Spatio-Temporal Data Mining (Part I)

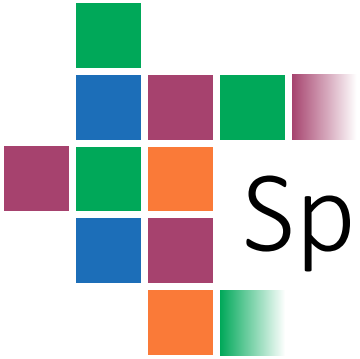
Donato Malerba and Annalisa Appice

Goals (Part I)

- We will
  - Introduce spatio-temporal data mining
  - Explain the main challenges
    - Complexity of input data
    - Correlation
  - Provide pointers to relevant literature

Outline (Part I)

- Spatial modeling, temporal modeling, ST modeling
- Spatial relationships
- Cross- and auto-correlation
- Spatial dependence
- Statistical approach
- Relational approach
- STDM Tasks (just an introduction, Part 2 follows)



Spatio-temporal Data Mining

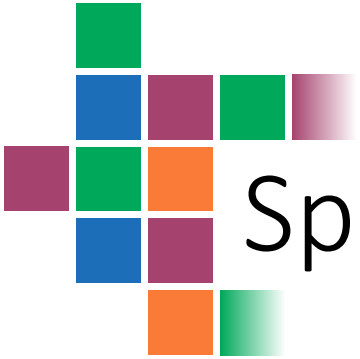
- STDM is the process of discovering interesting and previously unknown, but potentially useful patterns from spatio-temporal data.

Input data makes the difference w.r.t. classical data mining approaches!

Spatio-temporal Data Mining

A spatio-temporal object is characterized by at least:

- One spatial property
- One temporal property



Spatio-temporal Data Mining

- Complexity of ST objects is one of the main challenges of STDM.
- Spatio-temporal data are complex because they include discrete representations of continuous space and time.

Modeling Spatial Information

Two major approaches to conceptual modeling of space:

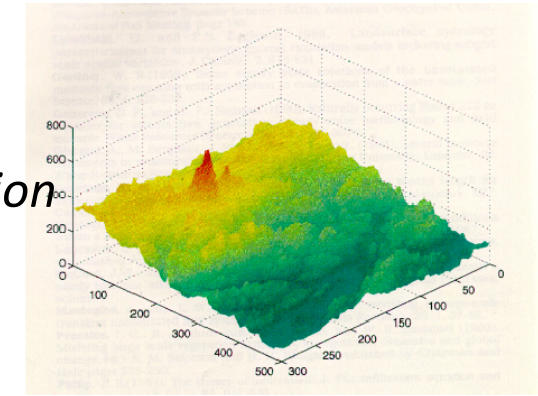
- Field-based model
- Object-based model

Field-based Model

- The world is seen as a continuous surface over which features vary.
- Spatial variation is defined by a number of Field Functions:

  $f: R^n \rightarrow$ *Attribute Domain*

  - *Examples: elevation, temperature, precipitation*



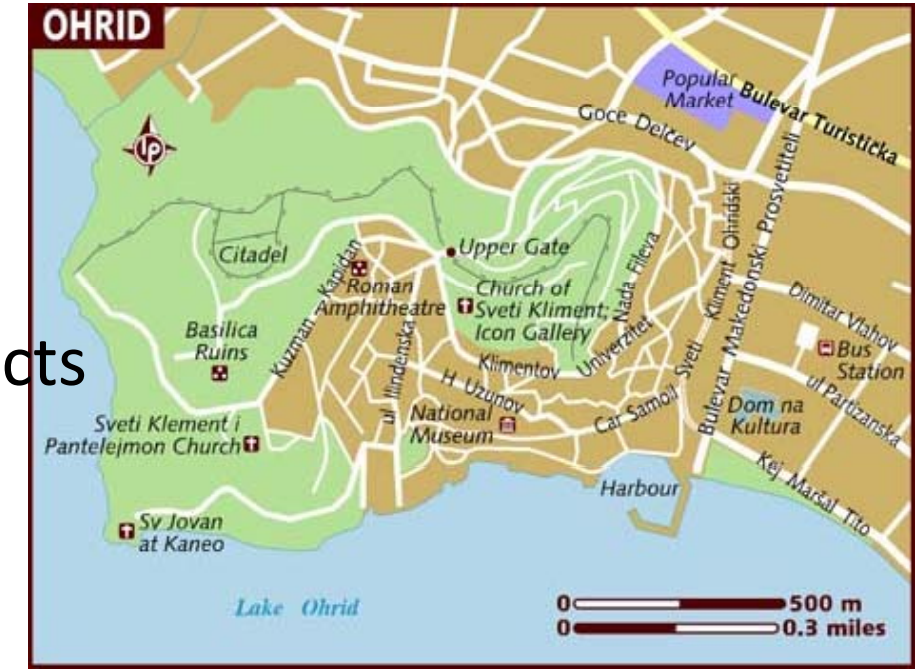
- Field Operations
  - Examples, addition(+) and composition(o).

$$f + g : x \rightarrow f(x) + g(x)$$

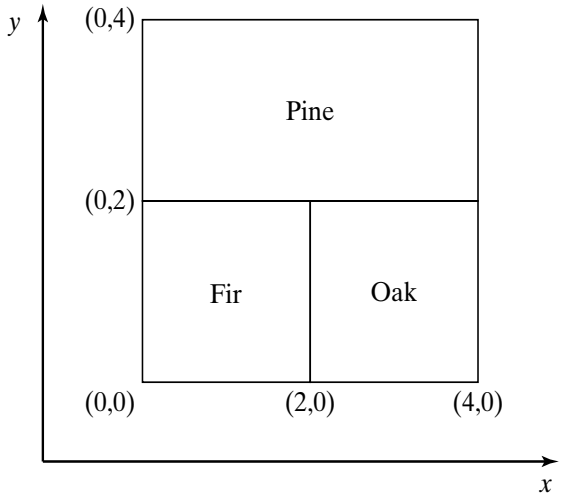
$$f \circ g : x \rightarrow f(g(x))$$

Object-based Model

- The world is seen as a surface littered with distinct, identifiable and relevant things or entities, called objects, which exist independent of their locations.
- Objects can be:
  - Zero-dimensional or punctual
  - One-dimensional or linear
  - Two-dimensional or surfacic
- Operations on spatial objects
  - Topological
  - Directional
  - Distance-based



Field-based vs. Object-based



(a)

Object Viewpoint of Forest Stands

Area-ID	Dominant Tree Species	Area/Boundary
FS1	Pine	[(0,2),(4,2),(4,4),(0,4)]
FS2	Fir	[(0,0),(2,0),(2,2),(0,2)]
FS3	Oak	[(2,0),(4,0),(4,2),(2,2)]

(b)

Field Viewpoint of Forest Stands

$$f(x,y) = \begin{cases} \text{"Pine,"} & 2 \leq x \leq 4\,;\, 2 < y \leq 4 \\ \text{"Fir,"} & 0 \leq x \leq 2;\, 0 \leq y \leq 2 \\ \text{"Oak,"} & 2 < x \leq 4;\, 0 \leq y \leq 2 \end{cases}$$

(c)

Modeling Temporal Information

Two major approaches to conceptual modeling of time:

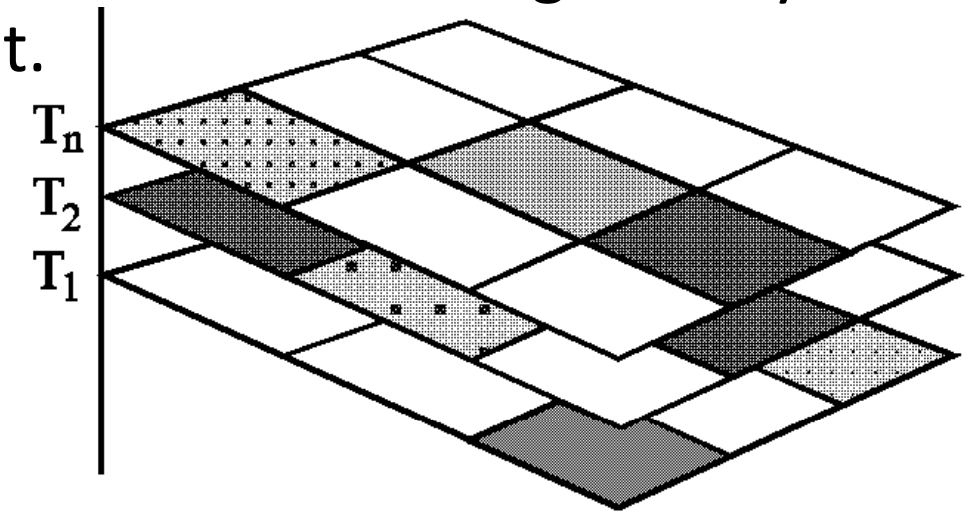
- Temporal snapshots (time series)
- Temporal change (delta/derivatives)

Temporal snapshots

Spatial layers of the same theme are time-stamped.

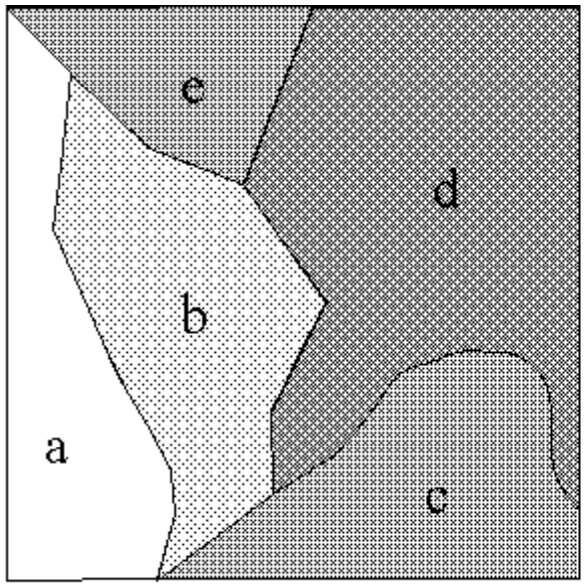
This model shows the states of a geographic distribution at different times without explicit temporal relations among layers.

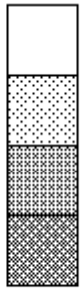
Changes occurring within the time lag of any two layers are not explicit.



Temporal change

A spatial layer at a given start time together with incremental changes occurring afterwards.



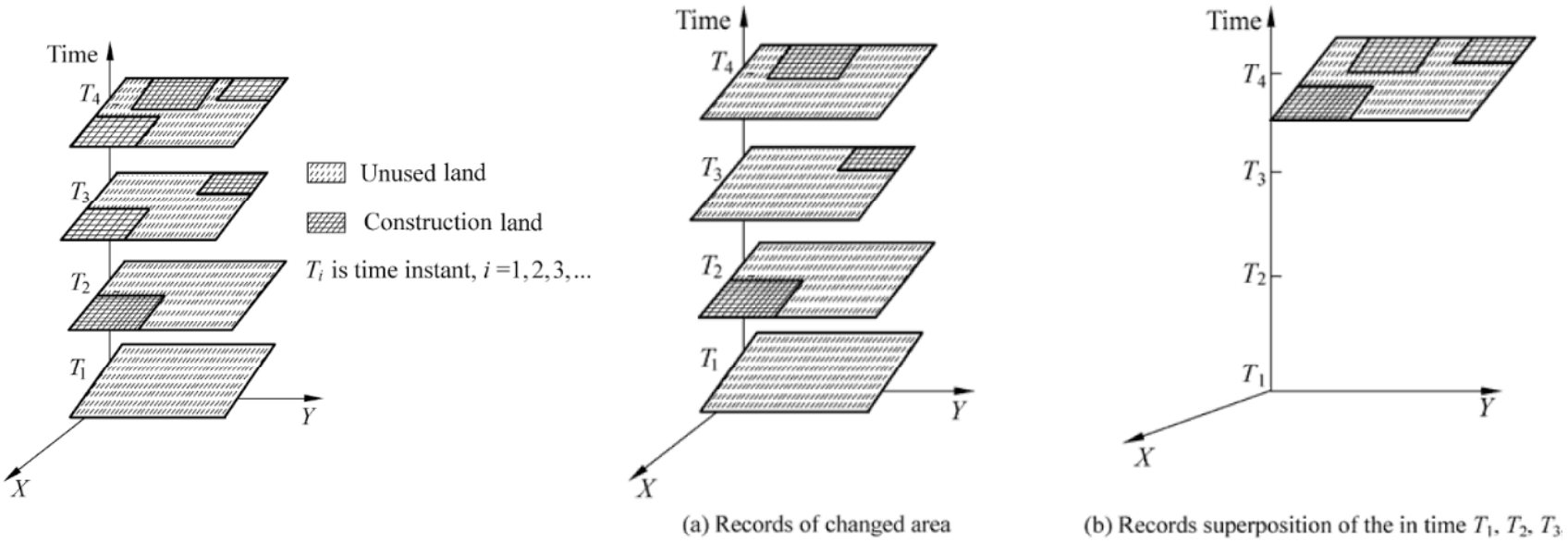


T_1	T_2	T_3	T_4
0	0	0	0
0	1	1	1
0	0	1	1
0	0	0	1

0: Unburned
1: Burned
a, b, c, d, e: Polygon ids

Temporal change

Superposition to get a snapshot view.



(a) Records of changed area

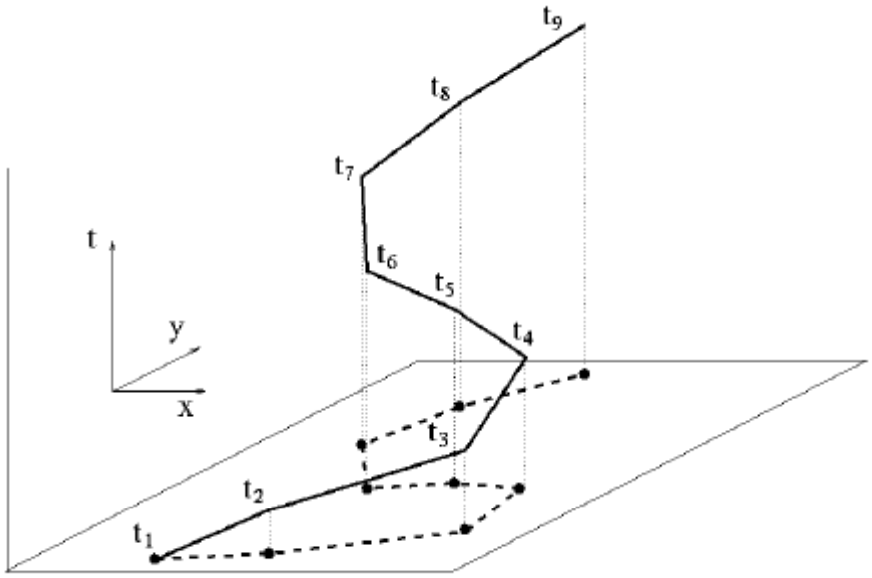
(b) Records superposition of the in time T_1, T_2, T_3

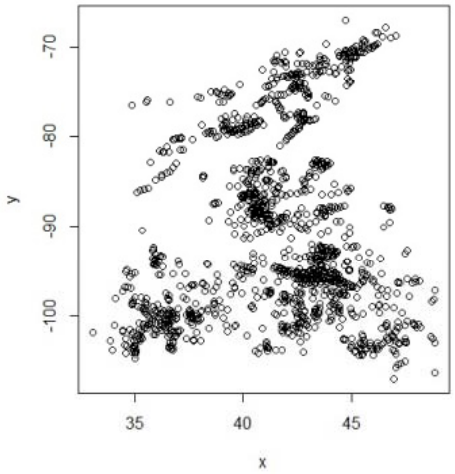
ST data taxonomy (Shekhar et al., 2015)

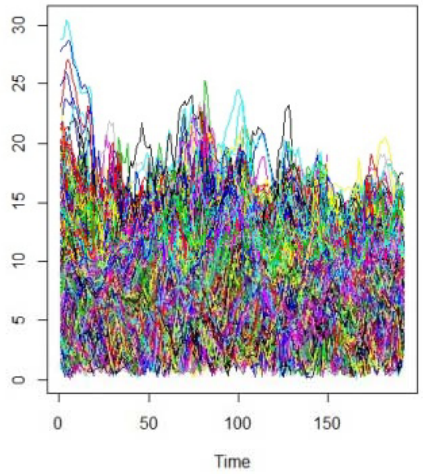
A spatial layer at a given start time together with incremental changes occurring afterwards.

Spatial Data		Temporal Snapshot	Temporal Change
Object Model	points	Point trajectories Spatial time series	Displacement/motion Speed/acceleration
	lines	Line trajectories	Motion, extension/rotation,
			deformation, split/merge
	polygons	Polygon trajectories	Motion, extension/rotation, deformation, split/merge
Field Model	Regular, irregular	Raster time series	Change across raster snapshot

Point trajectory
Spatial time series



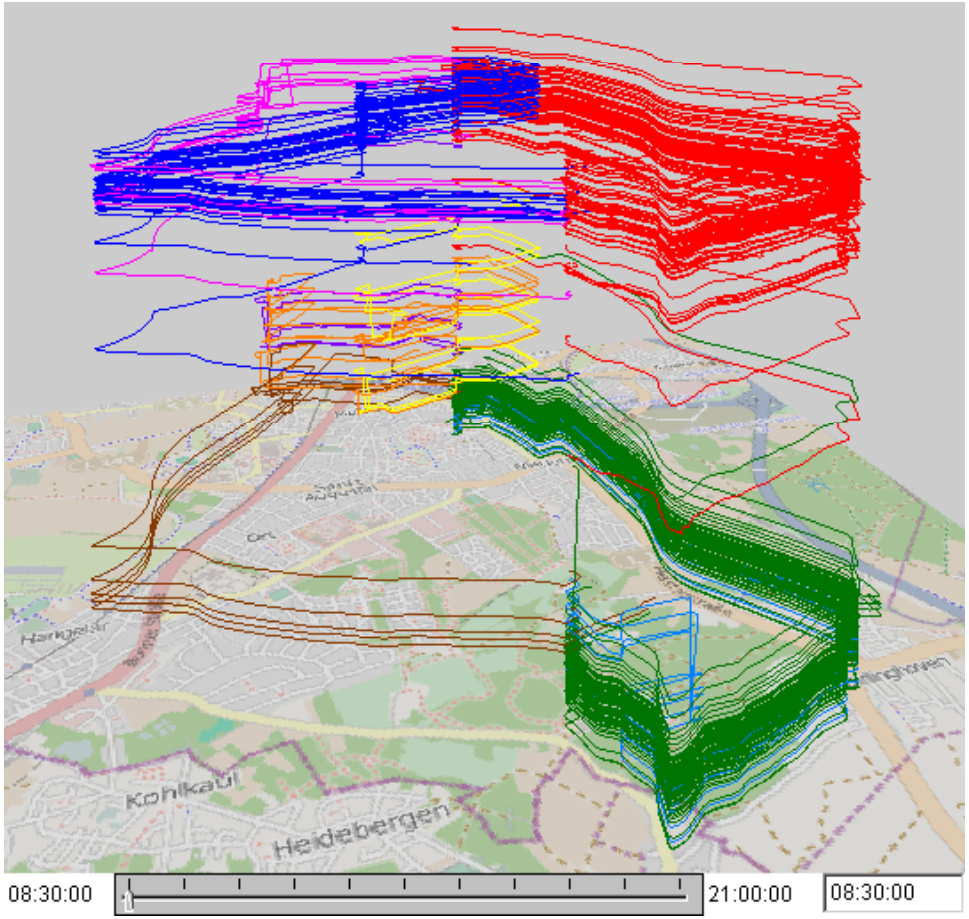




Pfoser et al., 2001.

Pravilovic et al., 2014

Line trajectories

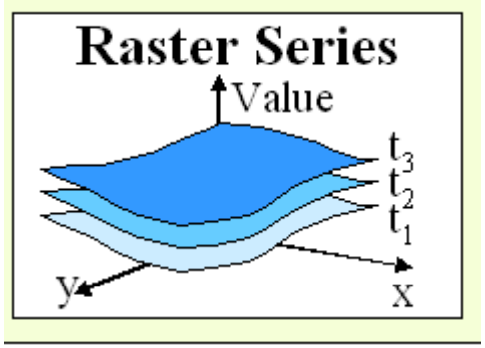


Andrienko et al., 2013

Polygon trajectories

Apeltauer et al., 2015

Raster time series



ST data taxonomy (Shekhar et al., 2015)

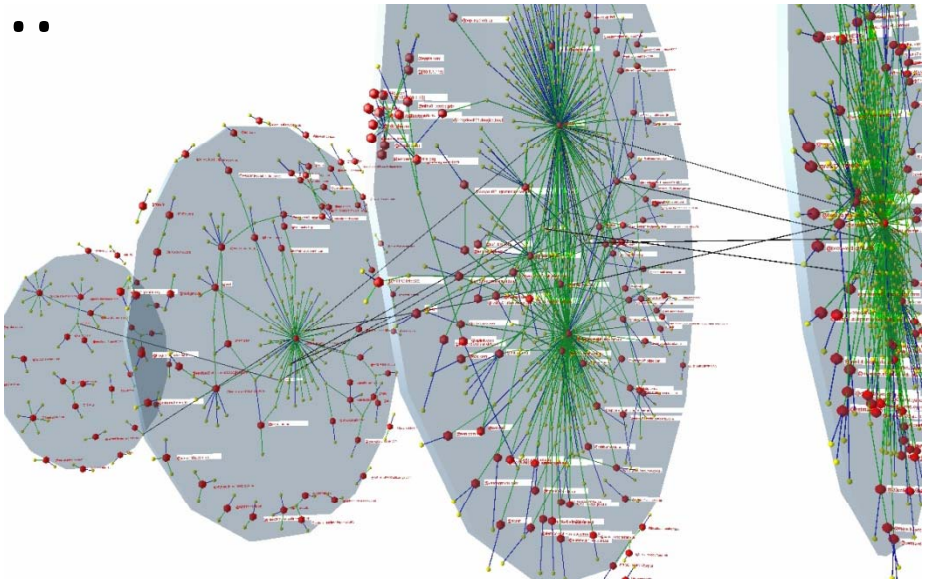
A spatial layer at a given start time together with incremental changes occurring afterwards.

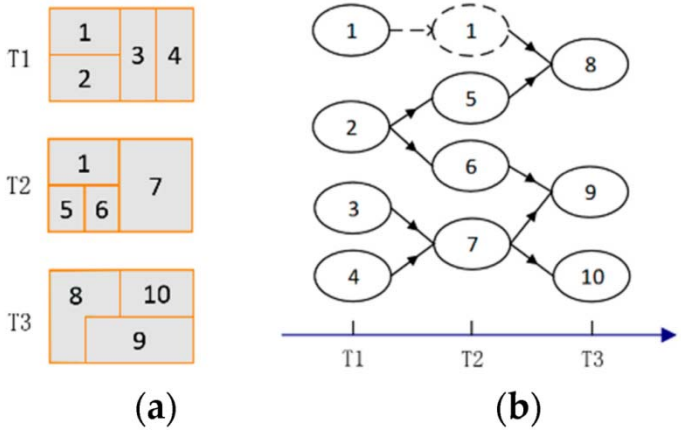
Spatial Data		Temporal Snapshot	Temporal Change
Object Model	points	Point trajectories Spatial time series	Displacement/motion Speed/acceleration
	lines	Line trajectories	Motion, extension/rotation,
			deformation, split/merge
	polygons	Polygon trajectories	Motion, extension/rotation, deformation, split/merge
Field Model	Regular, irregular	Raster time series	Change across raster snapshot

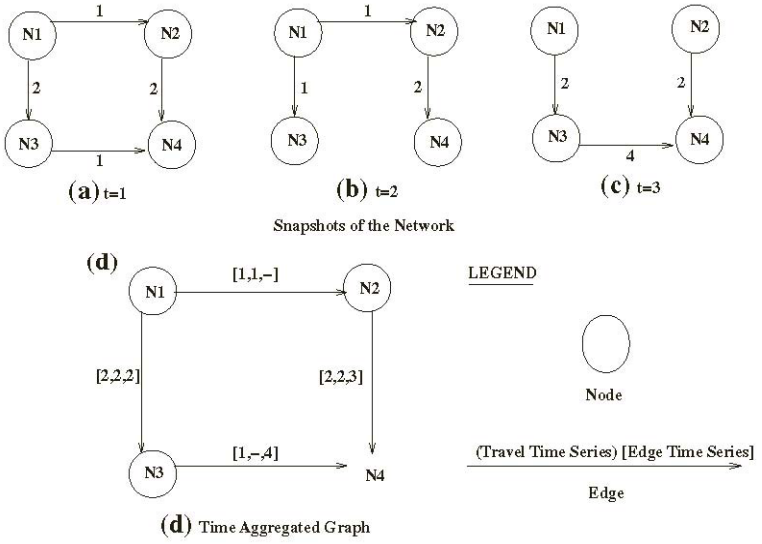
Further types …

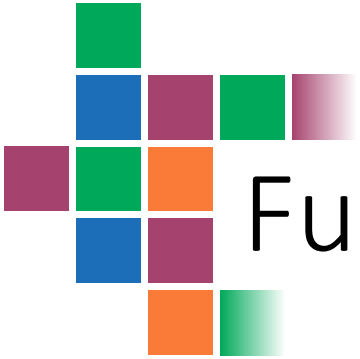
- Spatial network model
  - Data are graphs

  Spatio-temporal data are
  - Time expanded graph
  - Time aggregate graph
  - Network flow







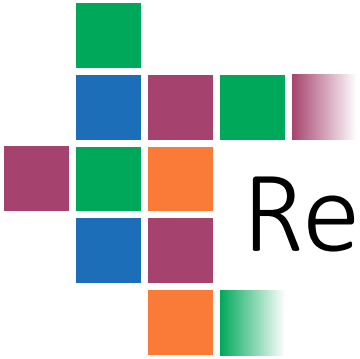


Further types …

- The combination of temporal changes with spatial network model corresponds to
  - Addition / removal of
    - Nodes
    - Edges

Data Attributes

- Non-spatiotemporal attributes: characterize non-contextual features of objects
  - Name, population, unemployment rate
- Spatial attributes:
  - spatial location (e.g., longitude and latitude, elevation)
  - spatial extent (e.g., area, perimeter)
  - shape or geometry
- Temporal attributes:
  - Timestamp (of spatial object, raster layer or spatial network snapshot)
  - Duration (of a process)



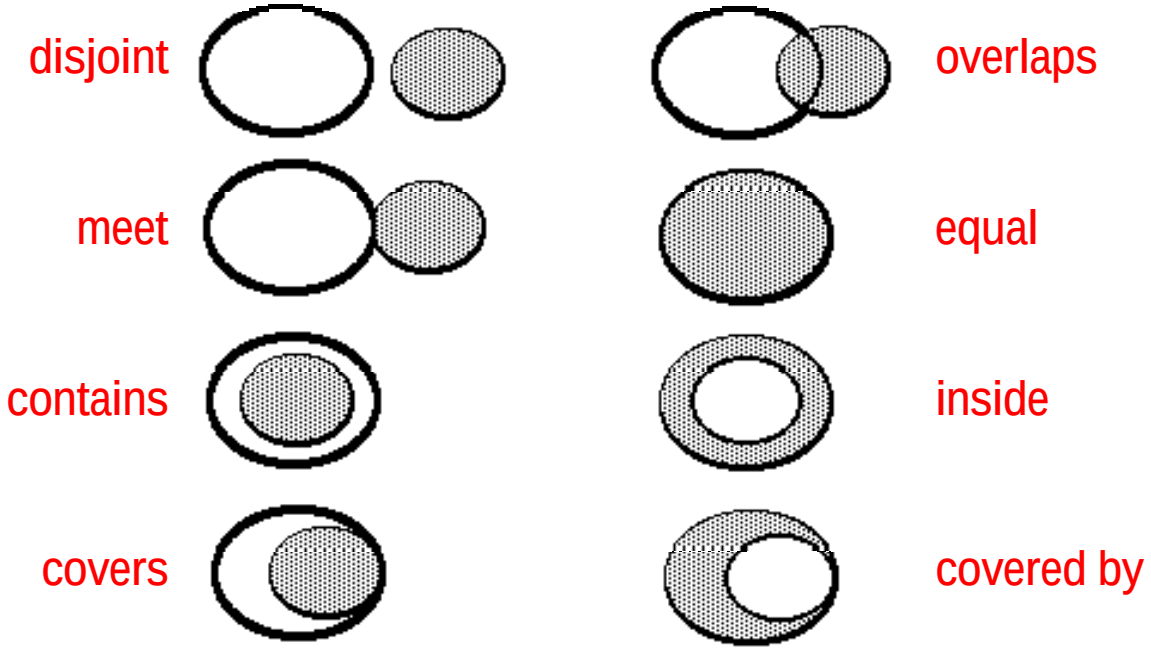
Relationships

- Spatial
  - Topological
  - Distance
  - Directional
- Temporal:
  - Before/after
  - During
- Spatio-temporal:
  - Attraction / Repulsion
  - Extension/expansion/rotation/deformation/split/merge

Topological relationships

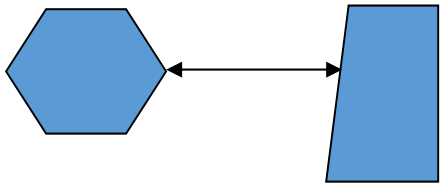
- Topological:
  - Invariant under homomorphisms (rotation, translation & scaling)
  - Semantics defined by the 9-intersection model

For regions

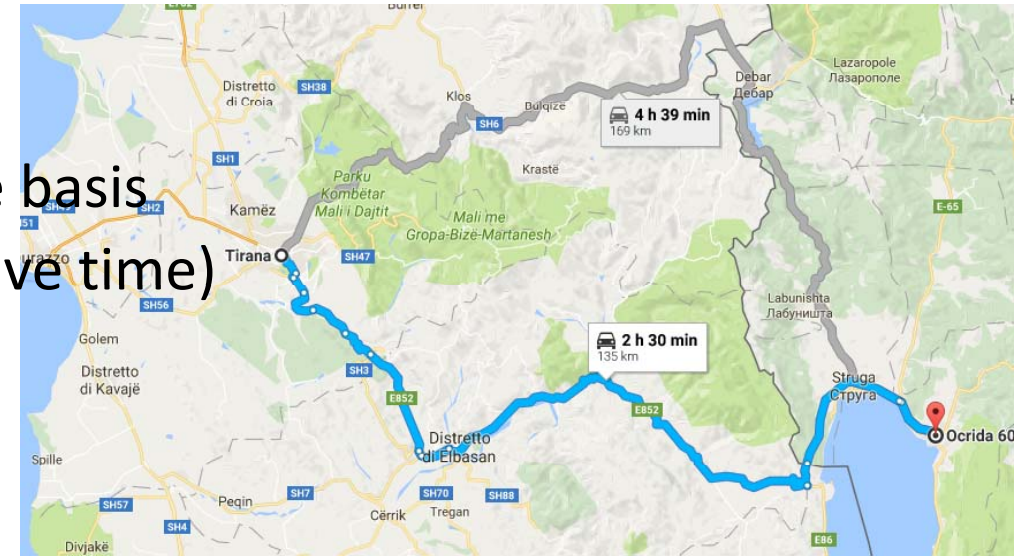


Distance relationships

- Metric
  - Euclidean distance between two points
  - For polygons it's an aggregate function (e.g., minimum)

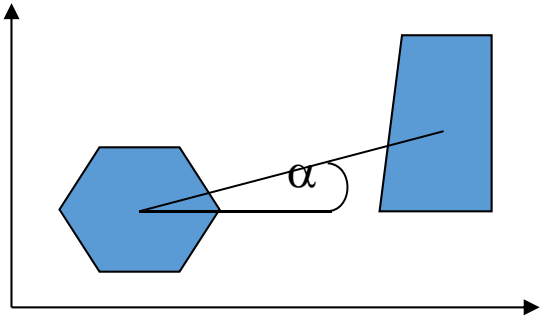


- Non-metric
  - Typically defined on the basis of a cost function (e.g. drive time)

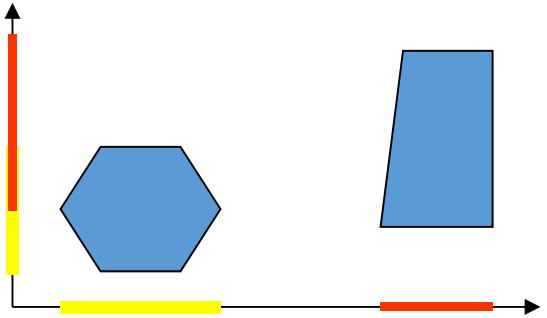


Directional relationships

- Based on an angle

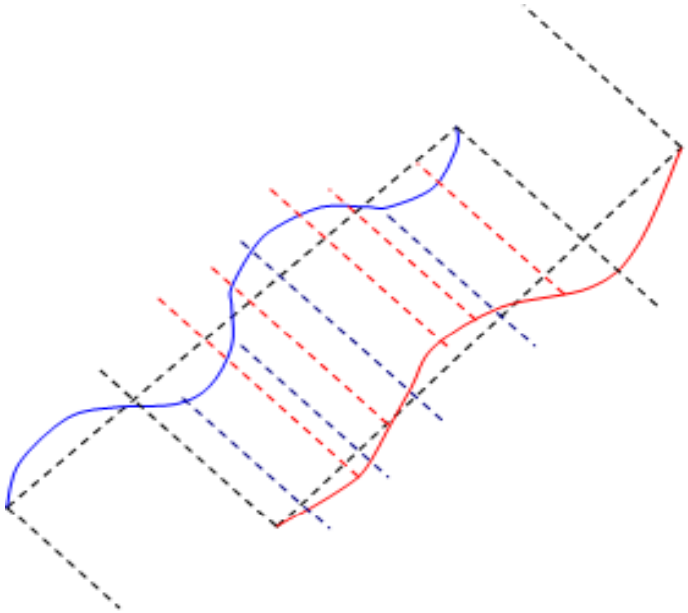


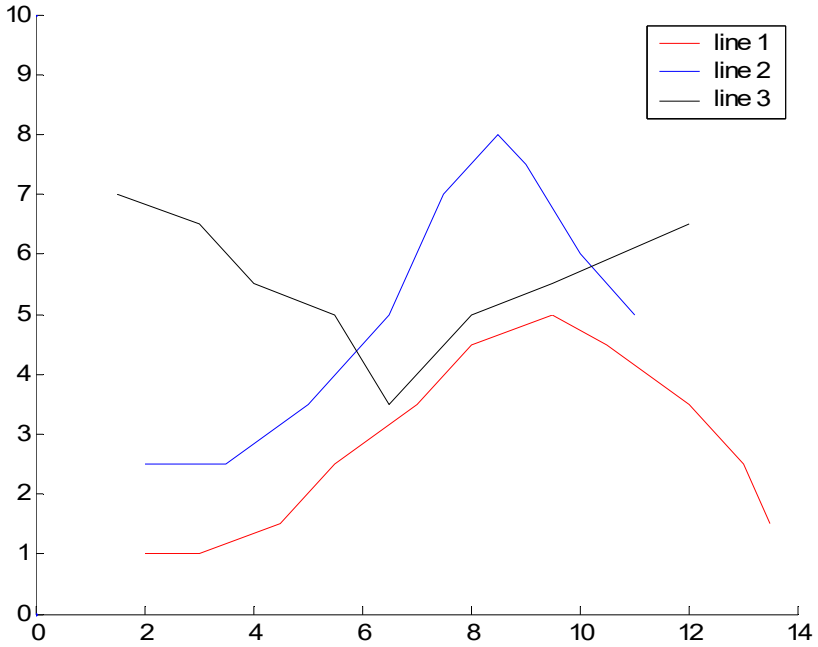
- Based on the extension of Allen's algebra

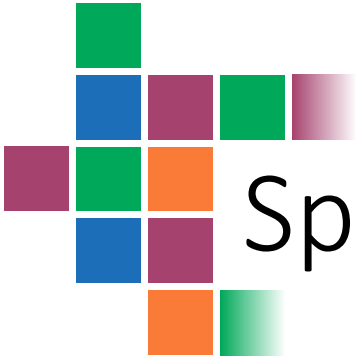


Hybrid relationships

- Line parallelism (combination of a topological + distance relationship)





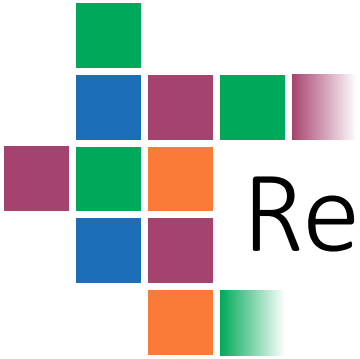


Spatial joins

- In a spatial DB, different spatial relationships ρ implicitly define spatial joins between two layers R_i and R_j

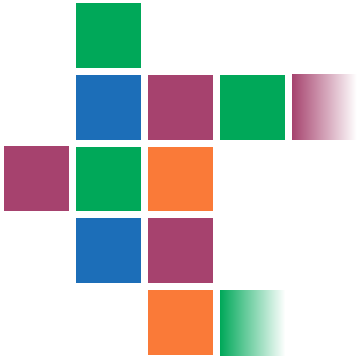
$$R_i \bowtie_\rho R_j$$

- Efficient computation of spatial relations is fundamental to perform spatial analysis
- Spatial indexing is fundamental



Relevant info in relationships

- Spatial, temporal and spatio-temporal relationships may convey relevant information on possible dependencies (statistical correlations) between non-spatiotemporal attributes



Spatial Relationships vs. Statistical Correlation

- The network of data defined by implicit spatial relationship ➔ the spatial dependencies

  distance of houses from main roads is a necessary condition for autocorrelation of burglaries

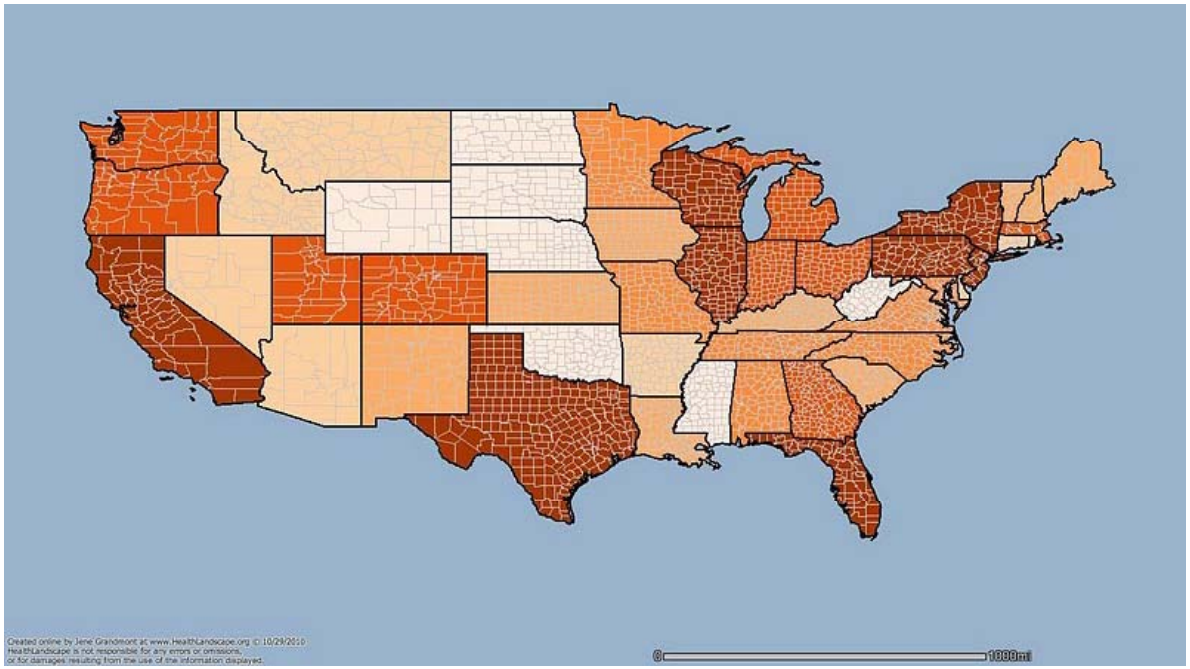
- Spatial correlation ➔ selection of relevant spatial relationships

  cross-correlation between price level of houses and the quality of services available in the nearby shows that the distance between houses and services is an important spatial relationship

Spatial Cross-correlation

- Correlation between two distinct attributes across space

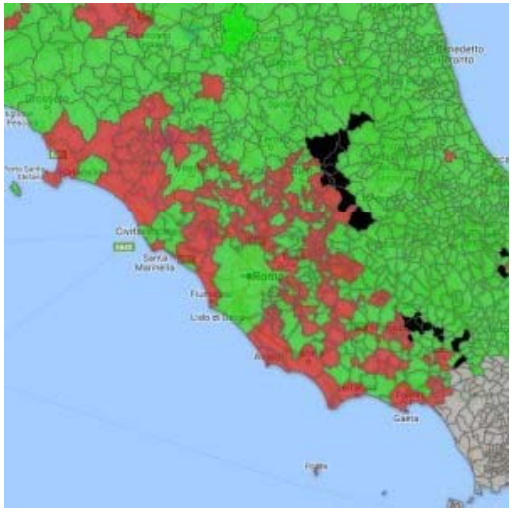
Number of Confectionery Manufacturing Establishments (Chocolate and Non-Chocolate), 2008

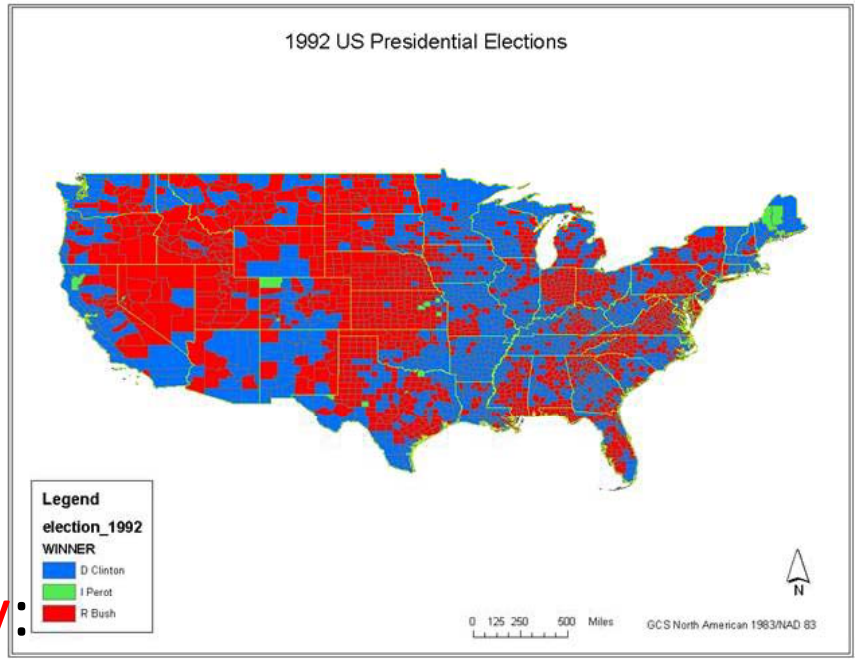


Courtesy of “The Health Foundation”

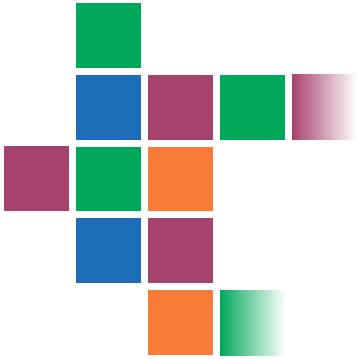
Autocorrelation

- Correlation of an attribute with itself across space





Tobler's first law of geography: Everything is related to everything else, but nearby things are more related than distant things.



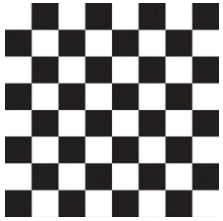
Autocorrelation: positive vs. negative

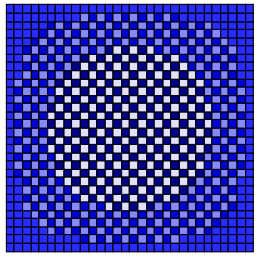
- Positive: in v, the attribute Y takes a value similar to those in N(v).
- Negative: in v, the attribute Y takes a value different from those in N(v).

Positive autocorrelation has received most of the attention, largely because few empirical examples of global negative autocorrelation have been found in spatial and social phenomena.

Negative Spatial Autocorrelation

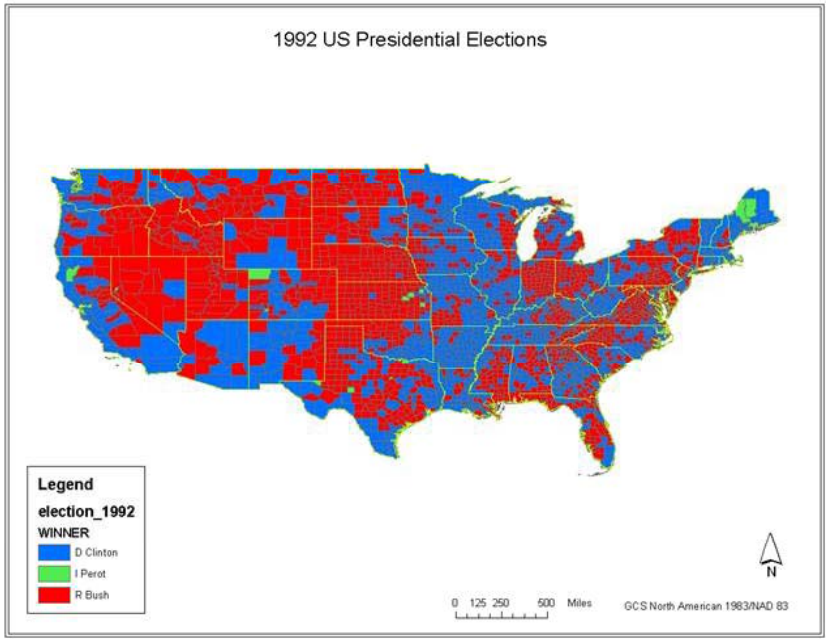
- It is an artifact of poorly devised areal units (Goodchild, 1986)
- It cannot exist in a continuous context.

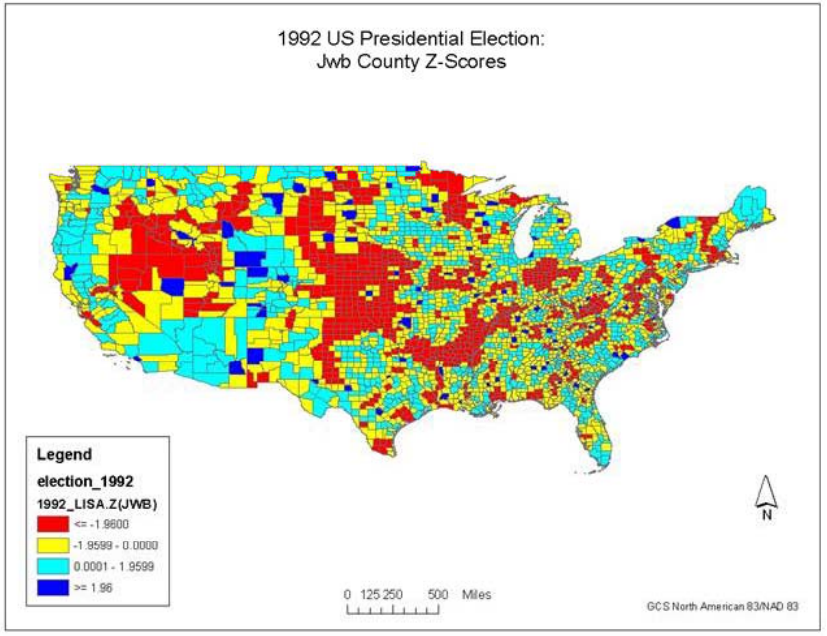




Negative Spatial Autocorrelation

Generally geographic distributions contain a mixture of both *local positive* and *local negative* spatial autocorrelation, with the positive tending to globally dominate the negative correlations.

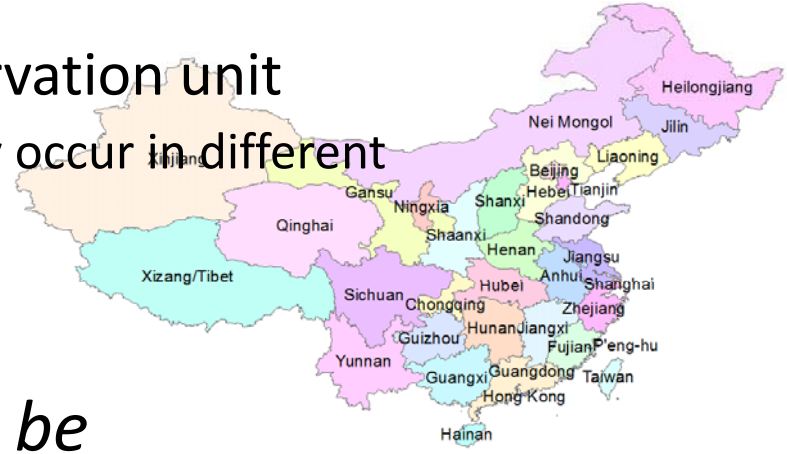




Measures of spatial autocorrelation

- Global Measures
  - A single value which applies to the entire data set
    - The same pattern or process occurs over the entire geographic area
    - An average for the entire area

- Local Measures
  - A value calculated for <u>each</u> observation unit
    - Different patterns or processes may occur in different parts of the region
    - A unique number for each location

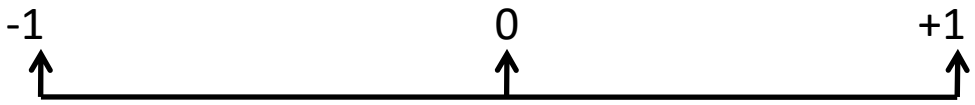


An equivalent local measure can be calculated for most global measures

Moran's I

$$I = \frac{N\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij})\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

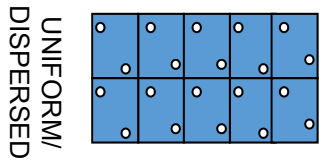
- The most common global measure of spatial autocorrelation
- Use for points or polygons
- Use for a continuous variable (any value)

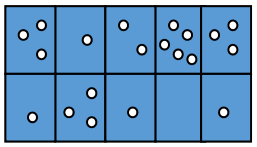


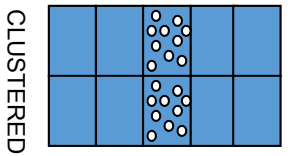
high negative spatial autocorrelation	no spatial autocorrelation*	high positive spatial autocorrelation

Can also use it as an index for dispersion/random/cluster patterns.

Dispersed Pattern	Random Pattern	Clustered Pattern



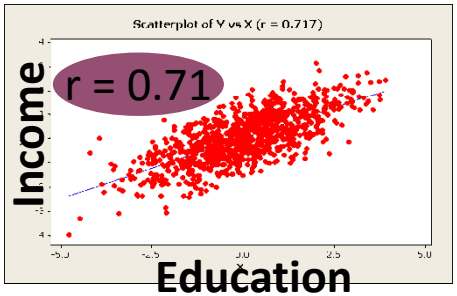




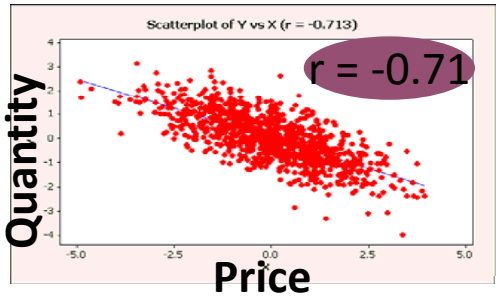
Moran's I vs. corr. coeffifient

Correlation Coefficient *r*

- Relationship between two variables



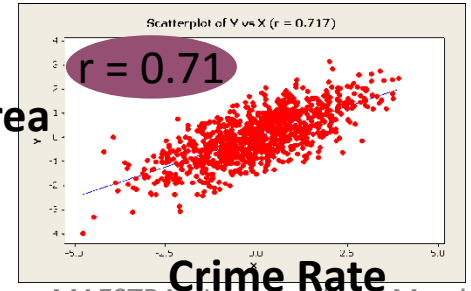
or



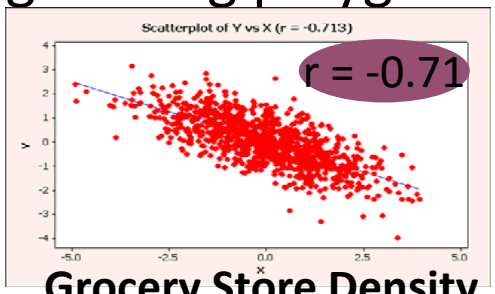
Moran's I

- Involves one variable only
- Correlation between variable, X, and the "spatial lag" of X formed by averaging all the values of X for the neighboring polygons

Crime in nearby area



Grocery Store Density Nearby



Local Indicators of Spatial Association (LISA)

- The statistic is calculated for **each** spatial unit in the data

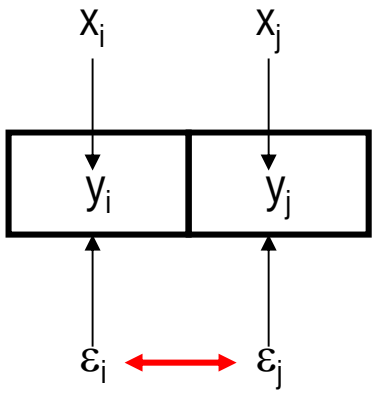
- For each spatial unit, the index is calculated based on neighboring units

Local Moran's I $$I_i = z_i \sum_j w_{ij} z_j$$

Spatial Dependence

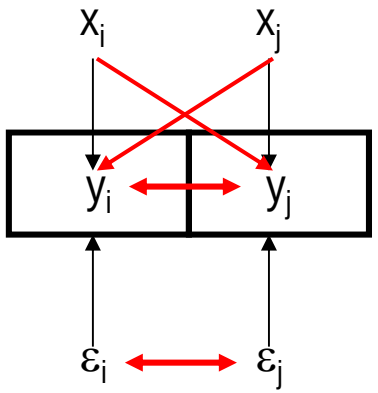
- Error
- Explanatory (non-target) variables
- Response (target) variables

Spatial error



Spatial lag

Spatially lagged explanatory variables



Spatially lagged response variables

Violated Assumptions

- Spatial error: error terms are uncorrelated
- Spatial lag: observations are independent (as well as error terms are uncorrelated)
- Due to this violation, in ordinary least squares regression, the correlation coefficients may be higher than they really are

➔ exaggerated precision (lower standard error)

➔ they are more likely found “statistically significant”

Violated Assumptions

"Anyone seriously interested in prediction when the sample data exhibit spatial dependence should consider a spatial model"

(LeSage & Page, 2001)

Models Developed in Statistics

1. spatial lag model: autocorrelation
2. spatial error model: correlation of errors
3. spatial cross-regressive model: cross-correlation.

L. Anselin, A., Bera, A. (1998): Spatial dependence in linear regression models with an application to spatial econometrics.

General Spatial Model

$$y = X\alpha + \beta D_1 y + \gamma D_2 X + u$$

- y: vector of observations of the dependent variable
- X: matrix of observations of the independent variable
- α: strength of local influence
- β: strength of autocorrelation
- γ: strength of cross-correlation
- D_1, D_2: spatial weight matrices (or *neighborhood matrices*)

General Spatial Model

$$u = \lambda D_3 u + \varepsilon$$

- u: error
- λ: strength of error (auto-)correlation
- D_3: spatial weight matrix
- $\varepsilon \sim N(0,\sigma^2)$; (*homoskedasticity*)

General Spatial Model

$\beta = 0$, $\gamma = 0$, $\lambda = 0$: **classical linear model**

What are the determinants of the price of a house?

Price = Sq. mt. + Age + Median Income + Dist. to Metro + error

General Spatial Model

$\gamma = 0$, $\lambda = 0$: **spatial lag model**

How do property appraisers determine the value of a property?

Price = D_1 * Price + Sq. mt. + Age + Median Income + Dist. to Metro + error

General Spatial Model

$\beta = 0$, $\gamma = 0$: **spatial error model**

What are the determinants of the price of a house?

Price = Sq. mt. + Age + Median Income + Dist. to Metro + error

error = D_3*error + ε

Spatial error can occur when:

1) Spatially correlated omitted variables

2) Spatially correlated aggregate variables

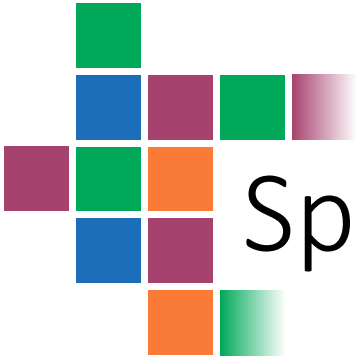
3) Spatially correlated errors in variable measurement

General Spatial Model

$\beta = 0$, $\lambda = 0$: **spatial cross-regressive model**

What are the determinants of the price of a house?

Price = Sq. mt. + Age + D_2*Age + Median Income + Dist. to Metro + error

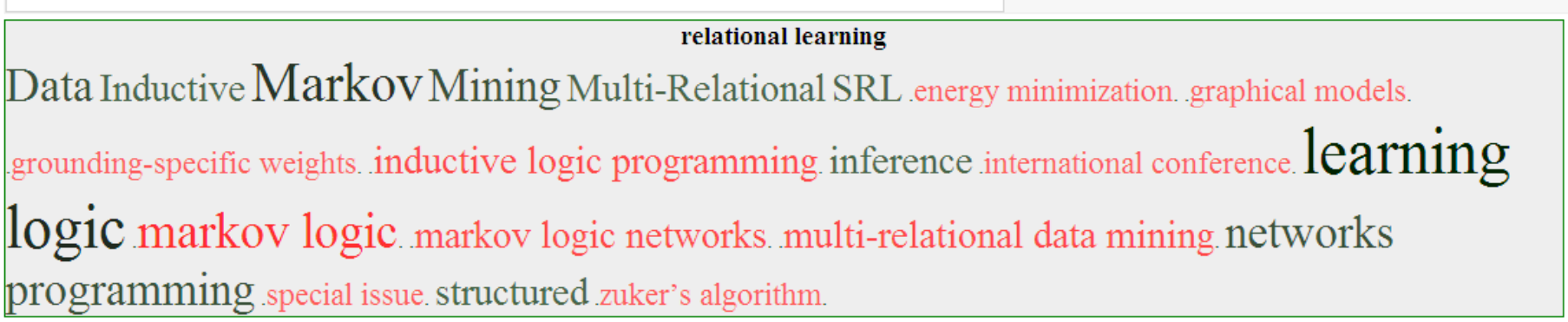


Spatial Modeling: Limitations

- D has to be carefully defined
- How can D express the contribution of different spatial relationships?
- Spatial dependencies are all handled in a pre-processing or feature extraction step
- All spatial objects involved in the spatial phenomena (rows of X) are uniformly represented by the same set of attributes

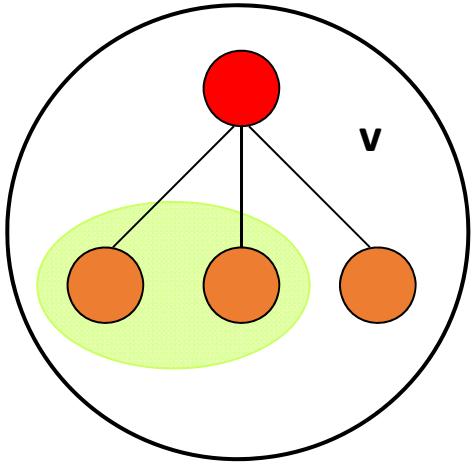
Relational Approach

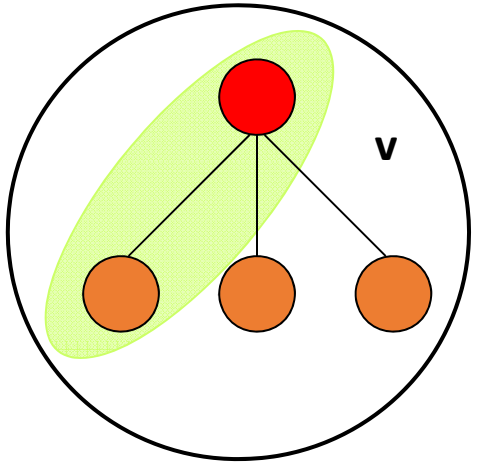
- Relational mining algorithms can be directly applied to various representations of network data, i.e. collections of interconnected entities.
- Investigated under many names



Relational Approach

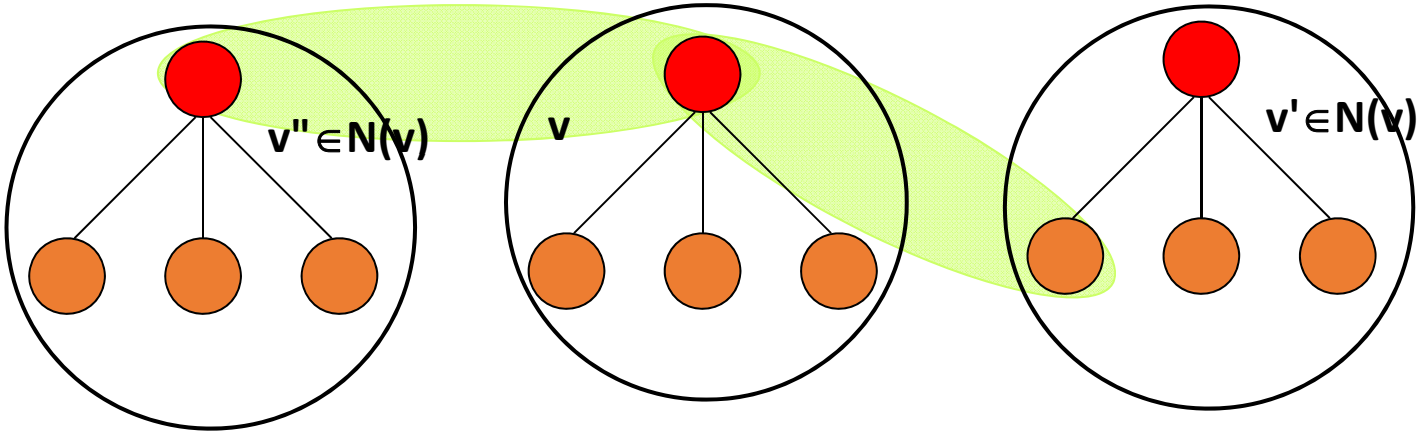
- Relational mining algorithms exploit two sources of correlation:
  - Local correlation, between attributes of each unit of analysis

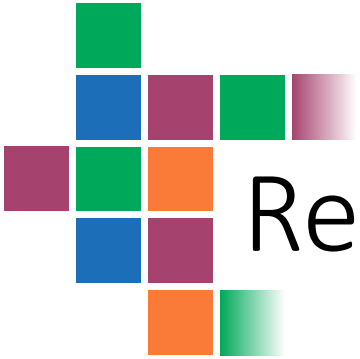




Relational Approach

- Relational mining algorithms exploit two sources of correlation:
  - Within-network correlation, between attributes of various units of analysis

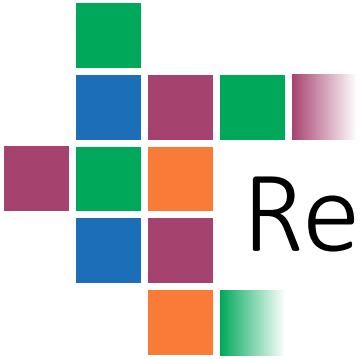




Relational Approach

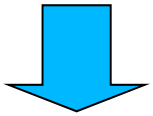
➔ Relational mining can consider various forms of correlation which bias learning in spatial domains.

➔Relational patterns reveal spatial dependencies and which of the many spatial relations are relevant.

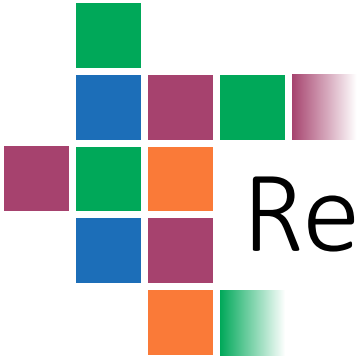


Relational Approach

- Spatial database as a kind of networked data where:
  - entities are spatial objects and
  - connections are spatial relationships



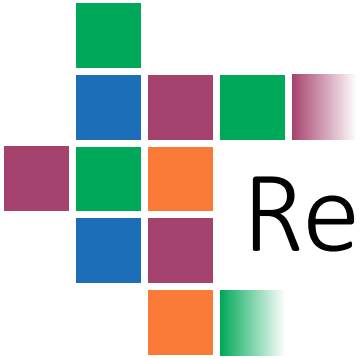
the application of relational mining methods is straightforward (at least in principle)



Relational Approach

- Many spatial data mining systems are based on general-purpose relational mining algorithms.

- W. Klösgen, M., May (2002). Spatial subgroup mining integrated in an object-relational spatial database.PKDD
- A. Appice, M. Ceci, A. Lanza, F.A. Lisi, D. Malerba (2003): Discovery of spatial association rules in geo-referenced census data: A relational mining approach. Intell. Data Anal.
- M. Ceci, A. Appice, D. Malerba (2004): Spatial Associative Classification at Different Levels of Granularity: A Probabilistic Approach. PKDD
- D. Malerba, A. Appice, A. Varlaro, A. Lanza (2005): Spatial Clustering of Structured Objects. ILP
- D. Malerba, M. Ceci, A. Appice (2005): Mining Model Trees from Spatial Data. PKDD
- M. Ceci, A.Appice, D. Malerba (2007): Discovering Emerging Patterns in Spatial Databases: A Multi-relational Approach. PKDD

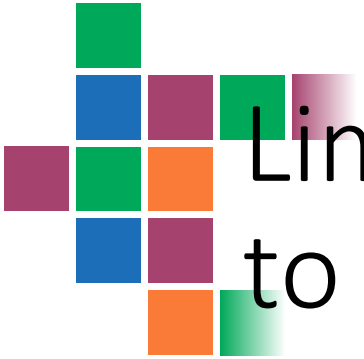


Relational Approach

- D. Malerba, M. Ceci, A. Appice (2009): A relational approach to probabilistic classification in a transductive setting. Eng. Appl. of AI 22(1): 109-116
- R. Frank, M. Ester, A.J. Knobbe (2009): A multi-relational approach to spatial classification. KDD.
- A. Appice, M. Ceci, D. Malerba (2014): Multi-Relational Model Tree Induction Tightly-Coupled with a Relational Database. Fundamenta Informaticae, 129(3): 193-224

REVIEWS

- D. Malerba (2008). A relational perspective on spatial data mining. IJDMMM 1(1), 103–118
- Donato Malerba, Michelangelo Ceci, Annalisa Appice: Relational Mining in Spatial Domains: Accomplishments and Challenges. ISMIS 2011: 16-24



Limits of a relational approach to STDM

- Pre-computation of spatial relations
- Computationally demanding
- Not practical for georeferenced streaming data

STDM Tasks

Classical attribute-oriented generalization

- Prediction of an aspatial attribute
  - Classification
  - Regression
  - Interpolation
  - Forecasting

+

  - Clustering
  - Outlier detection
  - Summarization

STDM Tasks

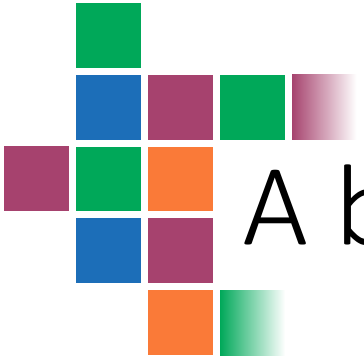
Spatial-oriented generalization

- Prediction of spatial properties
  - Location (also relative)
  - Geometry (polygon)
  - Relationships (co-location)

Temporal-oriented generalization

- Prediction of temporal properties
  - When (also relative) in a spatial context
  - Duration

and their combination … (e.g. trend detection)



A basic spatio-temporal pattern?

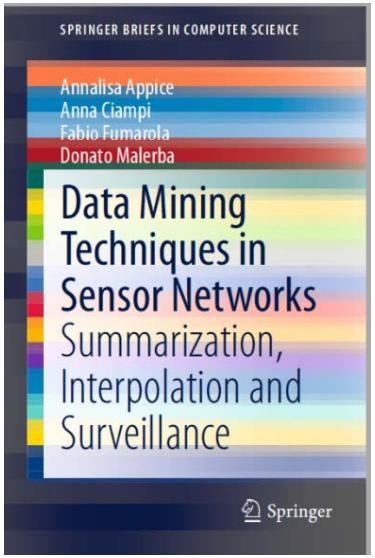
Which pattern to use as common ground for these tasks?

Analogy with frequent patterns which are:

- Local to a subset of transactions
- Useful for both predictive and descriptive tasks
  - Association analysis
  - Associative classification
  - Clustering based on association rules

Analogously, we look for a local pattern (both in space and time), which accounts for both spatial and temporal autocorrelation, useful for various tasks.

Trend Cluster



A trend cluster is a triple:

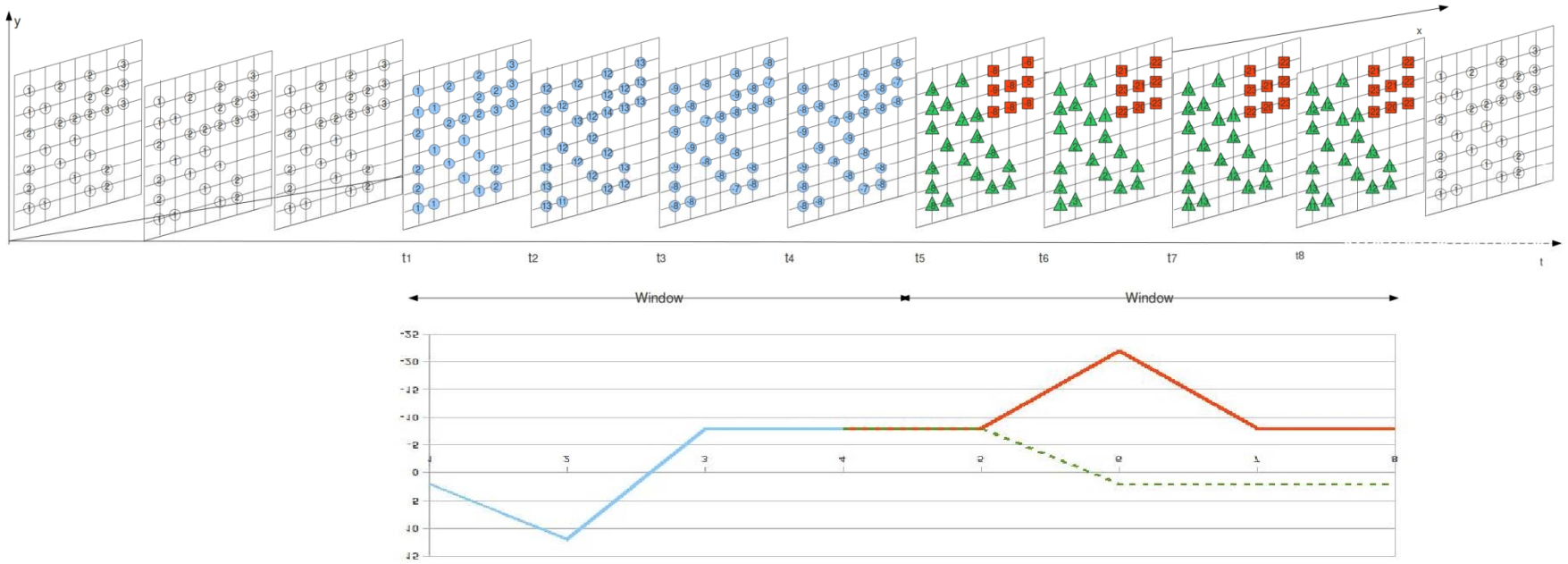
[W; C; T]

where:

1. W is a *time horizon* along which data are collected;
2. C is a *spatial cluster* of sensors which transmit values whose temporal variation is similar along W;
3. T is the *time series* which represents the trend of the clustered measures as they are collected at the transmission time points in W.

Trend Cluster Discovery

- Trend clusters discovered window by window.



Count-based window model

Trend clusters

Trend clusters effective pattern for several tasks.

- *Efficient mining algorithm*
- *Good trade-off between accuracy and time complexity*
- *Accounts for both spatial and temporal correlation*
- *Systems available for the research community*

See Part 2 by Annalisa Appice

Aknowledgements

- Learning Techniques in Relational Domains and their Applications
  - PRIN 2009 project funded by the Italian Ministry of University and Research (MIUR).
- EMP^3: Efficiency Monitoring of Photovoltaic Power Plants
  - funded by Fondazione Cassa di Risparmio di Puglia
- Vi-POC: Virtual Power Operation Center
  - Start-up Project funded by Italian Ministry of University and Research (MIUR).
- MAESTRA: Learning from Massive, Incompletely annotated, and Structured Data
  - FET Open Project funded by EU Commission.