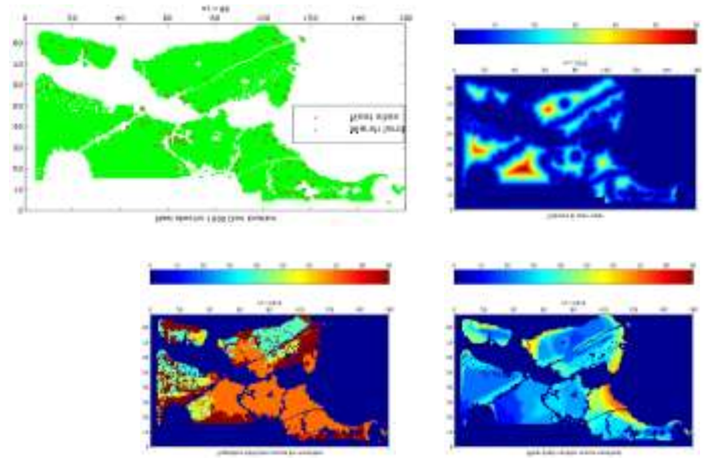
Actual Sites

Pixels with actual sites

Prediction 1

Prediction 2. Spatially more accurate than Prediction 1

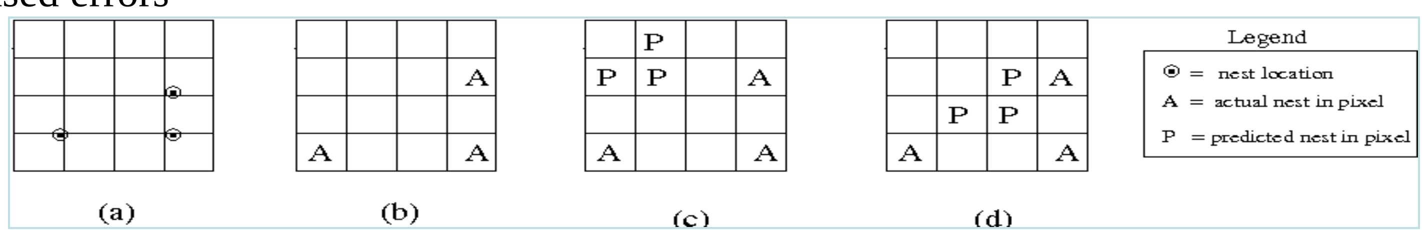
Space/Time Prediction



- 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

$$\boxed{\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$$

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



Outline

- Transportation domain
- A transportation dataset
- Data mining issues
  - Spatio-temporal networks
  - Spatial outliers
  - Hotspots
  - Co-occurrences
  - Location prediction
- Summary

Data Mining Challenges in Transportation

- Identify Limitations of Transportation Knowledge
  - Calibration of simulation parameters, e.g.
    - Day-time population distribution, traffic distribution
  - Non-equilibrium dynamics over space and time
  - Extreme events, e.g. evacuation, conventions, …

- Articulate value of data mining (DM)
  - Value of novel data sets
    - Lab.-based vs. on-road emissions or mpg
    - Context – weather, ambient temperature, vehicle to vehicle
    - Simulator estimated routes vs. gps-tracks
    - Volunteer information – pot-holes, speed, …
  - Value of novel data analysis or visualization techniques
    - anomalies

- Evaluate and evolve current DM
  - May current DM deliver value?
  - Are assumption of classical DM reasonable?
  - How can be improve current DM technique?

Data Mining and Transportation

- Potential value of data mining in transportation
  - Data driven discoveries to complement model driven ones
  - Hypothesis generation to complement hypothesis testing
  - Computational scalability
  - Conceptual scalability – models of gps-tracks
  - Which problems ?
    - Extreme events, …

- Potential value of transportation to data mining
  - Expose limitations, e.g. independence assumption
  - New challenges: e.g. spatio-temporal networks, …
    - New pattern families