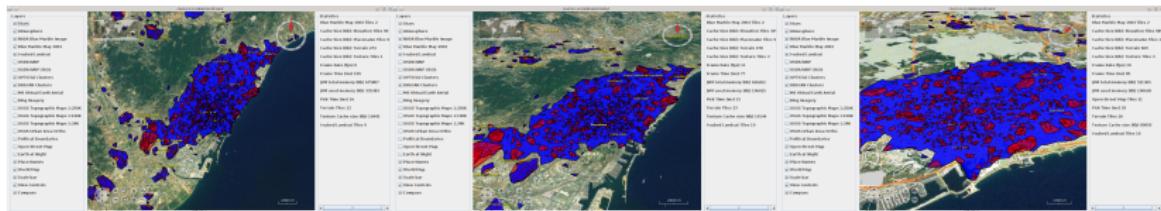


# Visualizing Micro-blogging Data using Clustering and GIS

Joana Simões<sup>1</sup>

<sup>1</sup>Eurecat, Centre Tecnologic de Catalunya

May 11, 2015



# Table of Contents

- 1 Introduction
- 2 Extracting Patterns
- 3 Putting Geographic Information into Context
- 4 A Case Study
- 5 Final Remarks

# Introduction

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



# Introduction

- Nowadays social media is a major source for information sharing.
- A proxy for human presence: powerful representations of the distribution of social media users within the territory.



# Introduction

- Nowadays social media is a major source for information sharing.
- A proxy for human presence: powerful 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.



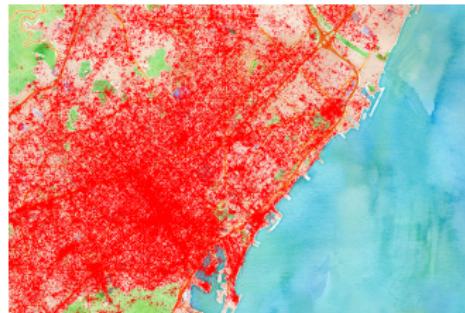
# Twitter

- Users on Twitter generate over 400 million tweets everyday.



# Twitter

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



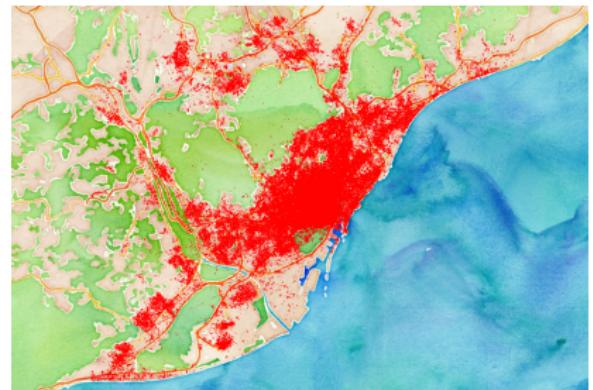
# Twitter

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



# Motivation

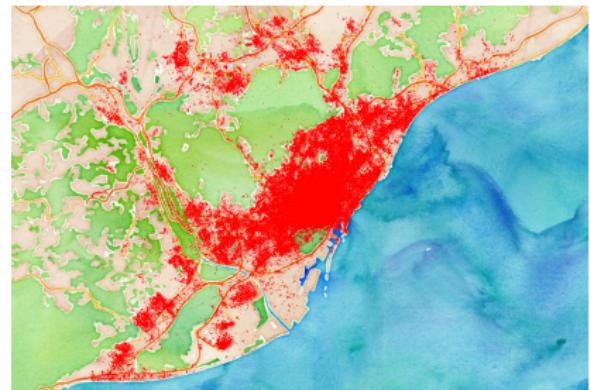
How to assimilate these large datasets and extract some relevant information?



# Motivation

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

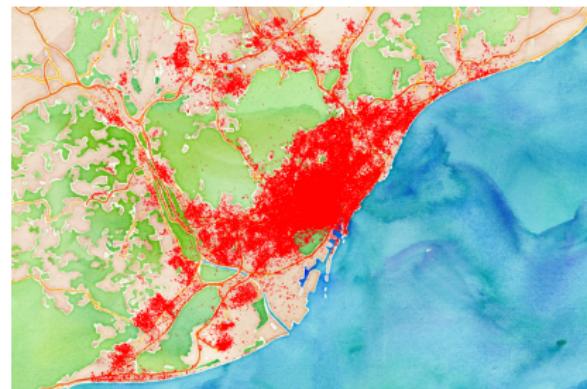
- Technological Challenge +  
**Representation Challenge**



# Motivation

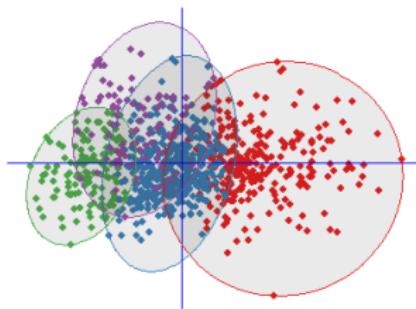
How to assimilate these large datasets and extract some relevant information?

- Technological Challenge + **Representation Challenge**
- Data mining/ Machine Learning



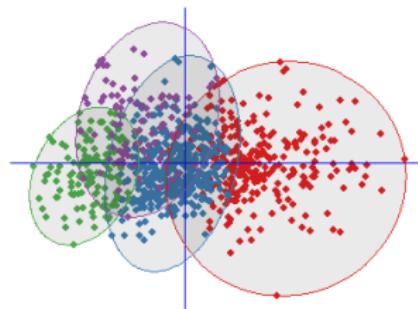
# Clustering

- Widely used due to its segmentation and summarization features.



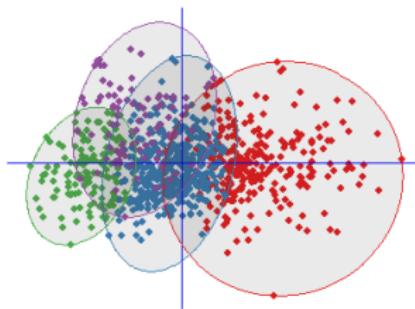
# Clustering

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

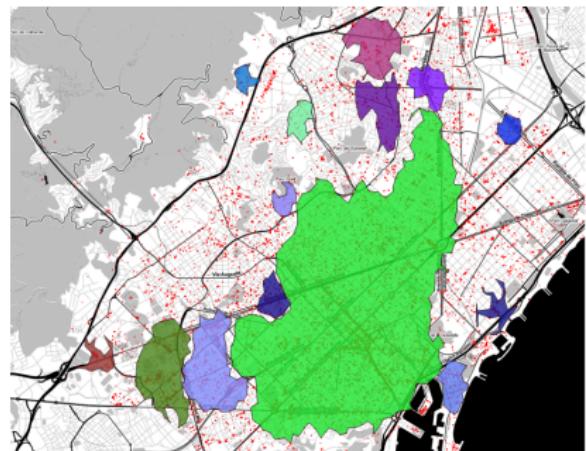


# Clustering

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

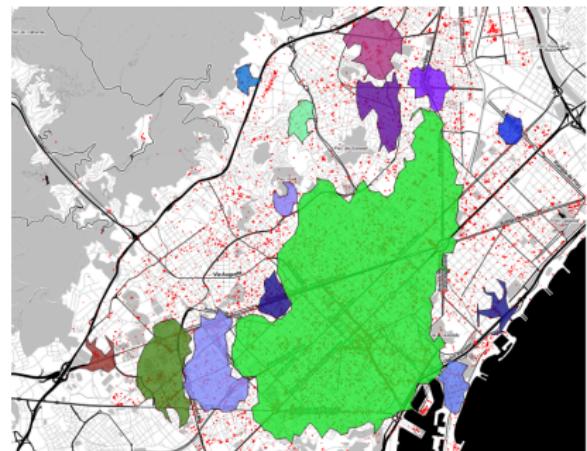


# DBSCAN



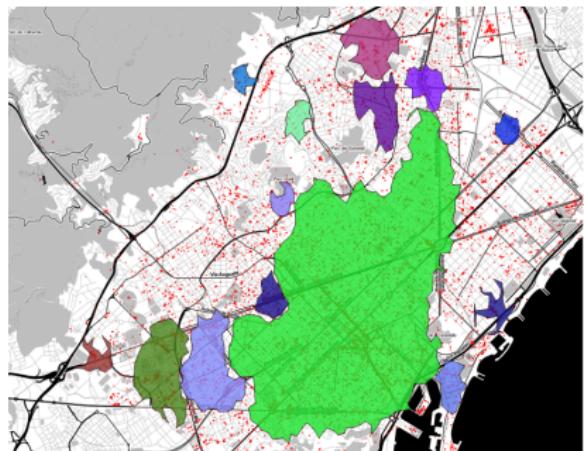
# DBSCAN

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



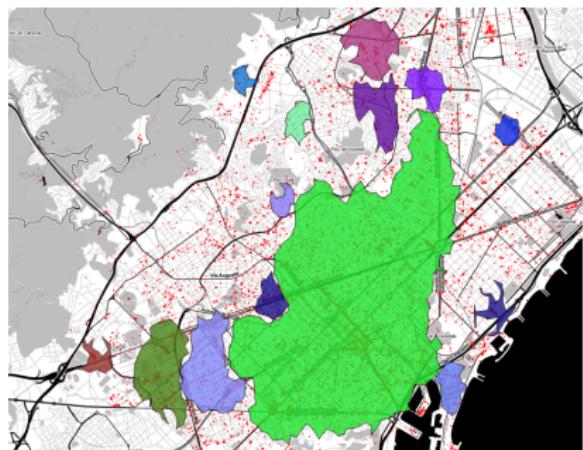
# DBSCAN

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



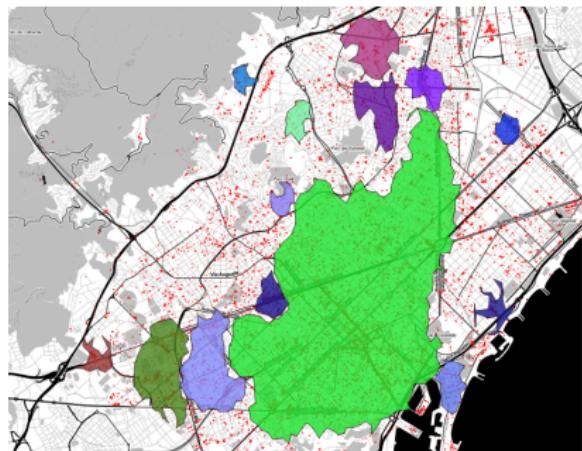
# DBSCAN

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

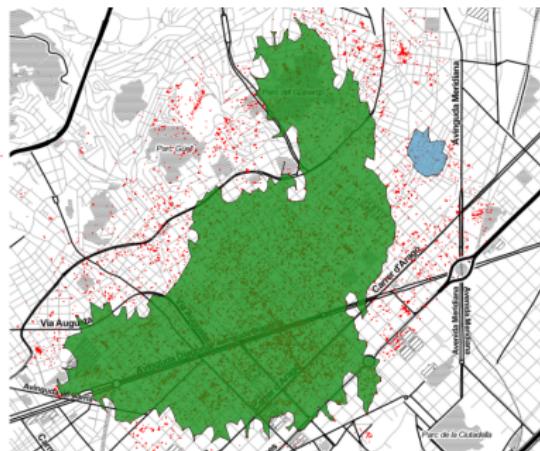


# DBSCAN

- 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

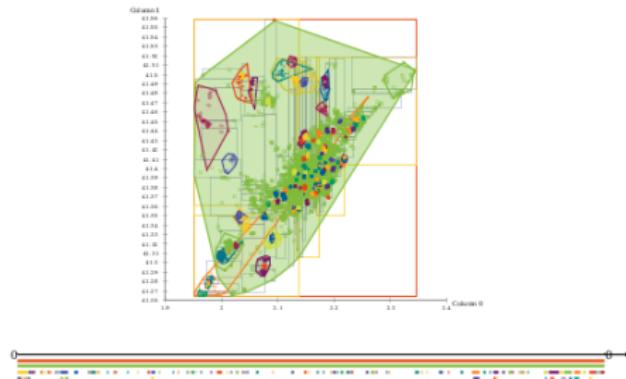


Less strict combination of parameters



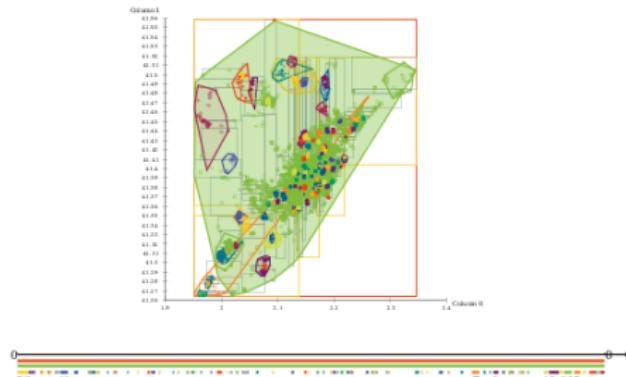
Strict combination of parameters

# OPTICS



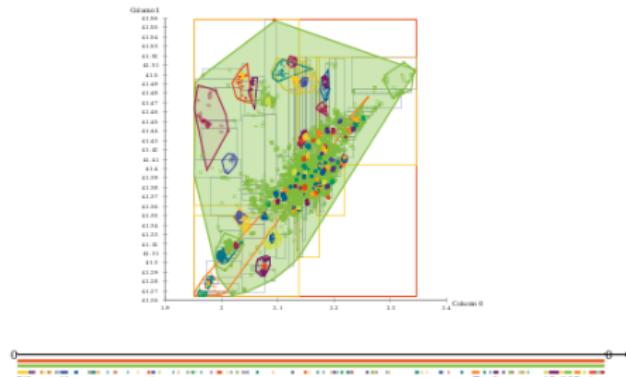
# OPTICS

- Generalization of DBSCAN.



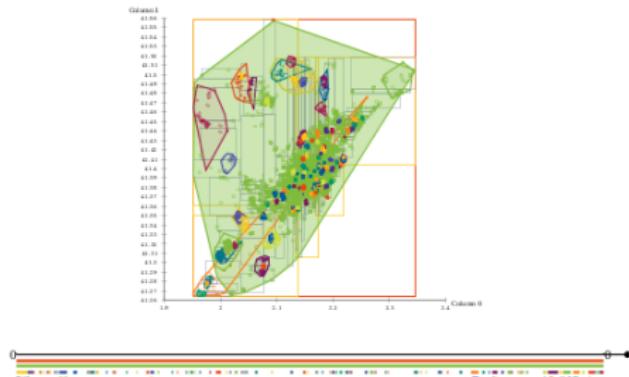
# OPTICS

- 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.



## Opticsxi: detecting Clusters



Opticsxi: detecting Clusters

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



## Opticsxi: detecting Clusters

- In the ordering clusters appear as *valleys*, separated by *noise regions*(peaks).
- We can define the *start* and *end* of a cluster, based on a threshold ( $\text{xi}$ ).



# Opticsxi: detecting Clusters

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



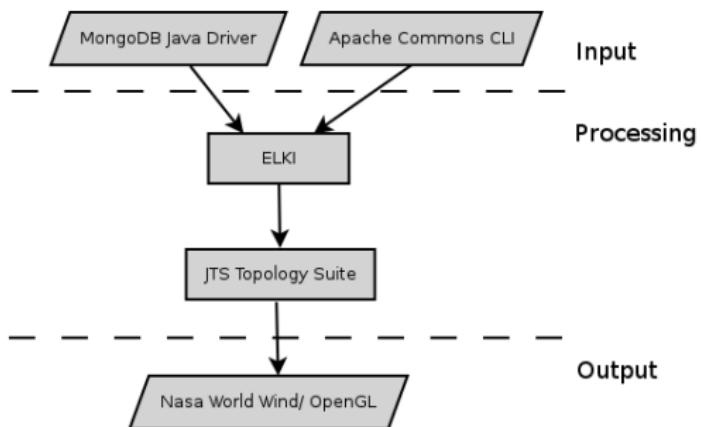
# GIS: Putting Geographic Information into Context

- Visualization is a key element to understand and interpret the results of data mining.
- What about Geospatial information...?
- WorldWind is an Open Source virtual globe developed by NASA that accesses a number of remote sensing datasets.
  - OpenGL.



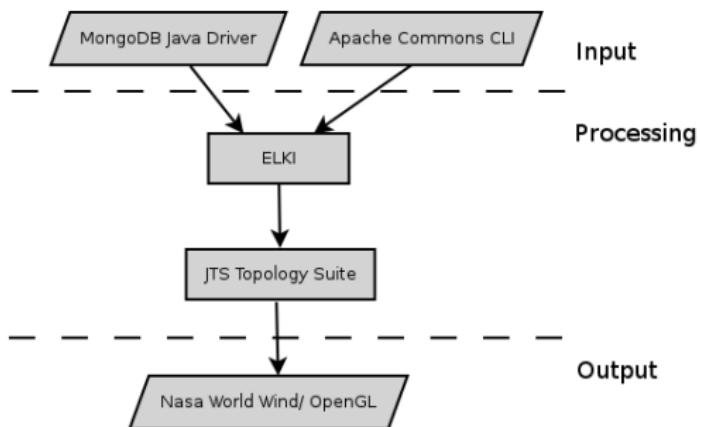
# Cluster Explorer

- Java App based on FOSS.



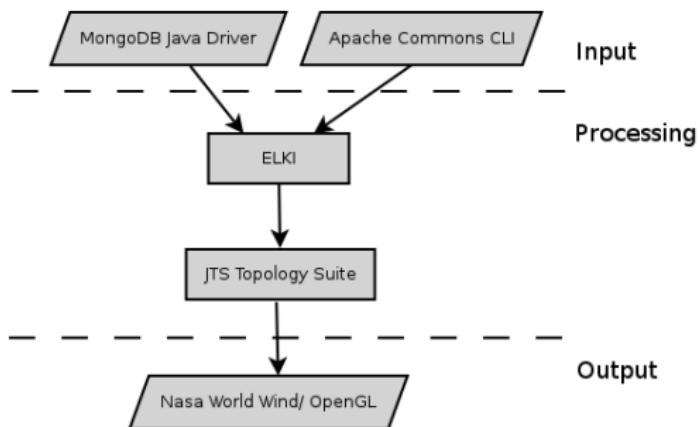
# Cluster Explorer

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



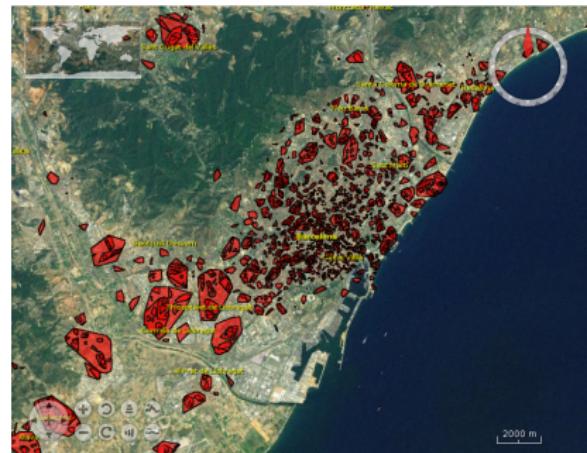
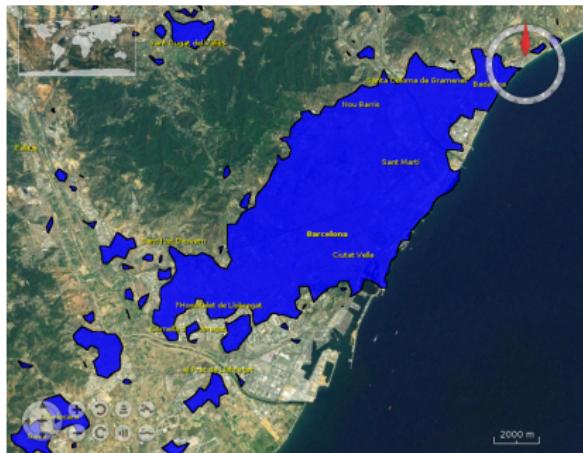
# Cluster Explorer

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

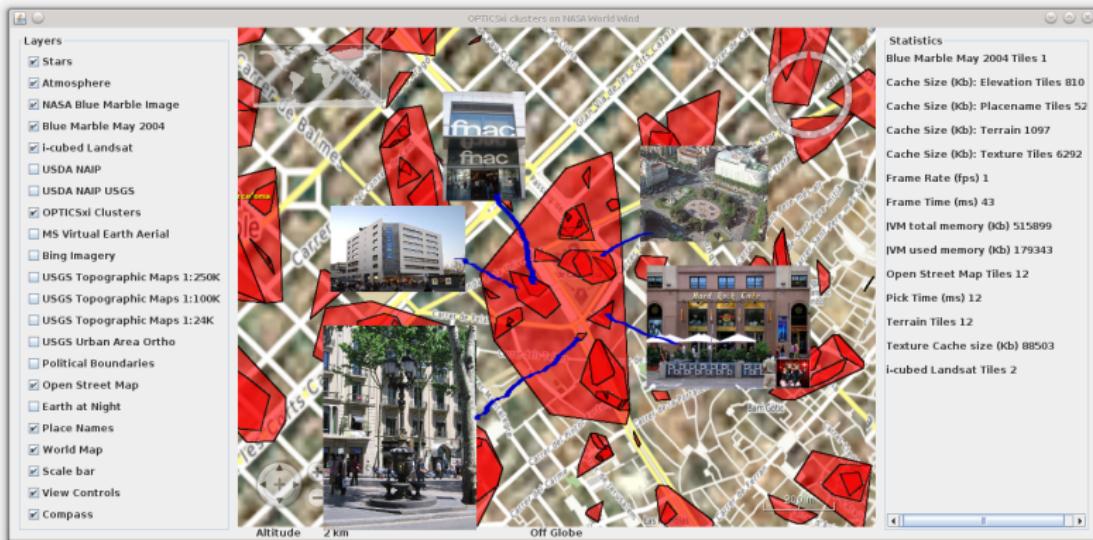


# A Case Study

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



# GIS and the value of Location-Analysis



## Final Remarks

## Final Remarks

- Flat cluster partition: summarizes the dataset and is easy to interpret.

## Final Remarks

- Flat cluster partition: summarizes the dataset and is easy to interpret.
- Hierarchical cluster partition: allows to look at city at multiple scales.

## Final Remarks

- Flat cluster partition: summarizes the dataset and is easy to interpret.
- Hierarchical cluster partition: allows to look at city at multiple scales.
- OPTICS: allows to detect clusters in areas of varying density.

## Final Remarks

- Flat cluster partition: summarizes the dataset and is easy to interpret.
- Hierarchical cluster partition: allows to look at city at multiple scales.
- OPTICS: allows to detect clusters in areas of varying density.
- 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.

## Final Remarks

- Flat cluster partition: summarizes the dataset and is easy to interpret.
- Hierarchical cluster partition: allows to look at city at multiple scales.
- OPTICS: allows to detect clusters in areas of varying density.
- 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.
- We hope to have demonstrated some potentialities arising from the integration between spatial data mining and GIS technologies, using FOSS.

# Thank You!

