# Proposal Framework
# to Spatial Data Mining

LATIN AMERICAN GEOSPATIAL FORUM

Instituto Brasileiro de Geografia e Estatística
Diretoria de Geociencias
Fátima Ferrão dos Santos

# Agenda

- ✓Introduction
- ✓Geo Knowledge Discovery
- ✓ Spatial Data Mining
- ✓ Energy, Oil and gas: some applications
- ✓Integration Framework
- ✓ Results
- ✓ Next steps

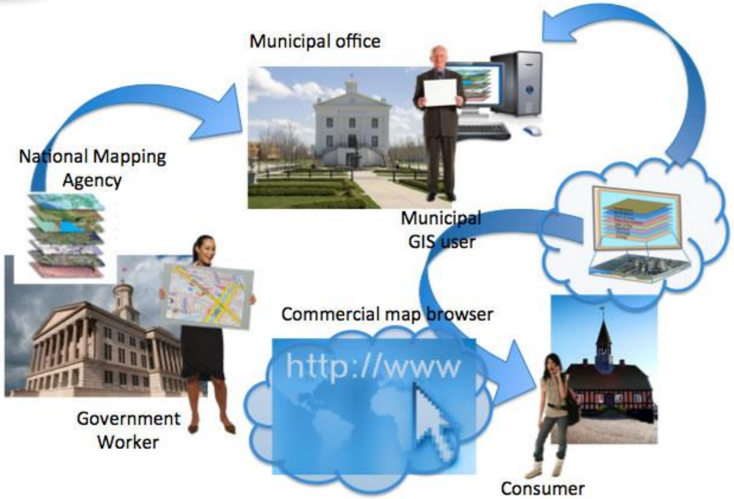
# Crescent use of geographic data

Defense and Intelligence

Energy Industry Oil and Gas

Risk and Disasters Management and Response

Social Networks



Electronic Government

# Data, Information and Knowledge

**Data**

**renda - Montly income**

**Despesas -Monthly expenditure**

**Information**

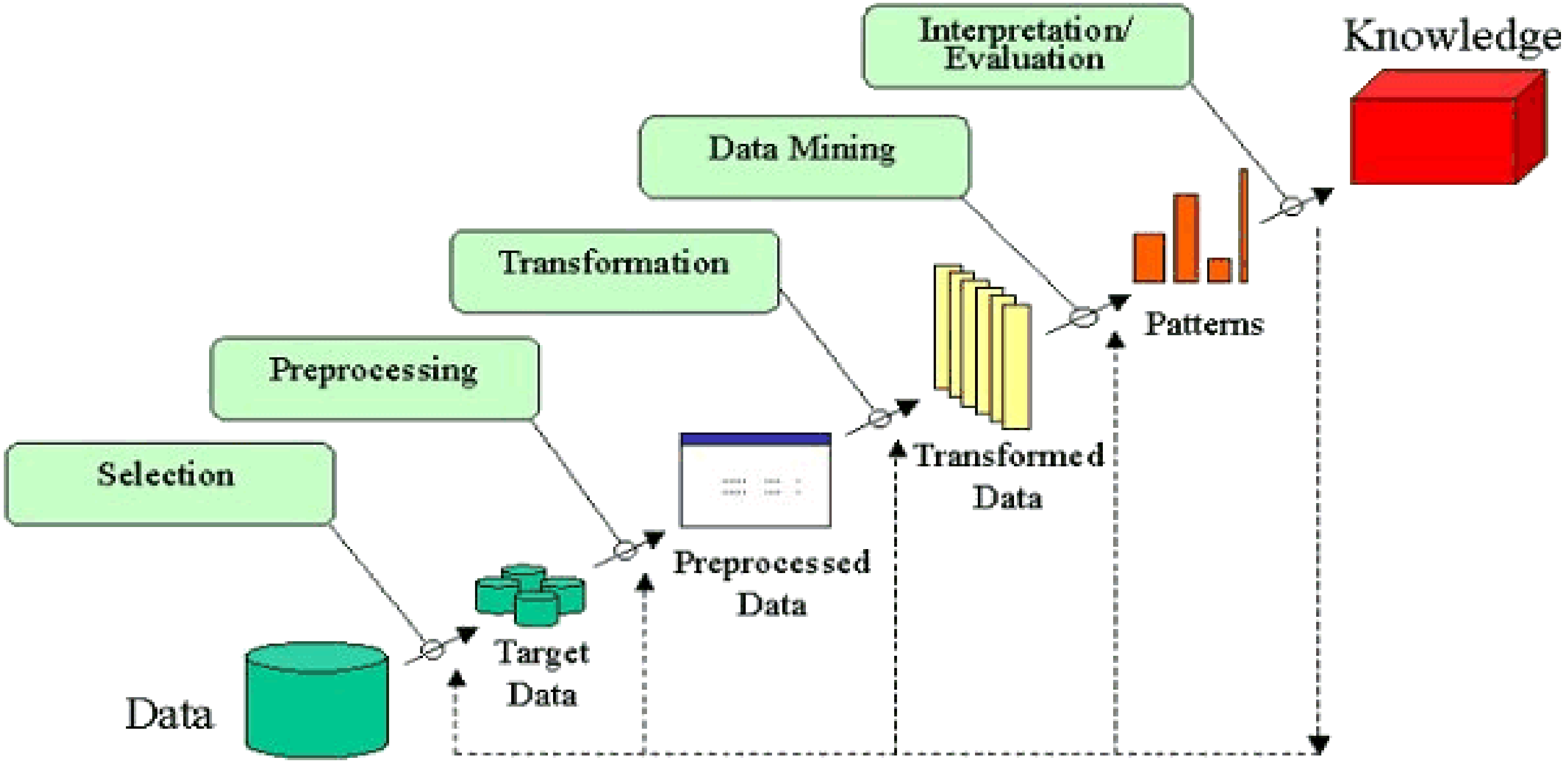
$$CME = \frac{(renda - despesas)}{renda} * 100$$

**Knowledge (rule association)**

**If *CME* < 35% e**

**census sector = XYZ → 99% of households achieved target ‘literate children’.**

# Knowledge Discovery in Databases (KDD)
# Geo Knowledge Discovery (GKD)



Source: Fayad et. al, 1996, 'From Datamining to knowledge discovery in databases'.

One of the tasks of the knowledge discovery process, which empowers the abilities to extract new, insightful information within large geographic databases and to formulate knowledge.

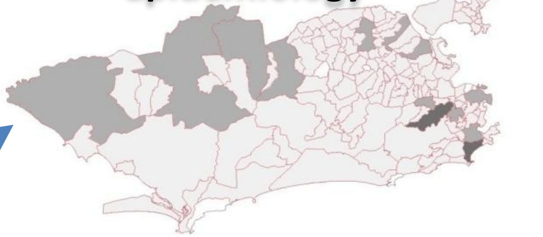
# Knowledge Discovery in Databases (KDD)
# Geo Knowledge Discovery (GKD)

Overall process that empowers the abilities to extract new, insightful information embedded within large heterogeneous databases and to formulate knowledge

### Energy

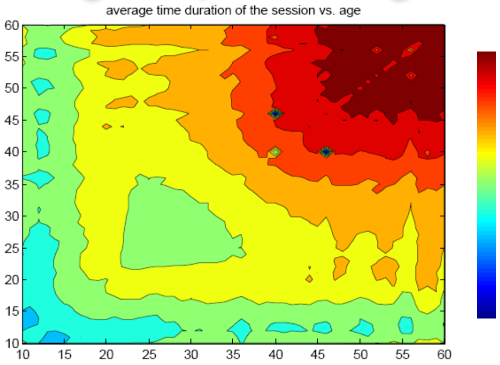
Help Management, cut costs,
Spatial rate analysis

### Epidemiology



Clustering  districts by: sex, HIV positive and age, 2005 (dark gray is the max value)

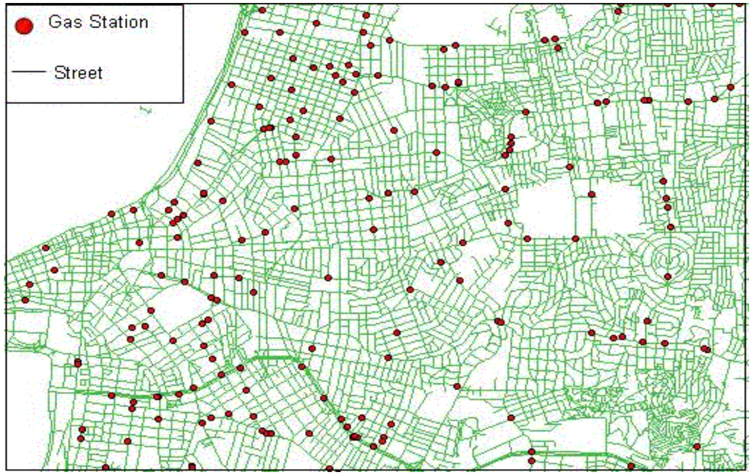
### Logistics / Marketing



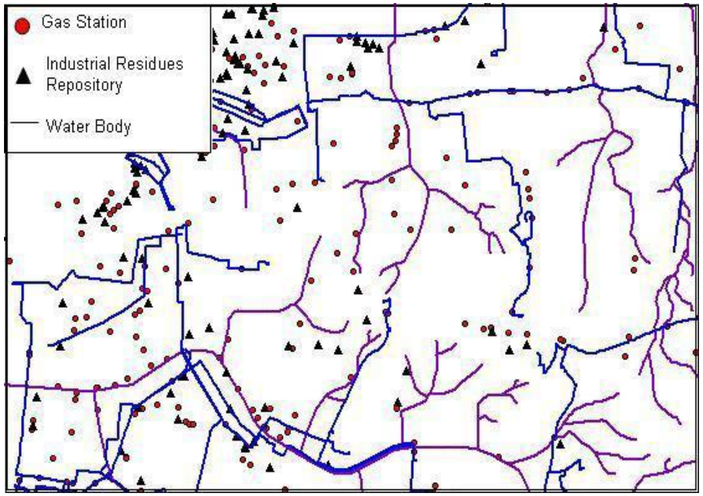
Average time duration of session by age (social networks)

# Information

Many relationships are well known geographic domain associations that may hind the discovery process and produce a large number of patterns without novel and useful knowledge.

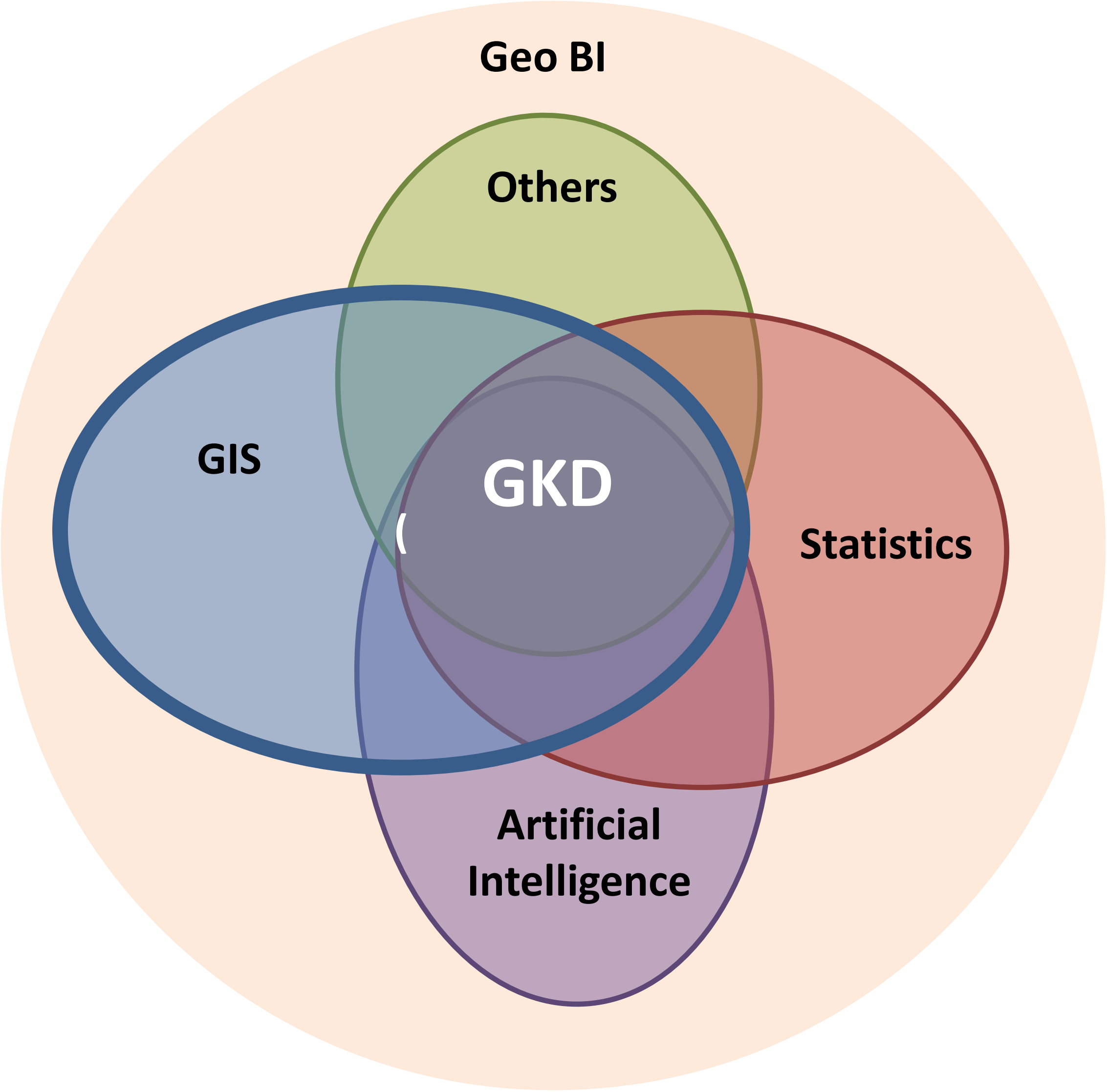


Trivial information: every gas station interceps street



Non-trivial information

# Knowledge Discovery in Geo Databases



# Energy: some applications

Integrate customers energy patterns with consumer bying behaviors

Reduce marketing costs: Identify potential consumers of green energy

green energy

Spatial rate analysis

Network control centers (outliers detection)

# An Experiment

✓Think of a number between 1 and 10.

✓Multiply this number by 9.

✓Work out the checksum of the result, i.e. the sum of the numbers.

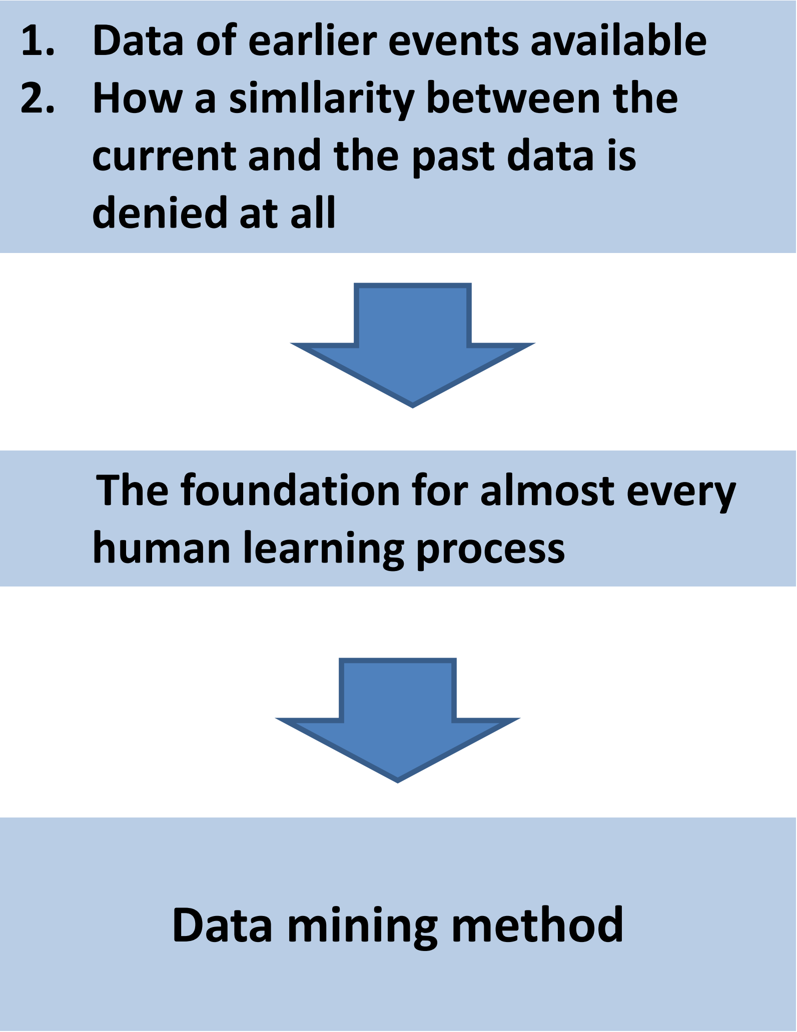
✓Multiply the result by 4.

✓Divide the result by 3.

✓Deduct 10.

**Do you believe in coincidences?**

# Data Mining



1. Data of earlier events available
2. How a simllarity between the current and the past data is denied at all

The foundation for almost every human learning process

Data mining method

# Data Mining

*"Every time I have seen a glass fall from a height of more than 1 meter it has broken"*

**Even without knowing the physical description, we and the method of data mining are able to generate an estimate of situations and forecasts**

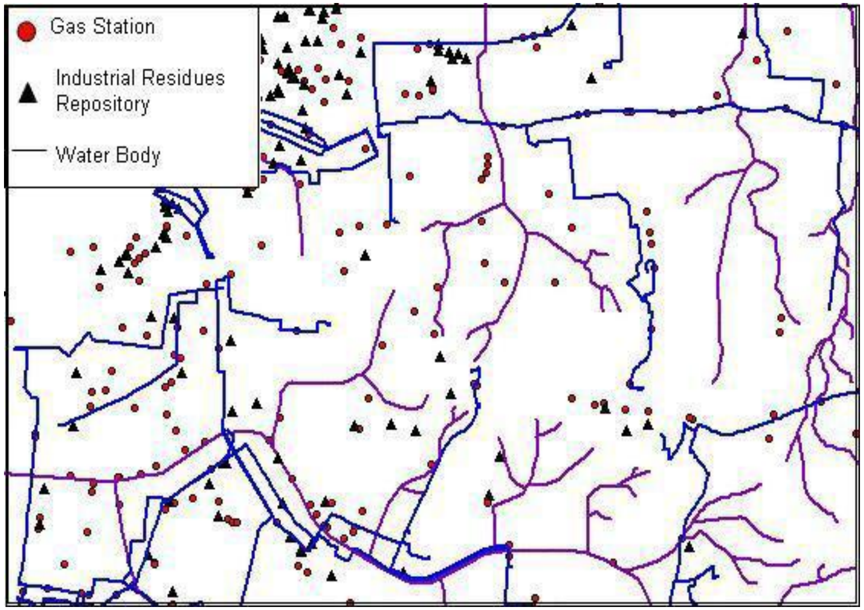
# Classical x Spatial Data Mining

IBGE
Instituto Brasileiro de Geografia e Estatística

- ✓ Classical data mining algorithms often make assumptions about the ndependence of data samples and identical distributions.

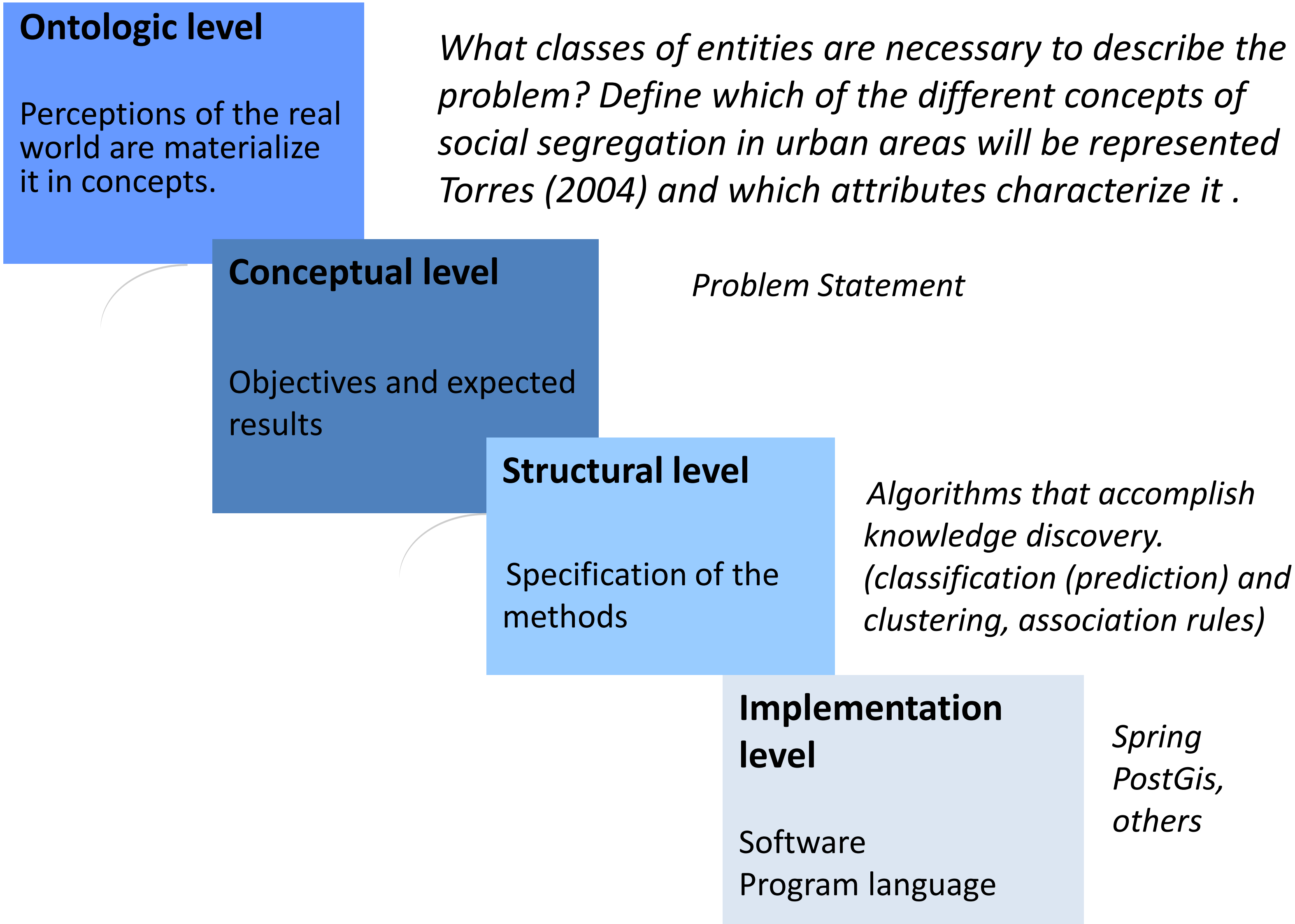
- ✓ The assumption about the independence of spatial data samples is generally false.

- ✓ Spatial data often has high autocorrelation among nearby features.



***"Everything is related to everything else but nearby things are more related than distant things (Tobler 1979)"***

# Proposal Methodology



**Ontologic level**

Perceptions of the real world are materialize it in concepts.

*What classes of entities are necessary to describe the problem? Define which of the different concepts of social segregation in urban areas will be represented Torres (2004) and which attributes characterize it .*

**Conceptual level**

Objectives and expected results

*Problem Statement*

**Structural level**

Specification of the methods

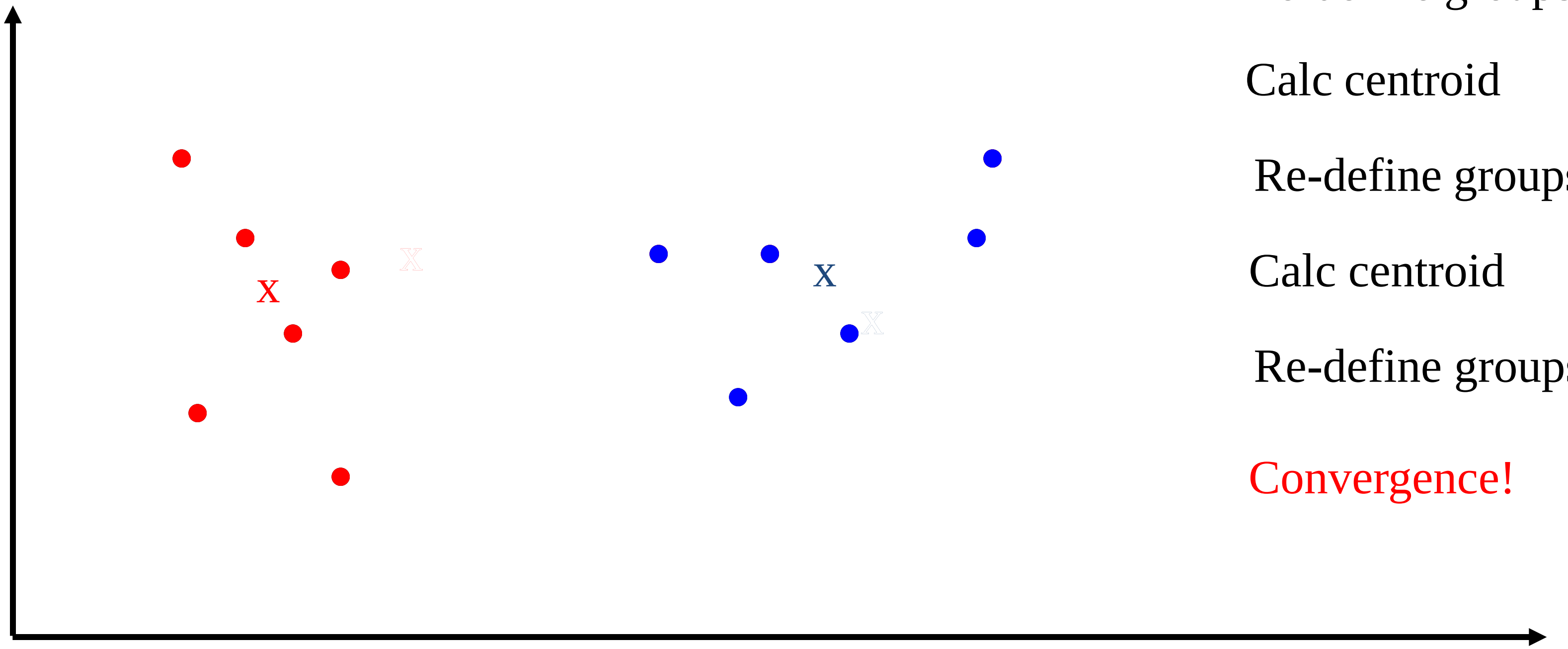
*Algorithms that accomplish knowledge discovery. (classification (prediction) and clustering, association rules)*

**Implementation level**

Software
Program language

*Spring*
*PostGis,*
*others*

# Conceptual level: clustering



Select seeds

Re-define groups

Calc centroid

Re-define groups

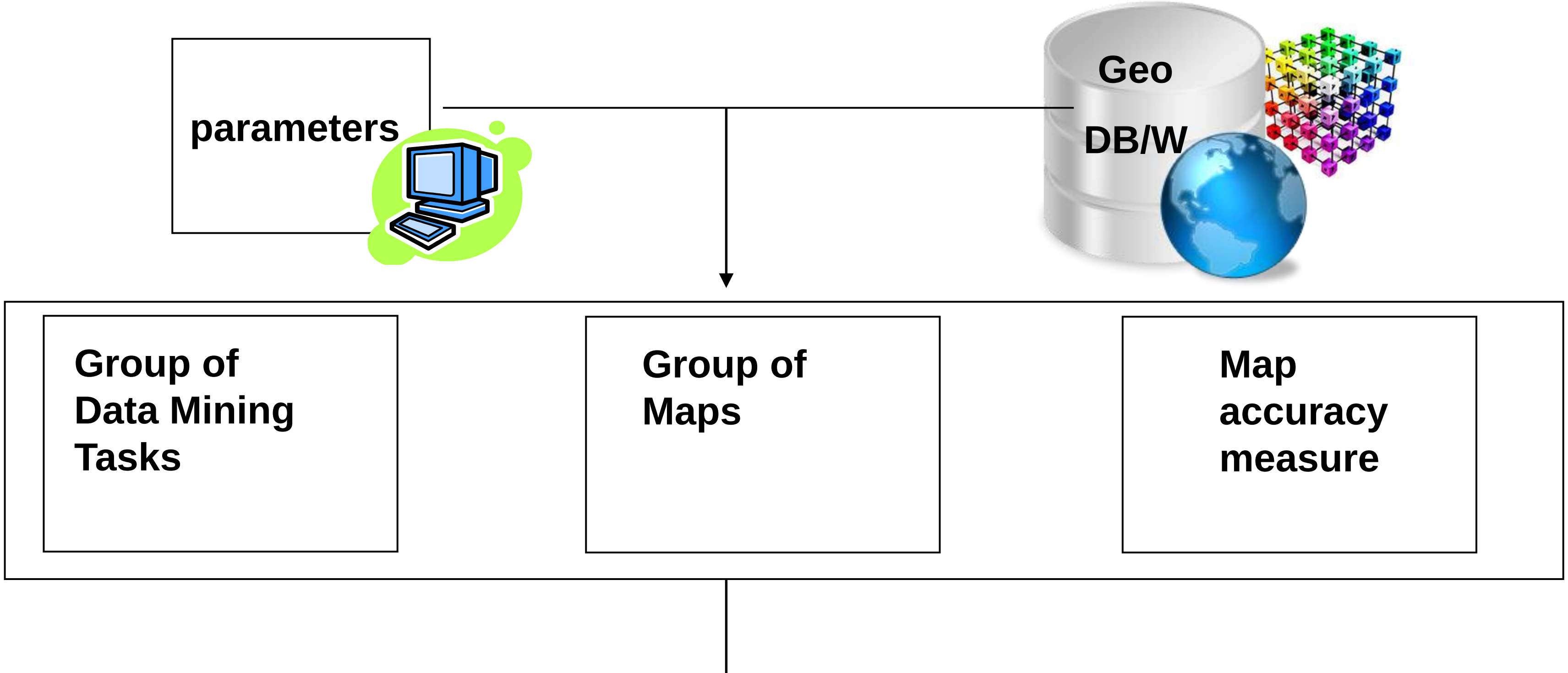
Calc centroid

Re-define groups

Convergence!

The task of grouping objects into meaningful subclasses (clusters) so that the members of a cluster are as similar as possible whereas the members of a different clusters differ as much as possible.

# Integration Framework



parameters

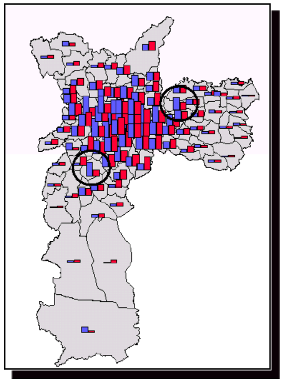
Geo

DB/W

Group of Data Mining Tasks

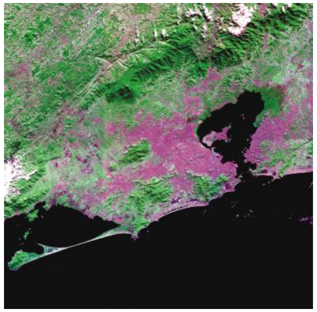
Group of Maps

Map accuracy measure



Geographic visualization of

the *patterns* discovered

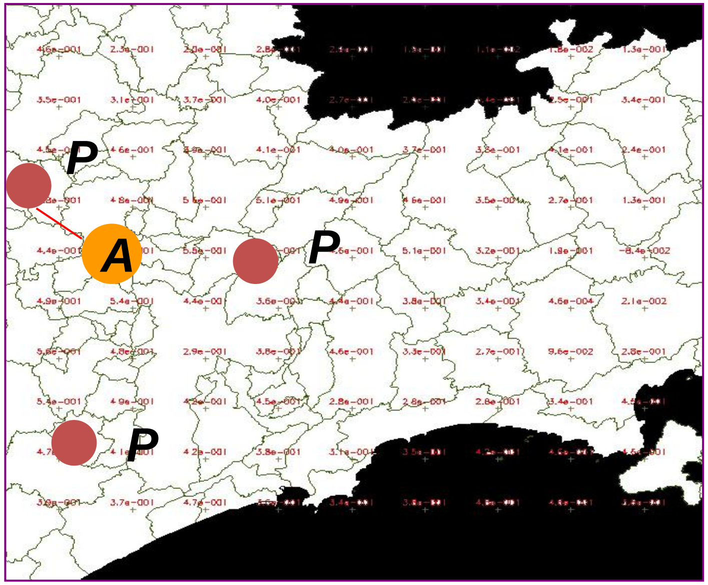


# Map Accuracy Measure

Usually, for a classification problem, the way we measure **classification accuracy** is calculating the percentage of correctly classified objects.

**Spatial accuracy** must measure how far the predictions are from the actual cases.

Our proposal measure of spatial accuracy is the average distance from real cases to the closest predicted occurrences.



$$I = \frac{\sum_{i}^{n} I_i}{n}$$

where:

*$I_i$ = d ($A_n$, $A_n$ nearest (P))*

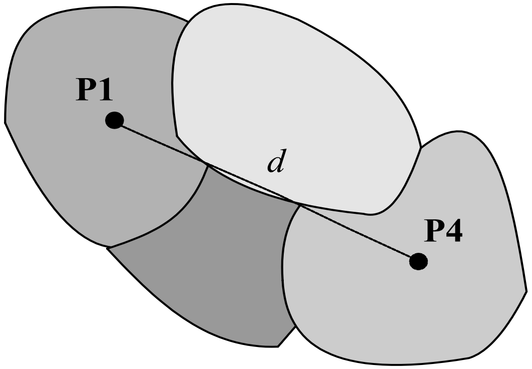
*An – actual cases*

*P – map layer of predicted cases*

*n – number of actual cases*

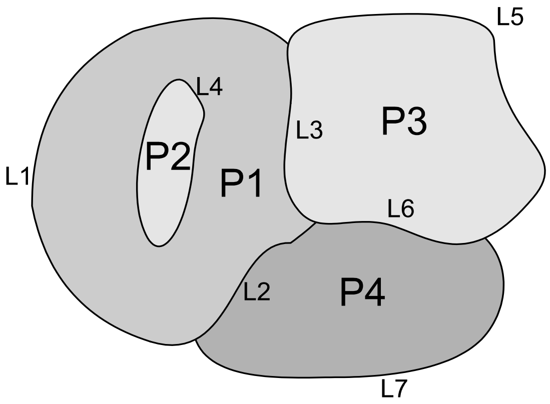
# Spatial Autocorrelation Measure

Concept from spatial statistics which explicitly models spatial autocorrelation.



$$w_{14} = 0 \qquad \textit{If } d > \text{max distance}$$

$$w_{14} = 1 \qquad \textit{If } d \leq \text{max distance}$$



Contiguity Matrix

$$W = \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} \\ w_{21} & w_{22} & w_{23} & w_{24} \\ w_{31} & w_{32} & w_{33} & w_{34} \\ w_{41} & w_{42} & w_{43} & w_{44} \end{bmatrix}$$

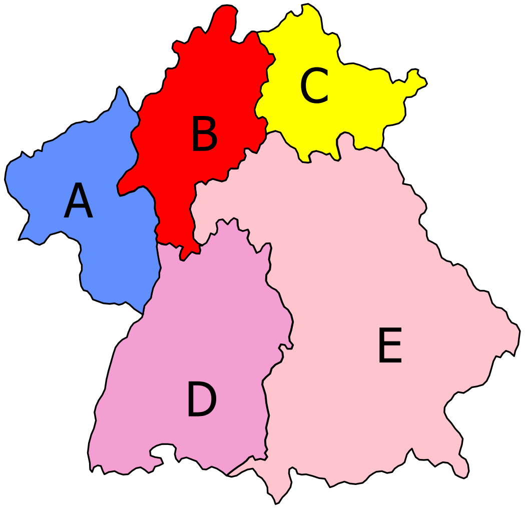
$w_{ij}$ : “distance” between object ***i*** and ***j.***

# Spatial Autocorrelation Measure

All spatial autocorrelation measures are crucially dependent on the design of the contiguity matrix W.

The design of the matrix determines “what constitutes a neighbourhood of influence”

|   | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 0 | 1 | 0 | 1 | 0 |
| B | 1 | 0 | 1 | 1 | 1 |
| C | 0 | 1 | 0 | 0 | 1 |
| D | 1 | 1 | 0 | 0 | 1 |
| E | 0 | 1 | 1 | 1 | 0 |



wij = 1, if Oi and Oj share an edge

# Spatial Autoregressive Model

In spatial regression the spatial dependencies of the error term or the dependent variable are directly modeled in the regression equation.

Assume that the dependent values y*i* are related to each other, i.e.

y*i* = f (y*j*)

Then the regression equation can be modified as:

$$\rho Wy + \beta X + \varepsilon$$

***SAR equation***

*W* - neighbourhood relationship contiguity matrix

*p* - parameter that reflects the strength of spatial dependencies between the elements of the dependent variable.

After having introduced the correction term $\rho Wy$ , the components of the residual error vector $\varepsilon$ are now assumed to be generated from independent and identical standard normal distributions.

# Influence Index

$$IF = \frac{\left(y_i - \overline{y}\right)\sum_{j=1}^{n} W_{ij}\left(y_j - \overline{y}\right)}{\sum_{j=1}^{n}(y_j - \overline{y})^2}$$

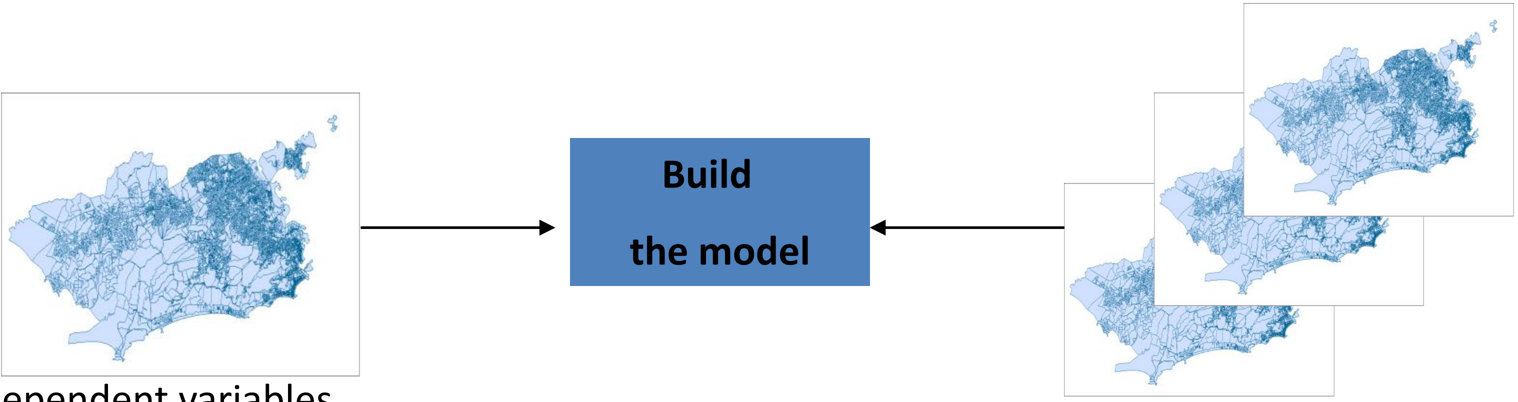
Influence Index

where:

- $n$ number of areas,
- $y_i$ value for the attribute in the area “i”,
- $\overline{y}$ average in the region,
- $w_{ij}$ weights considering areas $i$ e $j$. (design of contiguity matrix)

# Location Prediction

Our aim is to build a model for location prediction. With the values of independent variables we want to predict the location of occurences.



Build

the model

Dependent variables

Independent variables

Two  important aspects of the model:

1. Data samples may exhibit spatial autocorrelation and

2. Map-similarity measure is a combination of classical and spatial accuracy measure. Independent variables map pixels to real numbers. Dependent variables (location prediction) map pixels to binary domain.

   The goals of the model is to predict and to analyse the effects of including one spatial autocorrelation term.

# Location Prediction

An uniform grip was imposed and different types of measurements were considered (1)
Domain knowledge is crucial in deciding which attributes are important.

| | | | |
|---|---|---|---|
| | | | |
| | | | A |
| | | | |
| A | | | A |

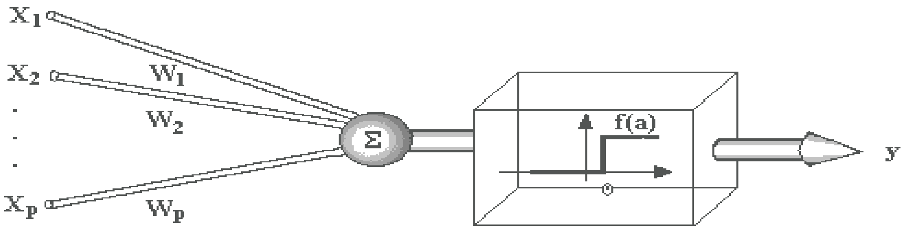
(a)

| | | | |
|---|---|---|---|
| | P | | |
| P | P | | A |
| | | | |
| A | | | A |

(b)

| | | | |
|---|---|---|---|
| | | | |
| | | P | A |
| | P | P | |
| A | | | A |

(C)



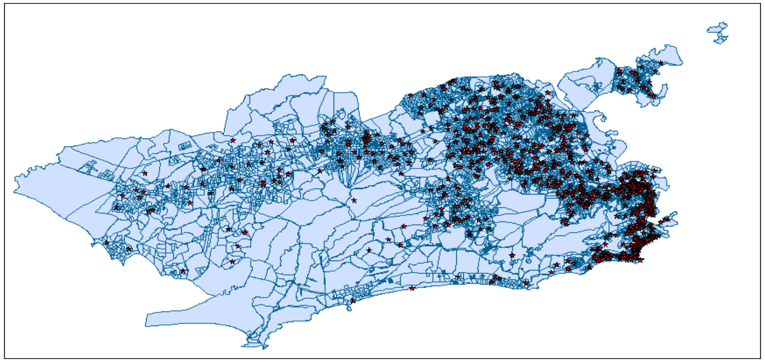
Supervised learning representation

(a) Actual cases (b) location prediction 1 (c) better location 2

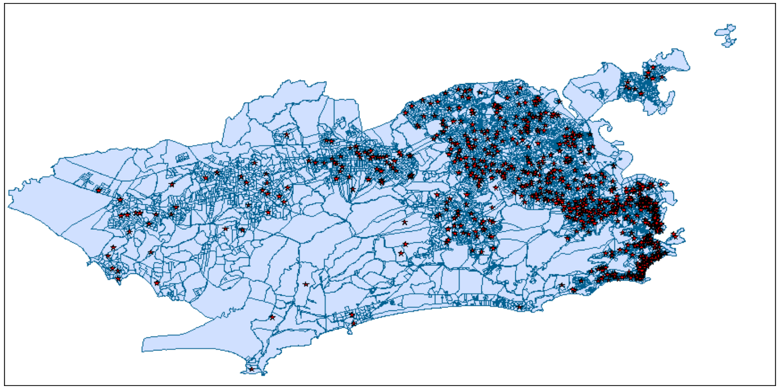
(1) Method for Identification of relevant variables. (Santos, F. 2008)

# Location Prediction

Usually, the model is built using a portion of the data, the learning or training data, which is, then, tested on the remaining of the data called the testing data.



Learning data



Testing data

**Build**

**the model**

Logistic regression

ANN Feedfowared Backpropagation

Others

**Evaluate**

**the model**

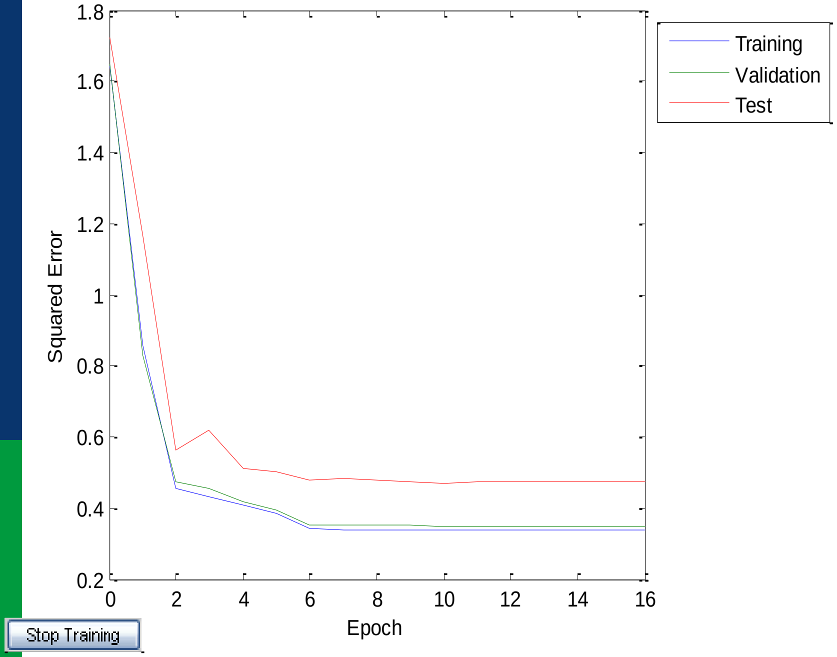
**Map accuracy Measure**

# Location Prediction

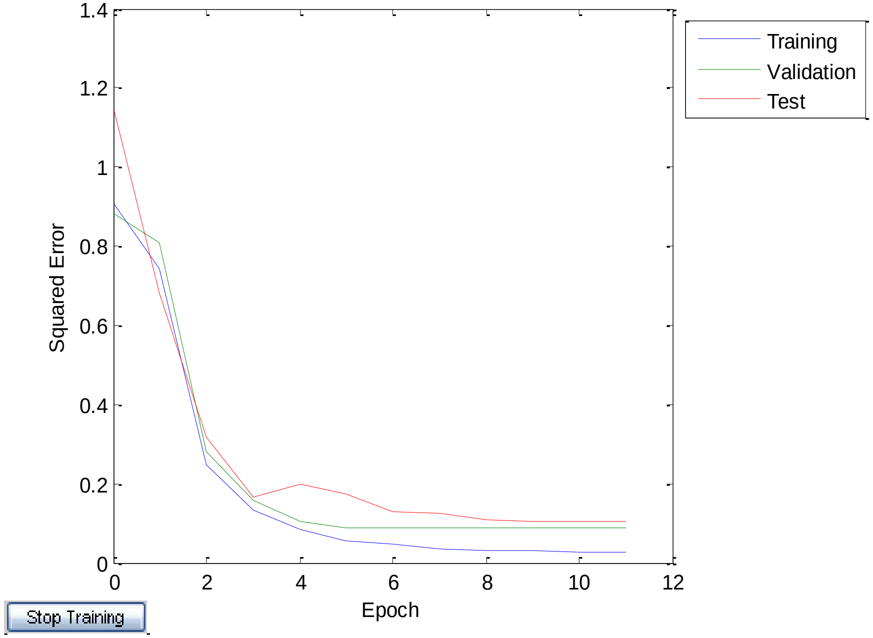
Classical and spatial prediction was compared on the data.

.

Classical



Spatial

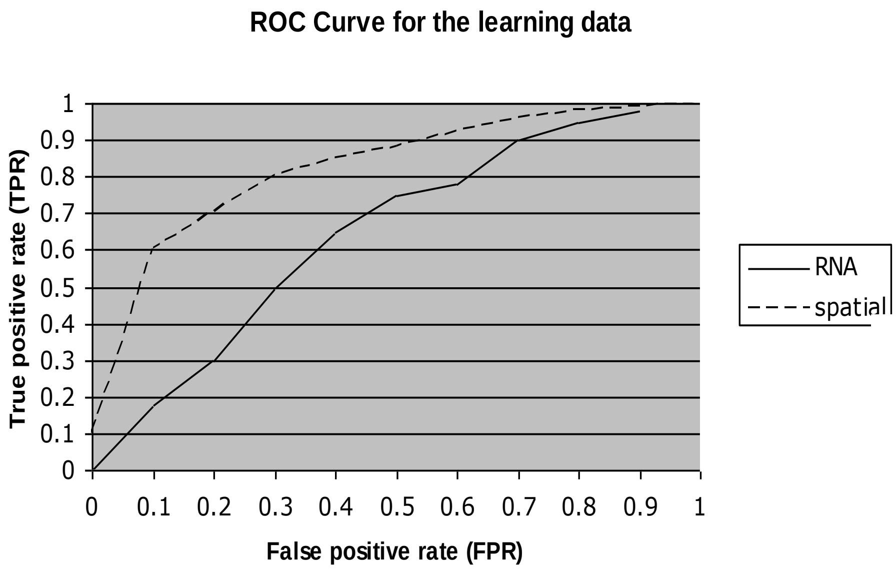


Squared Error Levenberg-Marquardt backpropagation MATLAB
Hidlayer:1 (tansig)

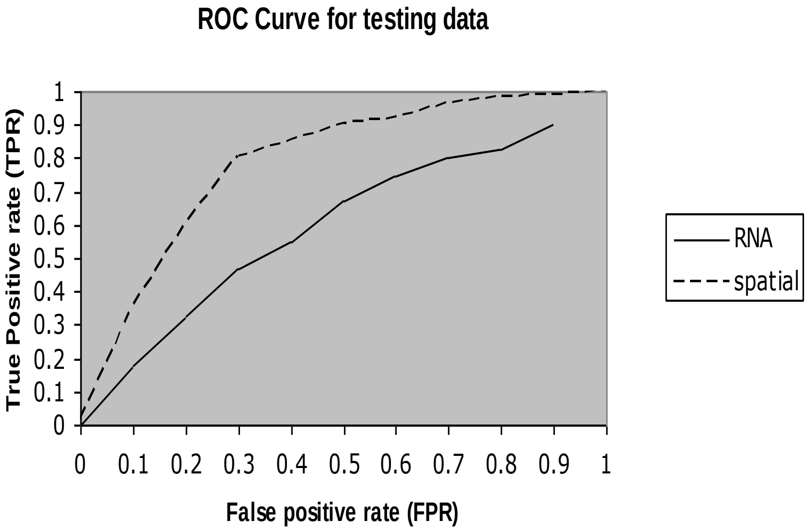
| Data Set | | Spatial (ANN) | Classical (ANN) | SAR |
|---|---|---|---|---|
| Learning | Spatial accuracy | 16.90 | 49.20 | 15.95 |
| Testing | Spatial accuracy | 19.21 | 42.52 | 19.30 |
| Learning | run-time (sec) | 90 | 15 | 17565 |

# Classical x Spatial Data Mining

ROC curves was considered to compare classification accuracy of classical and spatial regression.

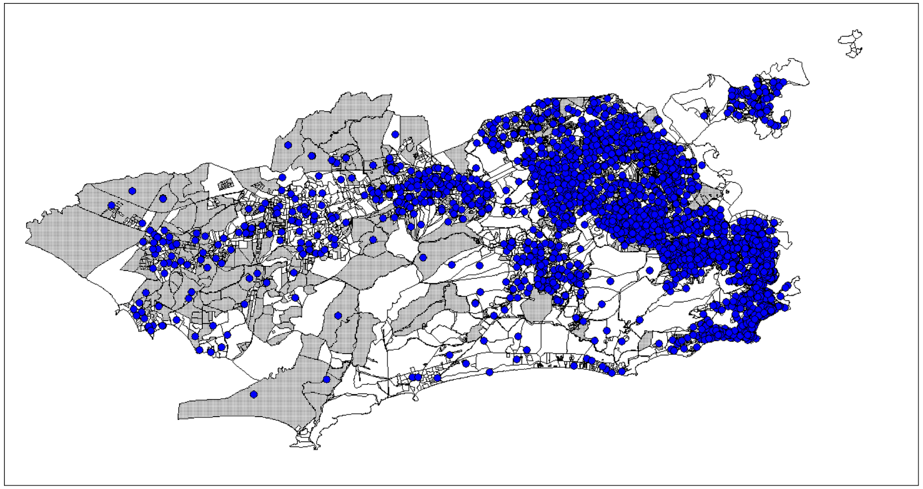


Learning data (year diagnoses 2005)

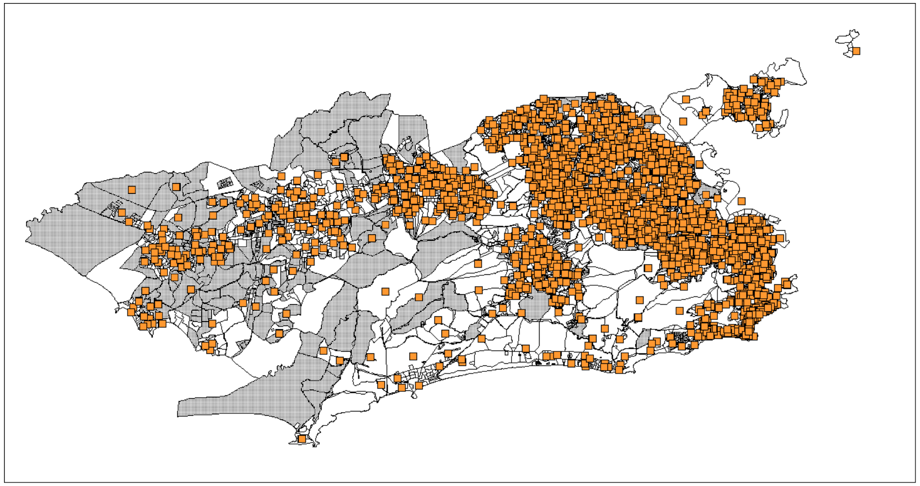


Testing data (year diagnoses 2006)

# Location Prediction

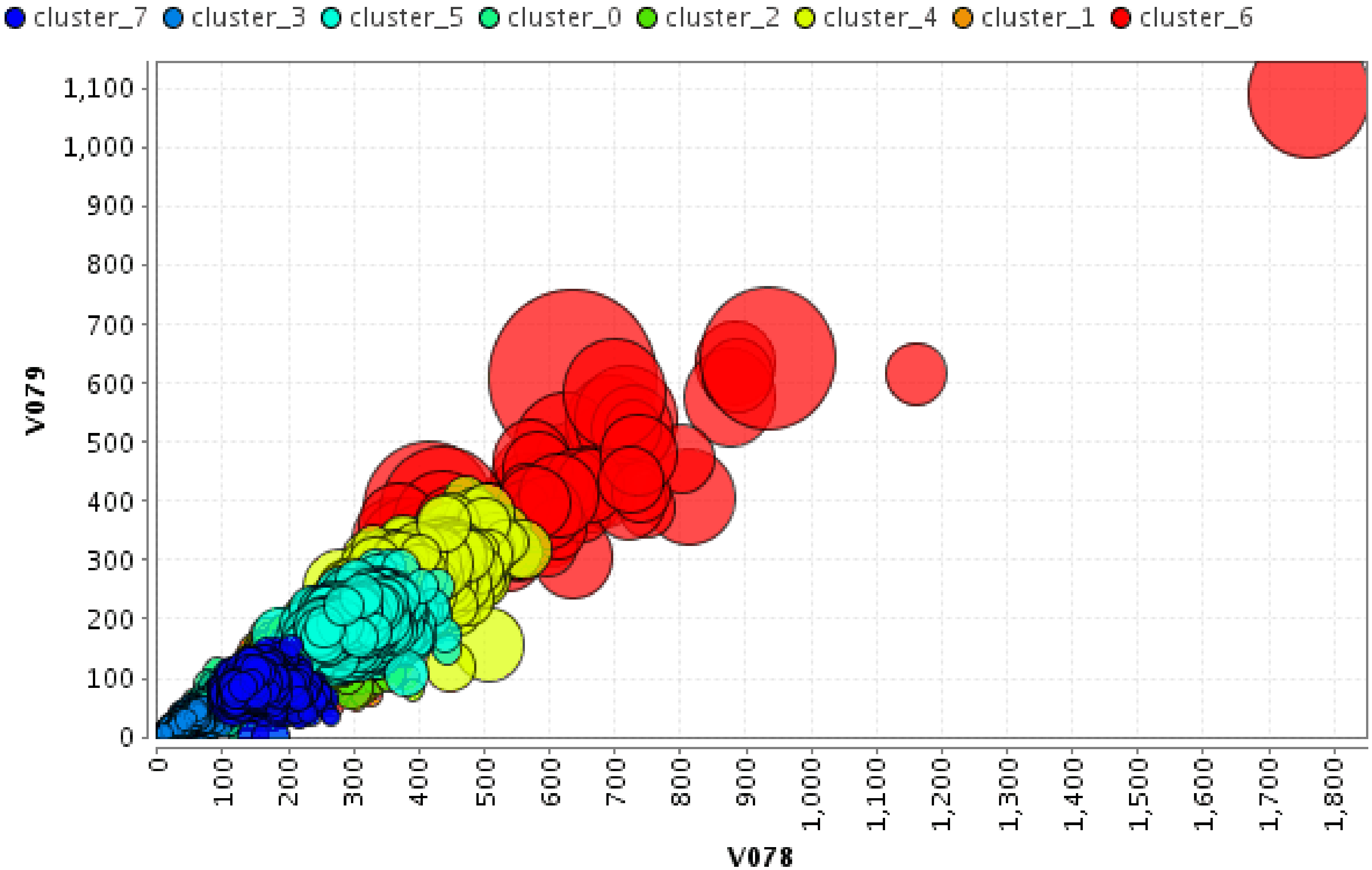


Current cases, 2010.



Predicted, 2011

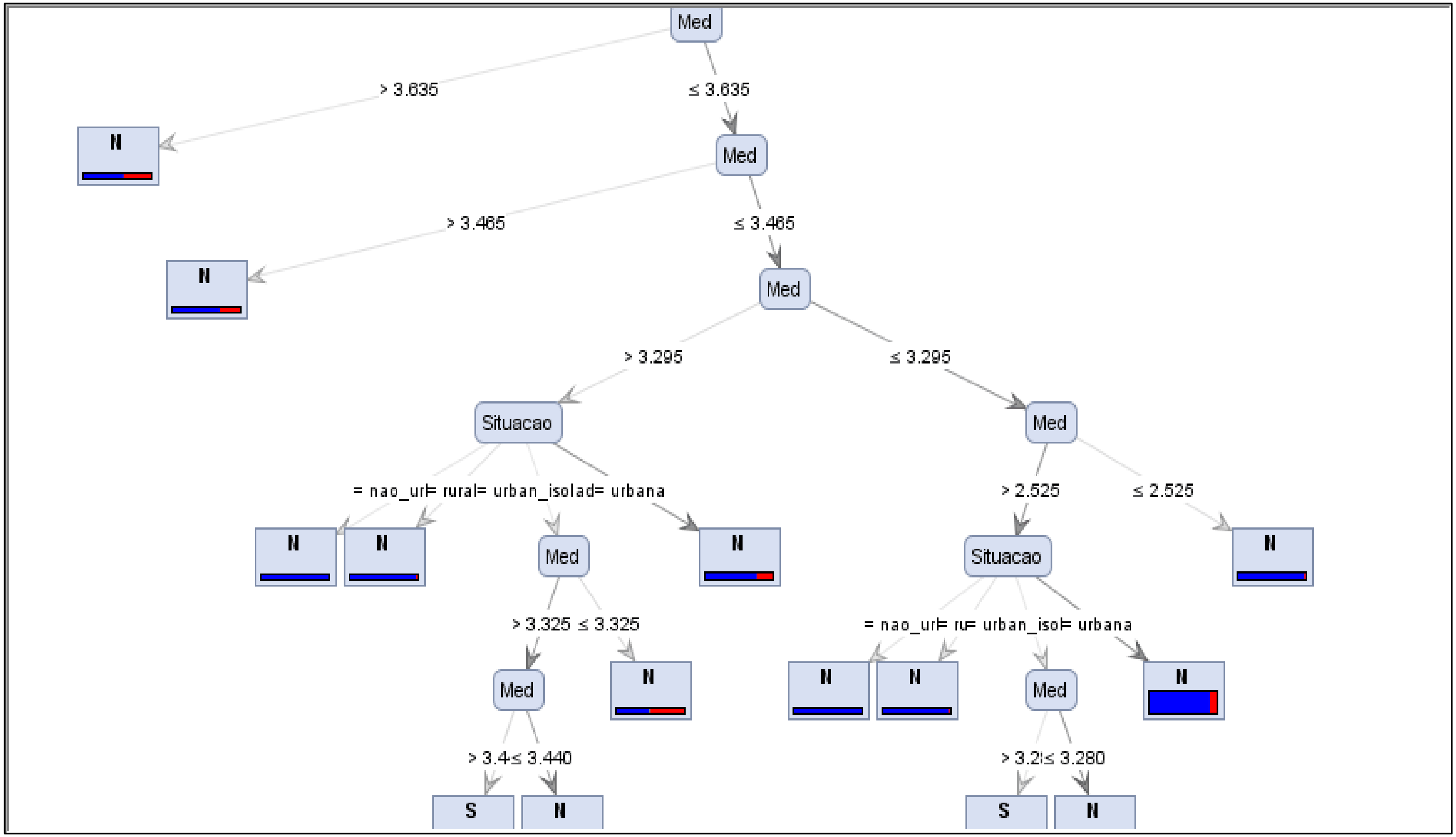
# Results



Clustering of census tracts with literate responsible and spouses age 10 or older (Census 2010)

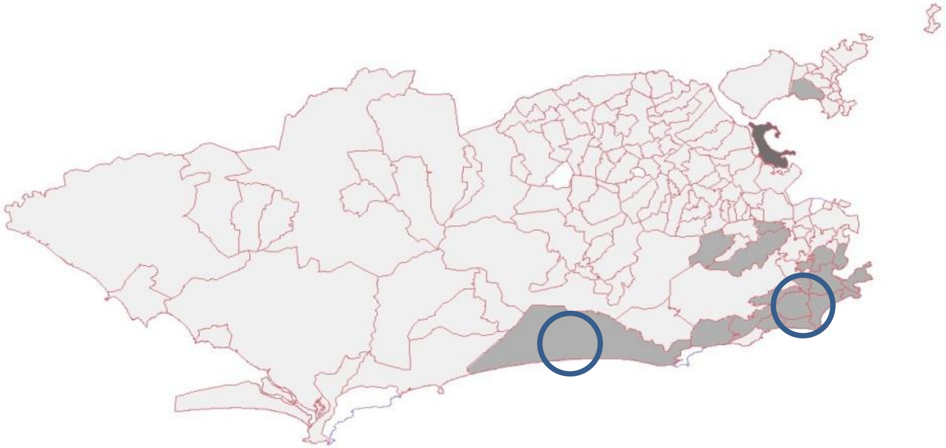
# Results

Decision Tree: persons/house, situation and type of sector Census 2010



# Results

## Spatial association rule



Target objects and final target regions (gray)

I1 ➔ I2

Support – tuples with all elements of  I1
Confidence – tuples with all I1 and I2

High proportion of occurences   ==>
total number of  woman  = very low   ^
total number of  enterprises = very low  ^
distance to the ocean = very low

# Conclusion

- ✓ Results had shown that spatial data mining had a better performance than the classical ones.

- ✓ The proposed framework is not dependent on a specific tool or software.

- ✓ The challenge facing in this research is aimed to the integration of Spatial Data Mining, Statistics, GIS, Geographic visualization and their association with geographic databases.

- ✓ It is in this context of attempting to build bridges between visualization, statistics, and data mining communities that our proposal is based on.

- ✓ Geographic visualization influences the way we interpret and analyze the data and, consequently, the construction of knowledge.

# Next Steps

- ✓ To improve the research to association rules.
- ✓ To develop a method to reduce “well-known” spatial association rules.
- ✓ To research other ways to measure *map accuracy*.
- ✓ Realize more spatial data mining on SDI

The characteristics that make geospatial data special have been acknowledged in many ways and one of the mainly important initiatives is the development of spatial data infrastructures (SDI). GKD and DM in spatial data infrastructures to date has been ad-hoc. Contributed data has not been coupled with contributed tools for data analysis and modeling. Data mining and knowledge discovery methods have not been implemented to deal effectively with new technology.

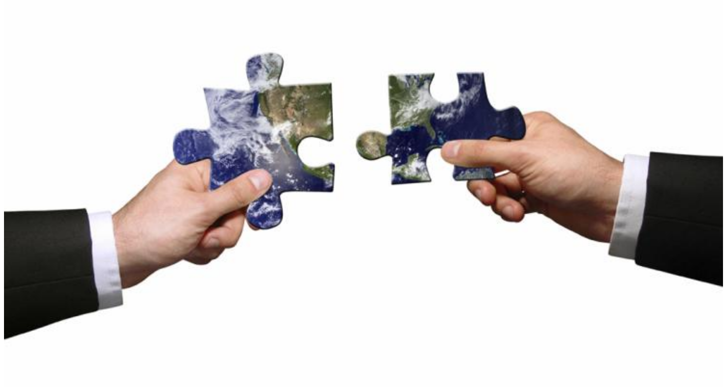
- ✓ Tools to extract patterns from spatio-temporal objects

Transformations among geographic objects over time are complex. The scales and granularities for measuring time can also be complex, preventing a simple "dimensioning up" of space to include time.

**Thank you!**

Fátima Ferrão dos Santos

fatima.santos@Ibge.gov.br



# Bibliography

IBGE, http://www.censo2010.ibge.gov.br/, jul/2012.

Fayad et. al, 1996, *From Datamining to knowledge discovery in databases*.

MATLAB, The MathWorks, Inc., http://www.mathworks.com/company, dez/2011.

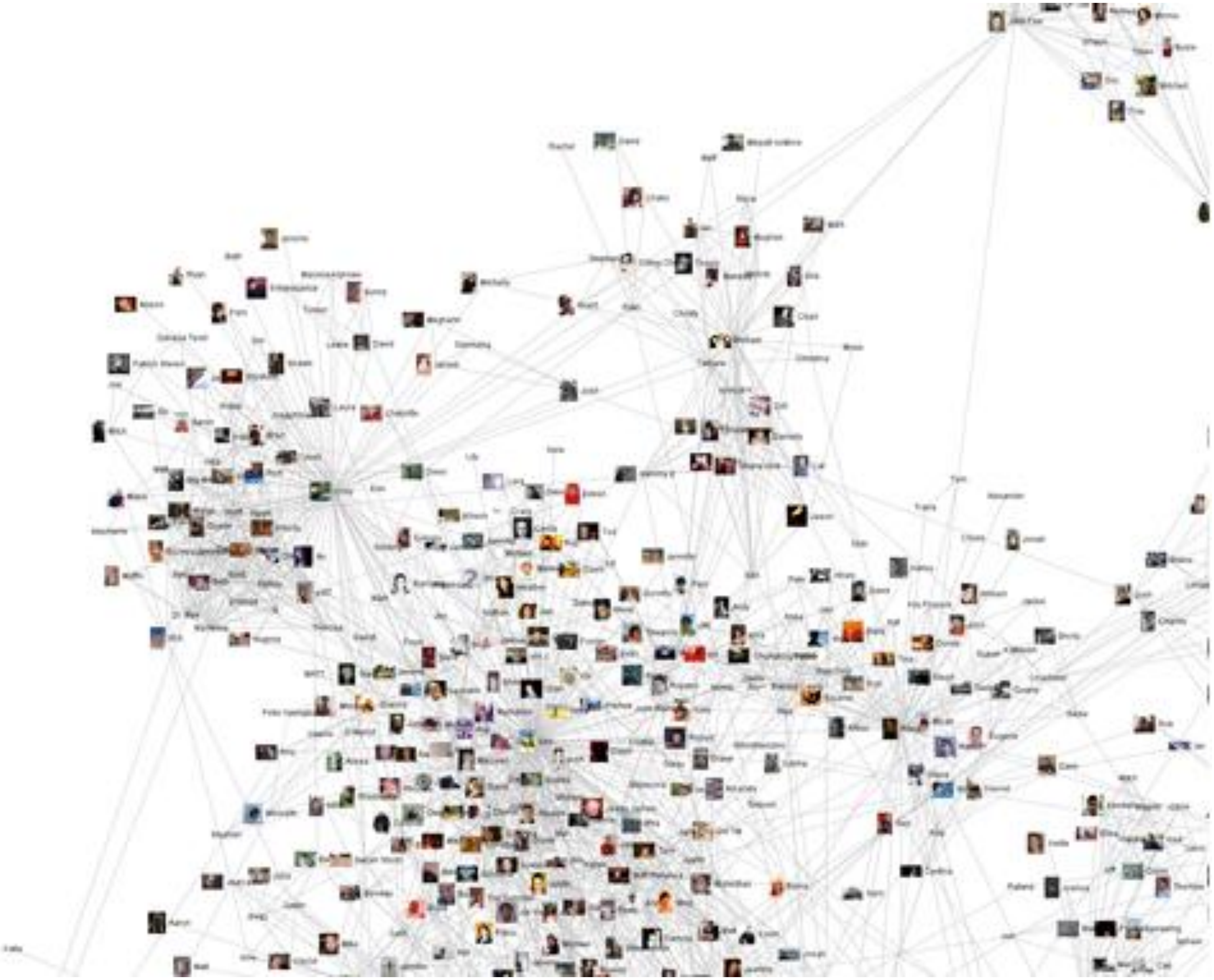
Camara et. al, 1996, SPRING: Integrating remote sensingand GIS by object-oriented data Computers & Graphics, 20: (3) 395-403, May-Jun 1996.

Santos, F.F. and Ebecken F.F., 2007, *Knowledge Discovery based on the integration of KDD and GIS*, Statistics for Data Mining, Learning and Knowledge Extraction, pp 45-47, Aveiro, Portugal

Santos, F.F., *Knowledge Discovery on Health databases based on the integration of geographic data mining and complex networks*, 2008.

Torres, H., *Segregação residencial e políticas públicas*: São Paulo na década de 1990, Revista Brasileira de Ciências Sociais, v. 54, p.41-56, 2004.

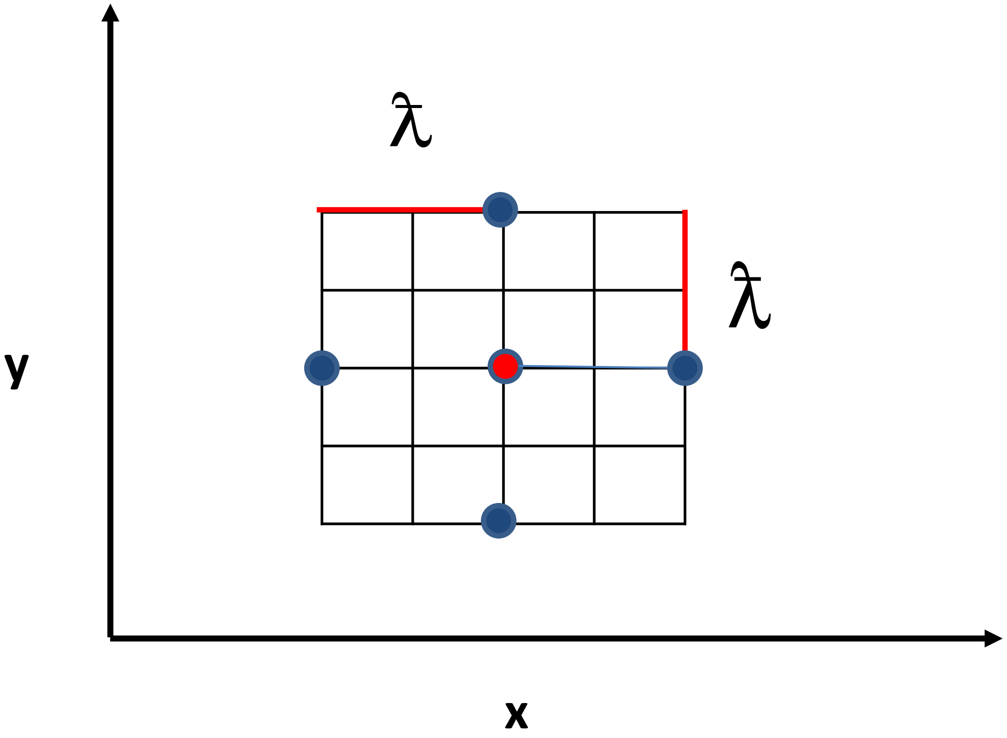
# Networks



IBGE consumers

# Proposal 1

If we have two explanatory value x and y. Its neighbourhood may include :



$(x + \lambda , y)$

$(x - \lambda , y)$

$(x , y + \lambda)$

$(x, y - \lambda)$

Then we evaluate the map-similarity measure on each parameter value tuple in this neighbourhood. The neighbour with the highest value of map-similarity measure is chosen. This process is repeated and it ends when no neighbour has a higher value of map-similarity measure. A local maxima has been found.

# Spatial Autoregressive Model

The benefits of modeling spatial autocorrelation are many:

1. The residual error will have much lower spatial autocorrelation, that is systematic variation.

2. If the spatial autocorrelation coefficient is statistically significant then it will quantify the presence of spatial autocorrelation.

SAR equation

As in the case of classical regression, the SAR equation has to be transformed via the logistic regression function for binary dependent variables. The estimates of $p$ and $\beta$ can be derived using maximum likelihood theory or Bayesian statistics.

The Bayesian approach using sampling based MCMC methods (Matlab) is computacional expensive (with slow convergence) and it is a non-trivial task to decide what ‘priors’ to choose and what are the appropriate analytic expressions for the conditional probability distributions.

# The virus HIV epidemic

All notified cases of persons 13 years old or over, with one year diagnosis in the period 1982 to 2005 (17.000 cases) (Fiocruz).

1. The exploratory data analysis was fulfilled according to:

- sex and
- category: homo/bisexuals, heterosexuals, injecting drugs users, blood and ignored (Join United Nations Programme on HIV/AIDS – UNAIDS (1999).

2. Considering spatial analysis, two cuts were established:

- incidency rate temporal evolution according to the municipal district tradicional division into neighborhoods.

3. Several sanitary, socio economic indicators , such as population size and poverty concentration measured by the rate of heads of the family with month income lower than two  minimum wage.