SELF-ORGANIZING MAPS (SOM)

Advanced Spatial Analytics
24.1.2019
Jaakko Madetoja
(slides also by Olga Špatenková)

Today

- Self-organizing maps:
  - Theory
  - Example
  - Exercise

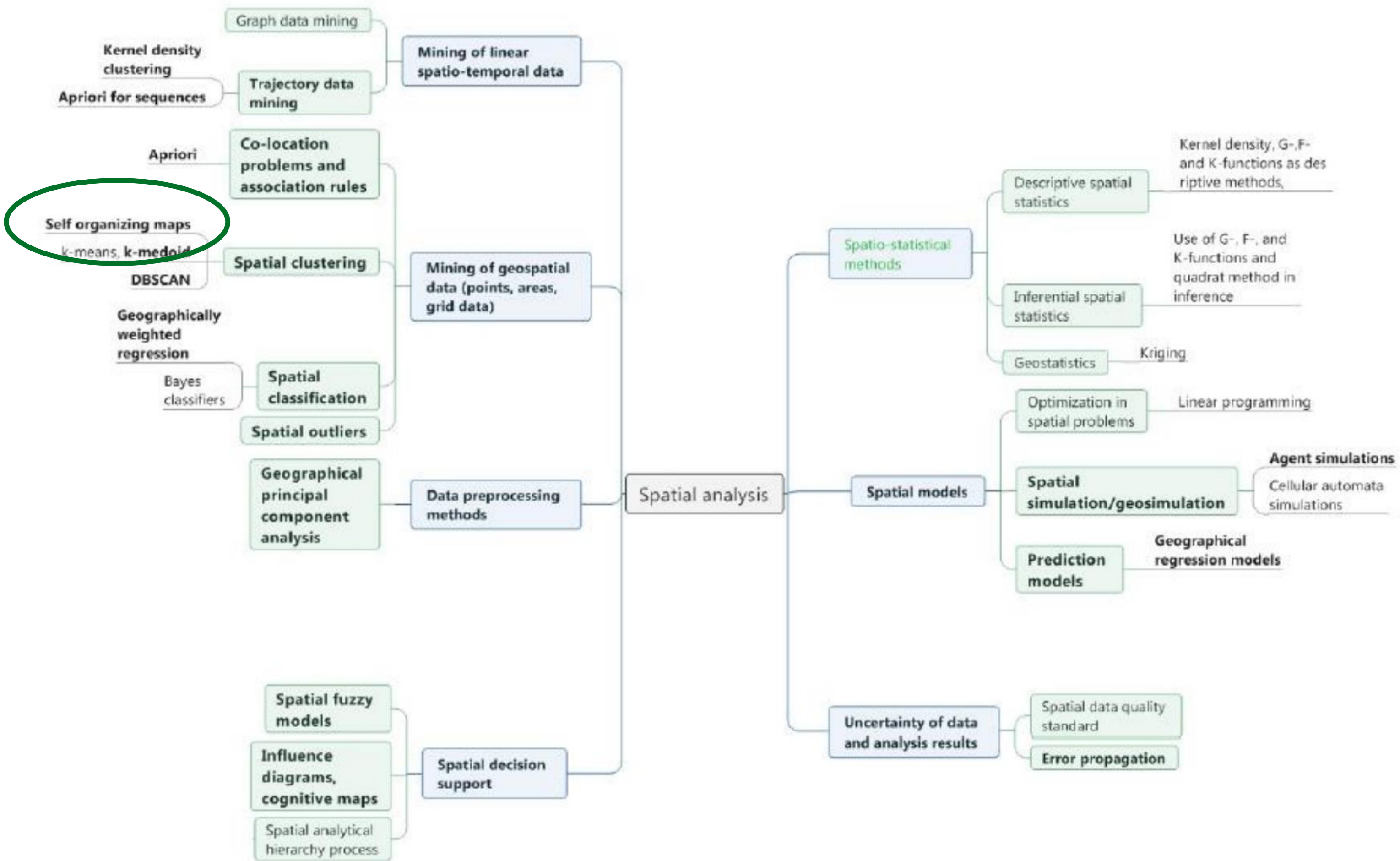
Learning goals

- After this lecture, you are able to
  - Explain how training an SOM and mapping values to it works
  - Explain how you can use SOM for clustering and finding correlations in the data set

Self-organizing Maps (SOM)

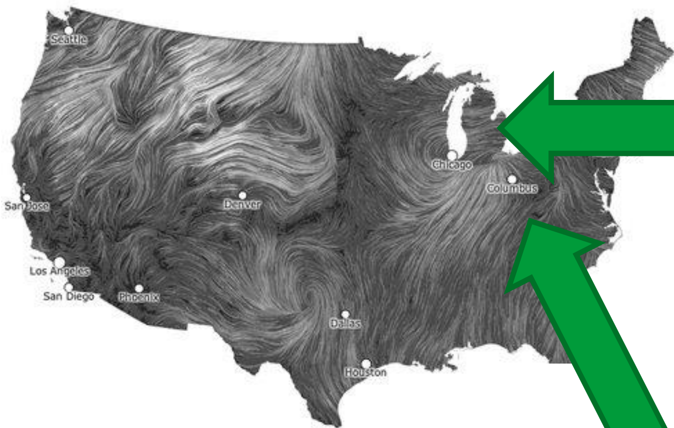
- References
  - **Špatenková, Olga: Discovering Spatio-Temporal Relationships: A Case Study of Risk Modelling of Domestic Fires, 2009, http://lib.tkk.fi/Diss/2009/isbn9789522482334/isbn9789522482334.pdf**
  - T. Kohonen: Self-Organizing Maps, 2nd edition, Springer Verlag 1997
  - R. Silipo: Neural Networks, in Berthold&Hand (eds), Intelligent Data Analysis, 2nd edition, Springer Verlag 2003
  - J. Vesanto: SOM-based data visualization methods, Intelligent Data Analysis, 3:111-126, 1999
  - E. L. Koua, M-J. Kraak: Alternative visualization of large geospatial datasets. The Cartographic Journal, 41:217–228, 2004

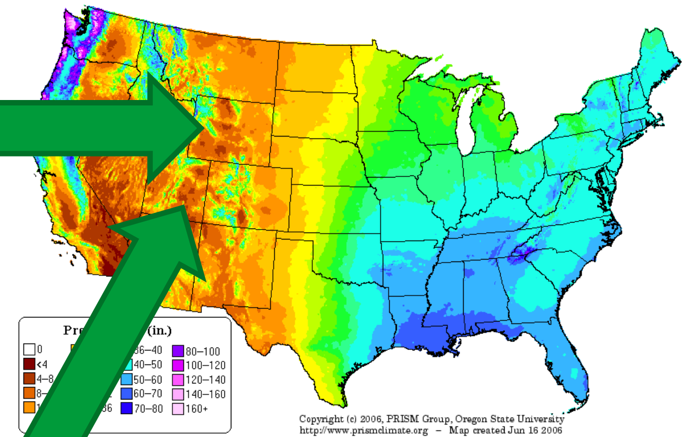
The big picture

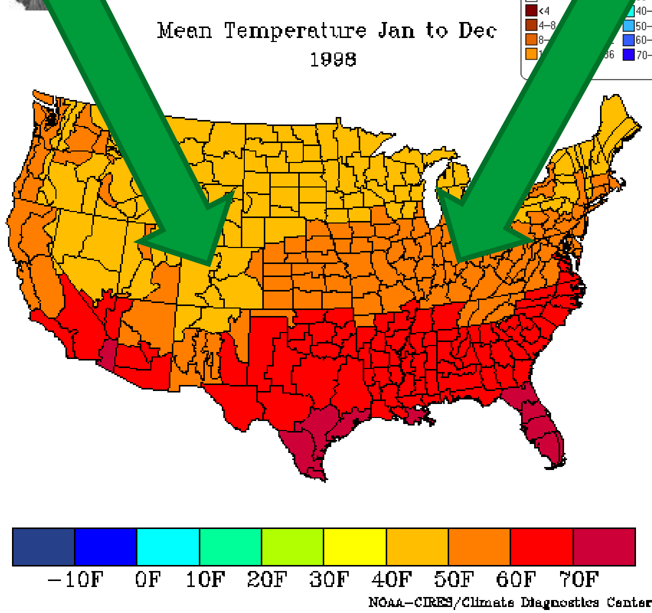


Introduction

Multivariate data: wind, precipitation, temperature







Task: describe the relationships between the attributes

Self-organizing Maps (SOM)

- Simplest possible description: SOM is a method to organize multivariate data and can be used to visualize different attributes

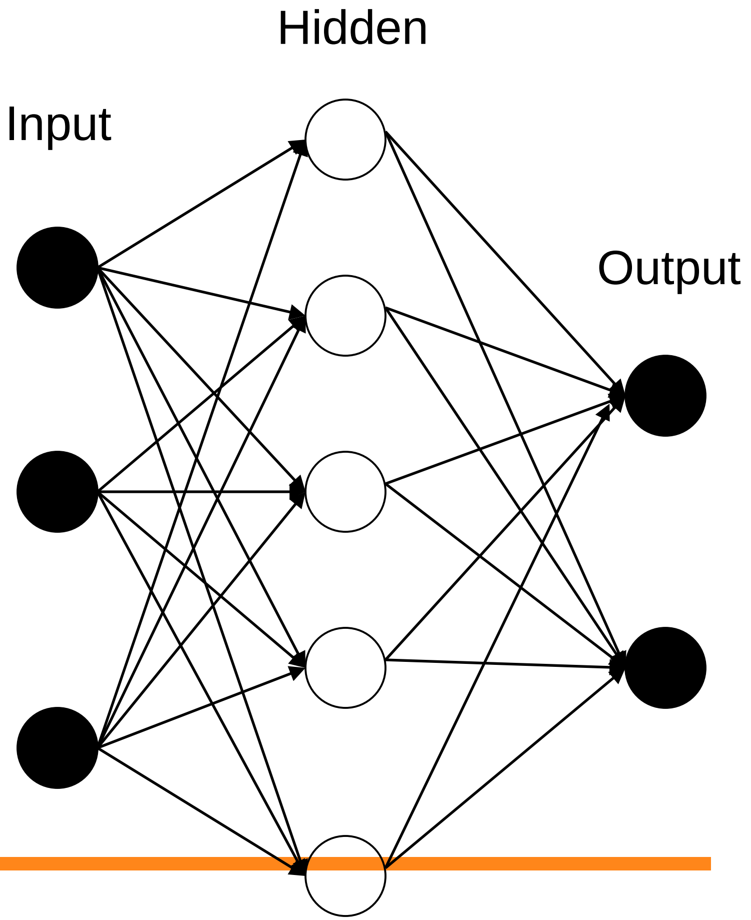
- SOM is an artificial neural network capable of distinguishing similarity patterns

- It is not a map in a traditional (cartographic) sense

- Some background next

Artificial Neural Networks

- Inspiration from biological NN
- Neurons (processing elements), connections
- Adaptive system – by weights (strength of connections)
- used to model complex relationships between inputs and outputs or to find patterns in data



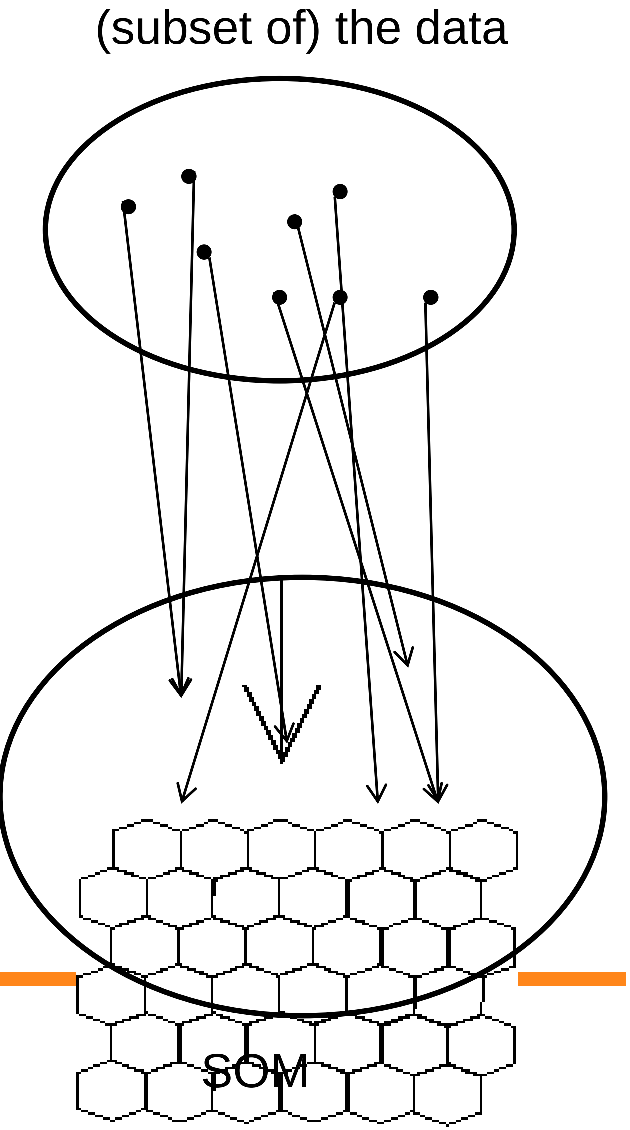
ANN Applications

- Inferring a function from observations

- Classification
- Pattern recognition
- Compression
- Clustering
- Function approximation
- Time series prediction
- etc.

SOM

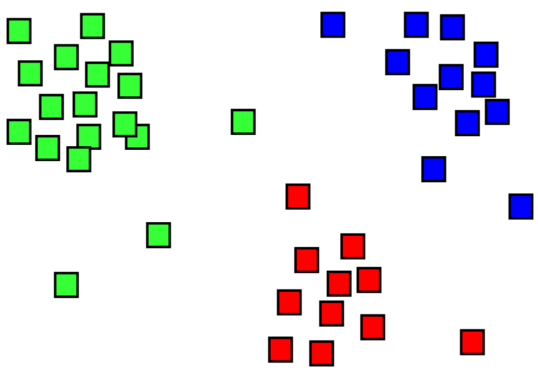
- Unsupervised neural network (no outputs defined)
- Maps multidimensional data onto a two-dimensional lattice of cells: each data object will be mapped to one cell (also called neurons)
- Each cell has the same dimensions as the data

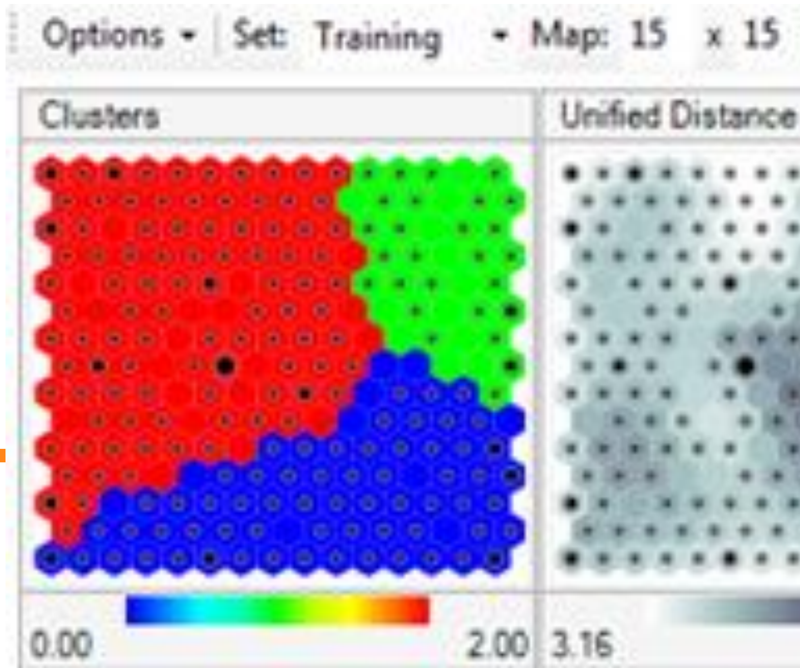


Topology

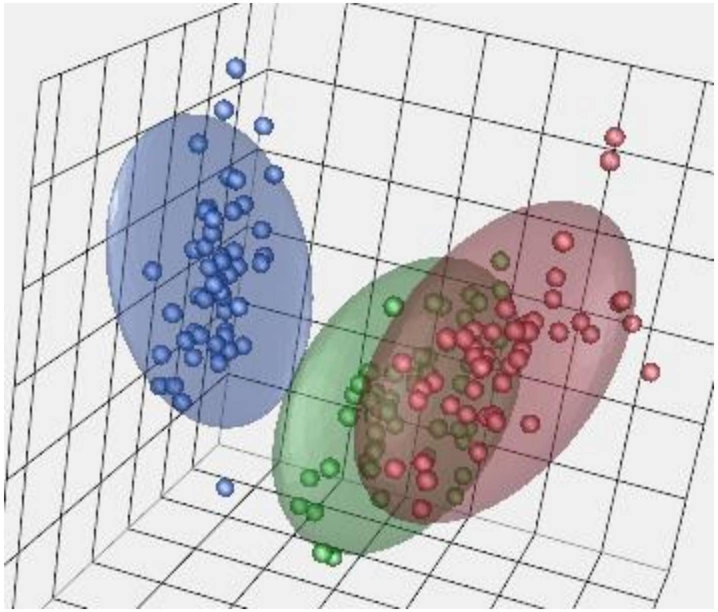
- SOM preserves topology and similarity patterns existing in the original space

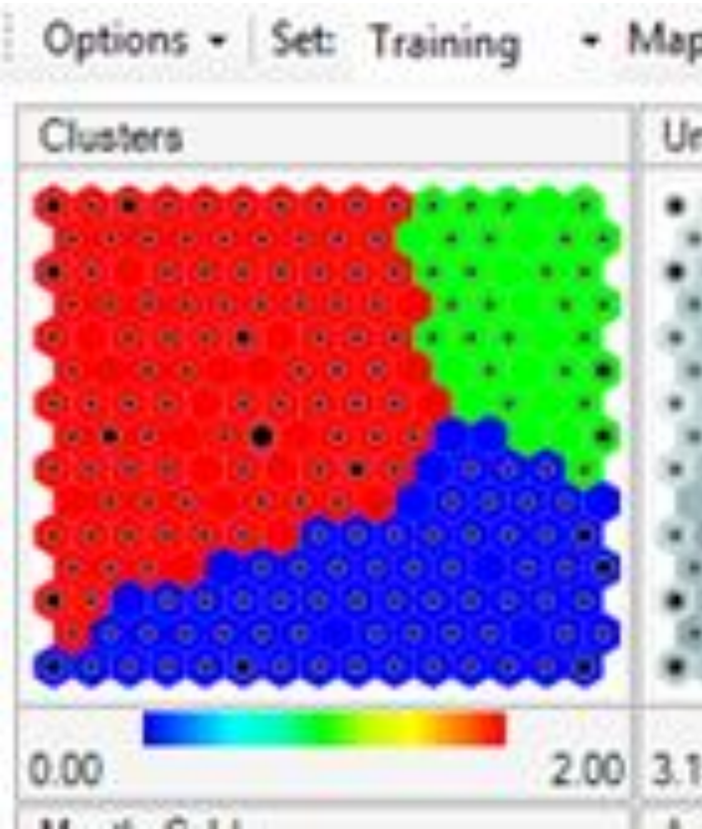
close by data items are mapped close to each other in SOM





Topology 3D

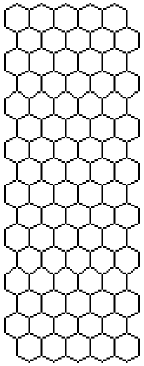




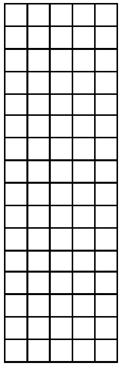
Lattice

the most common

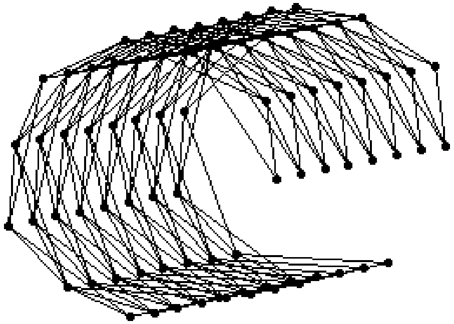
Hexagonal lattice



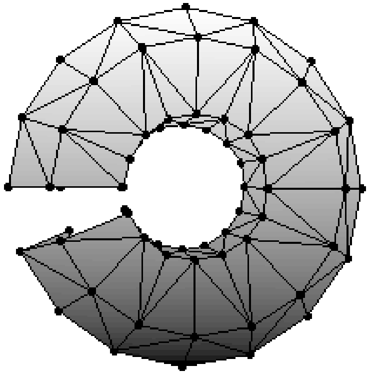
Rectangular lattice



Cylinder shape



Toroid shape



SOM Algorithm

- Training
  - Map construction based on input sample data

- Mapping
  - Automatic classification of a new input

Training

The training utilizes competitive learning:

1) Initialize the neuron values (called weights)
   - can be random or using for example principal component analysis
2) Pick a data (vector) sample and find the closest weight. This neuron is called Best matching unit (BMU)
3) The weight of the BMU and its neighbors are changed to be closer to the data sample
4) Pick another data sample and continue from step 2

The neighborhood shrinks with each iteration: at the beginning more neurons are affected and later only few weights are changed.

Updating the weights

Formula used in updating the weights (step 3 in previous slide):

old weight for the neuron

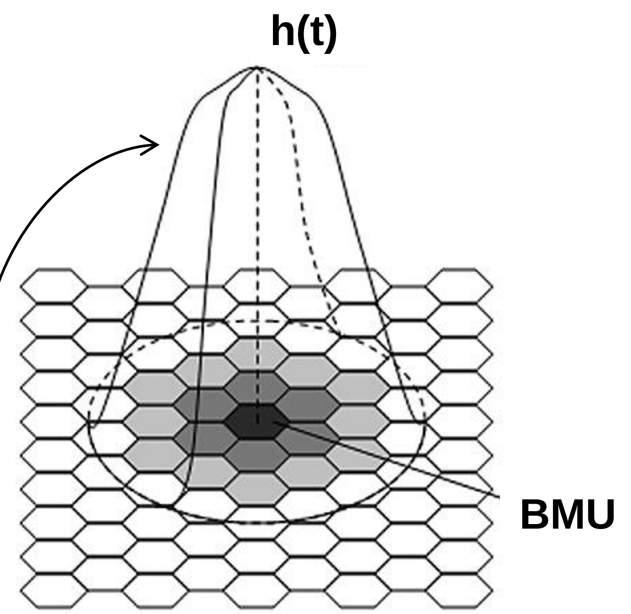
data vector (i.e. attribute values)

$$w(t+1)=w(t)+h(t)\alpha(t)(D(t)-w(t))$$

new weight for the neuron

neighborhood function

monotonically decreasing coefficient: makes sure that the neighborhood shrinks over time



Training and mapping

- Result of training:
  - All neurons (or cells) represent a model of input data (remember that each neuron has same attribute space as the data)
  - Close by neurons have similar attribute values in attribute space
- Mapping:
  - New input data is automatically classified to single winning neuron
  - Some data items can mapped to the same neuron and some neurons can have a situation with no data mapped to them

SOM Quality

- Since the projection to 2D lattice reduces dimensionality, information is lost during the process
- Balance between data representation accuracy and data set topological accuracy
  - average quantization error between data vectors and their BMUs on the map: how well the SOM represents the data
  - topological error measure: percentage of data vectors for which the first- and second-BMUs are not adjacent units (often called topographic, which seems incorrect)

Visual Representation of Clusters

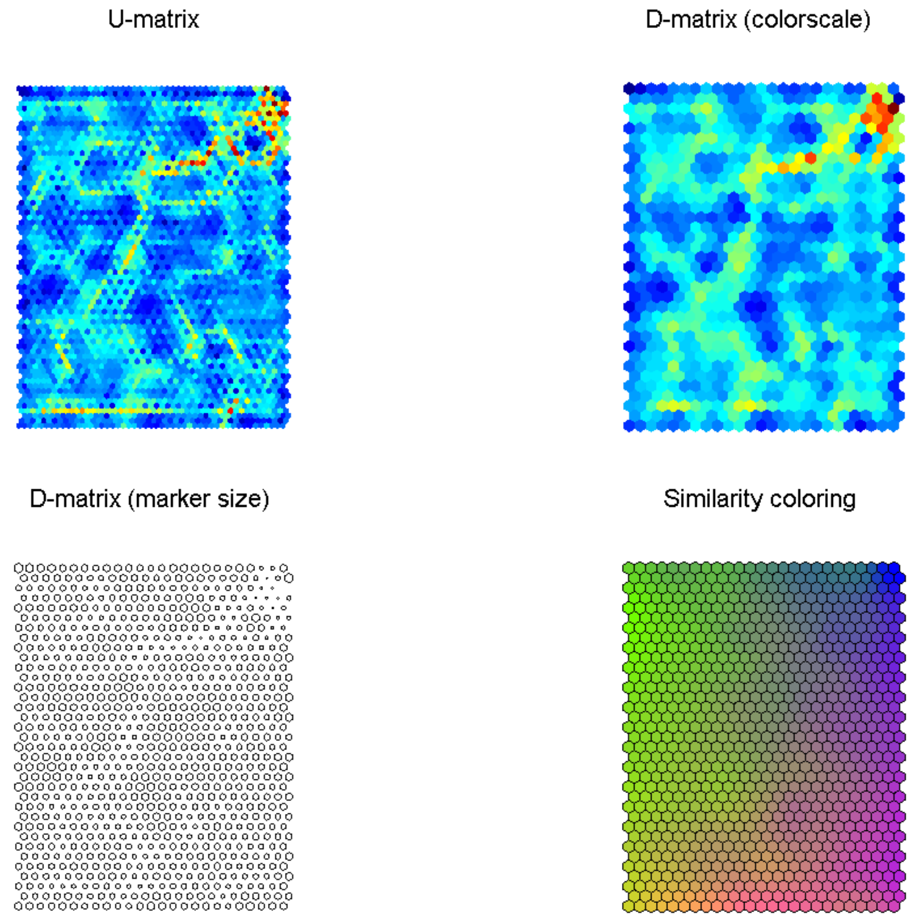
U-matrix: distance between a neuron and its neighbors

D-matrix: average of these distances

How can D-matrix be utilized:

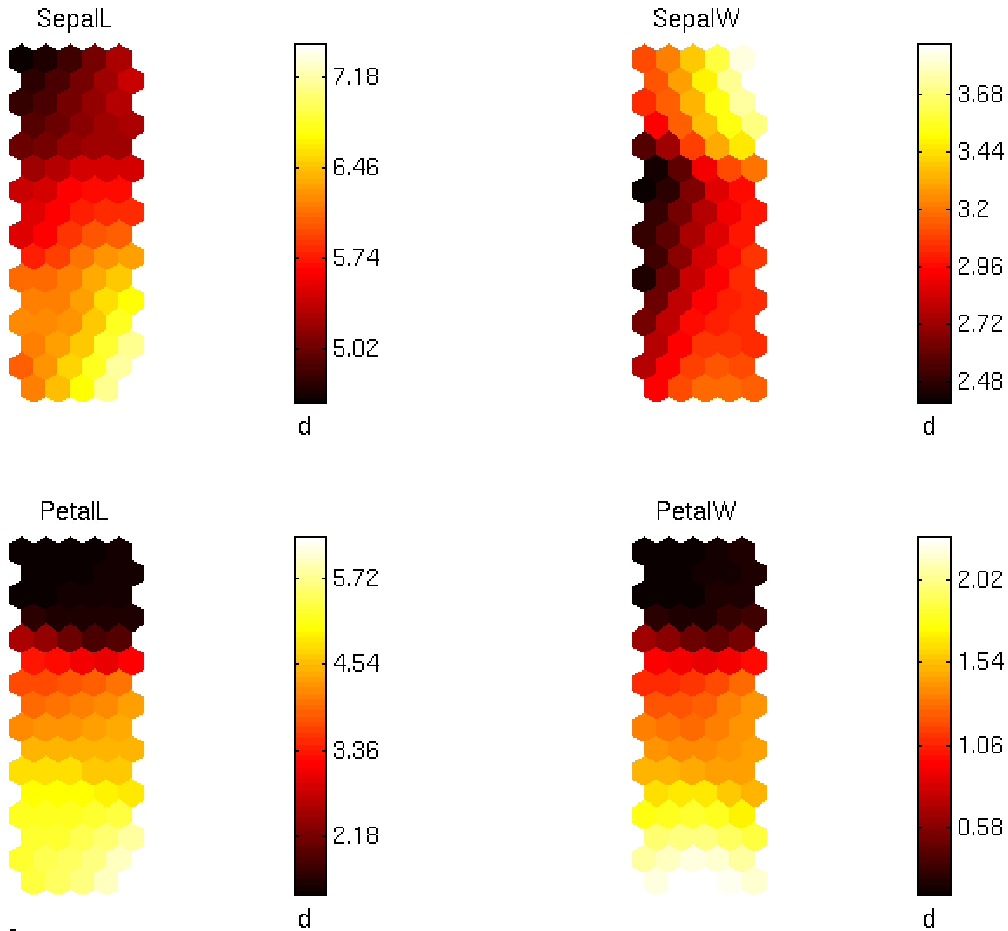
- Small values close to each other: a cluster
- A line of big values: a border between clusters

However, generally clustering is done by applying another algorithm (e.g. k-means) for the neuron weights



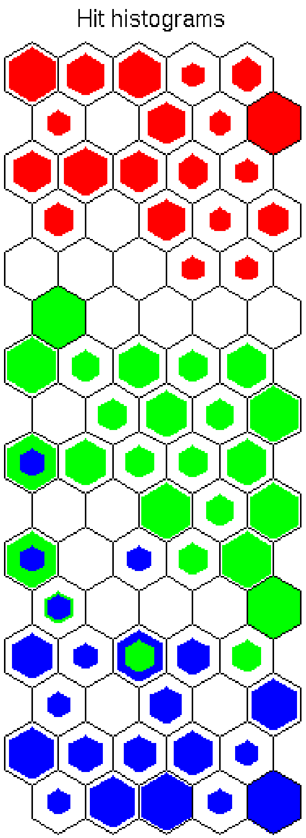
Visual Representation of Variables

As SOM has the same variables (or dimensions) as the input data, we can visualize them to view and describe the data

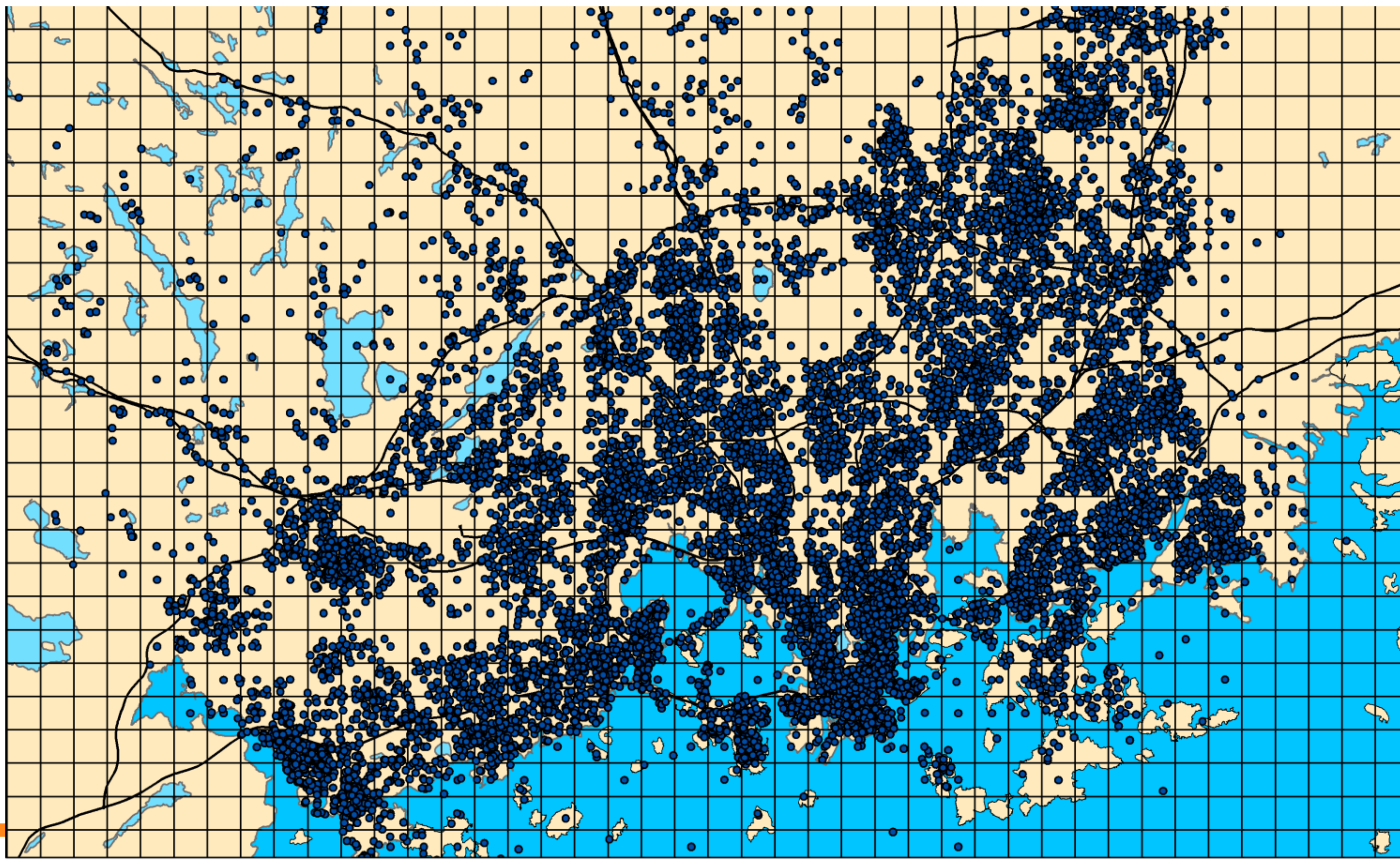


Visual Representation of Data Projections

Histogram tells how many data items have been classified to a neuron



Example: Incidents in Helsinki Metropolitan Area



Locations for fire & rescue service missions
in the Helsinki metropolian area 2004-2006

0 2,5 5 10 Kilometers

N

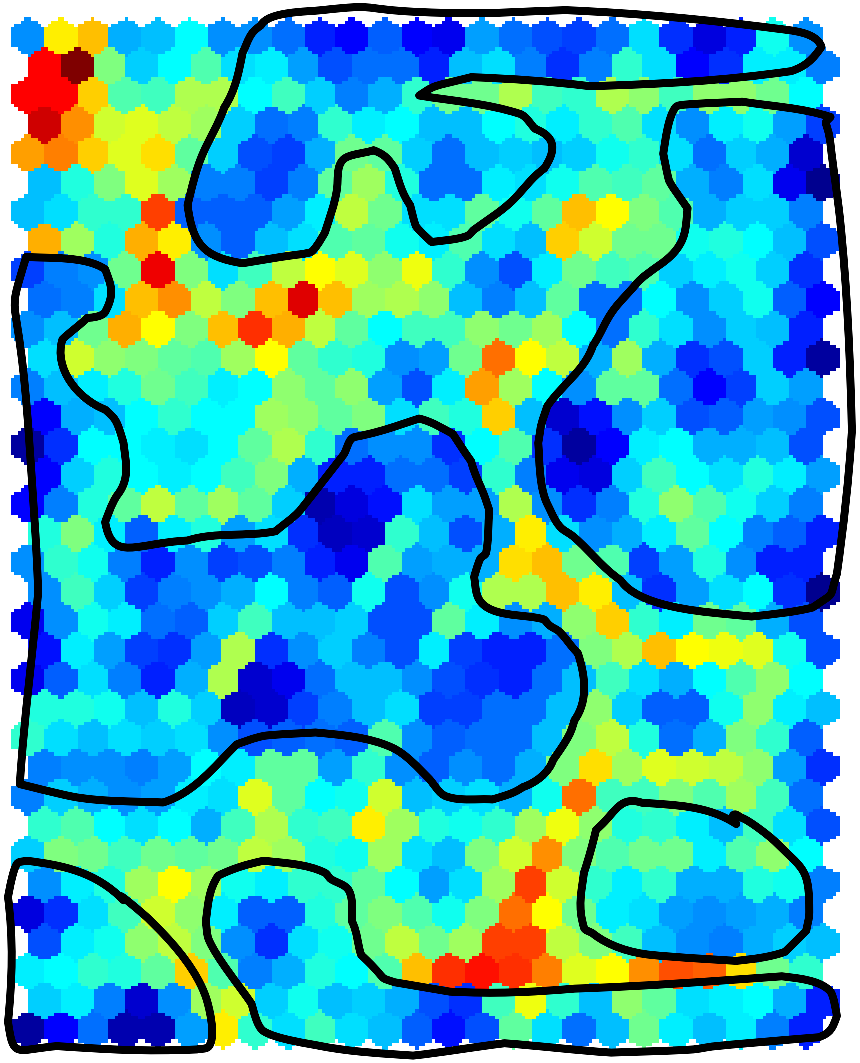
TKK-GiP 2007
Data source: PRONTO fire & rescue service missions, Helsinki 2004-2006, YTVSeutuCD background data

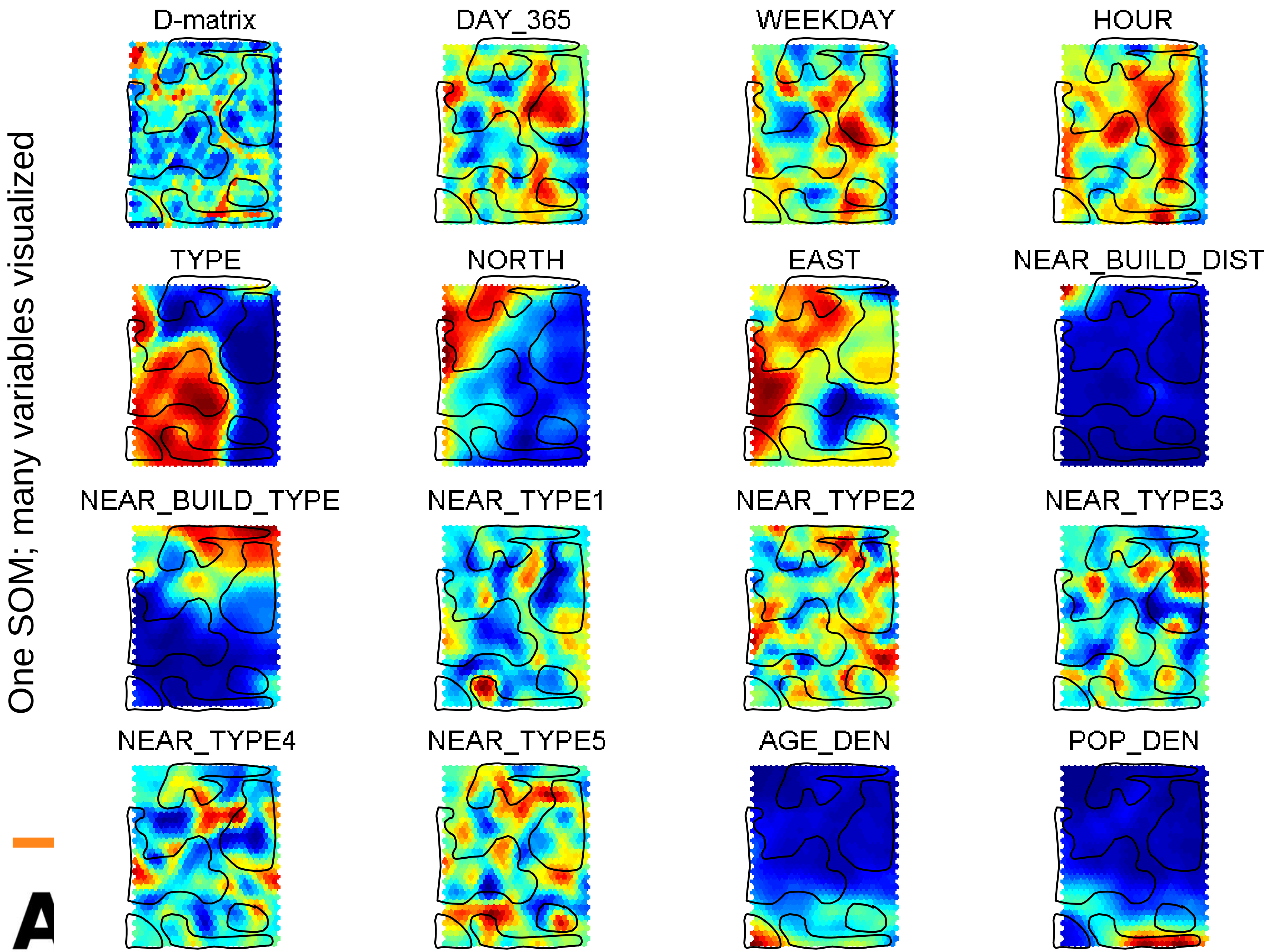
Data

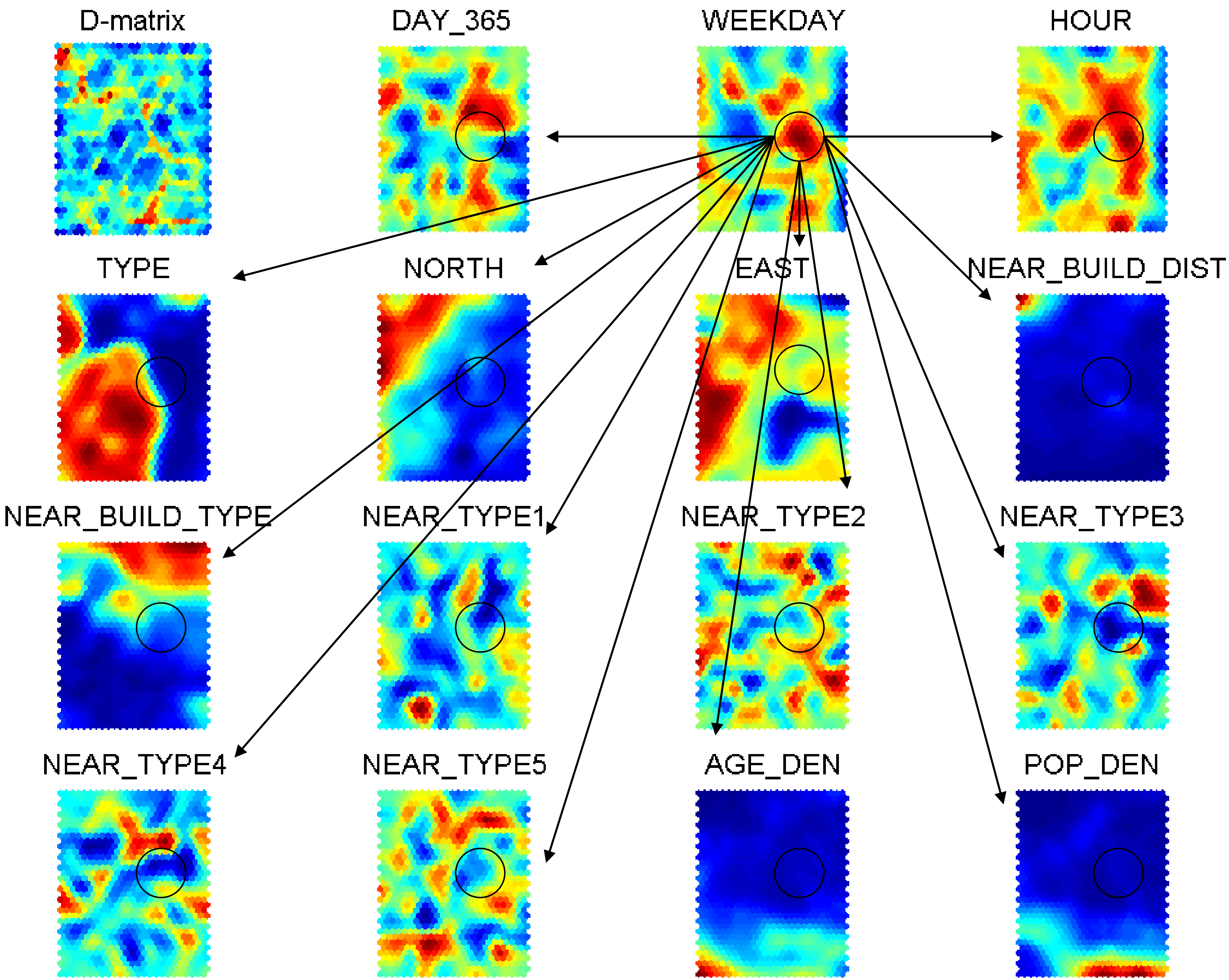
- Rescue incidents
  - Day in the year
  - Day of the week
  - Hour of the day
  - Type of the incident
  - X coordinate
  - Y coordinate
  - Type of the five nearest incidents
- Background information
  - Distance to the nearest building
  - Type of the nearest building
  - Population density
  - Age density

D-matrix

clusters




One SOM; many variables visualized
D-matrix
DAY_365
WEEKDAY
HOUR
TYPE
NORTH
EAST
NEAR_BUILD_DIST
NEAR_BUILD_TYPE
NEAR_TYPE1
NEAR_TYPE2
NEAR_TYPE3
NEAR_TYPE4
NEAR_TYPE5
AGE_DEN
POP_DEN


D-matrix
DAY_365
WEEKDAY
HOUR
TYPE
NORTH
EAST
NEAR_BUILD_DIST
NEAR_BUILD_TYPE
NEAR_TYPE1
NEAR_TYPE2
NEAR_TYPE3
NEAR_TYPE4
NEAR_TYPE5
AGE_DEN
POP_DEN

Conclusions

- Spatio-temporal multivariate data analysis for
  - Clustering
  - Data characterization
  - Correlation hunting

PROS x CONS

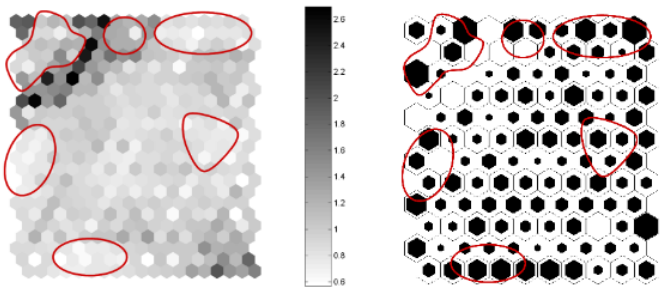
- Mathematical basis
- Easy visual interpretation
- Treats all attributes at the same time (also spatial and temporal)
- Preserves topology and data distribution in the input space

- Subjective
- Time consuming pre-processing
- Details obscured
- Missing connection between the software used (SOM toolbox for Matlab) and a map
- Not a spatial model

Your turn!

- Your task: Interpret the given SOM (from Špatenková, 2009)
- Team 1: Describe clusters
- Team 2: Characterize morning and evening fires
- Team 3: Describe clusters
- Team 4: Identify differences between evening, day and night fires

Team 1



(a) (b)

Figure 8.2: Clusters identified from the distance matrix as light areas delimited by darker colours (a) and corresponding data histogram (b) of the SOM for the incident dataset.

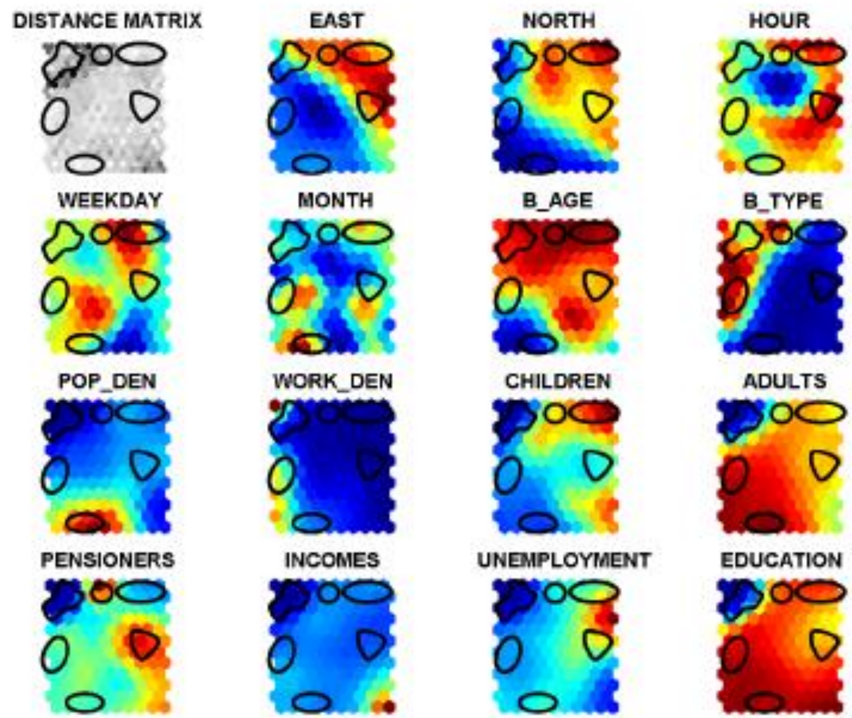


Figure 8.3: Location of the identified clusters in the component planes. The colour scale goes from dark blue (low values) through green (medium values) to dark red (high values).

Team 2

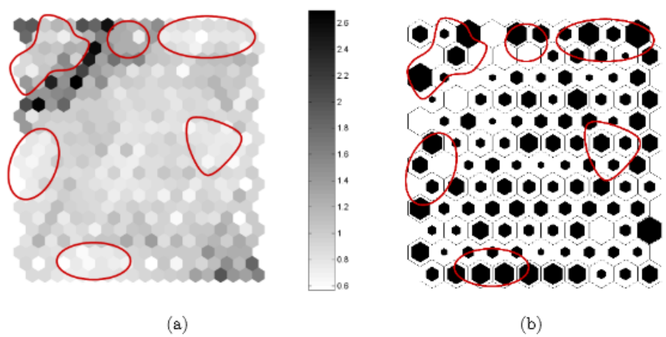


Figure 8.2: Clusters identified from the distance matrix as light areas delimited by darker colours (a) and corresponding data histogram (b) of the SOM for the incident dataset.

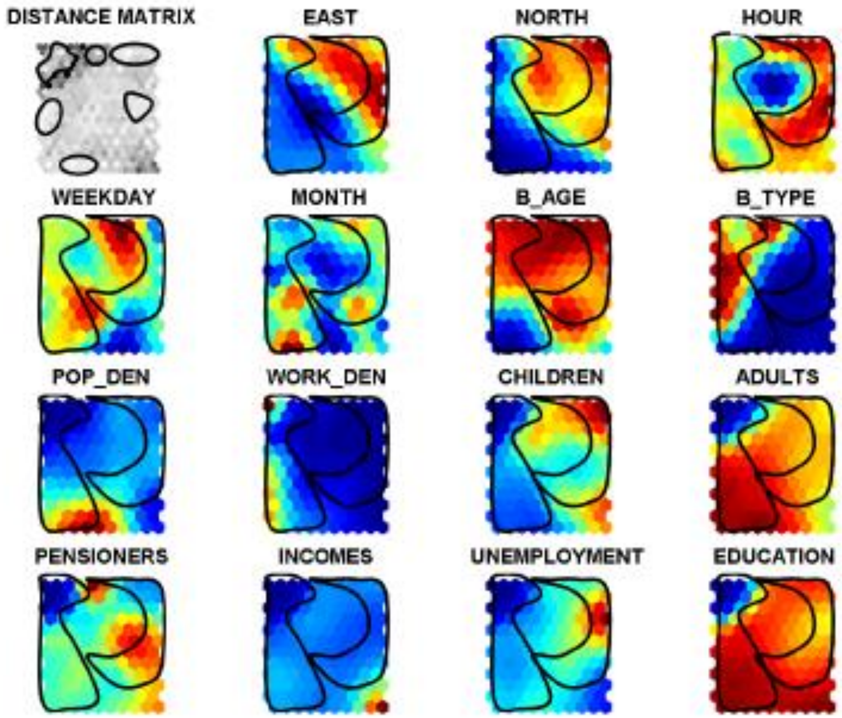
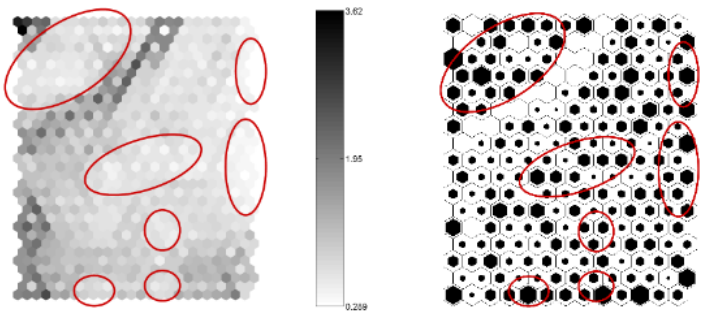


Figure 8.4: Discovering relationships between the attributes from the component planes. The green and red colours in the HOUR component plane indicate morning and evening fires, respectively. The characteristics of these incidents can be found from the remaining component planes. The colour scale goes from dark blue (low values) through green (medium values) to dark red (high values).

Aalto University

Team 3



(a) (b)

Figure 8.5: Clusters identified from the distance matrix (a) and corresponding data histogram (b) of the SOM for the grid representation of the data.

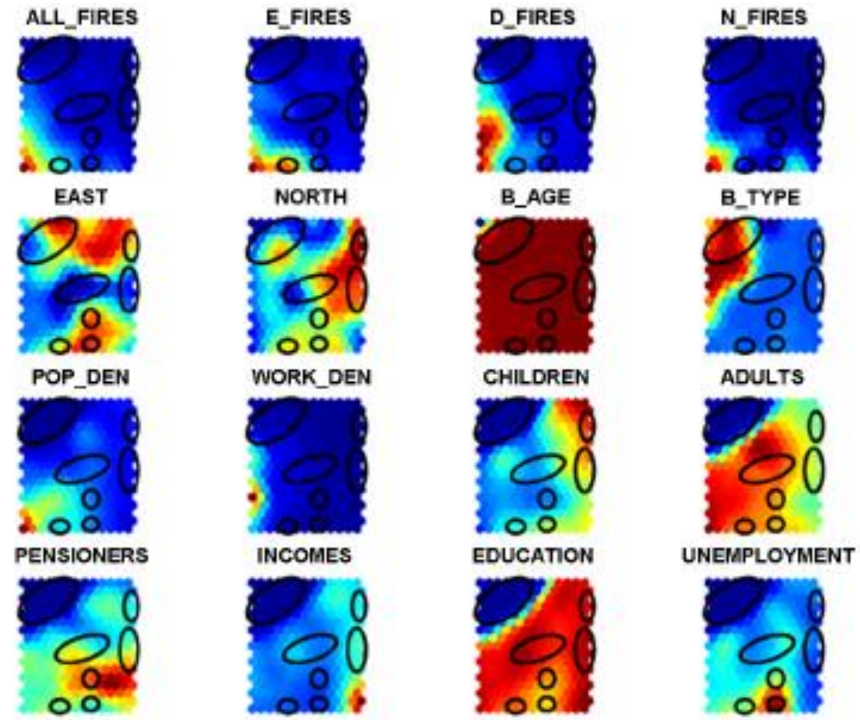


Figure 8.6: Location of the identified clusters for the grid representation of the data in the component planes. The colour scale goes from dark blue (low values) through green (medium values) to dark red (high values).

Team 4

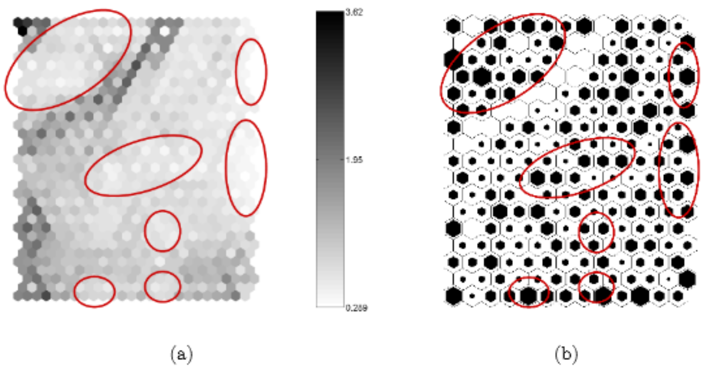


Figure 8.5: Clusters identified from the distance matrix (a) and corresponding data histogram (b) of the SOM for the grid representation of the data.

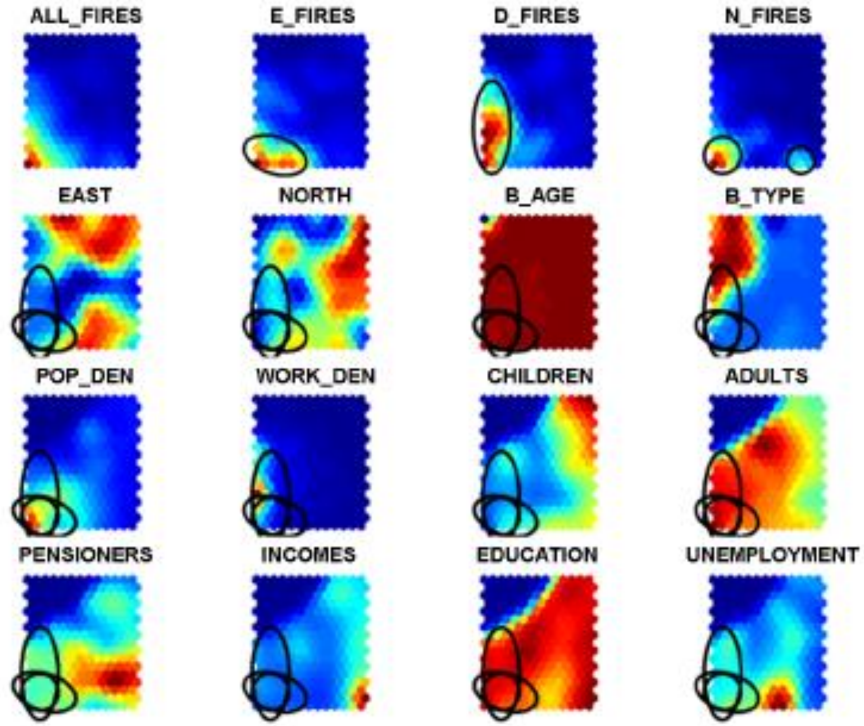


Figure 8.7: Identifying differences between e-fires, d-fires, and n-fires in the grid representation from the component planes. The colour scale goes from dark blue (low values) through green (medium values) to dark red (high values).

Thank you!