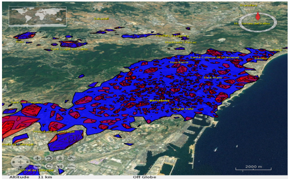
Visualizing Geolocated Tweets

A Spatial Data Mining Approach

Joana Simões [1]

[1]Eurecat, Centre Tecnologic de Catalunya

September 8, 2015



Why we *love* Twitter

- Nowadays social media is a major source for information sharing.

Why we *love* Twitter

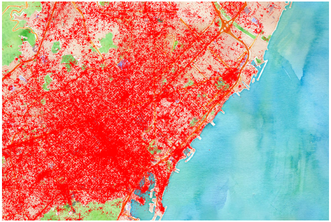
- Nowadays social media is a major source for information sharing.
- They provide accurate representations of the distribution of social media users within the territory.

Why we *love* Twitter

- Nowadays social media is a major source for information sharing.
- They provide accurate representations of the distribution of social media users within the territory.
- Due to its willingness in sharing data, Twitter has been a prime *playground*, for researchers and practitioners around the world.

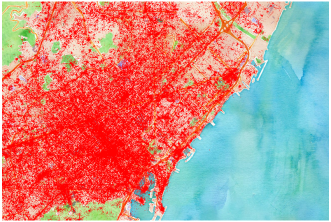
Some Context

- Users on Twitter generate over 400 million tweets everyday.



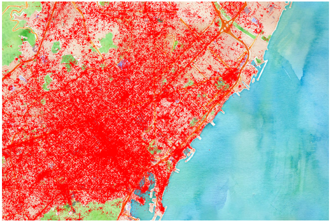
Some Context

- Users on Twitter generate over 400 million tweets everyday.
- Approximately 1% of all Tweets published on Twitter are geolocated.



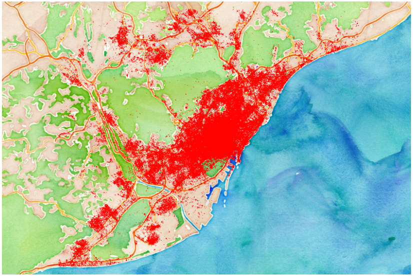
Some Context

- Users on Twitter generate over 400 million tweets everyday.
- Approximately 1% of all Tweets published on Twitter are geolocated.
- This (*Big*?) amount of data is not easily assimilated by the "human-eye".



Motivation

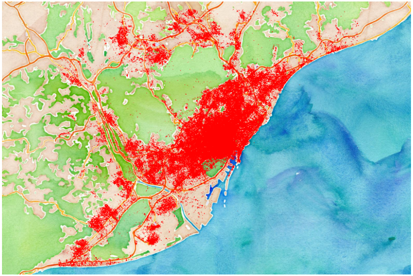
How to assimilate these large datasets and extract some relevant information?



Motivation

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

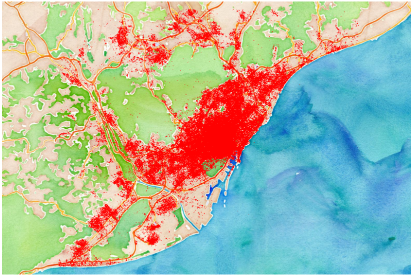
- Technological Challenge + **Representation Challenge**.



Motivation

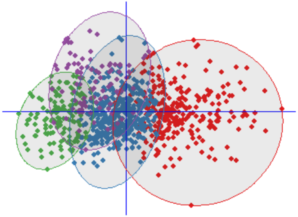
How to assimilate these large datasets and extract some relevant information?

- Technological Challenge + **Representation Challenge**.
- Machine Learning, GIS.



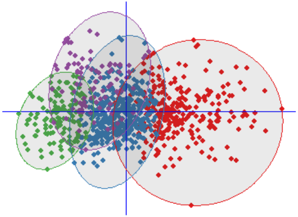
Clustering

- Unsupervised, descriptive, method.



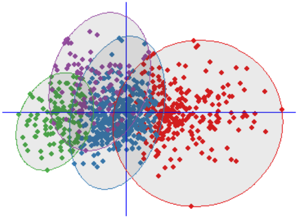
Clustering

- Unsupervised, descriptive, method.
- Widely used due to its segmentation and summarization features.

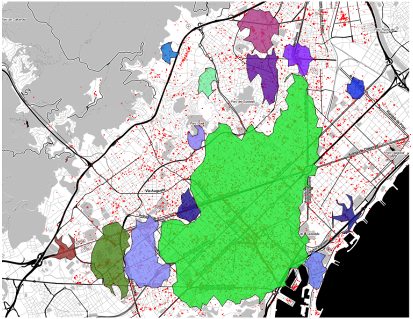


Clustering

- 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.

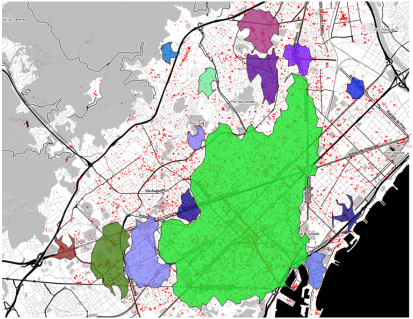


DBSCAN



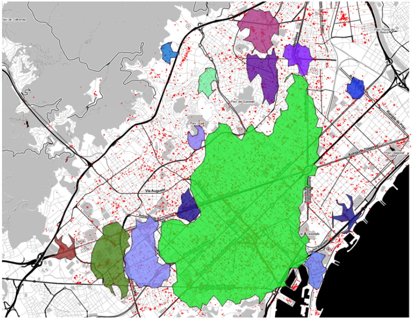
DBSCAN

- Detects an a priori unknown, number of clusters with an arbitrary shape.



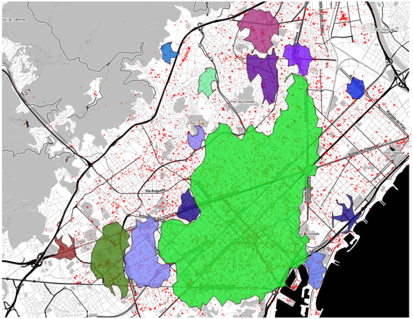
DBSCAN

- Detects an a priori unknown, number of clusters with an arbitrary shape.
- Density-based clustering.



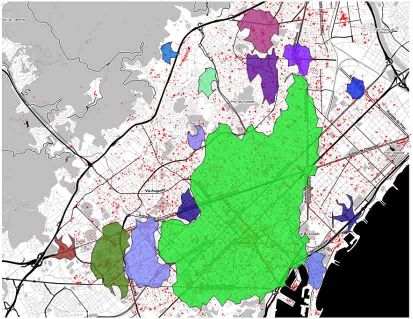
DBSCAN

- Detects an a priori unknown, number of clusters with an arbitrary shape.
- Density-based clustering.
- A *dense-region* is defined by global parameters:

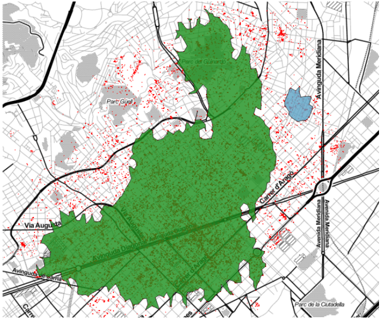


DBSCAN

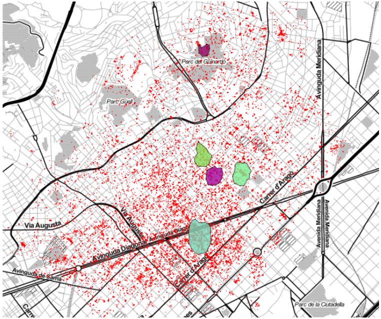
- Detects an a priori unknown, number of clusters with an arbitrary shape.
- Density-based clustering.
- A *dense-region* is defined by global parameters:
    - epsilon: neighbourhood radius.
    - minPts: minimum number of points required to form a dense region.



The Bottleneck: Global Parameters



Less strict combination of parameters



Strict combination of parameters

The Bottleneck: Global Parameters (cont.)

The Bottleneck: Global Parameters (cont.)

- Larger epsilon, lower minpts: larger clusters that include a great part of the points, but it fails to expose the details of the highly dense zones in the city centre.

The Bottleneck: Global Parameters (cont.)

- Larger epsilon, lower minpts: larger clusters that include a great part of the points, but it fails to expose the details of the highly dense zones in the city centre.
- Lower epsilon, larger minpts: detects these highly dense zones, but leaves out other *potential* clusters that are less dense, in the city outskirts.

OPTICS

OPTICS

- Generalization of DBSCAN.

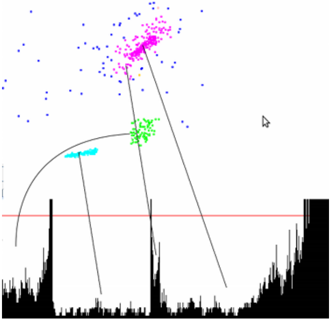
OPTICS

- Generalization of DBSCAN.
- Ordering of the database, such as points that are spatially closest become neighbours in the ordering.

OPTICS

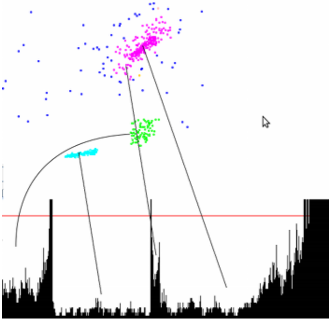
- Generalization of DBSCAN.
- Ordering of the database, such as points that are spatially closest become neighbours in the ordering.
- It does **not** produce a strict partition of the data.

OPTICS (cont.)

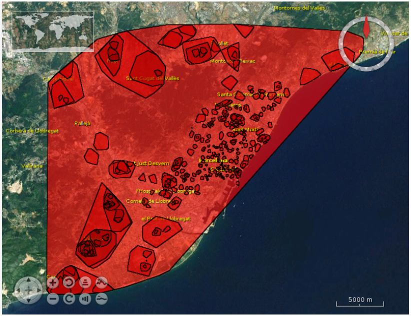


OPTICS (cont.)

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

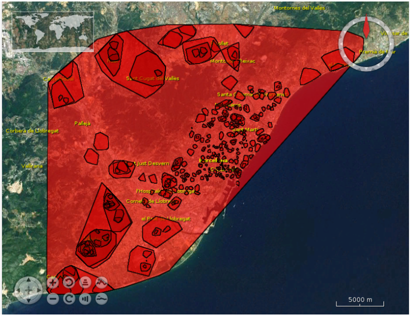


Opticsxi: detecting Clusters



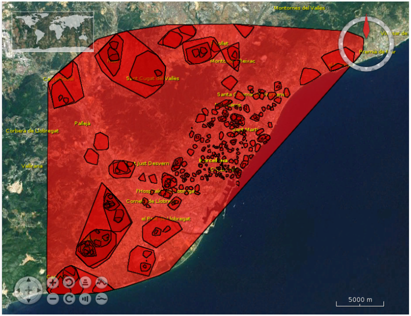
Opticsxi: detecting Clusters

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



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

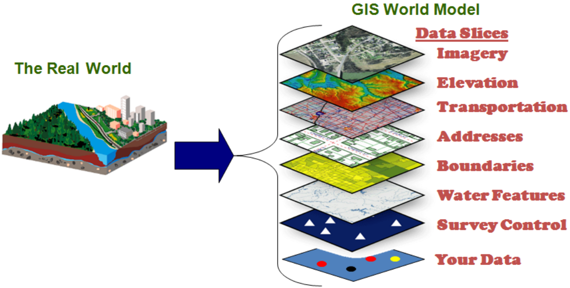
- Visualization is a key element to understand and interpret the results of data mining.

GIS: Putting Geographic Information into Context

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

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

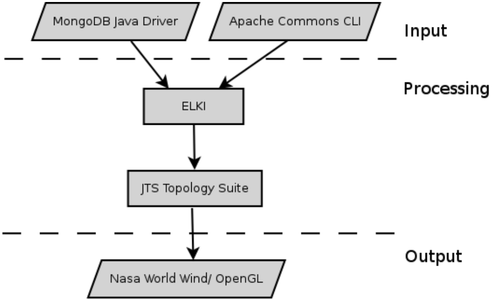
- WorldWind is an Open Source virtual globe developed by NASA that accesses a number of remote sensing datasets.

NASA WW

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

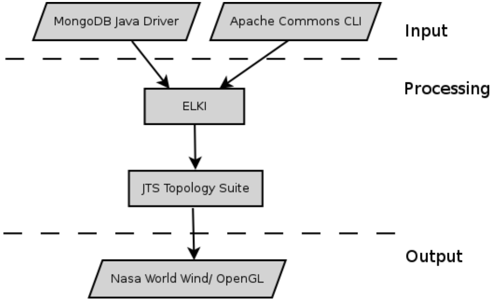
Cluster Explorer

- Java App based on FOSS.



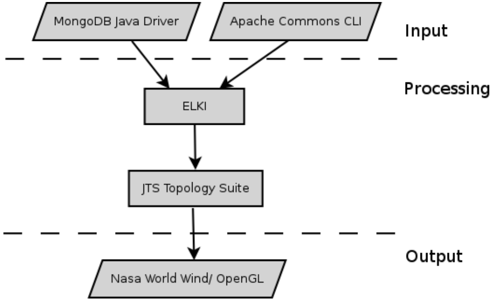
Cluster Explorer

- Java App based on FOSS.
- Generates clusters based on DBSCAN, OPTICSxi (or both).



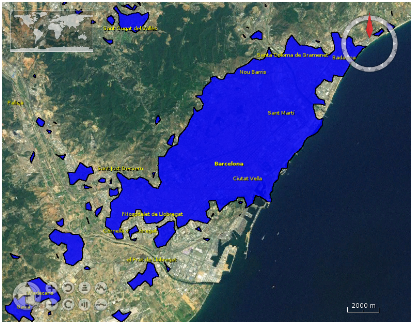
Cluster Explorer

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

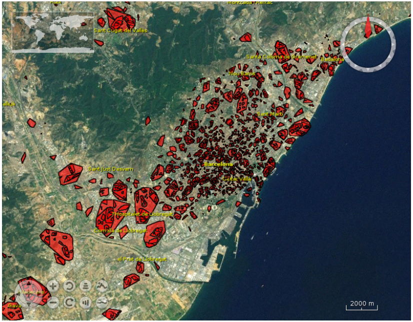


A Case Study

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

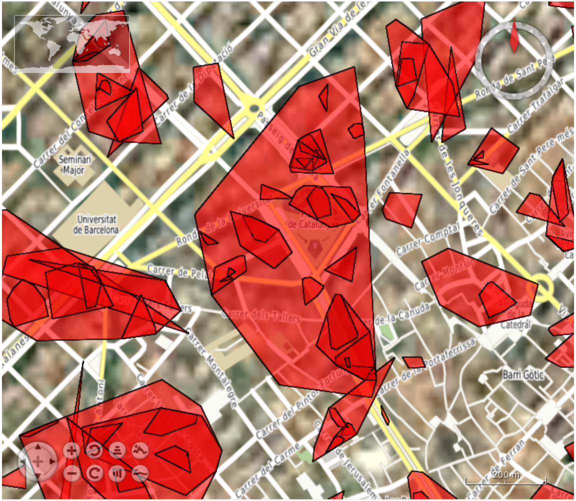


DBSCAN run

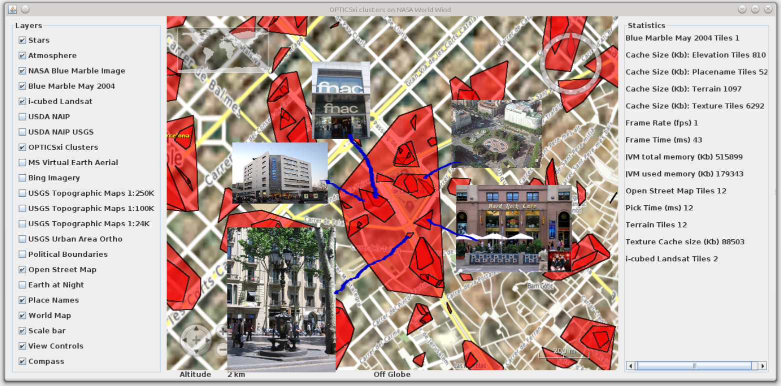


OPTICSxi run

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)
- Hard rock cafe (L2)
- Central gardens (L2)

Conclusions

Conclusions

- Flat cluster partition: good for summarizing the dataset and easy to interpret.

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

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

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.
- 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.

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.
- 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.
- The reproducibility of this approach is encouraged by the use of FOSS technologies.

Thank You for Listening!

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