Geo-statistical Exploration of Milano Datasets

*Irene Celino and **Gloria Re Calegari***

CEFRIEL

The 13th International Semantic Web Conference 2014
Riva del Garda, Italy
19 – 23 October 2014

Semantic Statistics workshop
19th October 2014

Research goal

Long term goal

- Compare data related to the same city but obtained from heterogeneous sources. Do they provide the same 'picture' of the city?

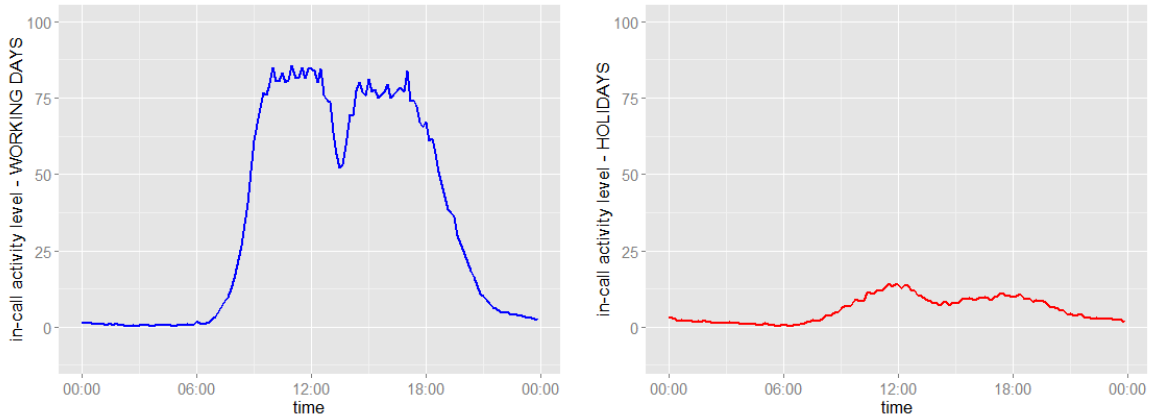
- Can we update a dataset (expensive and time consuming, updated once every 5/10 years) with another dataset (cheaper and always up to date)?

Presentation target

- Comparison of two heterogeneous datasets referring to the same city (Milan) to discover if they have any intrinsic correlation

Datasets available

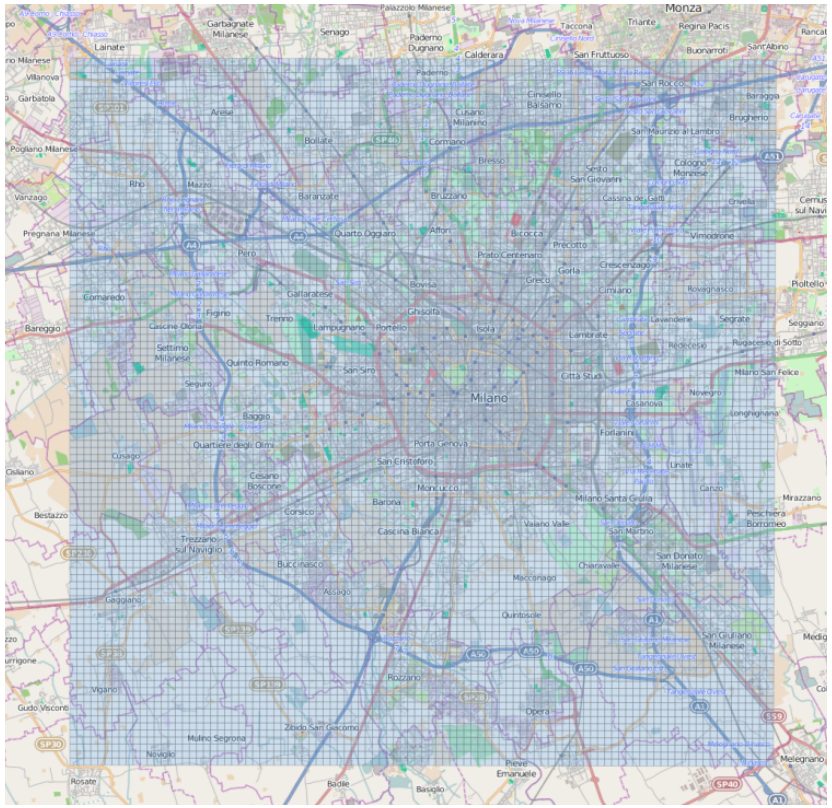
- Telecom (phone activity data) provided for their “Big Data Challenge”
  - Milan + surroundings
  - Nov-dic 2013
  - Grid of 10.000 cells
  - Activity recorded every ten minutes
  - A footprint for each cell



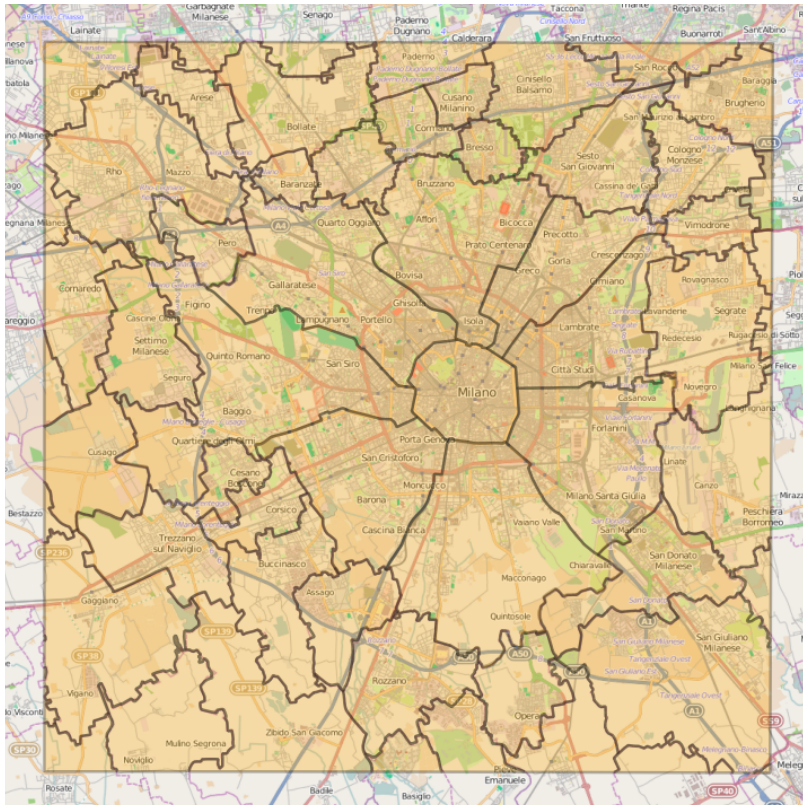
- ISTAT - Italian National Statistical Institute
  - demographic data of 2011 and 2001
  - divided by sex, age and nationality

Analysis performed using data in their original format (GeoJSON, csv) and serialization in RDF format at the end of the analysis.

Datasets available – different spatial granularity



Telecom – grid of 10.000 cells (250 x 250 m each)



ISTAT – 50 surrounding towns + 9 internal Milano districts

Pre-processing of data required.

Mapping of Telecom data into municipality granularity.

Methodology of analysis

Goal of analysis: do the two datasets have the same intrinsic meaning?

Unsupervised clustering to group data in each dataset:

- k-means (euclidean distance)
- Adaptive k-means (cosine distance)

Comparison of the two clustering results to find correlation between the two datasets.

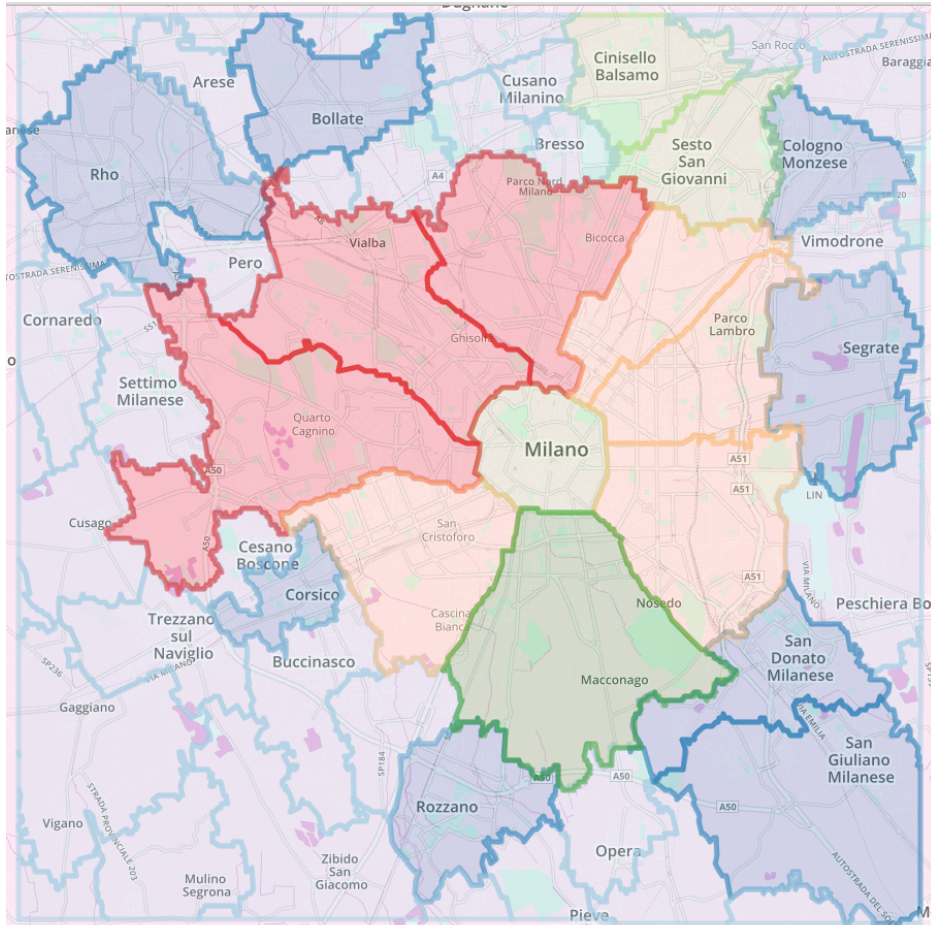
Validation of the comparison:

- Rand Index
- Kappa Index

(the closer to 1 the index is, the stronger the correlation between the data clusterings)

Experiments- Telecom vs ISTAT (range on total poulation)

Telecom - Kmeans 6 classes



ISTAT 2011 - range 6 classes

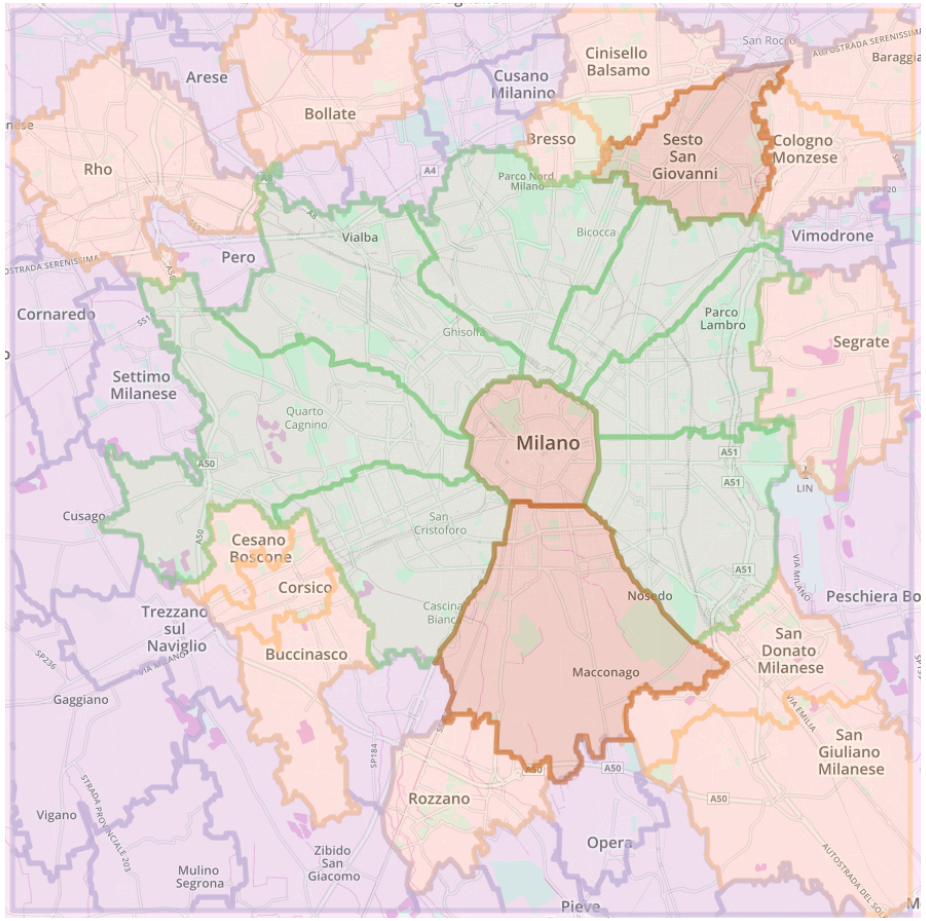
Rand Index	Kappa Index
0,23	0,23

Only partial correlation

Try to add information to total population (feature vector with distribution of populatin divided by sex, age and nationality)

Experiments - Telecom vs ISTAT (2011 fine-grained population)

Telecom - Kmeans 4 classes



ISTAT 2011 – Kmeans 4 classes

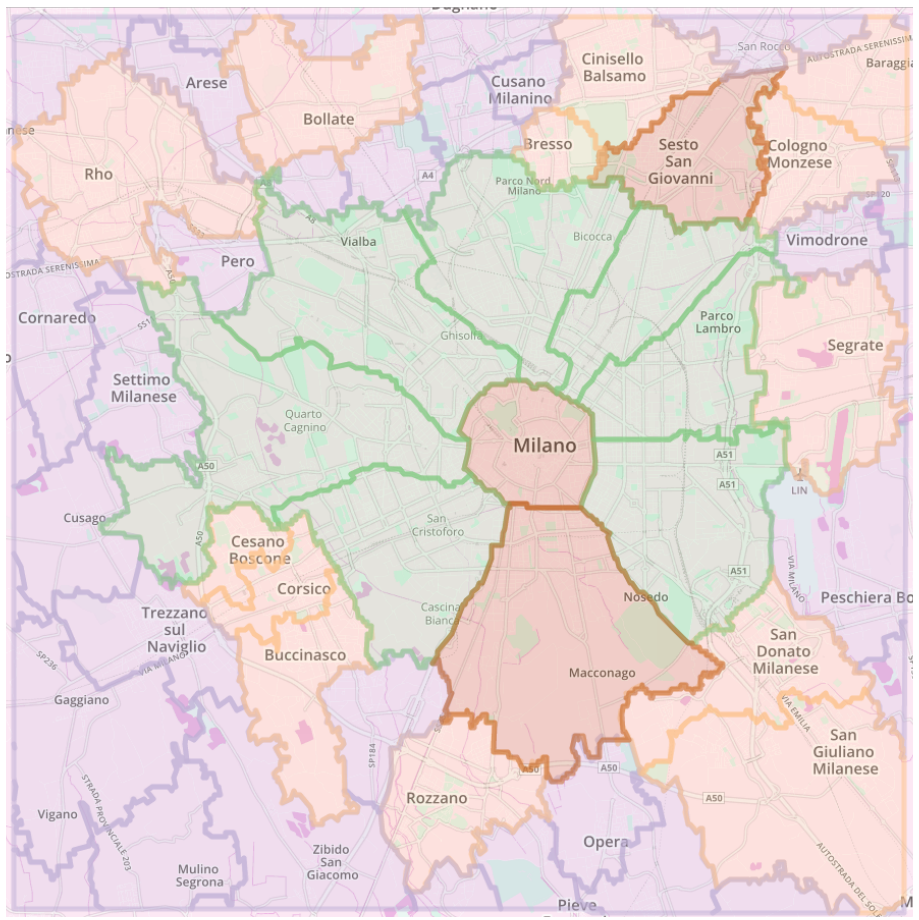
Rand Index	Kappa Index
0,77	0,81

Good correlation

Try to add historical information (2001 datasets). Does adding temporal dynamic improve corrrelation?

Experiments- Telecom vs ISTAT (2011+2001 fine-grained population)

Telecom - Kmeans 4 classes

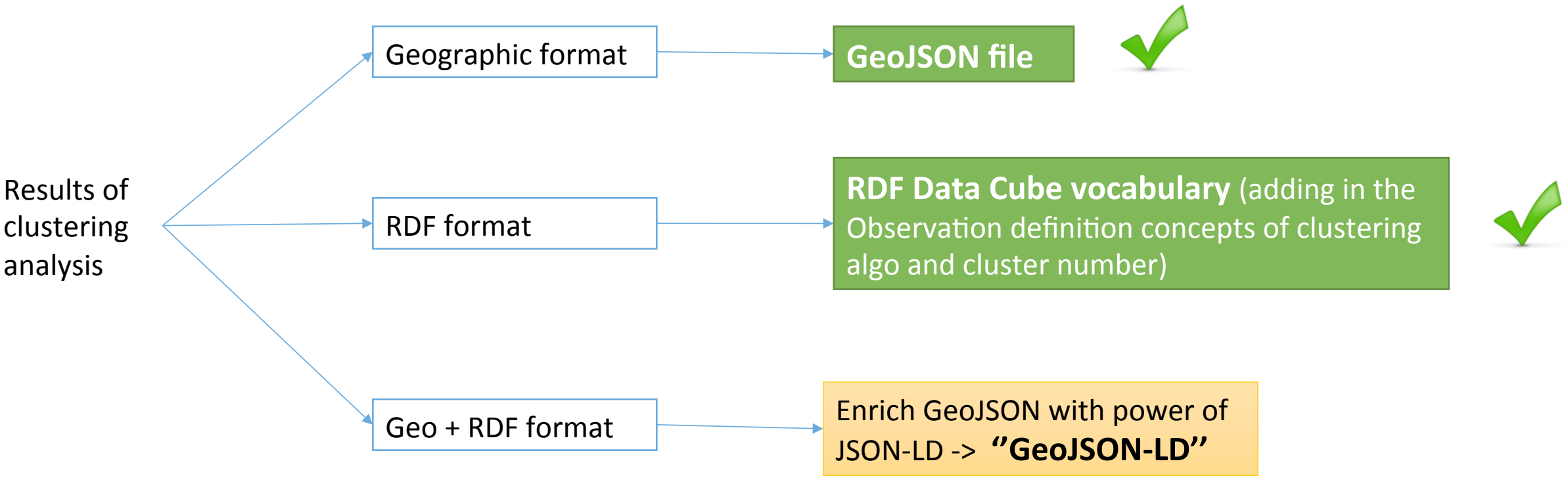


ISTAT 2011 + 2001 – Kmeans 4 classes

Rand Index	Kappa Index
0,77	0,81

Good correlation. The same as the previous test.

Clustering results serialization



‘’GeoJSON-LD’’

‘’GeoJSON-LD’’ is obtained by adding the ‘@context’ prefix to the GeoJSON file.

The prefix specify how to interpret GeoJSON tags as RDF resources

@context prefix

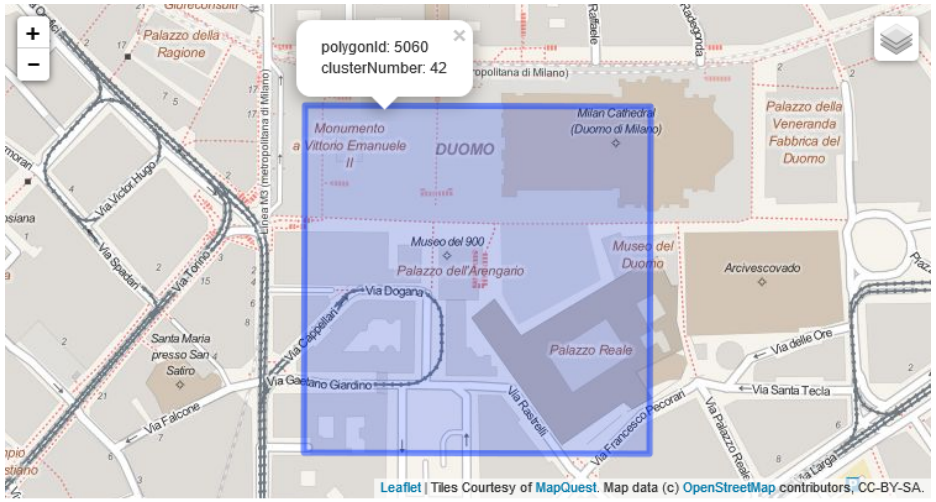
```
{
    "@context": {
        "qb" : "http://purl.org/linked-data/cube#",
        "geo" : "http://www.w3.org/2003/01/geo/wgs84_pos#",
        "sf" : "http://www.opengis.net/ont/sf#",
        "rdfs" : "http://www.w3.org/2000/01/rdf-schema#",
        "ex" : "http://example.org#",
        "type" : "@type",
        "Feature" : "qb:Observation",
        "FeatureCollection" : "qb:DataSet",
        "features" : { "@reverse": "qb:dataSet", "@container": "@set"      },
        "coordinates": { "@id": "geo:lat_long", "@container": "@set" },
        "Polygon": "sf:Polygon",
        "polygonId": "rdfs:label",
        "properties": "ex:properties",
        "geometry": "ex:location",
        "clusterNumber": "ex:clusterNumber"
    },
    …
```

GeoJSON file

```
    "type": "FeatureCollection",
    "crs": {
        "type": "name",
        "properties": { "name": "urn:ogc:def:crs:OGC:1.3:CRS84" }
    },
    "features": [
        {
            "type": "Feature",
            "properties": { "polygonId": 5060, "clusterNumber": 42 },
            "geometry": {
                "type": "Polygon",
                "coordinates": [
                    [
                        [ 9.188868698831616, 45.464410737578497 ],
                        [ 9.191874719535541, 45.46440572774091 ],
                        [ 9.191867544215187, 45.46229045910362 ],
                        [ 9.188861635922354, 45.462295468573735 ],
                        [ 9.188868698831616, 45.464410737578497 ]
                    ]
                ]
            }
        }
    ]
}
```

‘’GeoJSON-LD’’ pros and cons

Correct interpretation of the geographic information in the GeoJSON-LD file



GeoJSON-LD is correctly interpreted as GeoJSON by GIS

Problems in the interpretation of the RDF information.

‘’GeoJSON-LD’’ does not interpret correctly a lists of lists (polygon coordinates representation)

GeoJSON list of lists

```
"geometry": {
      "type": "Polygon",
      "coordinates": [
            [
                  [ 9.188868698831616, 45.464410737578497 ],
                  [ 9.191874719535541, 45.46440572774091 ],
                  [ 9.191867544215187, 45.46229045910362 ],
                  [ 9.188861635922354, 45.462295468573735 ],
                  [ 9.188868698831616, 45.464410737578497 ]
            ]
      ]
}
```

RDF serialization

```
_:c14n0 <http://www.w3.org/2003/01/geo/wgs84_pos#lat_long>
"4.5464405727740 91E1"^^<http://www.w3.org/2001/XMLSchema#double> .
_:c14n0 <http://www.w3.org/2003/01/geo/wgs84_pos#lat_long>
"4.54644107375785E1"^^<http://www.w3.org/2001/XMLSchema#double> .
_:c14n0 <http://www.w3.org/2003/01/geo/wgs84_pos#lat_long>
"9.188861635922354E0"^^<http://www.w3.org/2001/XMLSchema#double> .
_:c14n0 <http://www.w3.org/2003/01/geo/wgs84_pos#lat_long>
"9.188868698831616E0"^^<http://www.w3.org/2001/XMLSchema#double> .
_:c14n0 <http://www.w3.org/2003/01/geo/wgs84_pos#lat_long>
"9.191867544215187E0"^^<http://www.w3.org/2001/XMLSchema#double> .
```

Conclusion and future works

- Identification of a correlation between phone activity data and demographic information
  - Further tests using different datasets (land use, Point of interests)
  - Find efficient method for handling the different granularity levels of the datasets
  - Method for handling temporal aspect of phone activity data (we lost temporal information during clustering process)

- Find a method to overcome ‘’GeoJSON-LD’’ serialization problem

Thank you! Any question?

Further details at: http://swa.cefriel.it/geo/

Irene Celino and Gloria Re Calegari

CEFRIEL