

Geographical concentration of S&T activities

Lionel Villard

Main goals

Analyzing the **distribution of S&T** (here through patents and publications) **activities** and measuring the **aggregation effects** by identifying the existing geographical spaces where a high density of activity takes place.

The ambition is to look at clustering effects as they happen and not by considering **administrative borders that widely differ between countries**.

This suppose three steps :

- extraction of the **geographical informations** (toponyms, building names, postal codes...)
- **geocoding** of these information
- building boundaries to identify the **clusters** where the activities takes place

Geocode process

- Key elements to geocode addresses

- Data pre-processing

- Geocode extracted information

- How to generalize the geocoding process (work in progress)

Identifying clusters boundaries

- Common problems

- Two main propositions in RISIS

- A simple overview : algorithms and geocoded data

- A method based on a combination of two sequential approaches

- Main advantages of this method

- Parameters and java interface

Exemples of uses (nano s&t databases)

- Collaborations between clusters

- Temporal and dynamic characteristics

Future challenges

Key elements to geocode addresses

Identifying the **name of the cities** is a key element in the address for the geocoding process. When there is information at a lower scale, we use a dictionary (**postal codes** or **buildings name**) to geocode the address.

A fast toponyms disambiguation can be done by identifying the names of states for:

- **toponym type** : a city with a state's name
- **homonyms** : cities with the same name in the same country

US, Canada, Australia, Japan, Germany

Data pre-processing

- **Data gathering:**

data acquisition, extraction of the geographical information in the address (toponyms and postal codes);

- **Data standardization:**

data cleaning and identification of the best candidates at each scale

- country name or country codes at the country level;
- states or prefectures for the regional level;
- or place names and postal codes for the local scale.

Geocode extracted information

- **Matching** using place names and postal codes (even buildings names) dictionaries : comparison of the geographical data extracted to those of the database GeoNames. We have enriched the toponyms by the information of GeoNames (coordinates, but also other missing information). All ambiguous situations have been left side.
- **Geocoding based on a webservice** :
for the 9% left aside, we have proposed the addresses to a geocoding web service based (among other) on the Google engine (still 1.19% addresses not covered).

<i>Countries with more than 10 000 author's addresses</i>	<i>Harnonised country</i>	<i>Number of addresses</i>	<i>Addresses geolocalised</i>	<i>%</i>
Total for all the 166 countries		2 176 376	2 153 142	98,93%
UNITED STATES	US	471352	471322	99,99%
CHINA	CN	268630	268488	99,95%
JAPAN	JP	216934	215834	99,49%
GERMANY	DE	138001	137994	99,99%
FRANCE	FR	109136	109118	99,98%

How to generalize the geocoding process (work in progress)

GeoPy : a generic interface to query different Web Services (google, yahoo place, ... open street map...)

Main benefits :

- Ability to use open source web service (open street map, geonames) with less querying limitations
- Possibility to change the data sources (benefit from different coverages and accuracy)
- Better accuracy (localisation at the buildings level with open street map and google)
- Integration with other software (Platform CM...)

Geocode process

- Key elements to geocode addresses

- Data pre-processing

- Geocode extracted information

- How to generalize the geocoding process (work in progress)

Identifying clusters boundaries

- Common problems

- Two main propositions in RISIS

- A simple overview : algorithms and geocoded data

- A method based on a combination of two sequential approaches

- Main advantages of this method

- Parameters and java interface

Exemples of uses

- Collaborations between clusters

- Temporal and dynamic characteristics

Future challenges

Common problems

- **Dataset specific geographical distribution:** each dataset has his own geographical concentration;
- **Country specific boundaries:** each country has his own administrative boundaries and definition of municipality;

Two main propositions in RISIS

- **Administrative-based approaches:**

with a common definition of what is an urban area, these methods take in account the population concentration and some others demographic characteristics, and **merge administrative units** (mostly specific to countries). The boundaries produced can be used to map several variables in different contexts, and to characterise uniformly the new areas (OECD, 2012)

- **Bottom up approaches:**

project the geographical information of the field studied and build boundaries based on the **specific geographical distribution** of the data (IFRIS/ESIEE). Catch the local geographical concentration of the activities.

A simple overview : algorithms and geocoded data

Main families of algorithms that can be used with geographical data (M. Ouattara, 2010):

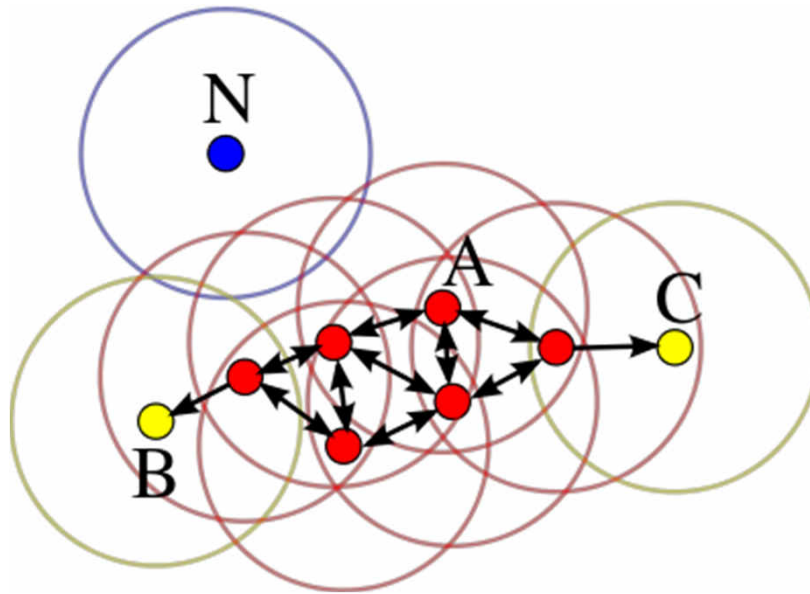
- **Hierarchical :**
algorithms try to group objects or divide them in subgroups;
- **Partition based methods :**
like k-means, objects shared some characteristics;
- **Density-based algorithms :**
boundaries of the clusters are built strictly by analyzing the geographical distribution of the activities.

A method based on a combination of two sequential approaches

1 / Identification of the **initial clusters** with a **density-based algorithm** (DBScan, 1996) that is able to identifying the area where the activities are concentrated. The clusters are defined by two parameters fixed before the calculation: all points of a cluster are surrounded by at least X points in a circle with a diameter of Y km.

Where are located the area in which activity is the most intense?

DBScan (*Density-Based Spatial Clustering of Applications with Noise*, M. Ester, HP. Kriegel, J. Sander & X. Xu, 1996)



Points A are core points

Points B and C are density-reachable from A and thus density-connected and belong to the same cluster

Point N is a noise point that is neither a core point nor density-reachable ($MinPts=3$ or $MinPts=4$)

Main advantages

- does not require to specify the **number of clusters** a priori
- can find **arbitrarily shaped clusters**. It can even find a cluster completely surrounded by (but not connected to) a different cluster.
- notion of noise, and is **robust to outliers**

Main disadvantages

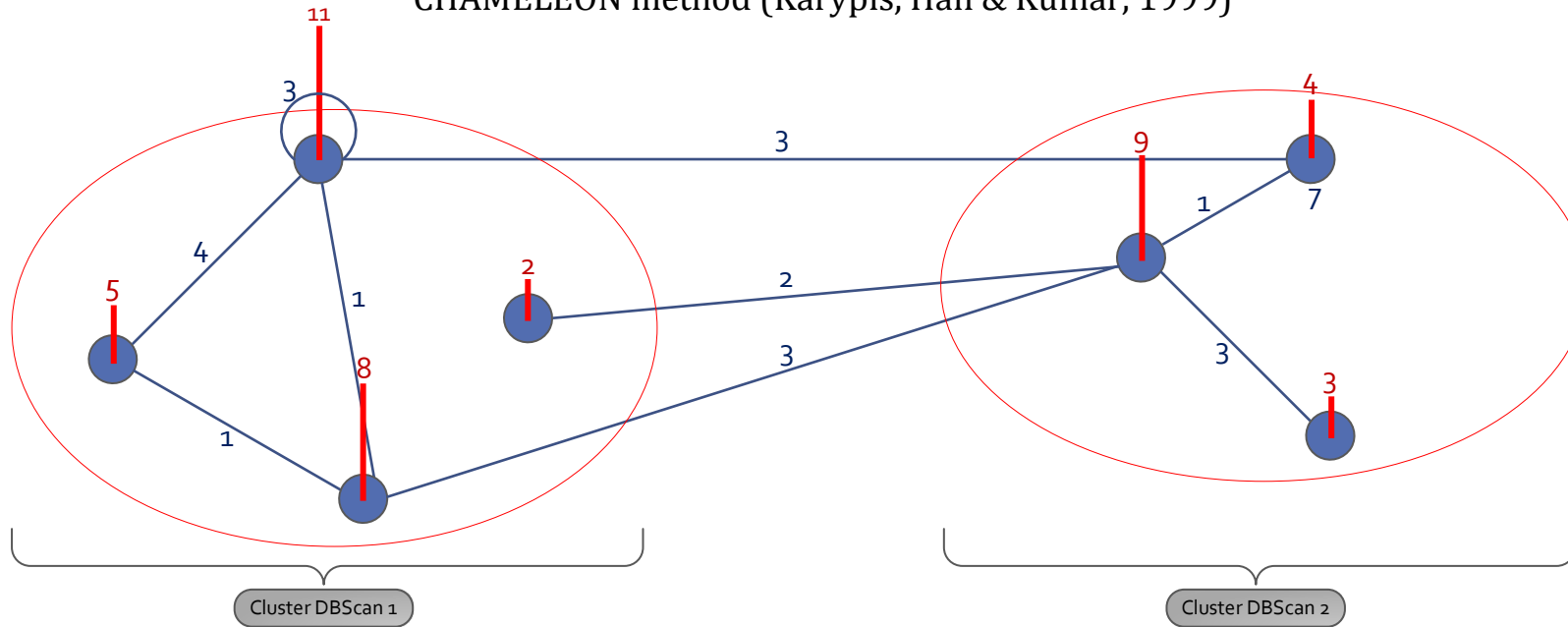
- not entirely deterministic: **border points** that are reachable from more than one cluster can be part of either cluster
- the quality of depends on the **distance measure** ("Curse of dimensionality" for high-dimensional data)
- cannot cluster data sets well with **large differences in densities**

2 / In a second step, we compare two different dimensions of the relation between the initial clusters:

2.1 How intense are the **relations between the initial clusters** (less than 100 km between the centroids) ? *RI/Relative Interconnectivity*

2.2 Does the final cluster will have a similar **profil of collaborations** as the two initial clusters taken separately (to avoid large variations of density of links in the final cluster) ? *RC/Relative Closeness*

CHAMELEON method (Karypis, Han & Kumar, 1999)



A cluster is defined by

T_i : the number of nodes (with different geographical coordinates),
 E_i : the links between these nodes
 C_i : the value of the links is the number of collaborations between the 2 nodes connected by this link

The relations between 2 clusters

$E_{(i,j)}$: The number of links between these two clusters
 $C_{(i,j)}$: the total number of collaborations supported by these links

Relative Interconnectivity (measure connectivity coherence between clusters)

RI is the ratio between the total number of collaborations between the two clusters ($C_{(i,j)}$) and the average number of internal collaborations of the two clusters.

$$RI_{(i,j)} = \frac{C_{(i,j)}}{\frac{C_i + C_j}{2}}$$

Relative Closeness (measure the similarity of the collaboration profiles of the two clusters)

The relative closeness between 2 clusters ($RC_{(i,j)}$) is the ratio between the absolute closeness of the two clusters (ratio between the total collaborations observed between the two clusters and the number of links between these two clusters) and the average internal closeness of the two clusters (based upon the number of nodes of the 2 clusters, $T_i + T_j$).

$$Cl_i = \frac{C_i}{E_i}$$

$$Cl_{(i,j)} = \frac{C_{(i,j)}}{E_{(i,j)}}$$

$$RC_{(i,j)} = \frac{Cl_{(i,j)}}{\frac{T_i}{T_i + T_j} \times Cl_i + \frac{T_j}{T_i + T_j} \times Cl_j}$$

Main advantages of this method

The combination of a **purely geometric approach** (a fixed perimeter to measure a density) and the analysis of the **relations between spaces** makes possible to build clusters boundaries :

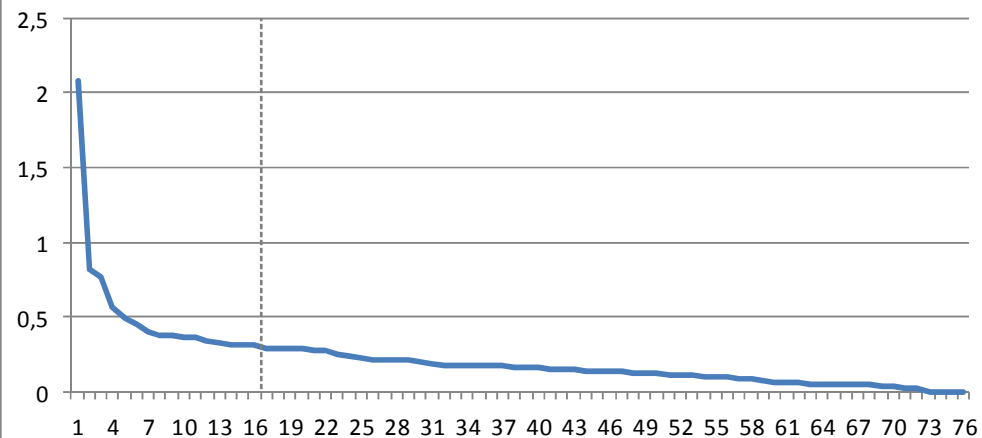
at a **micro scale** (clusters in a specific part of a town)
as well as at a **macro scale** (at a regional level for worldwide comparisons).

Example of local parameters selected for the databases on nanotechnologies (publications and patents)

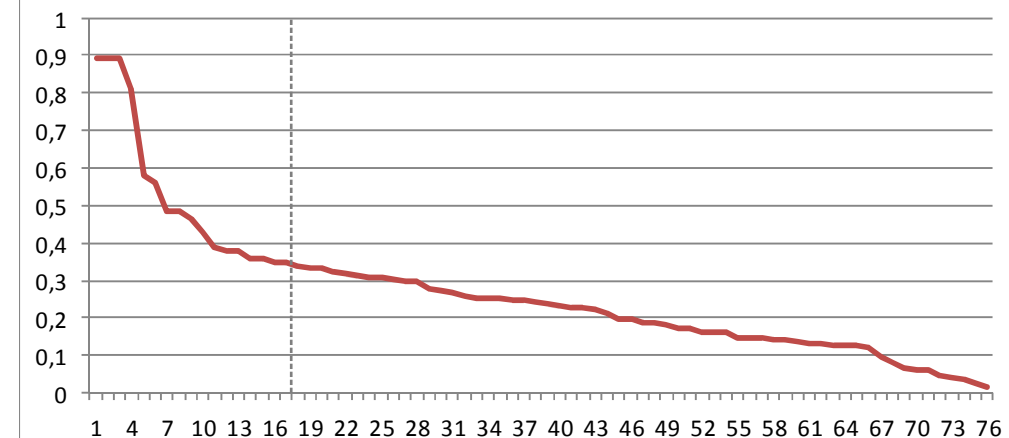
Criteria	Publications	Patents
Relative Interconnectivity	0.28	2
Relative Closeness	0.32	12.50
Minimal weight (addresses)	1 500	1 250
Maximum distance	25 Km	20 Km
Number of addresses analysed	2 129 107	1 523 093
Number of initial clusters (after DBScan)	313	186
Number of final clusters (after Chameleon)	295	155

Parameters for DBScan and Chameleon

RI (publications)



RC (publications)



A program in Java with an interface (Michel Revollo, 2014)

csv files inputs

- files with the different addresses depending of the sources (with temporal information, if needed)
- one file for all the geographical coordinates

csv files outputs

(clusterised coordinates and RI/RC value)

IDc	Latitude	Longitude	NbPub	ClustDBScan	ClustCham	Fusion
15768	19,3411999	-99,1486969	1821	17	17	Y
100333	19,3332996	-99,1166992	28	17	17	Y
100330	19,3288994	-99,1603012	7	17	17	Y
100341	19,3600998	-99,1062012	2	17	17	Y
15813	19,4284992	-99,1277008	3129	18	17	Y
15822	19,4500008	-99,1166992	38	18	17	Y
15811	19,4225006	-99,2027969	36	18	17	Y
15797	19,3999996	-99,1999969	8	18	17	Y
100365	19,4167004	-99,1832962	1	18	17	Y
100366	19,4167004	-99,0667038	1	18	17	Y
100368	19,4172001	-99,1568985	1	18	17	Y
15876	19,7514992	75,7138977	3307	19	19	N
16138	22,2854996	114,1579971	7814	20	20	N
16168	22,3167	114,1829987	4966	20	20	N
16180	22,3332996	114,1829987	240	20	20	N

Geocode process

- Key elements to geocode addresses

- Data pre-processing

- Geocode extracted information

- How to generalize the geocoding process (work in progress)

Identifying clusters boundaries

- Common problems

- Two main propositions in RISIS

- A simple overview : algorithms and geocoded data

- A method based on a combination of two sequential approaches

- Main advantages of this method

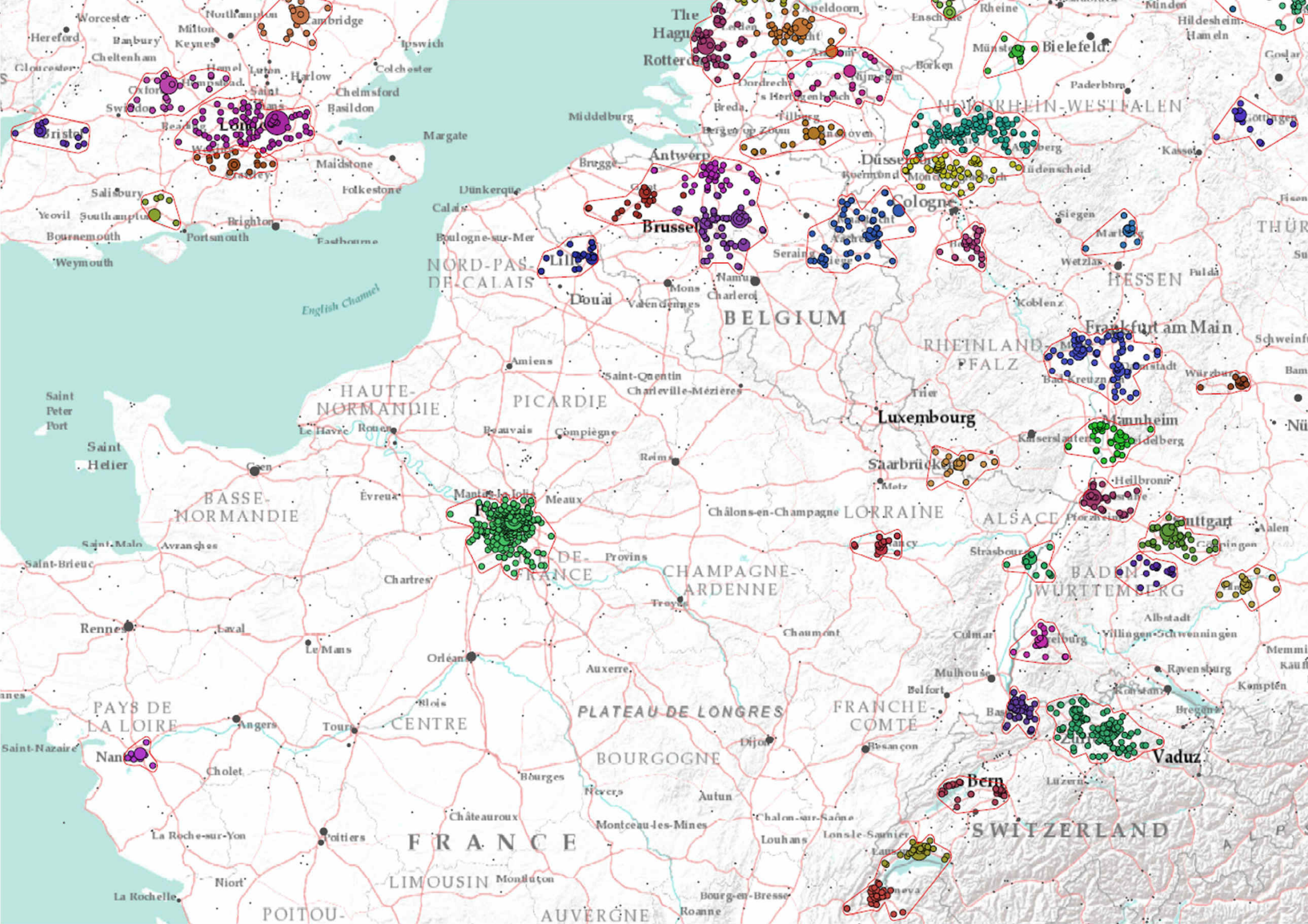
- Parameters and java interface

Exemples of uses

- Collaborations between clusters**

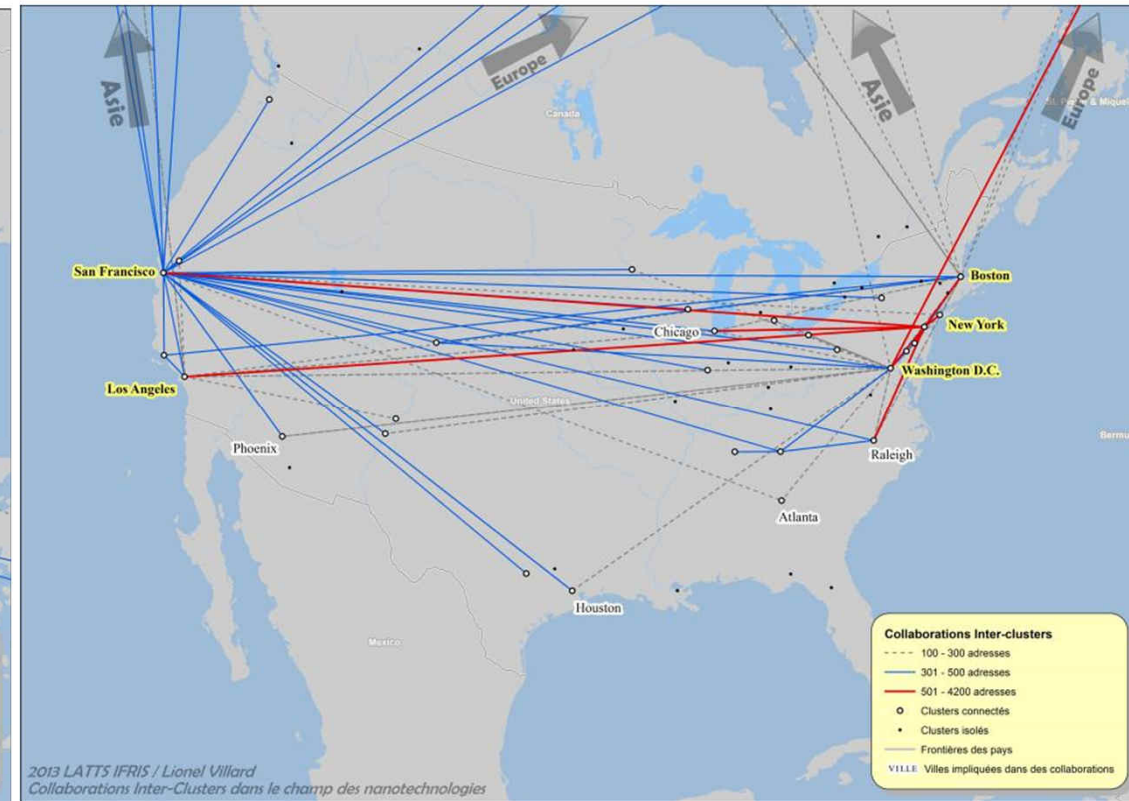
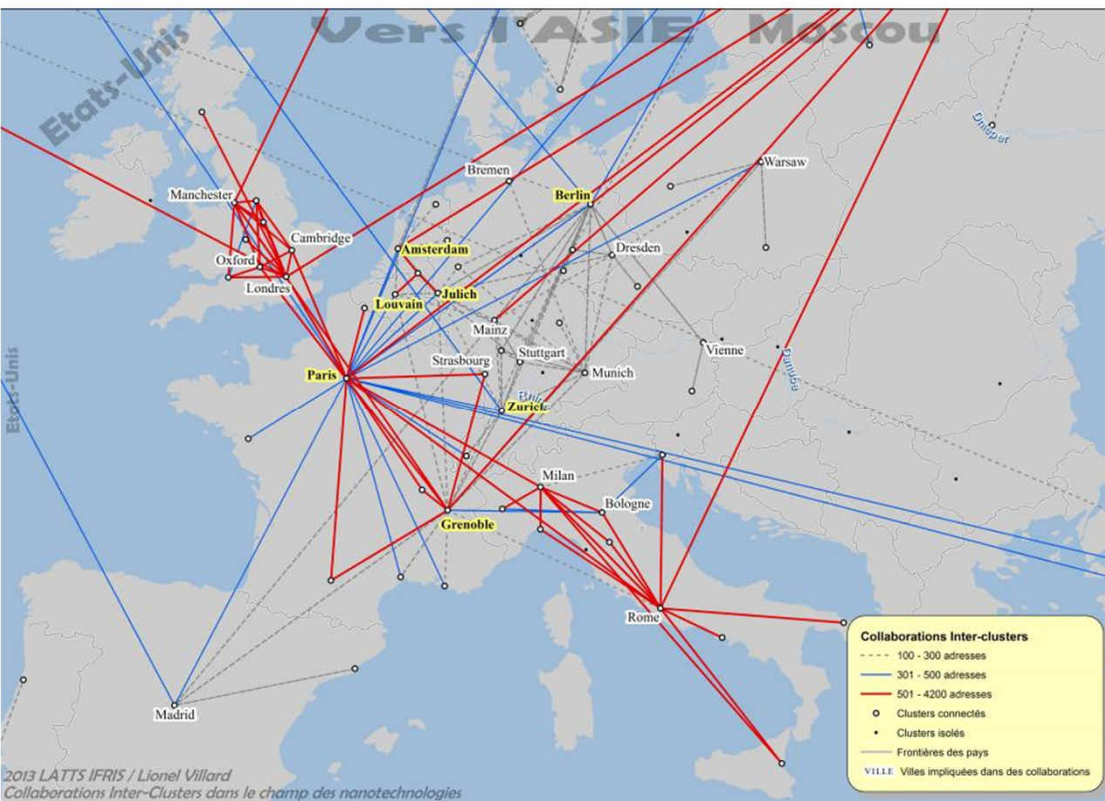
- Temporal and dynamic characteristics**

Future challenges



With a common definition of the unit of analysis, we can characterize and compare the different spaces where the activities are concentrated.

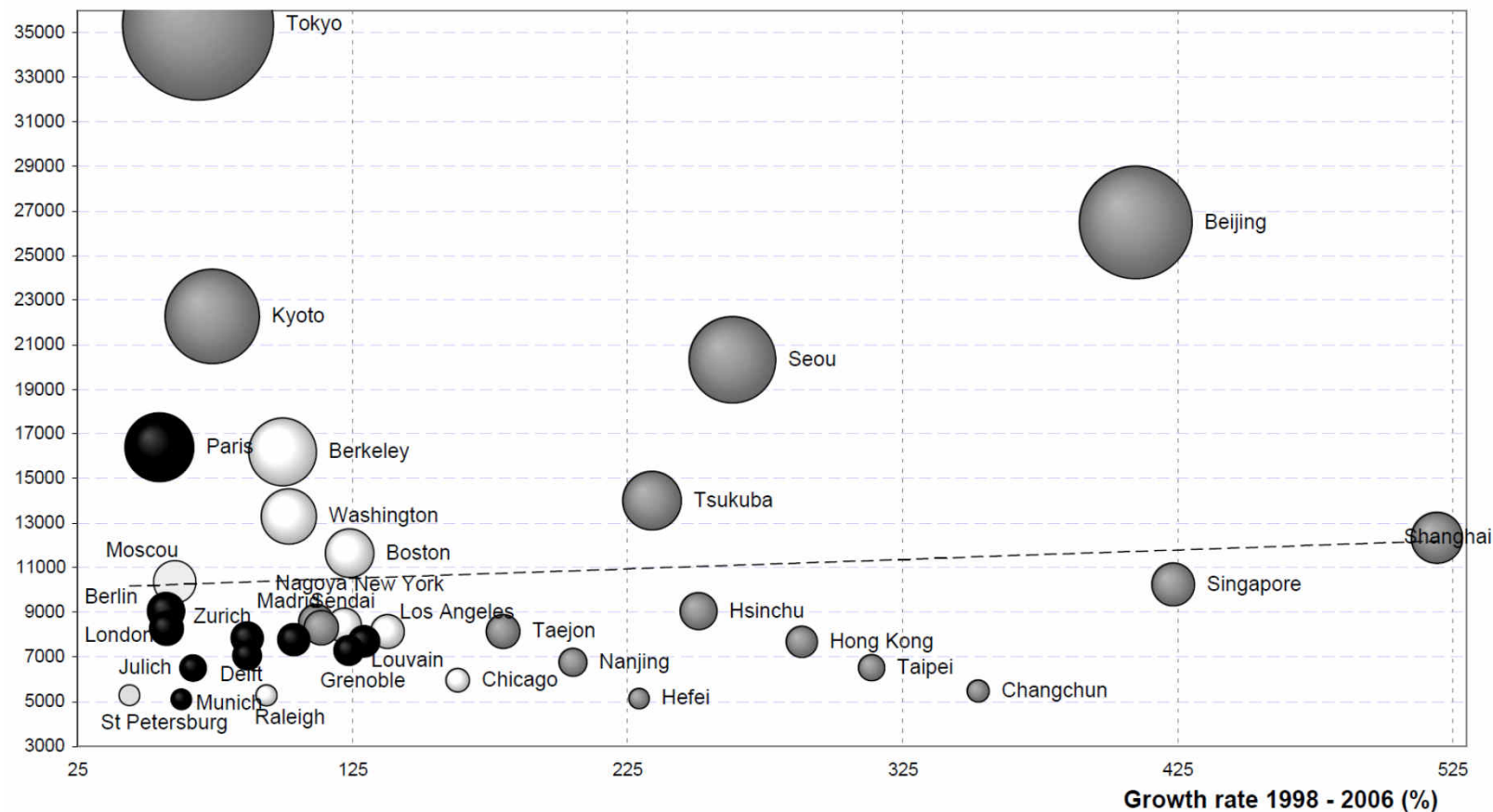
For example in terms of collaboration between continents and inside countries (network of co authors)



or to characterised clusters : link between size and rate of growth (A. Delemarle & al, 2009)

Table 5: Progression, Stability and Loss of Ground in Hierarchy by Size of Cluster and Geographical Area³⁰

Area	Size of Clusters								
	Small			Medium			Large		
	Progress	Stable	Lost Ground	Progress	Stable	Lost Ground	Progress	Stable	Lost Ground
Asia	26	1	6	7		2	5	2	
Europe	8	15	48			9			1
U.S. & Canada	17	10	21	1	2	1		2	1
Other	6	2	8			1			1



Geocode process

- Key elements to geocode addresses

- Data pre-processing

- Geocode extracted information

- How to generalize the geocoding process (work in progress)

Identifying clusters boundaries

- Common problems

- Two main propositions in RISIS

- A simple overview : algorithms and geocoded data

- A method based on a combination of two sequential approaches

- Main advantages of this method

- Parameters and java interface

Exemples of uses

- Collaborations between clusters

- Temporal and dynamic characteristics

Future challenges

Future challenges

for RISIS concerning the geographical aspect and the concentration of the activities:

- **Improving the geocode process** to be able to treat large dataset with heterogeneous sources (of other facilities).
- Do we want to **apply the OECD method** to other countries that are not covered by the actual Urban Areas?
- How to use of the **Urban Areas name to label the clusters** created in our datasets (when there is one)?

Estimating the overlap between our method and the OECD urban areas boundaries :

- Where are the perfect **overlaps**?
- Is there any OECD **urban areas that are split** with our methods?
- Is there **new places that emerge** outside the OECD urban areas boundaries in the covered countries?