A Data Scientist Exploration in the World of Heterogeneous Open Geospatial Data

Gloria Re Calegari, Irene Celino
CEFRIEL – Politecnico di Milano

Abstract
Open Geospatial Data abound today because digital technologies are more and more pervasive. The rich ecosystem of data producers and data consumers presents a multi-faceted picture of the environment, because heterogeneous data sources differ in terms of data complexity, spatio-temporal resolution and curation/maintenance costs.
In this paper, we present the challenges faced by a Data Scientist in exploring and analyzing heterogeneous Open Geospatial Data. This work is aimed at explaining the initial steps of a data exploration process, specifically aimed at discovering similarities and differences conveyed by diverse sources and resulting from their correlation analysis; we also explore the influence of spatial resolution on the dependence strength between heterogeneous urban sources, to pave the way to a meaningful information fusion.

Keywords
Open Geospatial Data, smart city, data diversity, spatial data resolution, correlation analysis

1 Introduction
The ever-increasing availability of Open Geospatial Data calls for smarter processing approaches. A Data Scientist is of course happy in front of such a wealth of knowledge about the space that surrounds us. Still the geospatial nature and the heterogeneity of such data provide interesting challenges for the data exploration.
Heterogeneous geospatial data describes the environment through the lenses of a multitude of information sources. Consequently, the collection, cleansing, curation and maintenance of specialized data sources results in a complex and expensive process, sometimes requiring manual intervention, sometimes involving error-prone automatic processing. The diverse origin of different datasets referring to the same environment represents a further challenge, because the data producers often have disparate objectives and goals that are reflected in the data itself. For a Data Scientists, therefore, the first step is to analyze available datasets in order to identify potential intrinsic dependence and inter-relationships between them. Since their heterogeneous provenance reflects specific and distinct perspectives, we need to investigate whether and how to reconcile the possibly diverging “pictures” of the urban environment those sources convey.
In this paper, we focus on the data exploration of diverse information sources that refer to the same environment; we present the issues and difficulties of such an analysis and we provide a possible best practice, based on the use of
free and open source software tools, to start a process to get a deeper understanding of an environment starting from the available Open Geospatial Data. While our current work focuses on urban environments – because we believe this kind of analysis is of paramount importance in the context of the so-called Smart Cities – the Data Scientist’s approach here introduced can be applied also in different spaces.

The remainder of the paper is as follows: Section 2 introduces the characteristics of Open Geospatial Data sources and the main challenges in their processing; Section 3 details the information sources about Milano used in this research; Section 4 illustrates our experiments of correlation analysis applied at different spatial resolution levels, and Section 5 concludes the paper with some perspectives on our future work.

2 Availability of Open Geospatial Data and Challenges in their Analysis

Open Geospatial Data about cities abound today. The sources of such information are constantly increasing, due to the pervasiveness of information and communication technologies in the so-called Smart City domain. In this section, without the claim of being comprehensive, we would like to give an overview of the possible urban-related datasets that can be found today and of some of the challenges in those datasets management, manipulation and analysis.

With the advent of the open data movement, with its call for transparency and knowledge sharing, a very large number of data sources has been made available on the Web, through a new generation of CMS systems able to give access to this wealth of information, often originating from public bodies and research activities. With special reference to urban information, local authorities have started publishing numerous datasets referring to the city environment: demographics and statistics from municipalities (e.g. distribution of population, family income, crime statistics), listing of local businesses from chambers of commerce, various levels of descriptions about the environment from an urban planning perspective (e.g. land use or land cover, cadastre information), and so on. Those data sources share the geographic aspect even if they are originated by different actors with different purposes.

Furthermore data coming from private businesses that used to be closed data sources are nowadays becoming available, sometimes even as open data. Examples of this kind of datasets are public utilities information, including telecommunication operators: as collateral effect of the mobile networking services, telco companies collect data about the phone activity over time and also over space (due to the positioning of transceiver towers). This kind of data can provide specific perspectives on what happens in our cities and can be a strong indicator of people presence and movement in the urban environment.

Moreover, the popularity of sensor technologies and the so-called Internet of Things (IoT) has led to the availability of massive real-time and streaming information, like climate sensors from environmental agencies (e.g. temperature, pressure, humidity and other ecosystem measures), smart meters and GPS traces from public utilities (e.g. energy consumption or public transportation position).

After the Web 2.0 boom, also user generated information about cities has
become ubiquitous. Crowdsourcing initiatives like OpenStreetMap\(^1\) have popularized the Volunteered Geographic Information paradigm of “citizens as sensors” (Goodchild, 2007) and have collected data about different kinds of points of interest in urban environments (e.g. monuments, restaurants, public services). Location-based social networks like Foursquare, Twitter, Flickr have also produced a stream of “check-ins” and geo-located information that represent the digital counterpart of human activities in the urban space.

Managing, processing and comparing those diverse Open Geospatial Data can be cumbersome for a data analyst. Besides common issues like dealing with data scale and improving data quality, we would like to highlight some challenges that emerge from the specific case of comparison between datasets referring to the same geospatial environment.

One issue arises from the \textit{varying spatial resolution} of information sources: being produced by diverse actors for different reasons, it is quite common that datasets are heterogeneous in terms of the geospatial extent they refer to. For example, population statistics could be at municipality level, land use information from cadastre could be at building level, and smart meters measures could refer to individual points (identified by latitude and longitude coordinates). This means that those sources are not immediately comparable.

Information sources can also refer to \textit{different time-frames}: population census is usually done every \(n\) years, while sensor information is potentially provided in real-time; some other data sources can be made available as historical dumps, while IoT data can have different frequency updates (every 10 minutes vs. once a day). Directly comparing those sources can lead to poor results, because connections and correlations could be traced between datasets that give different pictures about the environment.

Moreover, because of the time-frame, as well as because of the data provider, data sources can differ in terms of \textit{reliability}: apart from pure data quality issues, in managing and processing different datasets it is important to take into account whether and to what extent the source can be trusted; for example, while data from public authorities can be considered “official”, information deriving from user generated efforts or crowdsourcing campaigns can be less dependable.

For a data analyst, the first task is to prepare urban datasets for the subsequent elaboration. Besides common activities like data transformation, cleansing or normalization, we focus our attention on two specific cases that are particularly relevant in a Smart City scenario.

One issue when dealing with diverse data sources is the heterogeneity of their representation in terms of geospatial resolution. The first activity therefore is to \textit{uniform spatial resolution}, finding a trade-off between fine-grained and coarse-grained data. This implies interpolating, aggregating or splitting information in different datasets (Gotway & Oung, 2002) so that they become comparable in subsequent elaborations.

In today’s Smart Cities, information is often characterized by the temporal dimension, as in the case of IoT sensor data and real-time sources. While having both time and space dimensions enriches the city representation conveyed by those sources, time-series frequently consist of big data (Kitching, 2014), thus they require specific processing techniques. In the context of an exploratory analysis of urban data aimed at comparing different datasets, a fundamental activity is therefore \textit{data compression}: pre-processing large-scale...

\(^1\) www.openstreetmap.org/
time-series to get a more manageable compressed representation that can be used to seek relations with other sources. Several approaches to temporal data compression exist (Mitsa, 2010); the main idea is to reduce high-dimensional temporal sequences to compact representations that summarize the original data into a pattern. The simplest example is the computation of a temporal signature or footprint: a 1-dimensional vector in which each element represents the average behaviour of the measured quantity over a specific time interval. Common footprints represent daily patterns, possibly distinguishing between different days of the week; the interval granularity greatly depends on the nature of data (e.g. sensors' frequency of measure).

3 Overview of the Milano Datasets employed in our work

Our case study deals with the exploratory analysis of diverse urban datasets related to the municipality of Milano in Italy. The datasets used in the analysis are illustrated in Table 1: the open data about population demographics from Milano municipality\(^2\); the land use classification elaborated within the CORINE European initiative\(^3\) with CORINE multi-level taxonomy of land cover and made available as open data by Lombardy Region\(^4\); the Points of Interest (POIs) of the city provided by both Milano municipality and OpenStreetMap; two months of mobile call data records provided by the Telecom Italia mobile operator for their “Big Data Challenge”\(^5\) as Open Data.

<table>
<thead>
<tr>
<th>Domain (content)</th>
<th>Data Source</th>
<th>Data Format</th>
<th>Spatial Resolution</th>
<th>Reference Period</th>
<th>Volume (records)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics (population)</td>
<td>Milano Open Data</td>
<td>Shapefile</td>
<td>Census area</td>
<td>2011</td>
<td>10s</td>
</tr>
<tr>
<td>Urban Planning (CORINE land use)</td>
<td>Lombardy Open Data</td>
<td>Shapefile</td>
<td>Building-level resolution</td>
<td>2012</td>
<td>10 Ks</td>
</tr>
<tr>
<td>Mobile Telephony (call records)</td>
<td>Telecom Open Data</td>
<td>Tabular</td>
<td>City grid cells (250m x 250m)</td>
<td>2013 Nov-Dec</td>
<td>100 Ms</td>
</tr>
<tr>
<td>Points of Interest (POIs)</td>
<td>Milano Open Data</td>
<td>Shapefile</td>
<td>Points (lat-long)</td>
<td>2013</td>
<td>1 Ks</td>
</tr>
<tr>
<td>Points of Interest (POIs)</td>
<td>Open Street Map</td>
<td>Shapefile</td>
<td>Points (lat-long)</td>
<td>2014</td>
<td>1 Ks</td>
</tr>
</tbody>
</table>

Table 1: Characteristics of the used datasets.

*Telecom dataset* records every ten minutes the activity occurred in Milano in Nov-Dec 2013 (60 days), and maps it into a grid of squared cells of 250 m. Five different phone activities are stored: incoming and outgoing calls, incoming and outgoing SMSs (text messages) and Internet connection. Because of the size (100Ms data records) and complexity (including both temporal and space dimension) of the dataset, a number of pre-processing operations were required.

\(^2\) [http://dati.comune.milano.it/](http://dati.comune.milano.it/)
\(^4\) [https://www.dati.lombardia.it/](https://www.dati.lombardia.it/)
To reduce Telecom dataset size and to take into account the spatial information, we decided 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, like the one shown in Figure 1.

![Figure 1: Telecom footprint of call-in activity. Concatenation of weekdays and holidays activities.](image)

In particular, we averaged the values recorded in the same cell and at the same time during all the day of the period under analysis, obtaining 144 values for each of the 5 phone activity types (720 values total). We successfully performed all these computationally intensive operations using R free software for statistical computing.

As Figure 1 reveals, we treated weekend and weekday data separately, obtaining a double footprint, one describing weekday activity and the other the holiday behaviour.

Actually, as literature confirms (Reades, Calabrese, Sevtsuk, & Ratti, 2007), weekend and weekdays data usually displays high differences in both shapes and magnitudes.

CORINE 2012 dataset provides data about the land use types of Milano territory and it classifies them by using CORINE multi-level taxonomy. Land use types can range from more general definition of residential, agricultural, industrial, wild areas to a more specific characterization as hospitals, roads, railways, construction sites and so on.

The dataset has a fine grained spatial granularity, actually the land use is defined at a building-level resolution.

After analyzing the distribution of the different land uses in the Milano area, we selected the level of taxonomy more suitable for our case study (by selecting specific categories or by grouping some classes together). We decided to start analyzing the two most general types of land use that could identify and feature a metropolitan area: the residential and the agricultural areas.

As regards demographics information, the dataset provides information about the number of inhabitants for each of the 6079 census areas of Milano municipality. A census area is a partition of the territory defined by the Italian National Institute of Statistic (ISTAT). Each census area is made up of one or

---

6 http://www.r-project.org/
more blocks and each block is composed of contiguous buildings surrounded by mobility infrastructures or natural barriers like streams. The sizes of census areas are different and the median value is 12000 m$^2$.

Lastly, points of interest provided by the two sources (OpenStreetMap through the Overpass API and Milano municipality open data portal), are slightly different in terms of categories: they both have POIs about transports, schools and sport facilities, but in addition OpenStreetMap provides information about shops and amenity places of the city. On the other hand, data coming from Milano municipality are “official”, whereas the OSM dataset, being user generated data, may be less reliable or incomplete.

As Table 1 reveals, besides content heterogeneity, the datasets also differ in terms of spatial resolution and that’s why a pre-processing phase was required in order to make them comparable.

Since population, Telecom and land use datasets have a too fine-grained spatial resolution (census areas, square cells of 250 m and even building level resolution) and POIs datasets consist of punctual information, we needed to find a suitable uniform spatial resolution to continue with the data exploration and analysis. We decided to choose two different spatial resolution levels for our analysis.

Firstly we mapped all datasets into the 88 district subdivision (Figure 2) defined by the local municipality (more specifically by the “PGT - Piano di Governo del Territorio”). Since these districts, named “Nuclei di Identità Locale” (NIL), can be considered as areas connected by mobility infrastructures/services that share common commercial activities, parks, services and meeting places, this mapping offers both a meaningful and a manageable spatial resolution. This district-level granularity is neither too coarse-grained to invalidate the datasets mapped to their level, nor too fine-grained to make a manual evaluation of analysis results impossible.

However, since we are interested in understanding how the adopted spatial resolution affects the data analysis, we also mapped all the datasets into a more fine-grained resolution, a grid of 3538 square cells of 250 m, which is the original resolution of Telecom data records (Figure 3). Even if this uniform subdivision of the space is, beyond any doubt, less meaningful and harder to interpret, we would like to find out if results are in some way similar and comparable between the two resolutions.

![Figure 2: 88 Milano districts – NIL.](image1)

![Figure 3: Grid of 3538 squared cells.](image2)
Since both the cell grid and the NIL resolution levels are more coarse-grained than the original resolutions of CORINE, demographic and Telecom data (building-level, census area and the grid itself respectively), we mapped the data in the new spatial granularities. This operation was quite easy, since we performed these interpolations using QGIS\(^7\), a free and open source Geographic Information System. We intersected the layers describing the target resolution levels (the cells grid and the NIL subdivision) with the original data shapefiles (census area, building-level resolution and grid). Then we interpolated the original data values consequently: for example if a census area resulted to overlay 60% on a cell and 40% on another adjacent cell, its population value was split and attributed 60% to the first cell and 40% to the second one. It stands to reason that this approach can lead to an approximation, but we can consider it appropriate for our case study.

A different kind of mapping was required for POIs data, as they are data points described by latitude and longitude pairs. In this case, we simply computed POIs density in each district and in each cell.

Once all the datasets are represented at the same spatial resolution levels, another best practice is to perform data normalization, which becomes more relevant when we are dealing with districts, as they have different surface areas.

As regard Telecom data, at cell-level we firstly adopted a normalization-in-time methodology on the cells data records, which consists of dividing each value by the average activity of all cells at that given time. When we moved to NIL level we enriched the normalization process by calculating a sort of phone activity density, by dividing the normalized-in-time information by the NIL surfaces in order to take into account the different districts areas, so to have also a normalized-in-space data.

As regards population and POIs, we normalized the data by the area of each district/cell, obtaining the population/POI’s density for each NIL and cell.

As we dealt with percentages, normalization process was not required for CORINE data.

To sum up, all the datasets have been adapted to both resolution levels and the experiments illustrated in Section 4 were repeated for both granularities.

4 Data Exploration Experiments with Milano Datasets: Correlation Analysis

Open Geospatial Data provide information about the city in different forms and with different levels of complexity. The simplest example consists in the variation of some quantity over the urban space: if the city environment is for example divided in \(N\) neighborhoods, a dataset can be a 1-dimensional structure (a vector of size \(N\)), in which each element represents the magnitude of some quantity in each neighborhood (e.g., density of population, average temperature, income indicators).

Whenever several datasets have such a form, or can be summarized by a single variable (interpolating or averaging some quantity), a best practice is to analyze each of those dataset to understand data distribution, to identify outliers and so on. Then, in order to identify possible correspondences between different datasets, we can apply statistical analysis and compute comparison metrics. This kind of correlation analysis is described below.

\(^7\)http://www2.qgis.org/it/site/
If we consider datasets' pairs, i.e. we compare a variable with another variable, we apply bivariate analysis to the two datasets to determine if there is any relationship between them. To measure whether and how those two variables simultaneously change together, different correlation indexes can be computed. The most common measure of dependence between two variables is the Pearson's product-moment correlation coefficient (Pearson, 1895), defined as the covariance of the two variables divided by the product of their standard deviations:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(N - 1)s_x s_y}$$

where $\bar{x}$ and $\bar{y}$ are the sample means and $s_x$ and $s_y$ are the standard deviations of the two variables.

Pearson's $r$ is between -1 and 1 and indicates the strength of the linear relationship between the two variables.

We applied bivariate analysis on all the datasets presented in Section 3 and we measured the relationships between them by computing the Pearson's $r$ index. We performed the analysis at both NIL and cell level, with the aim of discovering if some analogies exist between results obtained from datasets with different spatial resolution.

Since we needed 1-dimensional vectors, we pre-processed the two n-dimensional datasets, CORINE and Telecom, as follows. Regarding CORINE land use, we decided to separately use data about residential and agricultural areas; in this way, we got two vectors, one representing residential use and the other one representing agricultural use of Milano districts and cells in terms of percentages.

On the other hand, for Telecom data we decided to apply Principal Component Analysis (Jolliffe, 2002), a statistical procedure that converts a set of possibly correlated observation into a set of values of linear uncorrelated variables (the principal components), through an orthogonal transformation. PCA is used for dimensionality reduction, as just a few principal components can be able to explain the overall variance of the problem. From our analysis, the first principal component of the telecom dataset explained about the 90% of our dataset’s variance, we considered it a good approximation for representing the whole dataset as a 1-dimensional vector.

Figure 4 lists the results of all the pairwise comparisons performed at NIL level. The plots on the diagonal show data distributions which turn to be left-skewed for all the datasets.

Analysing Pearson's $r$ coefficients, we found that the highest positive correlations exist between phone activity and residential land use (0.83) and between phone activity and POIs (0.84 and 0.85). Those results suggest that a correlation between those datasets can actually exist. The respective plots in the lower-diagonal parts of the scatter-plot-matrix also show that data fits quasi-linear models.
Figure 4: Scatter plot matrix with correlation between datasets at NIL space granularity level: Pearson’s r correlation coefficients are above the diagonal, dataset distribution is on the diagonal and the plots of correlation are below the diagonal.

On the other hand, as can be expected, we observed negative correlation between the agricultural land use and all the other datasets (Figure 4). We interpret this result as a sign that human actions (like those recorded by mobile activity or by presence of POIs) are inversely related to agricultural areas. Then we compared these indexes with the results obtained using the cell spatial resolution level. The Pearson's r (Figure 5) at cell-level are all lower than the ones at NIL-level. The higher Pearson's r values are again between Telecom data and residential areas (0.66) and between Telecom and POIs from OpenStreetMap (0.68). We can notice that the correlation values at cell-level involving the population are the ones closest to the NIL-level results and the correlation between population and residential areas is even the same (0.76). These results suggest us that the choice of the suitable resolution level can have a significant impact on the correlation results.

The 250 m squared cell maybe is so small that there cannot be enough information to characterize it properly or the data available are not enough punctual and precise.
Since we found that both at NIL and at cell level there was a good correlation between the average phone activity and the population density (0.57 and 0.52 respectively), we deepened this analysis by investigating the change in correlation values at different times of the day. In particular we wondered if the correlation could change during the day accordingly to the everyday human behaviour pattern (get up, go to work, come back home in the evening).
For this reason we extracted the incoming and outcoming call activities at six different hours (at 4 am, 8 am, 12 am, 4 pm, 8 pm and 12 pm) and we computed a 1-dimensional vector for the average mobile activity at each specific hour.
We did these experiments using the datasets at both resolution levels and taking separately weekday and weekend activities. Figure 6 plots the results obtained.
At first sight, we can see that weekday and weekend profiles are different: while the former has peaks of correlation in the morning (8 am) and in the evening (after 8 pm), the latter reaches the maximum values at mealtimes (at 12 am and 8 pm).
Figure 6: Plot of Pearson’s r coefficient obtained by comparing the population density with the weekend and weekdays in-out calls at different hours of the day (4am-8am-12am-4pm-8pm-12pm) – blue line – and with the average in-out calls activity – red line. Plots (a) and (b) refer to data at NIL-level whereas plots (c) and (d) at cell-level.

The weekday profile proves our hypothesis that phone activity is related to the everyday human behaviour pattern: the correlation between phone activity and population density is lower than the average during the day when people are at work and become higher when people come back home in the evening. In addition, in the weekend the average correlation is higher than in weekdays. This trend confirms that the amount of phone activity is correlated with the actual presence of people at home.

The shape of the profiles at the two different granularity level are quite the same and, as already seen above, correlation at cell level is overall lower than the one at NIL-level.

5 Conclusion
The proliferation of Open Geospatial Data related to the urban space calls for smarter solutions to process and mine such data. In this paper, we presented
our best practices for data exploration process and in particular we illustrated
our data analytics experiments on a set of heterogeneous datasets related to
the city of Milano. Our purpose is to identify dependence between diverse
sources, to understand if distinct datasets provide similar representations of
the city. We experimentally evaluated the correlations between datasets at
different levels by employing different spatial resolution (coarse-grained district
level vs. fine-grained cell level).
Our empirical results show that correlations between different sources exist,
even if their strength depends on spatial resolution. As a consequence,
investigations on urban data must take this evidence into consideration and
tailor experiments at the most suitable resolution level to obtain meaningful
results.
It is worth noting that, given the available information, we decided to perform
such experiments even if the sources do not refer to the same time-frames
(2011-2014). While this means that our results risk to be less significant, we
believe that the general trends showed by our analysis could be confirmed by
datasets referring exactly to the same period.
This further refinement analysis is on our agenda, as soon as we get urban
datasets homogeneous in terms of their temporal dimension.
Starting from the promising results in terms of correlation between
heterogeneous sources, we aim to extend our investigation toward a predictive
approach, wondering if it would be possible to learn the characteristic of a
dataset from the information given by multiple and diverse data sources. In
particular we are wondering if it would be possible to use one or more “cheap”
datasets – like often are open data – as proxy for more “expensive” data
sources. In other words, would it be possible to (semi)automatically generate or
revise an outdated dataset, which otherwise would require a costly human
work, on the basis of the content of other up-to-date information sources?
To reach this goal we need to move from our initial exploration analysis to
statistical learning experiments (Minasny, 2009) and to supervised machine
learning (Kotsiantis, 2007) techniques, making use of both regression and
classification algorithms from the simplest linear models to the more complex
and powerful Neural Network and Random Forest. Given the intrinsic spatial
nature of our datasets, spatial autocorrelation and dependencies between data
could also exist, thus the Moran’s I index is a very important indicator. In this
case the adoption of Spatial Autoregressive models could provide benefits to
the analyses.

References
✔ Kitching, R. (2014). The real-time city? Big data and smart


