Host institution: Faculty of Forestry, University of Agriculture in Krakow

Host: Piotr Wężyk, Assoc. Prof., Laboratory of Geomatics, Institute of Forest Resource Management
Al. 29 Listopada 46, 31-425 Krakow, Poland

Telephone number: +48-126625082 Fax.+48-124119715 web page
geo.ur.krakow.pl

STSM title: Organization of knowledge in large datasets: clustering algorithms and strategies

STSMS dates: July 3th-13th, 2016

Report of the visit

Data mining is an ubiquitous theme that is present in many fields of science: surely in pattern recognition, classification of images, detecting communities in networks, and even decision trees and neural networks.

Dr Wezyk at and University of Kracow, is actually engaged in a challenge launched by the European Commission on the mapping of high resolution satellite data and Lidar data. For “Big Data” a conceptual approach must be proven to have a practical impact. A typical modern and urgent topic is the problems encountered using “Big Dataset” , in this particular case the (Aerial) Laser derived set of all buildings in the city from very large point data sets (12 points per square meter on more than 300 km2). This requires to analyse millions of image objects and their topological relationships in order to create a computer generated map. One example can be given as a use case in the mapping of urban areas (in accordance with “Mapping guide for a European Urban Atlas”, EU Commission). The task requires many levels of knowledge: on the specific topics, on the management of images from satellites, and, last but not least, on the transformation of raw images into significant knowledge, to be used for further purposes (Wezyk et al 2008).

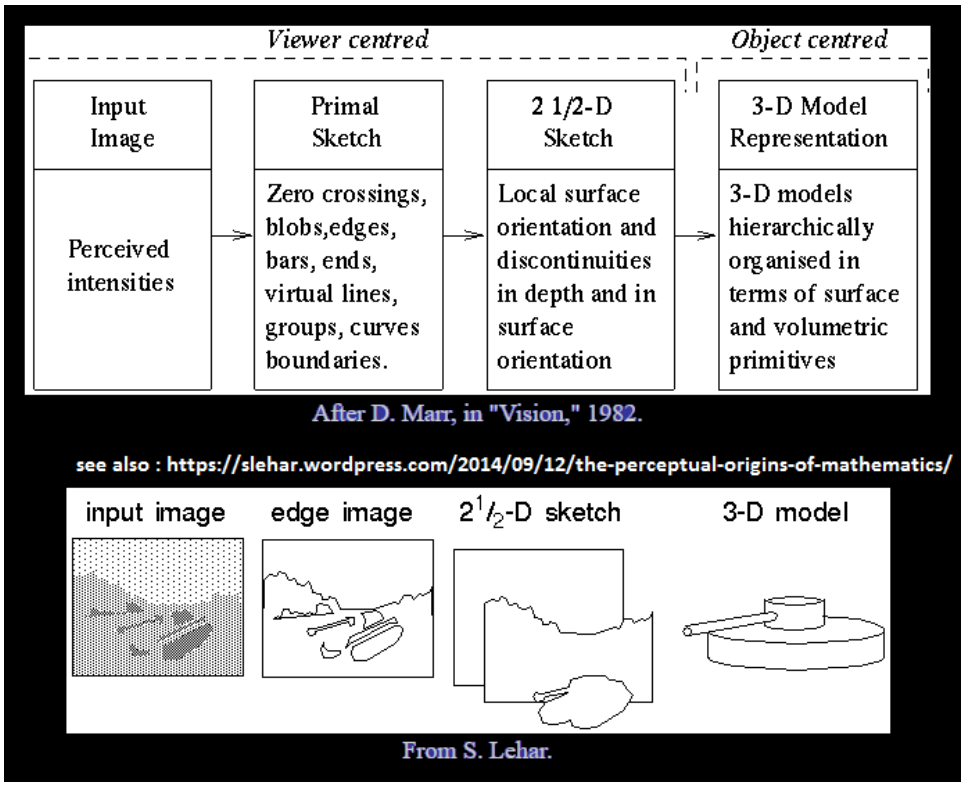


Fig. 1. The difficulty of pattern recognition

The conceptual choice:

In order to make clear the difficulty of the grouping object, let us explain the figure here below: from an initial images, the contours are at first identifying the zones with the same colors, in accord to a sharp change of the color gradient. However, the passage from Fig. 1(b) to Fig 1(d) is all but trivial for automatic computer detection. The concept of dealing with computer vision is in extenso designed by Marr (Marr, 1982). Essentially it is the primal sketch in which grouping of various regions inside the image becomes crucial. A common and trivial procedure for human vision but complex for computer vision.

On the issue of the production of maps, the Definiens/eCognition OBIA (Object based Image Analysis) on the images taken by the satellite already produces an identification of the contours of buildings, distinguishing them from all the rest. In turn, the buildings are divided accord to their size: small, medium, large (Fig. 3). Beside this classification, that may already raise issues[1], we focused on the problem of the grouping of buildings in accord to their geographical (Euclidean) distance, because also the experts are not always agreeing in the belonging of a building to a specific group.

In this perspective, the application of methods already used in complex networks becomes interesting.

In fact, actually, the problem of detection of communities in networks is a problem of classification, so the

1

The thresholds for the definition of each size have to be defined; buildings that are very close could be identified as being one building, so the image from the satellite has to be integrated with information drawn from the municipalities on the actual division of areas. However, questions like the best path for making a new road, or posing new paths for infrastructure or risk analysis, most need to know the best grouping of buildings, rather than their ownership

methods originally thought for social networks can be expanded to further classification problems, as soon as the object under examination can be represented through a network. This is the case with the problem of the best grouping of building in images taken from the satellite, combining Definiens/eCognition with methods for community detection that have already been experienced in Complex networks.

Let us explain in more details the following problem. Consider the picture here below, with 40% filling of black cells. This is a standard advice in the EU Manuals and has been so for the past 20 years. Each black cell corresponds to a building.

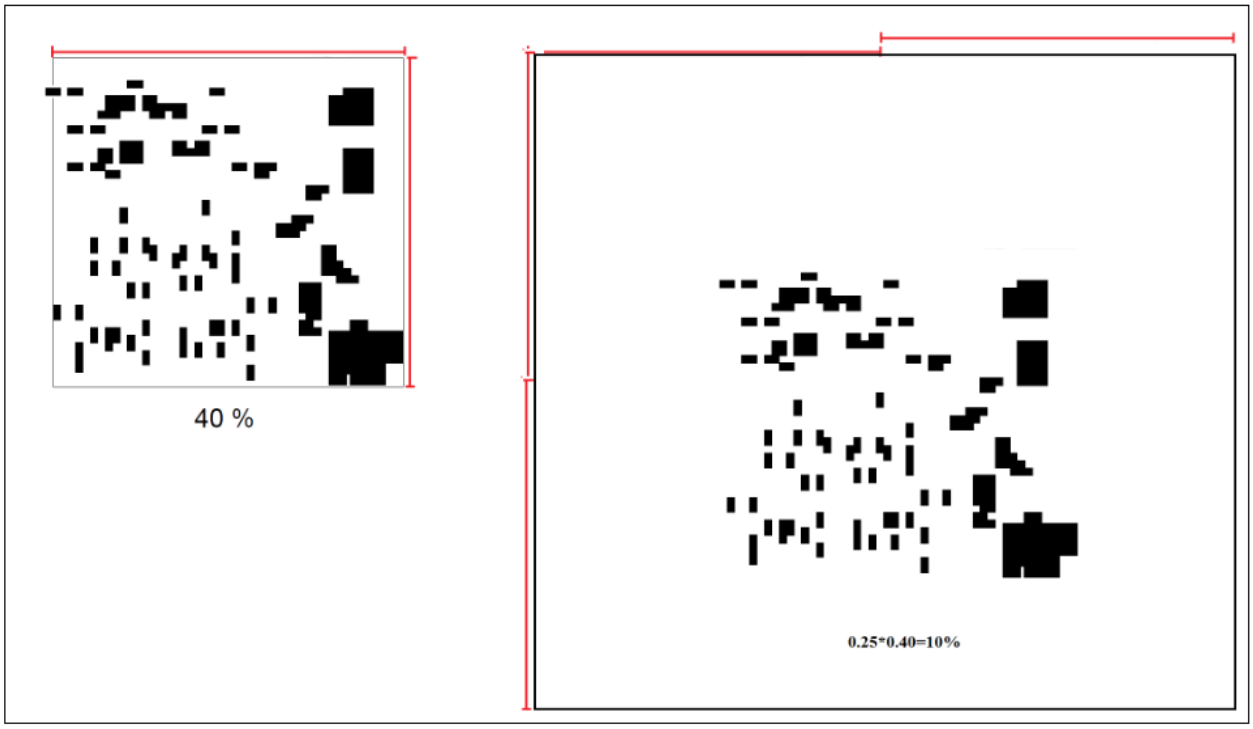


Fig. 2. Example from [1], page 8. Left side: the image. Right side: extending the frame with 2*2 makes a 10% density.

Definiens/eCognition is able to recognize each of the rectangles as a building, that belongs to one of the classes: small, medium, large. The basic classification problem already solved is the trivial one of building class large medium and small. The Fig 2 explains the weakness of the concept of density. With an extension of the frame, the density drops from 40% to 10% without any change in the structural elements. To base a grouping only on the characteristics of the structural elements a new conceptual approach is designed.

In order to proceed with the construction of a network, we enumerate the buildings:

The first 5 units are the large blue buildings (nodes 1 to 5 in the network, starting from the top); then the medium red ones (nodes from 6 to 10 in the network), and then the other 45 small yellow (Fig. 3).

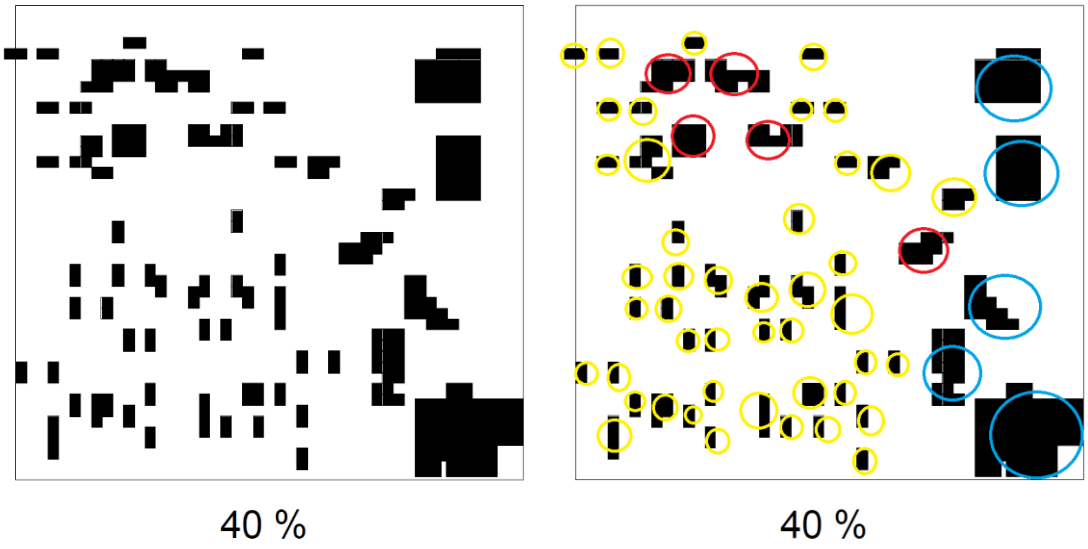


Fig. 3. Example from [1], page 8. Left side: the image. Right side: the grouping into large (blue), medium (red) and small (yellow) buildings.

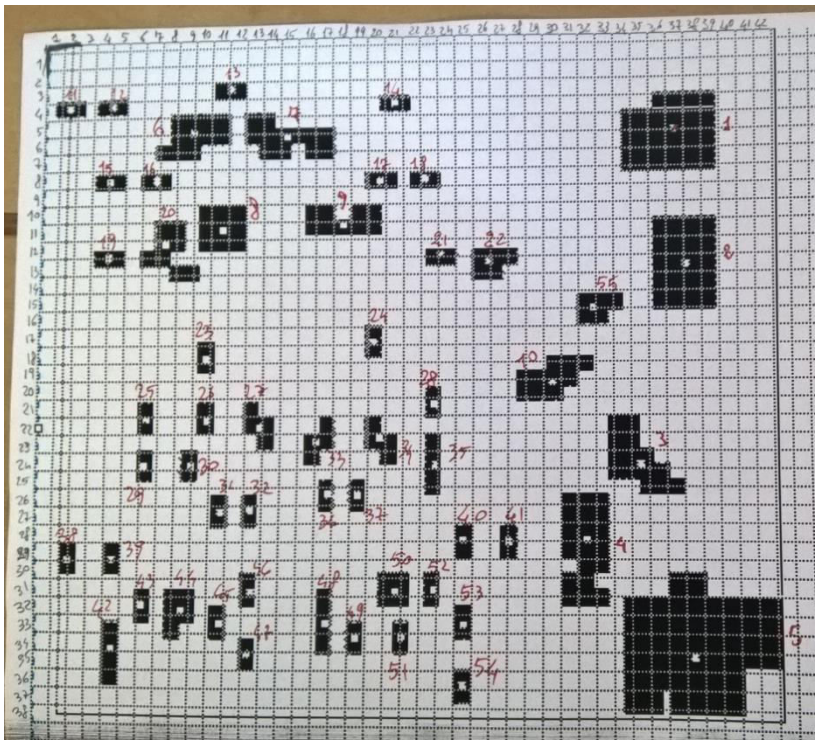


Fig. 4: the enumeration of buildings for their identification as network nodes.

As soon as the nodes have been defined, the links (undirected) were built assigning the Euclidean distance among the center of mass of the buildings (the white dot in the center of each black blocks in Fig. 3). Such links were pruned, so that only the links below a maximum distance were kept.

After the construction of the network, the algorithm of Blondel et al. was run on this example.

Keeping 20% of the weights in accord to the Euclidean distance (MATLAB commands: >>[a,b,c]=cluster_jl(grafo(0.1,0.2),1,1,0,0);listacomunita(c)) we get the following communities:

Community 1 : 3, 4, 5, 10, 28, 33, 34, 35, 36, 37, 40, 41, 48, 49, 50, 51, 52, 53, 54,

Community 2 : 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24,

Community 3 : 23, 25, 26, 27, 29, 30, 31, 32, 38, 39, 42, 43, 44, 45, 46, 47,

Community 4 : 1, 2,

Community 5 : 55.

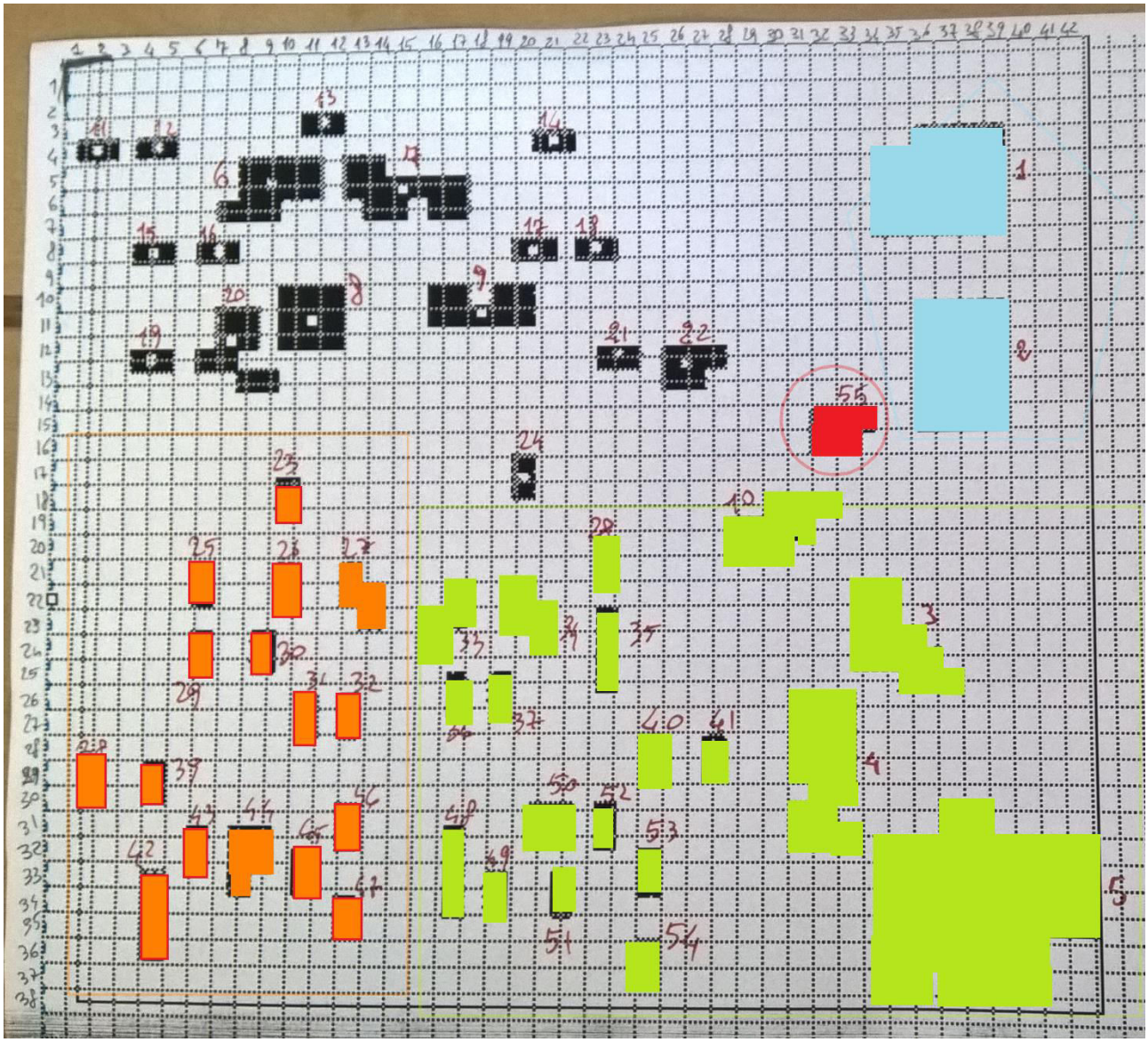


Fig. 5: the grouping of buildings: blue, red, black, orange, and green.

Of course, mixture of weights can be envisaged, and they are going to give raise to different communities, in the same way in which different communities are evidenced when the thresholds are changed.

We made several tests, but we are showing this particular one in the report, because the difference among the orange and the green part was not immediately visible/trivial at once; or, at least, it did not emerge from other analyses of the image. Actually, this partition is quite interesting because the alignment 27-32-46-47 corresponds to a potential road corridor, and it the visual inspection could not have detected it.

We made further tests. We report the test on the figure where the occupancy of the areas is at 10%:

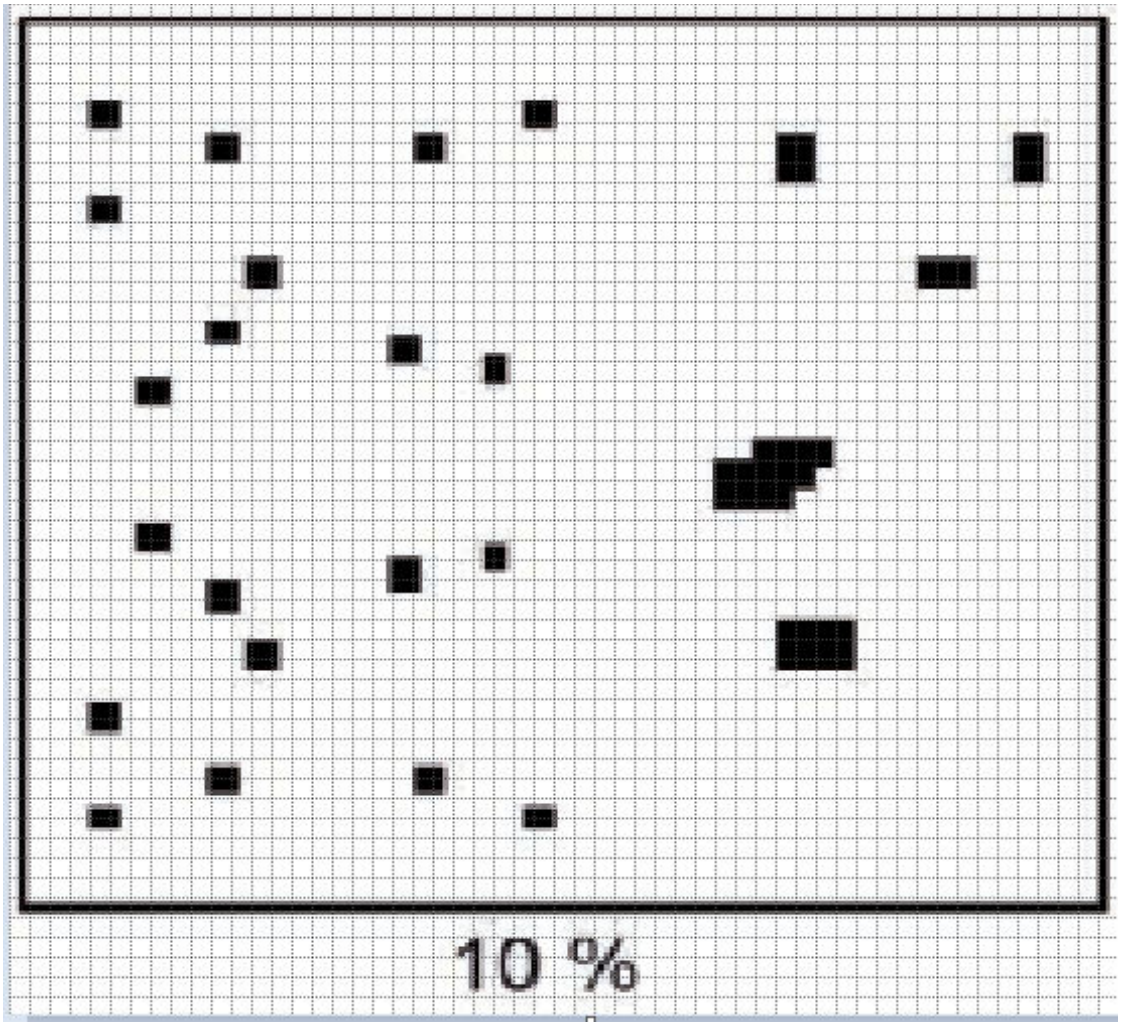


Fig. 5: occupancy of the areas at 10%

Again, we enumerate the cells and we fix coordinates. Running the program

we get:

Community 1 : 1, 2, 3, 4, 5, 9, 10, 11,

Community 2 : 17, 18, 19, 21, 22, 23, 24,

Community 3 : 12, 13, 15, 16,

Community 4 : 6, 7, 8,

Community 5 : 14, 20,

Community 6 : 25

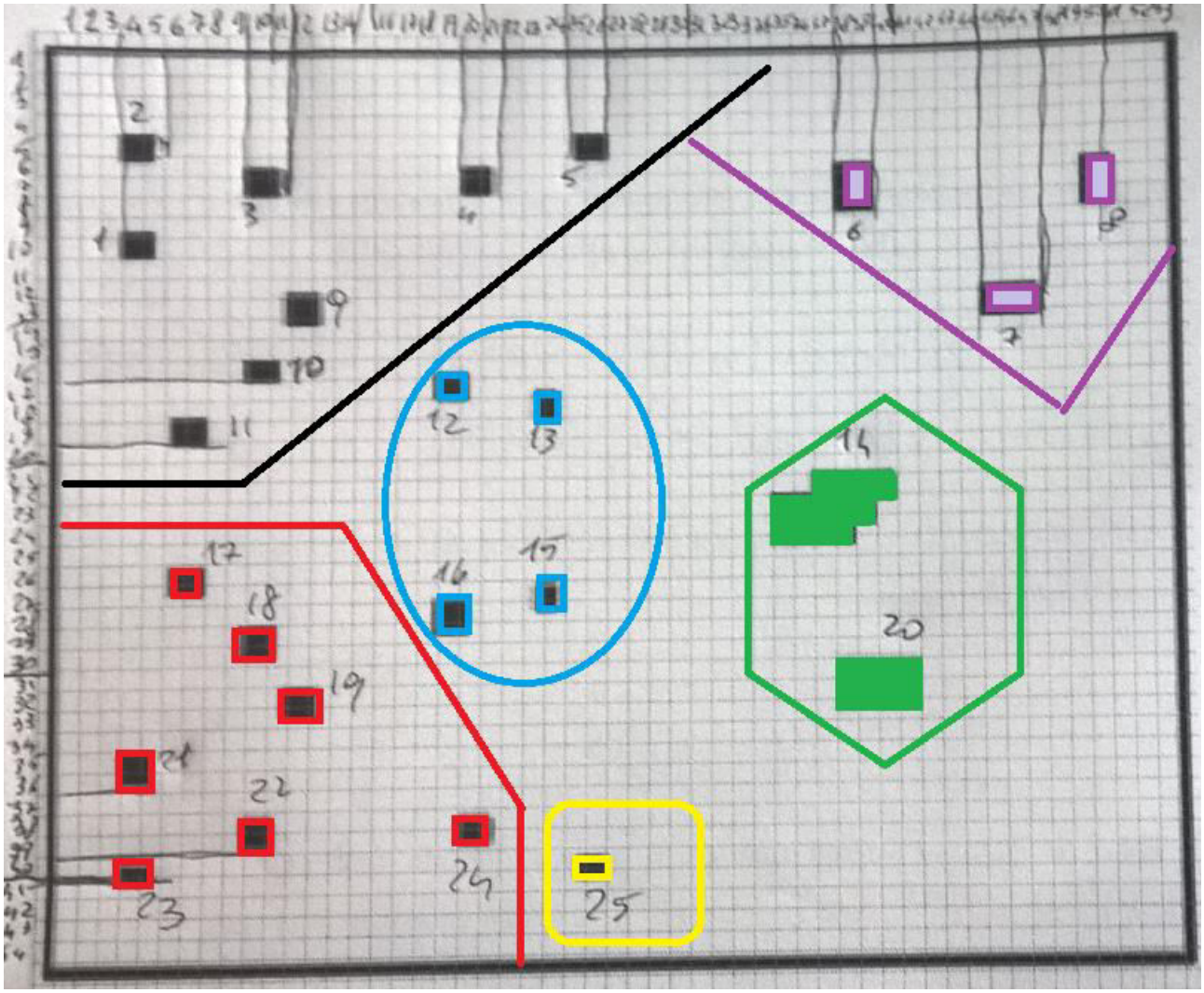


Fig. 6: detection of communities in the area filled at 10%

The result evidence the dis-alignment of the node 25 with respect to the nodes 17, 18, 19, and 24. In urban area maps, alignments most correspond to the existence of roads, railways, and tramways, or to the possibility of planning them.

Another part of the visit was consisting in the opposite contamination of fields: from data mining to complex networks.

Our attention was attracted by the paper Vieira et al. on the image from satellites that have to detect the readiness for harvest of the areas devoted to the plantation of sugar canes in Brazil. In this paper, the classification has to be in two subsets: ready/not ready for the harvest. A dataset checked

by experts has been used for training the algorithm in order to build a binary decision tree. J48 is a modification of the statistical classifier WEKA C4.5 algorithm, that became quite popular after ranking #1 in the Top 10 Algorithms in Data Mining pre-eminent paper published by Springer LNCS in 2008 [Wu et al., 2008].

Each level of such a decision tree depends on an explanatory variable, and the training set is used for determining the yes/no threshold.

In line with this interest on decision trees, the visit has started with the participation to the EUROXXVIII event in Poland, where it has been possible to meet many colleagues working on Data mining. Among the presented papers, we point out the paper [Da Silva et al., 2016] on the usage of the J48 algorithm.

The expansion of the decision trees for the detection of communities in networks is going to be examined further in the next future, since it seems quite promising.

Summing up, the visit has shown to be effective for a better integration of techniques and methods that have been developed for examining data in quite different disciplines. While images from satellite have always constituted a large source of data, the new social networks platforms, communication platforms, financial databases, etc. allow to consider the extraction of information from large datasets also in socio-economic environments. On the reversal way, not-strict classification used in complex networks clustering add new classification techniques to the already existing fuzzy set-based classification already known in data mining.

During the visit, there has been some interaction with students and postdocs. On July 11th, a crash course on MATLAB on main topics and key Toolboxes (interpolation, optimization, visualization), necessary for working on the subject. The pictures in the Appendix 2 show the seminar room.

In conclusion, the scientific interchange among different fields has been quite fruitful, and it has shown potentialities ok forward for future work on the topic. In conclusion, this first visit, and the exchange of information among different fields has shown potentialities that may be expanded in the future.

Cracow, July 13th, 2016

Giulia Rotundo

Bibliographic references

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks - Journal of statistical mechanics: theory and experiment, 10, P1008

Da Silva G, Barros M, Gomes Costa H, Oliveira A, Santos M, Evaluating Attributes Selection Techniques in the Classifiers Construction Process, EURO XXVIII Abstract

R. de Kok, P. Wezyk, (2008) Principles of full autonomy in image interpretation. The basic architectural design for a sequential process with image objects Object-Based Image Analysis. Blaschke, Th., Lang S., Hay, G.J. (Eds.). Series: Lecture Notes in Geoinformation and Cartography. Springer Berlin Heidelberg, ISSN: 1863-2246, 697-710.

Evans TS, Lambiotte R (2009) Line graphs, link partitions, and overlapping communities, Physical Review E 80 (1), 016105

Kramer, S. J48 – OpenTox (2011) Available online: http://www.opentox.org/dev/documentation/components/j48 http://www.opentox.org/dev/documentation/components/j48

Li A, W RS, Zhang S, Zhang XS (2005) Quantitative Function and Algorithm for Community Detection in Bipartite Networks

Mapping guide for a European Urban Atlas, Ref. RD-1 ITD-0421-RP-0003-C5 I 1.00 C5-Service Validation Protocol, EU Commission

https://cws-download.eea.europa.eu/local/ua2006/Urban_Atlas_2006_mapping_guide_v2_final.pdf

Ostaszewski M, Gawon P., Bouvry P (2015) Multi-objective Hierarchical Clustering of Complex Knowledge with Support of Ontology, Euclidean and Graph-based Distances, EUROXVII abstract invited in session "MC-84: Data Mining in Bioinformatics, stream Computational Biology, Bioinformatics and Medicine".
Quinlan JR, Yang Q, Yu PS, Zhihua Z, and David Hand et al(2008) Top 10 algorithms in data mining. Knowledge and Information Systems 14.1: 1-37.

Velickov S., Solomatine DP, Yu, Price RK (2000) Application of Data Mining Techniques for Remote Sensing Image Analysis, Proc. 4-th International Conference on Hydroinformatics, Iowa, USA http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.324.1810

Vieira MA, Formaggio AR, Rennó CD, Atzberger C, Aguiar DA, Pupin Mello MP, Object Based Image Analysis and Data Mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas, Remote Sensing of Environment, Volume 123, August 2012, Pages 553-562, ISSN 0034-4257, http://dx.doi.org/10.1016/j.rse.2012.04.011.

WEKA packages http://www.cs.waikato.ac.nz/ml/weka/

Wezyk P, de Kok R, Zajaczkowski G (2004) The role of statistical and structural texture analysis in VHR image analysis for forest applications. A case study on Quickbird data in the Niepolomice Forest

Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH, Steinbach M, Hand DJ, Steinberg D (2008) Top 10 algorithms in data mining, Knowl Inf Syst, 14:1–37

Zhang S, Wang RS, Zhang XS, Identification of overlapping community structure in complex networks using fuzzy -means clustering, Physica A: Statistical Mechanics and its Applications, Volume 374, Issue 1, 15 January 2007, Pages 483-490, ISSN 0378-4371, http://dx.doi.org/10.1016/j.physa.2006.07.023.

(http://www.sciencedirect.com/science/article/pii/S0378437106008119)

Zhang S, Hu G, Min W, A neurodynamic framework for local community extraction in networks, arXiv:1508.02177v1 [cs.SI] 10 Aug 2015

Appendix 1- List of MATLAB programs for the example

%the following function loads the data

```
function A=assegnaA
A=[5    37  22
13  38  20
24  35  11
28  32  15
35  38  51
5   9   6
6   15  7
11  11  9
11  18  7
20  30  5
4   2   2
4   5   2
3   12  2
4   21  2
8   5   2
8   7   2
8   20  2
8   23  2
13  5   2
12  8   6
12  24  2
13  26  2
18  10  2
17  19  2
21  7   2
22  10  2
22  13  2
21  22  2
24  7   2
24  9   2
27  11  2
27  13  2
26  17  3
22  19  3
24  23  4
26  17  2
26  19  2
29  2   2
29  4   2
28  26  2
28  27  2
34  5   4
32  7   2
32  9   3
33  11  2
31  13  2
35  13  2
33  17  4
34  19  2
31  21  3
33  22  2
31  23  2
33  25  2
```

```
36  25  2
16  32  4];
```

%the following function constructs the adjacency matrix from the data and cancels all the links which weight is above “livello”

```
function B1=grafo(livello)
A=assegnaA;
B=zeros(length(A));
for i=1:1:size(A,1)
    for j=i+1:1:size(A,1)-1
B(i,j)=sqrt((A(i,1)-A(j,1))^2+(A(i,2)-A(j,2))^2);
    end
end
m=max(max(B));
p=livello;
H=1-heaviside(B-p*m);
B1=B.*H;
```

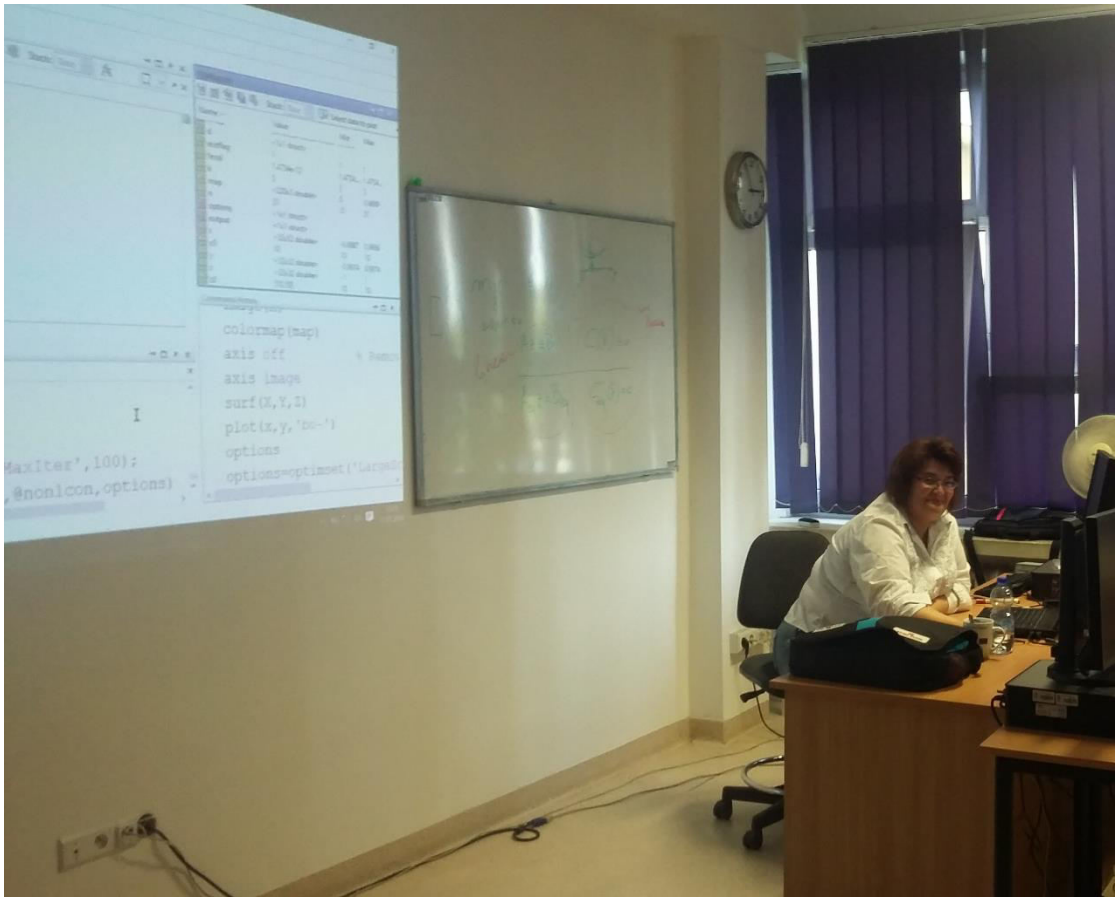
```
%the following function interacts with the algorithm Of Blondel et al. Written by Antoine Scherrer and prints the identity of the nodes belonging to the same community
function listacomunita(elenc)
elenco=cast(elenc,'int8');
for k=min(elenco):1:max(elenco)
    fprintf('Comunita %2d : ',k);
    for i=1:1:length(elenco)
    if elenco(i)==k
        fprintf(' %2d,',i);
    end
    end

    fprintf('\n');
end
end
```

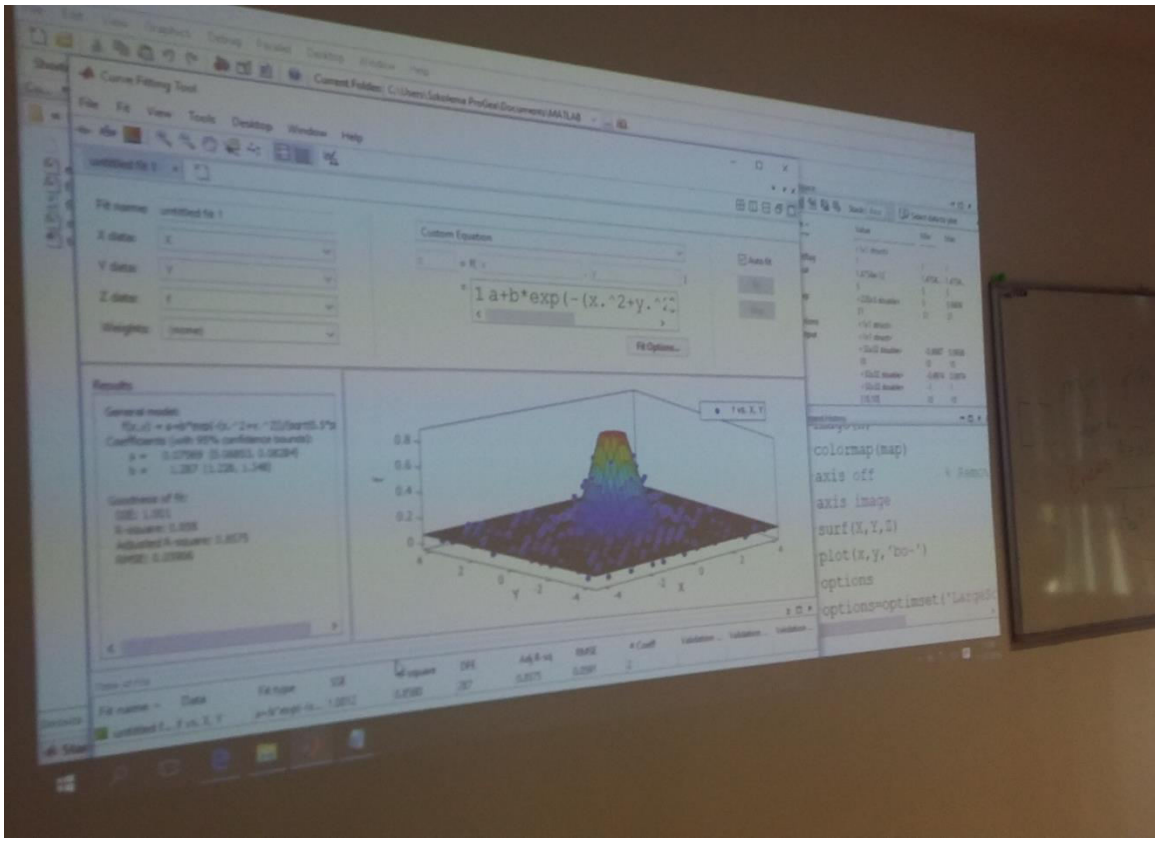
Test on the figure at 10% of occupancy of the area

```
function A=assegnaA1
A1=[10 4 2
 5  4  2
 7 10  2
 7 20 2
 5 25 2
 8 38 4
13 46 3
8 50 3
13 12  1
16 10  1
19  7  1
17 19 1
18 23 1
23 38 13
27 24  1
```

```
28 20 2
26  6 1
29 10 1
32 22  1
30 40 10
36 40  1
36  4 1
38 10 1
40  4 1
38 20 1
40 25  1];
```

Appendix 2 – computer room 11/07/2016




Curve Fitting Tool
File Fit View Tools Desktop Window Help
Custom Equation
1 a+b*exp(-(x.^2+y.^
colormap(map)
axis off
axis image
surf(X,Y,Z)
plot(x,y,'bo-')
options
options=optimset('