Geo-statistical Exploration of Milano Datasets

Irene Celino and Gloria Re Calegari

CEFRIEL – Politecnico di Milano, via Fucini 2, 20133 Milano, Italy
{irene.celino,gloria.re}@cefriel.com

Abstract. Data about cities is today available from a multitude of sources at high volumes and rates; a multi-disciplinary and cross-sectorial data processing and analysis has become more and more pervasive to better understand urban evolution. Statistical datasets, together with other (big) data related to cities, can be employed to semantically explore and characterize the urban space.
In this paper, we present our early work in processing, combining and contrasting different datasets about the city of Milano – population statistics and phone call records – and we briefly discuss about the challenges and opportunities related to the semantic representation of this kind of geo-statistical data.

1 City Data Availability

With the advent of digital computing and its increasing pervasiveness, in complex environments like cities multiple heterogeneous stakeholders – public bodies, private businesses and utilities, citizens – produce, consume and exchange digital information at an ever growing volumes and rates. While data spans across very different domains, those datasets offer different point of views on cities, thus providing different "reflections" of what happens in the environment.

In our work, we aim at analysing data related to the same city but obtained from heterogeneous sources: by comparing and contrasting those datasets we would like to understand whether they provide the same "picture" of the city. We focus our analysis on the city of Milano (and its metropolitan area), exploring two datasets: the open data about population demographics from ISTAT, the Italian Institute for Statistics[1], and two months of mobile call data records provided by the Telecom Italia mobile operator for their "Big Data Challenge"[2]. The used datasets are heterogeneous in content, format, granularity and volume, thus they require different processing methods (cf. Section 2); still they convey some "semantics" of the Milano territory, thus their statistics and analytics are worth a uniform semantic representation (cf. Section 3).

2 City Geo-statistical Data Processing

The main idea of our work is to try to use phone activity data as a "proxy" for other kinds of data. We would like to update a dataset, whose maintenance may

[1] Cf. http://www.istat.it/en/.

[2] Cf. http://www.telecomitalia.com/tit/en/bigdatachallenge.html.

be very expensive and time consuming, by using another different dataset which is cheaper and more easily available. In our case, data about the population density and distribution is provided by the ISTAT census. This data collection is costly and demanding and takes place once every 10 years, on average. On the other hand Telecom call data records are continuously generated as a side effect of the mobile service.

In this section we describe our datasets and the experiments we perform to understand if those datasets are correlated and so, if they can be used in a combined way to have an overall picture of the population distribution.

2.1 Available datasets and pre-processing

Since some of the datasets available in our analysis are big data (million records, cf. Table 1) we need a statistical tool to process them and produce a compressed representation. We decide to use R, a free software environment for statistical computing and graphics[3]. As regards data serialization, we decide to analyse them in their native format (which is not RDF) and, at a later time, to convert and publish the processing results in RDF (cf. Section 3).

Domain (content)	**Data Source**	**Data Format**	**Spatial Granularity**	**Reference Period**	**Volume (records)**
Statistics (population)	ISTAT	Tabular	Municipality/ district level	2001 & 2011	10s
Mobile Telephony (call records)	Telecom Italia	Tabular	City grid cells (250mx250m)	2013 Nov-Dec	100Ms

Table 1: Characteristics of the used datasets

Telecom dataset describes the phone activity occurred in the area of Milano for about two months. The Milano area is mapped into a grid of 10.000 cells (250 x 250 m) and, every ten minutes, for each cell of the grid, the number of incoming/outcoming calls, SMSs and Internet activity are recorded. To reduce the dataset size and to take into account the spatial information, we decide to compress all the data of each cell into a “footprint”, i.e. a summarizing data structure which records for each time slot of ten minutes the average activity of that cell, distinguishing between working days and holidays. The resulting data consists of one footprint for each cell and for each activity type.

As regards ISTAT information, the available data is about the number of inhabitants for each Italian municipality, divided by age, sex and nationality. We take into account only those districts that overlap with our Milan-grid. The granularity of this dataset is considerably more coarse-grained than the one of Telecom dataset (59 districts vs 10.000 cells).

[3] Cf. http://www.r-project.org/.

2.2 Geospatial information processing

Since in our analysis we need to take into account all the geospatial information, we use the free GIS application QGIS[4] to perform the required geospatial operations (merging, overlapping, intersection of map layers).

As regards the demographics dataset, we get the district geo-data as shape files from ISTAT as well. Since the municipality of Milano is too large compared to the other municipalities of the metropolitan area and data about the nine territorial divisions of Milano are available, we decide to account these 9 districts as they were municipalities as well. Therefore, the geospatial layer with the nine Milano districts is added to the one with municipalities by applying a GIS operation.

We exploit QGIS functionalities also to handle the different granularity levels, in order to make data comparable. Since the granularity of ISTAT and Telecom datasets are different, we map the Milano grid-cell into the municipality areas by assigning to each district the weighted sum of phone activities of the corresponding cells.

2.3 Analytics approach

Once all the available datasets are pre-processed as explained, we proceed with their analysis in order to understand if a correlation between the datasets exists. In particular we want to examine if the phone activity is correlated with the demographic information, so that we can use Telecom data to keep population distribution up to date in an efficient way.

We compare the clusters, obtained via simple grouping/subsetting or through clustering algorithms, between ISTAT and Telecom data. We adopt the K-Means unsupervised clustering algorithm which exploits the Euclidean distance as similarity metrics.

Then we compare the clustering results, measuring the correlation between groups using the Adjusted Rand Index [1], which represents a measure of similarity between two data clusterings, and the Kappa index [2], which takes into account the number of elements on the diagonal of a contingency table. In both cases, the closer to 1 the index, the stronger the correlation between the data clusterings.

2.4 Clustering Experiments

Experiment 1. Firstly we test if the phone activity is directly correlated with the number of inhabitants of a district. For this reason we classify the ISTAT 2011 dataset in a rough way, subdividing the 59 municipalities based on the total number of population using a set of thresholds. We group the districts in 6 classes and we compare them with the 6 Telecom data clusters calculated by the K-Means algorithm. The resulting Rand and Kappa indexes values (0,23 each) are quite far from the desired value of 1, indicating a low degree of correlation

[4] Cf. `http://www.qgis.org/it/site/`.

between this data (cf. Table 2). Even if we group municipalities in more or fewer classes than 6, the indexes don't improve.

Experiment 2. Since using only the total population doesn't seem to give meaningful and useful results, we try to analyse a more fine-grained ISTAT 2011 demographic dataset, which splits up population by sex, age and nationality. As this data is multi-dimensional, we decide to use clustering algorithms to group data. After trying various combination of algorithms and parameters we find out that the one which maximizes the correlation indexes with the phone activity is the K-Means with 4 clusters. In this case we have a considerable improvement in terms of Rand and Kappa indexes, which reach respectively the values of 0,77 and 0,81, as shown in Table 2.

Experiment 3. Then, we wonder if adding historical demographic information to the current ISTAT 2011 data can improve the correlation, because of the additional information about population evolution over time. However, the results obtained using both the 2001 and 2011 ISTAT data are the same as the previous experiment in terms of both Rand and Kappa coefficients (cf. Table 2) and in terms of clusters created.

Experiment	**Dataset 1**	**Dataset 2**	**Cluster dataset 1**	**Cluster dataset 2**	**Adjusted Rand Index**	**Kappa Index**
1	ISTAT 2011	Telecom	Range - 6 classes	K-means 6 classes	0,23	0,23
2	ISTAT 2011	Telecom	K-means 4 classes	K-means 4 classes	0,77	0,81
3	ISTAT 2001+2011	Telecom	K-means 4 classes	K-means 4 classes	0,77	0,81

Table 2: Cluster correlation indexes between ISTAT and Telecom datasets.

To sum up, the high values of Rand and Kappa indexes obtained in experiments 2 and 3 indicate that a correlation between demographic data and phone activity might exist. Those experiments' results can be visually explored on the maps provided at `http://swa.cefriel.it/geo/semstats2014.html`.

3 City Geo-statistical Data Representation

Urban datasets from different sources can be heterogeneous because they reflect and cover different aspects of the city. Still, they can share some commonalities: their spatio-temporal characteristics [3, 4]. Therefore, in representing the results of our geo-statistical data processing, besides making use of the RDF Data Cube vocabulary [5], we can adopt interoperable models for spatial and temporal information. Hereafter, we discuss opportunities and open issues related to the statistical and spatial dimensions and we focus on Semantic Web/Linked Data representations; on the other hand, we do not address the time dimension, for which we simply adopt the W3C Time Ontology [6].

3.1 (Geo)clustering information represented with RDF Data Cube

The statistical analyses on urban datasets illustrated in Section 2 mainly consist in applying clustering algorithms, therefore the statistical results we would like to semantically represent are the geographic clusters.

Since RDF Data Cube [5] is one of the most popular and comprehensive models to describe multi-dimensional statistical data, we adopt this vocabulary also for our clustering results and we create a possible definition of the `qb:MeasureProperty` for the cluster numbering, the `qb:AttributeProperty` for the clustering algorithm and their respective `qb:ComponentSpecifications`. This definition is generic and can be applied to any clustering scenario. We successfully employed it to translate to RDF the results of our Milano data processing, to enable a Linked Data-compliant publication of such data on the Web. Regarding the representation of our geospatial aggregation unit – the cell in the previous example – we re-used `sf:Polygon` class defined in GeoSPARQL [7]. To describe the geographic coordinates of the polygon vertices we employed the `gsp:asWKT` property, but several other alternative options are possible[5].

Both the clustering Data Cube definition and the respective RDF representation of a sample observation are linked from this submission's web page at `http://swa.cefriel.it/geo/semstats2014.html`.

3.2 Geo(clustering) information represented with GeoJSON(LD)

The geospatial dimension of our input datasets was available in GIS-specific format, as GeoJSON files. As explained in Section 2, we process the datasets in their native formats, thus it is natural to produce the data processing output using again a geospatial data interchange format; because of its simplicity and popularity, we decide to adopt GeoJSON[6] [9], which is supported by numerous GIS software packages as well as mapping services. In the GeoJSON serialization of the previous cell example, the observation is a `Feature`, its location is described as a `Polygon` and the observation's dimensions and measures are specified by the `properties` list of the `Feature`.

A pure GeoJSON serialization, however, would miss all the advantages of using Linked Data. An ideal solution would take the best from both worlds: the native geospatial expressiveness of GeoJSON and the machine-readability of RDF. A step towards this optimal solution can be achieved by leveraging JSON-LD [10], the JSON-based format to serialize Linked Data, standardized this year by the W3C.

An ordinary JSON document can be turned in a JSON-LD document by adding a `@context` that specifies how to interpret the JSON tags as RDF resources. We decide to include a suitable `@context` in our geospatial output documents. The resulting GeoJSON(LD) is correctly interpreted by GIS systems (cf. the example provided at `http://swa.cefriel.it/geo/semstats2014.html`).

[5] A comprehensive discussion on the RDF representation of geometries is included in the ISA Programme Location Core Vocabulary [8].

[6] Cf. `http://geojson.org/`.

On the other hand, interpreting this enhanced GeoJSON as JSON-LD still has open issues: polygon coordinates are expressed as a list of lists, which is not allowed in the current version of JSON-LD; the `properties` object in GeoJSON also adds an indirection step between the subject and the objects of the relevant predicates. Our attempt is also in line with the proposed GeoJSON-LD vocabulary[7]; however, this proposal is still in its infancy and introduces custom definitions that don't take into account the existing geo-spatial vocabularies.

4 Conclusions

In this paper, we presented the first steps of our geo-statistical exploration of heterogeneous datasets related to the Milano metropolitan area. We discussed some issues related to data processing and analytics results' semantic representation. At `http://swa.cefriel.it/geo/semstats2014.html` our early findings are available, while our future work will be devoted to extend our experiments to additional datasets and further analytics algorithms, as well as to improve the semantic representation of geo-statistical data.

Acknowledgments

This work was supported by the PROACTIVE project (id 40723101) co-funded by Regione Lombardia (POR-FESR 2007-2013).

References

1. Rand, W.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association **66**(336) (1971) 846–850
2. Cohen, J.: Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological bulletin (1968)
3. Janowicz, K., Scheider, S., Pehle, T., Hart, G.: Geospatial Semantics and Linked Spatiotemporal Data. Semantic Web **3**(4) (2012) 321–332
4. Grossner, K., Janowicz, K., Keßler, C.: Place, Period, and Setting for Linked Data Gazetteers. In JR. Mostern, H. Southall, M.B., ed.: Placing Names: Enriching and Integrating Gazetteers. Indiana University Press, Bloomington, IN (2014)
5. Cyganiak, R., Reynolds, D., eds.: The RDF Data Cube Vocabulary. W3C Recommendation (2014)
6. Hobbs, J.R., Pan, F., eds.: Time Ontology in OWL. W3C Working Draft (2006)
7. Open Geospatial Consortium: OGC GeoSPARQL – A Geographic Query Language for RDF Data. Technical report, OGC (2011)
8. Perego, A., Lutz, M., Archer, P., eds.: ISA Programme Location Core Vocabulary. EU ISA Programme Core Vocabularies Working Group (Location Task Force) (2013)
9. Butler, H., Daly, M., Doyle, A., Gillies, S., Schaub, T., Schmidt, C.: The GeoJSON Format Specification. Open standard, http://geojson.org/ (2008)
10. Sporny, M., Kellogg, G., Lanthaler, M., eds.: JSON-LD 1.0 – A JSON-based Serialization for Linked Data. W3C Recommendation (2014)

[7] Cf. `http://geojson.org/vocab`.