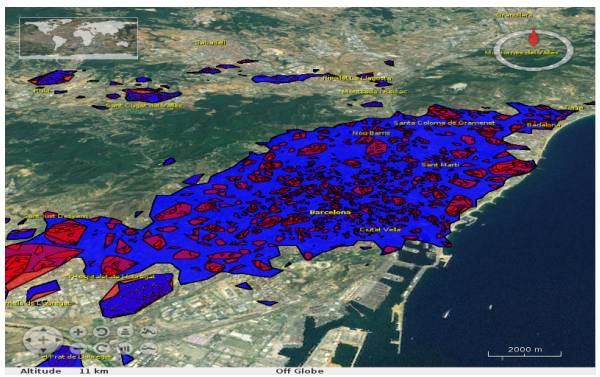
Visualizing Geolocated Tweets

A Spatial Data Mining Approach

Joana Simões [1]

[1]Eurecat, Centre Tecnologic de Catalunya

September 10, 2015

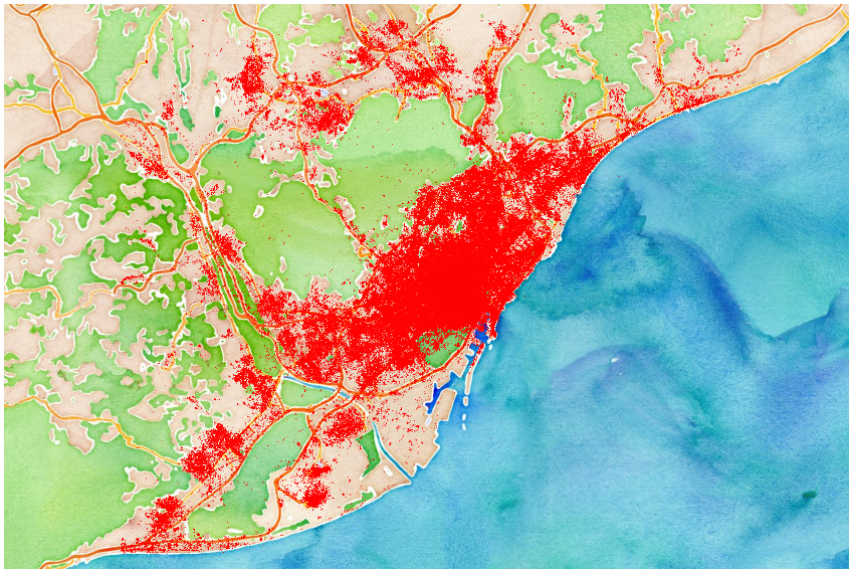


Why we *love* Twitter

- Due to its willingness in sharing data, Twitter has been a prime *playground*, for researchers and practitioners around the world.
- Users on Twitter generate over 400 million tweets everyday.
- Approximately 1% of all Tweets published on Twitter are geolocated.

Motivation

How to assimilate these large datasets and extract some relevant information?

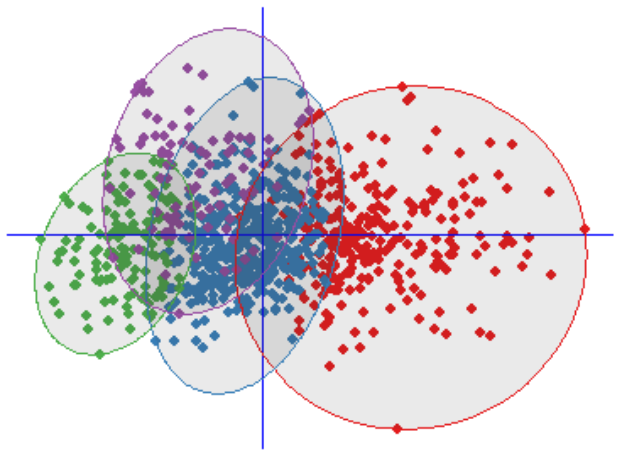


Approach

- Technological Challenge + **Representation Challenge**
- Machine Learning/ GIS

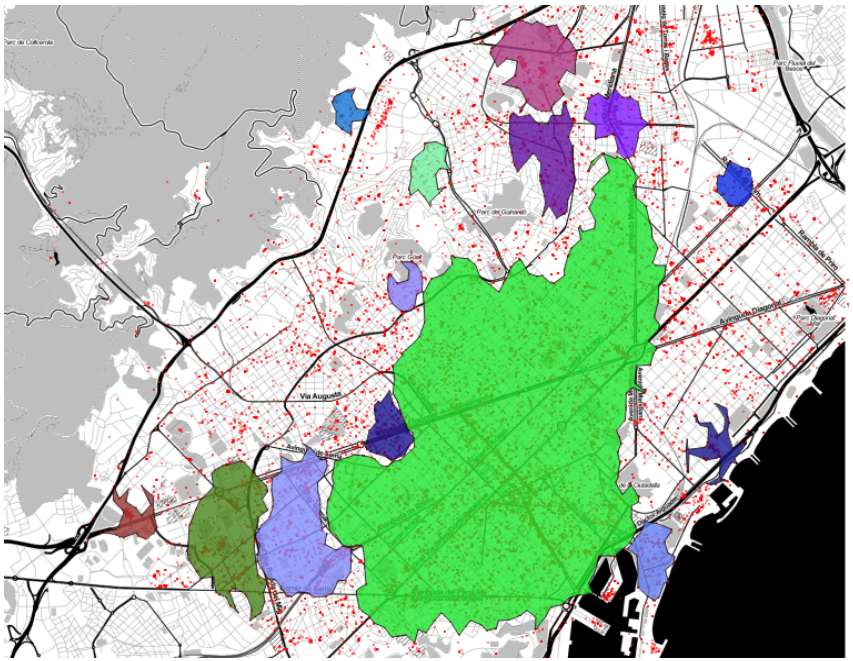
Clustering

- Unsupervised, descriptive, method.
- Widely used due to its segmentation and summarization features.
- Identifies groups of objects, which are similar between them, and distinct from the rest.

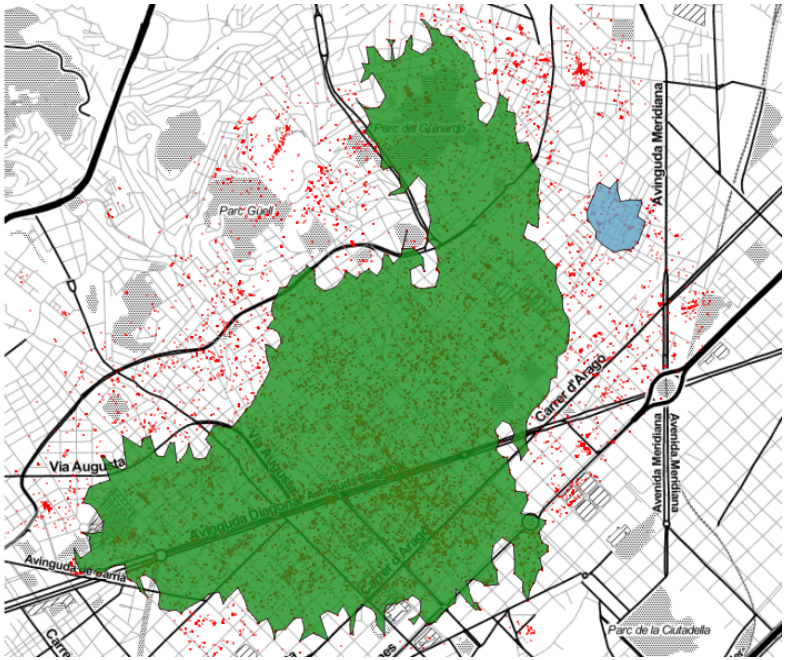


DBSCAN

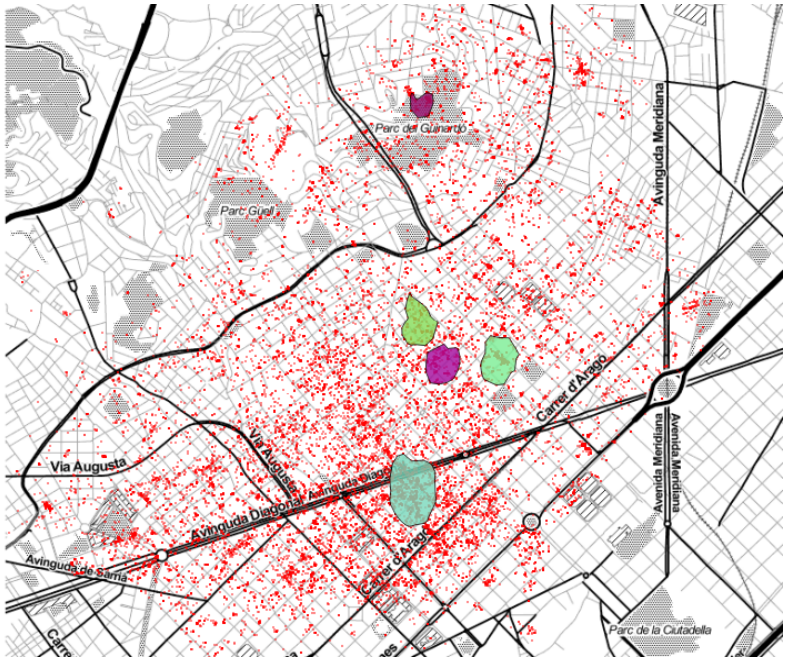
- Density-based clustering.
- A *dense-region* is defined by global parameters:
  - epsilon: neighbourhood radius.
  - minPts: minimum number of points required to form a dense region.



The Bottleneck: Global Parameters



Less strict combination of parameters



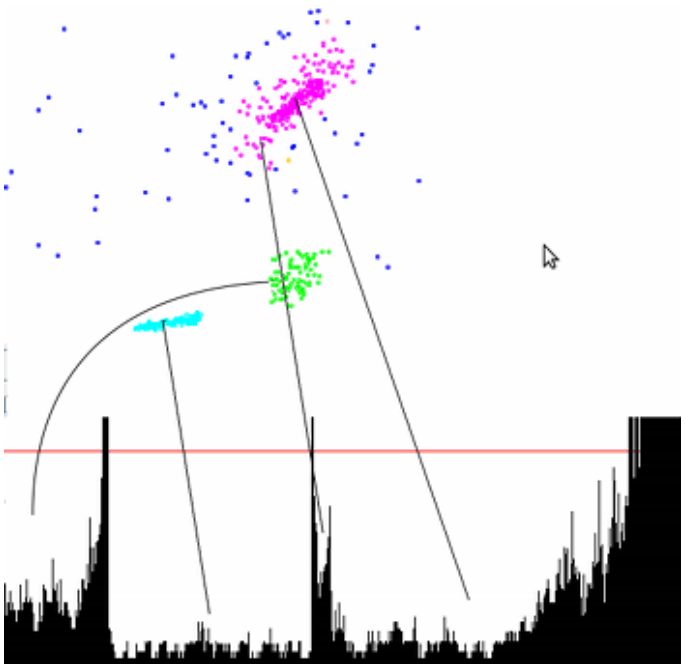
Strict combination of parameters

OPTICS

- Generalization of DBSCAN.
- Ordering of the database, such as points that are spatially closest become neighbours in the ordering.
- It does **not** produce a strict partition of the data.

OPTICS (cont.)

- In the ordering clusters appear as *valleys*, separated by *noise regions*(peaks).

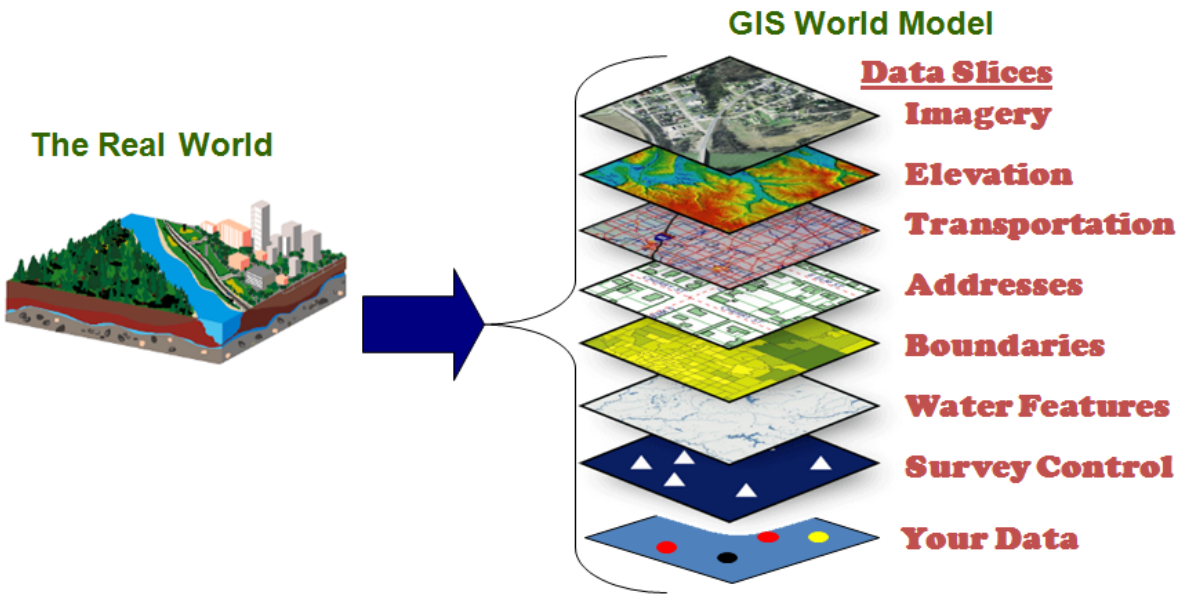


Opticsxi: detecting Clusters

- We can define the *start* and *end* of a cluster, based on a threshold (xi).
- Hierarchical partition.

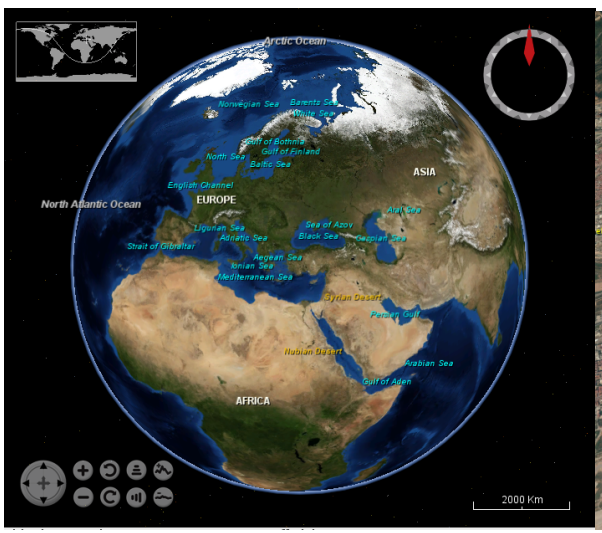
GIS: Putting Geographic Information into Context

- Visualization is a key element to understand and interpret the results of data mining.
- What about Geospatial information...?



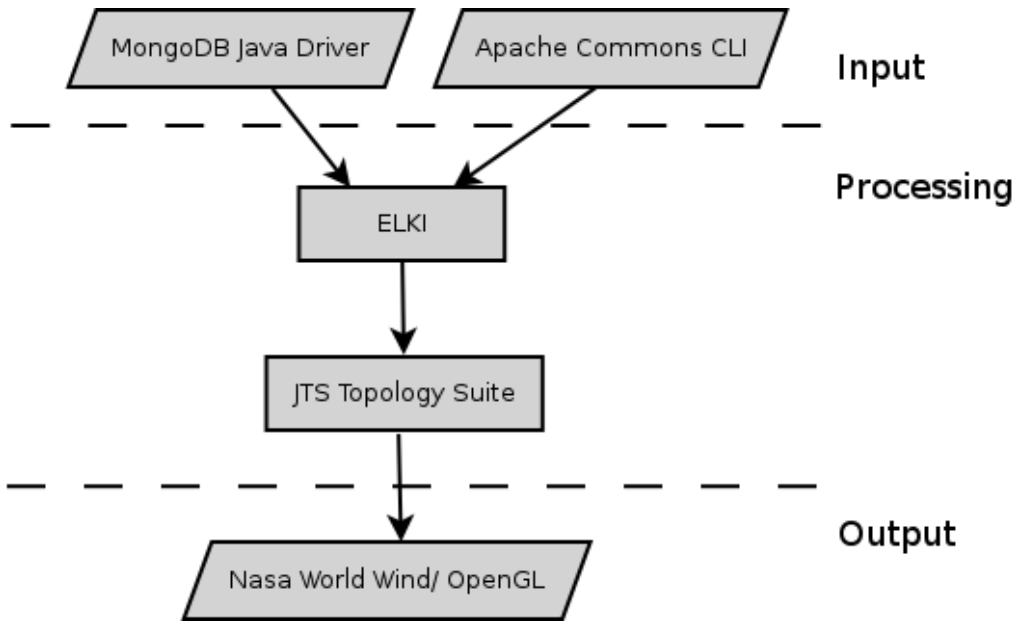
NASA WW

- WorldWind is an Open Source virtual globe developed by NASA that accesses a number of geographic datasets.
  - OpenGL.



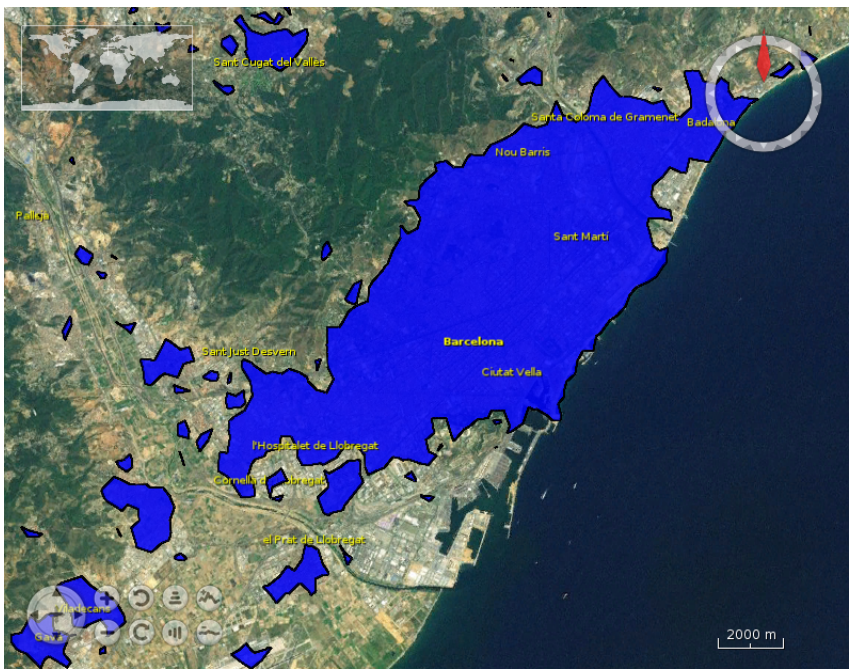
Cluster Explorer

- Java App based on FOSS.
- Generates clusters based on DBSCAN, OPTICSxi (or both).
- Displays the results on a virtual globe.

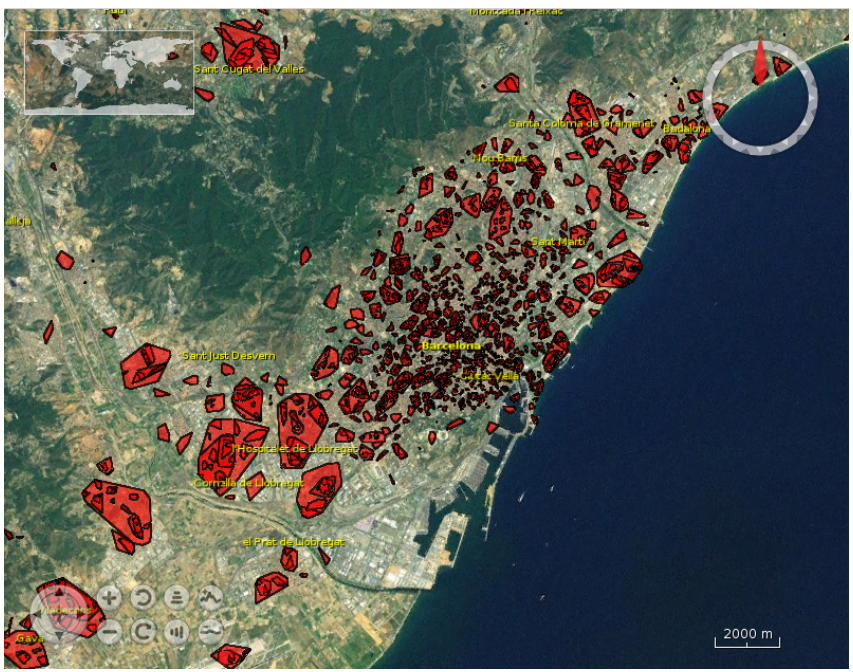


A Case Study

A set of geo-located Tweets in the city of Barcelona, over a period of five days

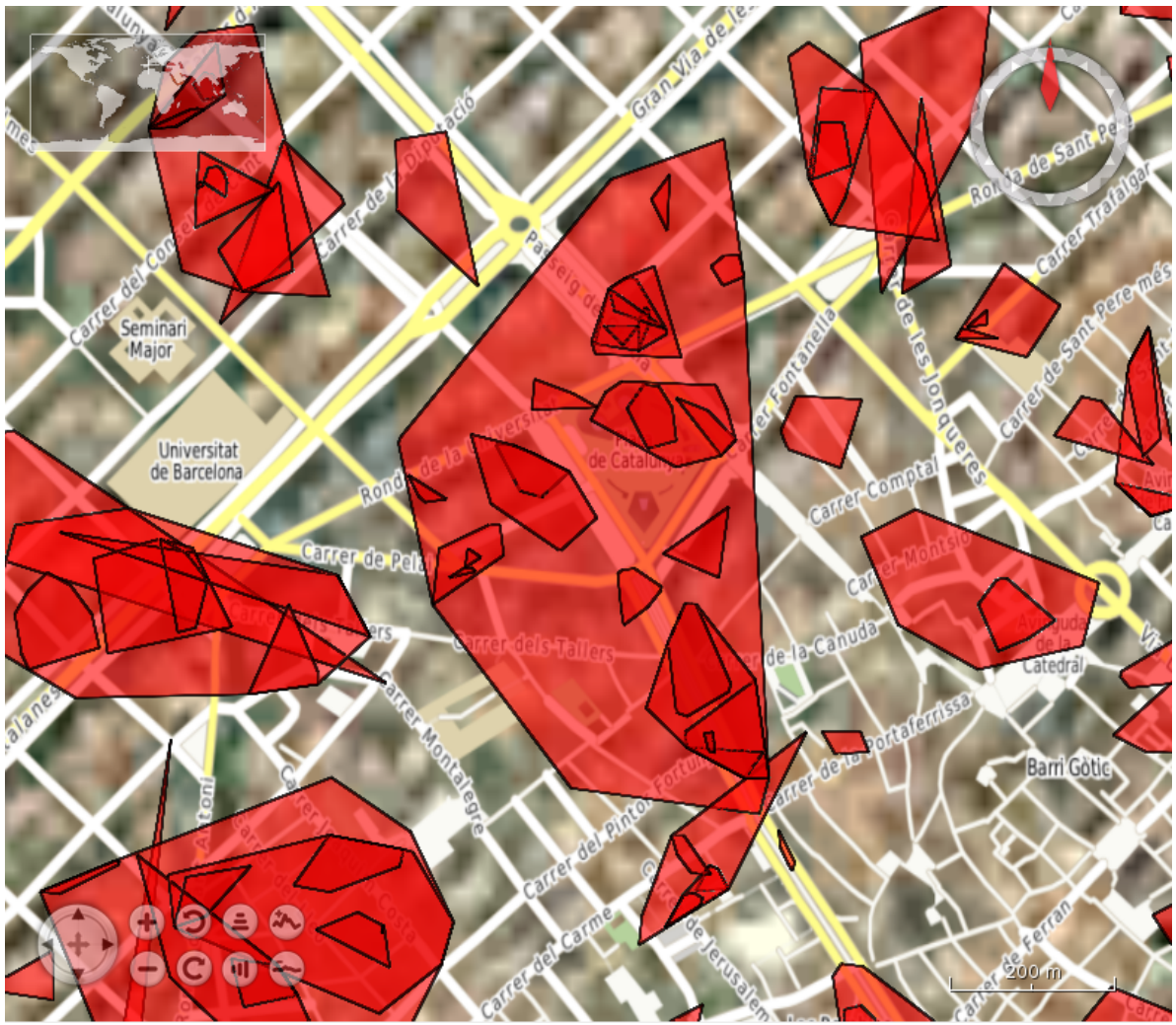


DBSCAN run

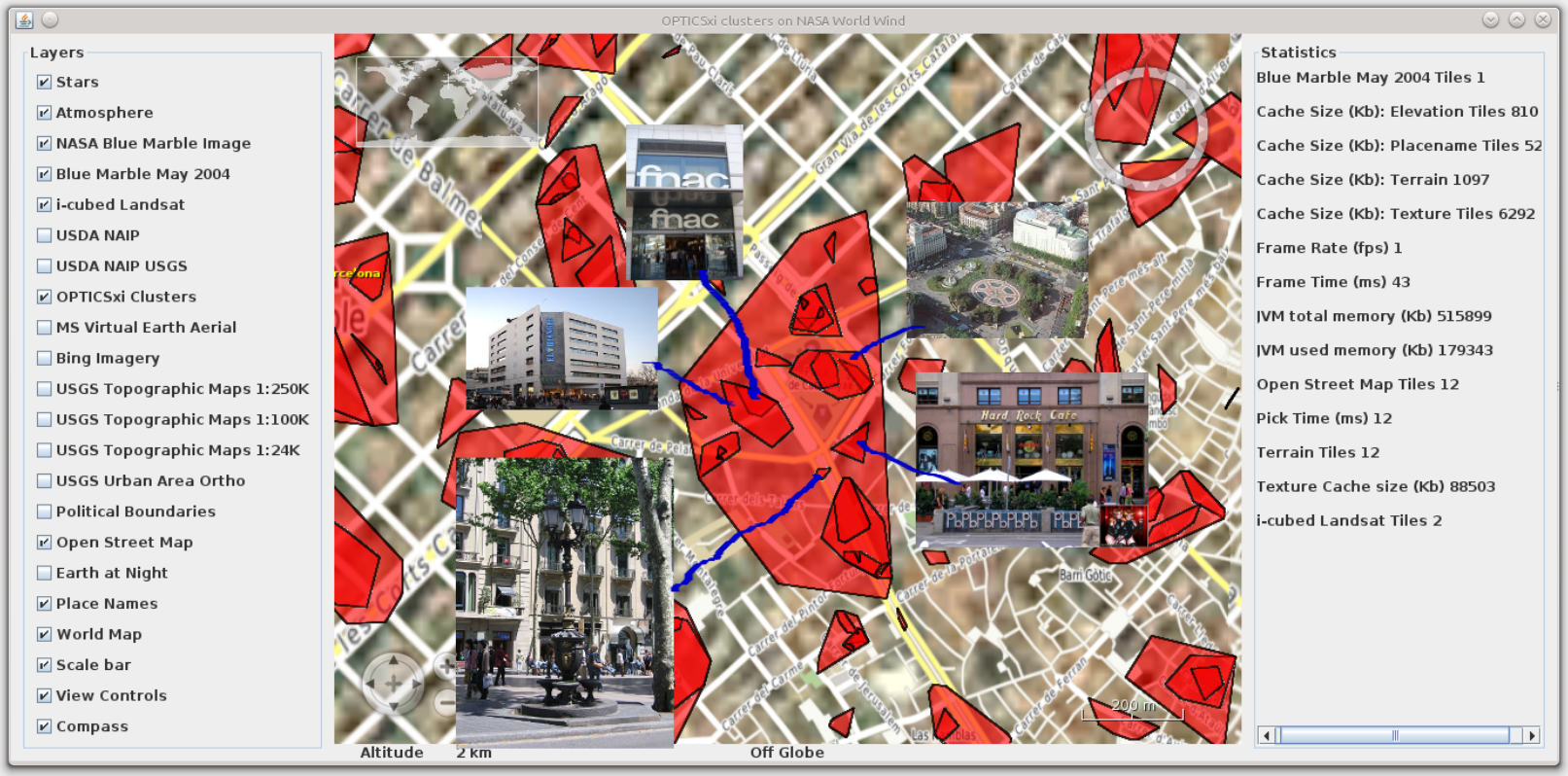


OPTICSxi run

GIS and the value of Location-Analysis (cont.)



GIS and the value of Location-Analysis (cont.)



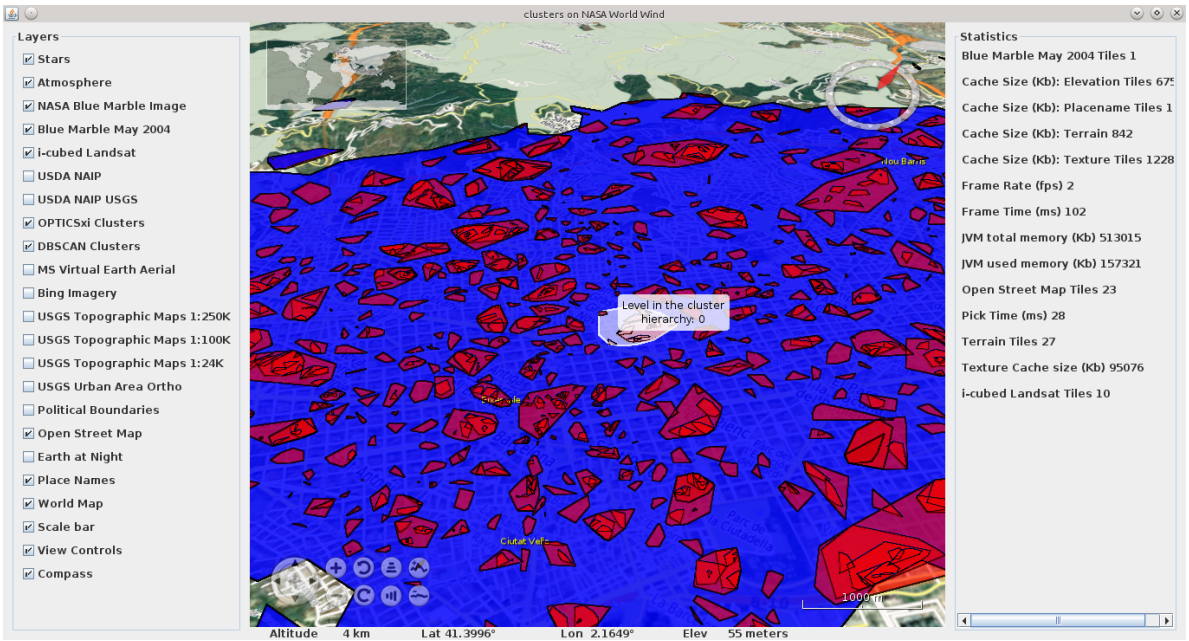
GIS and the value of Location-Analysis (cont.)

Identified clusters (hierarchical level):

- Plaza de Catalunya cluster (L1)
- Triangle Shopping Centre (L2)
- FNAC shop (L3)
- Font de Canaletes/Metro (L2)
- Hard rock cafe (L2)
- Central gardens (L2)

Conclusions

- Flat cluster partition: good for summarizing the dataset and easy to interpret.
- Hierarchical cluster partition: allows to look at the city through multiple scales.



Conclusions (cont.)

- Visualizations provided by the use of a virtual globe have proved to be a flexible and context enhancing tool, that was crucial for the interpretation of the results.
- Combining machine learning and GIS, can be a promising approach in the field of spatial data mining.

Thank You for Listening!

This presentation is available at:
`http://tinyurl.com/nd29g3f`