

POLITECNICO DI MILANO

Corso di Laurea Magistrale in Ingegneria Informatica

Dipartimento di Elettronica e Informazione



Land use identification using mobile phone data

Relatore: Prof. Emanuele Della Valle

Correlatore: Irene Celino

Tesi di Laurea di:

Kourosh Sheykhvand 780801

Soheil Behnam Roudsari 780238

Anno Accademico 2013-2014

Acknowledgements

We would like to express our gratitude to our supervisors Emanuele Della Valle, Irene Celino for the useful comments, remarks and engagement through the learning process of this master thesis. Furthermore we would like to thank Hadi Rouhani for introducing us to the topic as well for the support on the way.

Also Special thanks to our family, particularly our parents, thank them for their love, supports, and resolute belief in us. Special thanks goes to our best friends in Italy that support us in these three years.

Abstract

Land use identification is an essential requirement in the process of urban planning. The traditional approaches for identification and classification of type of uses in piece of land (e.g. industrial, residential, commercial, etc.) are too time consuming and too costly for the scale of modern large cities. To address these issues novel methods of land use identification-classification have been introduced which are based on digital data that is potentially indicative of land use types, example of such data includes phone or mobile phone activity data or location based services based data.

The goal of this thesis is to predict the actual land use of an area of 3000KM area in Milan, Italy based on the historical land use data from 2009 (Provided by CORINE) and call details records (CDR data) from 2013 (Provided by Telecom Italia). It is reasonable to expect that the spatial and temporal patterns of mobile calls coming in – going out from/to piece of land vary depending on the types of the land uses available in that land. For instance, in an area under industrial use the number of outgoing calls in evenings is probably much smaller with respect to that of work hours or compared to out-going calls from a domestic area in evenings.

We have divided the whole mentioned area into 10000 cells each of which being 250M*250M. For each cell we extracted a call out profile (2013) consisting of average call out over 60 days. We then measured the similarity of the profile of each cell to the ‘land use profile’ of different land uses estimated from land use data 2009. The land use profiles of specific land use are estimated as the average out call profile of a sub group of cells labeled with that land use in 2009. We have suggested a few different distance measures as well as a few different methods to extract the above mentioned cell’s subgroup. We have compared the predicted land uses of 49 samples cells which are mostly construction sites with the actual 2014 land use of them as a measure of prediction accuracy. Our suggested method that uses Mahanalobis as similarity

measure and two-class k-means clustering or cell subgroup selection predicts the actual land usage with more than 90% accuracy.

Riassunto

L'identificazione d'uso del suolo è un passaggio essenziale nel processo di pianificazione urbana. Gli approcci tradizionali di identificazione e classificazione di tipologie d'uso di terreni (es. industriale, residenziale, commerciale ecc) sono procedimenti lenti e costosi per le dimensioni delle moderne metropoli. Per evitare questi problemi, sono stati introdotti nuovi metodi di classificazione e identificazione dell'uso dei terreni basati su dati digitali che sono potenzialmente indicativi sull'uso dei terreni, esempio di tali dati include dati di attività telefoniche o di cellulari

L'obiettivo di questa tesi è di predire l'attuale identificazione d'uso del suolo di un'area di 3000km² a Milano (Italia) sulla base dei dati d'uso del suolo storici dal 2009 e sulla base dei dettagli delle registrazioni delle chiamate effettuate dal 2013 a questa parte (CDR data). È ragionevole aspettarsi che i modelli spaziali e temporali di chiamate mobili in arrivo e in uscita dal suolo in esame variano a seconda dei tipi di suolo disponibili in quell'area. Ad esempio, in una zona industriale il numero di chiamate in uscita durante la sera è probabilmente molto minore rispetto a quello delle ore lavorative o rispetto alle chiamate uscenti da una zona residenziale la sera.

Abbiamo diviso l'intera area indicate nel 10000 cellule ognuna delle quali dell'essere 250M * 250M. Per ogni cella abbiamo estratto un profilo di chiamata (2013) composto di chiamata media fuori oltre i 60 giorni. Abbiamo poi misurato la somiglianza del profilo di ogni cella alla 'utilizzo profilo di terra' di terreno diversi usi sulla base dei dati di uso del suolo 2009 I profili di uso del suolo di uso specifico territorio sono stimati come il profilo medio fuori di chiamata di un gruppo di sub cellule marcate con quella dell'uso del suolo nel 2009 abbiamo proposto alcune misure di distanza differenti, nonché un paio di metodi diversi per estrarre sottogruppo della cella di cui sopra. Abbiamo confrontato gli usi del suolo previsti su 49 campioni di cellule che sono per lo più cantieri con l'effettivo utilizzo 2014 terra di loro come

una misura di precisione di previsione. Il nostro metodo suggerito che utilizza Mahalanobis come misura di similarità e due di classe k-means clustering di selezione o sottogruppo di cellule prevede l'utilizzo effettivo terreno con oltre il 90% di precisione.

Table of Contents

Acknowledgements.....	i
Abstract.....	ii
Riassunto.....	iv
Table of Contents.....	vi
List of Figures.....	viii
List of Tables.....	x
Chapter 1 Introduction.....	1
1.1 Problem statement.....	1
1.2 Solutions.....	3
1.3 Structure of the Thesis.....	4
Chapter 2 Background.....	6
2.1 Urban Computing.....	6
2.1.1 Urban computing Framework and its goal.....	7
2.1.2 Data sources in urban computing.....	8
2.2 Land use Identification.....	8
2.2.1 Recent Land use identification methods from mobile phones.....	9
2.2.2 Problems in land use identification.....	10
2.3 Data Analysis of large data.....	11
2.3.1 Data Clustering Algorithms.....	12
2.3.2 Exclusive Clustering or Centroid-based clustering (e.g. K-means).....	14

2.3.3	Similarity Measurements	17
2.4	Data visualization with web application technology	20
Chapter 3	Problem Setting and Data Sources Exploration.....	26
3.1	Project proposal.....	26
3.2	Raw Data Sources	27
3.2.1	Telecommunication activity Data source.....	28
3.2.2	CORINE data source of 2009	32
3.3	Preprocessing of data	36
3.4	Naïve solution	38
3.4.1	Problem setting (Challenges).....	42
Chapter 4	Solution Space	44
4.1	General Structure of proposed Solutions	44
4.2	Solution 1: Comparison to Weighted Profile	45
4.3	Solution 2: Comparison to denoised clustered profile	48
4.4	Comparison and discussion.....	51
4.5	The proposed application	52
Chapter 5	Evaluation	57
Chapter 6	Conclusion and Future work.....	68
6.1	Future works.....	69
Bibliography	71
Appendix A	77
Appendix B	83

List of Figures

Figure 1 CORINE Land Cover Methodology (124 is the CLC code for airport).....	2
Figure 2 Telecom Big Data Challenge info graphics.....	2
Figure 3 The land use improvement from 2009 to 2013 in Piazza Gae Aulenti, Milano.....	3
Figure 4 A) Motivation and B) Goal of urban computing.....	6
Figure 5 General framework of urban computing [3].....	7
Figure 6 Clustering example by K-means [23].....	13
Figure 7 Cosine similarity values in different orientation	18
Figure 8 Level of correlation	18
Figure 9 Illustration of Euclidean distance (a) and Mahalanobis distance (b) where the contours represent equidistant points from the center using each distance metric. [28]	19
Figure 10 Responsive web application over mobile phone and PCs, the picture illustrated the proposed application usability on both mobile and pc devices.....	23
Figure 11 Basic example for leafletjs map which is used on mobile phone.....	23
Figure 12 Various data visualization methods by amCharts	24
Figure 13 MVC architecture schema [32]	25
Figure 14 SOFEA architecture.....	25
Figure 15 Milano grid and in-calls footprint for weekdays/weekends in cell id 6060	27
Figure 16 Cell 6060 activities during two weeks with the mean value	31
Figure 17 A) The call-in activity profile for the 4th to 17th of November B) The call-in activity profile between 18th to 31th December.....	32
Figure 18 Distribution of land uses.....	34
Figure 19 Distribution of land uses by taxonomy.....	35
Figure 20 Prevalent land uses	36
Figure 21 Usage of mobile phone during weekday and weekends for two different zones ...	37

Figure 22 Average call in activity data for cell 6060 for the whole two months	38
Figure 23 K-means++ clustering result over call-out activity by k=24 (which is all the land use type from CORINE)	40
Figure 24 K-means++ clustering over call-out data by k = 5 (the number of taxonomy)	41
Figure 25 K-means++ clustering output over 2xx and 1xx	42
Figure 26 General structures of proposed approaches	44
Figure 27 Conceptual model of solution 1	46
Figure 28 A) Calculation each land use signal in each cells based on assumption-2. It calculate these signal by weight of each land use which are in CORINE data B) profiles of all land uses, these profile are averages of same land uses in all cells	46
Figure 29 Conceptual model of solution 2	49
Figure 30 The difference percentage between outlier and inlier in each class of land use	50
Figure 31 Snipped code for Mahalanobis distance calculation in python. In this snipped code some of the basic calculation are removed.	51
Figure 32 Livinglanduse application architecture and data flow	53
Figure 33 Conceptual model for initializing phase	54
Figure 34 Translate retrieved data to json style. It retrieves weekday and weekend value of a cell in one day (10 min by 10 min)	55
Figure 35 Output of json file. It shows just weekday and week end value of first 30 min of the day for cell 6060	55
Figure 36 Load data which is provided from previous step to amchart	55
Figure 37 Plot the loaded data for a cell	56
Figure 38 Visualized all information about a cell. Such as cell id, CORINE information, prediction result, signal of a cell, profile signal of land uses	56
Figure 39 Google Maps Street View into a time machine	58
Figure 40 Evaluation mechanism to validate proposed approach accuracy	60
Figure 41 Analyzed land use for all the Milano	62
Figure 42 The improvement of the result base on the approaches	67

List of Tables

Table 1 Telecommunication activity Data set	28
Table 2 Milano Land use percentage data set information of 2009 from CORINE	32
Table 3 CLC code labels for each land use in 2009	34
Table 4 Matrix of CDR data for the 4th of November (weekday).....	37
Table 5 Main data structure for computation.....	38
Table 6 Output of solution 1	47
Table 7 Output of phase 1 (normalized data).....	49
Table 8 Output of solution 2	51
Table 9 Differences between solution 1 and 2.....	52
Table 10 Milano Ground Truth details in 2013	59
Table 11 Summary result of all the methods with respect to the base lines	61
Table 12 Random, Weighted random and Top 1 in 1 km result.....	63
Table 13 Top 1 in 5 cluster and 16 cluster and 24 cluster	64
Table 14 The result of the CWP	65
Table 15 The result of CDCP 1 which displays the first lowest distance and the CDCP 2 which displays the second lowest distance	66
Table 16 Comparison of solutions	69

Chapter 1

Introduction

In order to increase the welfare of citizens, urban planners are focusing on the control and design of the urban environment. Two main challenging process of the urban planning are:

- Characterization of land use
- Identification of landmarks

Urban planners require large amounts of data on urban land use and urban landmarks in order to make public policy decisions. These data are typically collected by direct observation or questionnaires that make a big effort to realize how citizens interact with the urban environment. This approach has some limitations such as the flexibility of citizens to provide such information or the cost of the handling questionnaires. Another method to realize the mentioned challenges is using GIS (Geographic Information Systems) which provide satellite imagery and street views. The problem raises here is lack of up to date data from GIS system, since the images are not captured frequently [1]. In order to overcome these issues, we are going to provide a cost effective approach to identify land uses and landmarks using mobile phone data which are provided by telecom Italia. In this chapter we discuss about the problem statements shortly and then we fully focused on it in the related chapter, then we discuss about the existed and proposed solutions, and finally we list the structure of this thesis work.

1.1 Problem statement

Generally as we mentioned, Urban computing is concerned with the (re)design of urban environment and exploiting land use information. To support this issue we are trying to propose **a method to collect and classify land use information** consists of census activities (e.g. the EU CORINE Land Use cover program [2]) **which are expensive to update**. Figure 1 displays

the airport installations: runways, buildings and associated land in Great Britain. Human activities leave several traces in the digital world that can be exploited to elicit and predict the land use in urban environments [1].

One of the new data sources relevant for the study of urban environments are cell phone records, as they contain a wide range of human dynamics information (ranging from mobility, to social context and social networks) that can be used to characterize individuals or geographical areas.

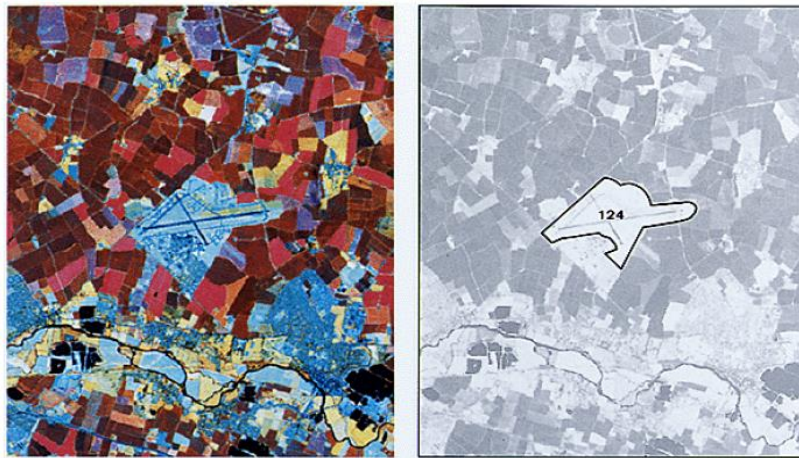


Figure 1 CORINE Land Cover Methodology (124 is the CLC code for airport)

The data provided within the dataset of the Telecom Big Data Challenge are geo-referenced, anonymous and for the territories of Milan and of the Autonomous Province of Trento. The dataset contains millions of records of data covering the period from November to December 2013. The dataset contains data pertaining to telecommunications, energy, public and private transport, social networks and events In Figure 2. You can see the size of the dataset available to the participants of the competition.

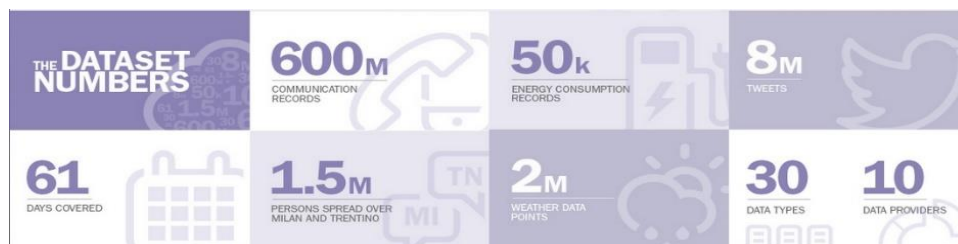


Figure 2 Telecom Big Data Challenge info graphics

Therefore, here by having these data sets we came up with an idea to detect the living land use automatically. For instance, Figure 3 depicted the change of land use from 2009 to 2013 in Porta Nuova-Garibaldi area.

But what about less evident changes in land use? Can we automatically detect the "living" land use from the analysis of streaming activity data? It is the main challenge we are going to deal with it in this thesis work. In the next section we are going to hint the solutions that we have in mind.



Figure 3 The land use improvement from 2009 to 2013 in Piazza Gae Aulenti, Milano

1.2 Solutions

There are some approaches to solve the problem described in previous section. Unsupervised learning or clustering is one of them which is used in recent works. The idea behind of these solutions is to find signature of each land use by clustering the signal of mobile usage that are actually signals of CDR. pure clustering approach is a fast and simple solution, but it could not be effective for our case study as we discussed in the problem setting chapter. So we considered these solutions as a naïve approach. Tuning the clustering method for specific situation is one of the solutions that can be come up at first sight. But in our proposed solutions we benefits from CORINE data that is information about land uses of Milan area in 2009. This information can be helpful if we assume that most of the cells and land uses have not been changed from 2009 to 2013. The reason to figure out this assumption is the nature and characteristic of Milano. Generally our solutions are based on two phases:

- Profile detection of each land use
- Similarity measure of each cell with all land use profiles

Our two proposed solutions have differences in these two phases. First solution benefits from weighted profiling for profile detection phase and adjusted cosine similarity for similarity measuring, but the second solution use denoising by clustering for profile detecting and Mahalanobis distance for similarity measuring. Chapter 5 explained these two solutions in detail.

Case study

The study we present has been done using CDR data collected from Milano for a period of the two last month from November 1st 2013 to January 1st 2014 for big data challenge telecom Italia¹. The data consist of 10000 cells (100*100)' CDR data. Moreover, for our purposed solution we used the data of the CORINE land use 2009, which is extracted from urban community and included the land use data for each cell.

1.3 Structure of the Thesis

This thesis is organized as follows:

- **Chapter 2** introduces the some background concepts in urban computing and its challenges, then discuss about the state of the art of the land use identification which is a very innovative idea. Then in this chapter we discuss about the data analysis concepts and its tools (such as clustering algorithms and similarity measurements) and then we introduce data visualization methods and tools which mostly used in our work.
- **Chapter 3** describes the problem setting and data sources; in the first section of this chapter we discuss about the idea of the project and then we move to the data sources which are used during the research. We completely discuss about the data exploration and feature selection in this chapter, then we go across the naïve solutions and finally we finish the chapter with introducing the main challenge.
- **Chapter 4** describes the main solutions over the problem which introduced in pervious chapter, the result of each solutions is represented and then comparison takes place. In

¹ <http://www.telecomitalia.com/tit/it/bigdatachallenge.html>

this chapter the proposed web application and its architecture through some code snippets are discussed.

- **Chapter 5** introduces the ground truth of the thesis work and the activities which were done to collect the information of the selected case study, and then discusses about the experiment evaluations to find out the correctness of each method and summarizes them.
- **Chapter 6** includes the conclusion of the work which is done during this thesis work, more over in this chapter we discuss about the future works and limitations which have effects on the result of the experiment.
- **Appendix A** lists the CLC codes for the readers who wish to know more details about the CORINE land use.
- **Appendix B** includes the satellite pictures of the case study.

Chapter 2

Background

This chapter is divided to the following sections: Urban computing concepts and the related work which have done (Section 2.1); Description of land use identification and its problems (Section 2.2); brief description of Data Analysis tools (Section 2.3); Data visualization tools (Section 2.4)

2.1 Urban Computing

Urban computing is a process of acquisition, integration, and analysis of big and heterogeneous data generated by a diversity of sources in urban spaces, such as sensors, devices, vehicles, buildings, and human, to tackle the major issues that cities face, e.g. air pollution, increased energy consumption and traffic congestion. Urban computing connects unobtrusive and ubiquitous sensing technologies, advanced data management and analytics models, and novel visualization methods, to create win-win-win solutions that improve urban environment, human life quality, and city operation systems. Urban computing also helps us understand the nature of urban phenomena and even **predict the future of cities**. Famous research companies such as Microsoft and Google have done different projects in this domain so far.

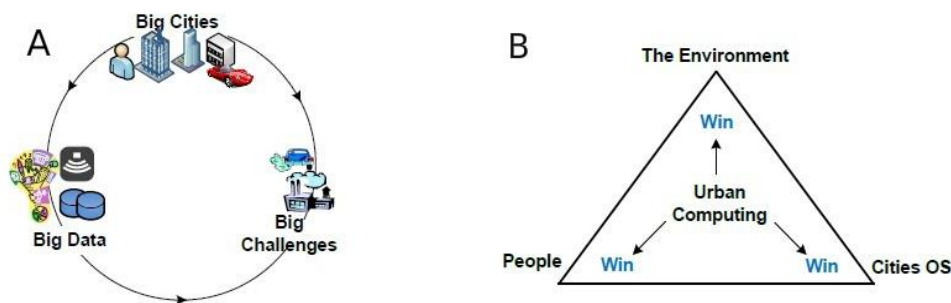


Figure 4 A) Motivation and B) Goal of urban computing.

Motivation

Motivation and goal of urban computing is depicted in Figure 4 A and B [3]. Urban computing is an interdisciplinary field fusing the computing science with traditional fields, like transportation, civil engineering, economy, ecology, and sociology, in the context of urban spaces.

2.1.1 Urban computing Framework and its goal

Figure 5 depicts a general framework of urban computing which is comprised of four layers:

- **Urban sensing:** In the urban sensing step, we constantly probe people's mobility, e.g., routing behavior in a city's road network, using GPS sensors or their mobile phone signals, or social media info such as twitter and Facebook.
- **Urban data management:** the human mobility and social media data are well organized by some indexing structure that simultaneously incorporates spatio-temporal information and texts, for supporting efficient data analytics.
- **Data analytics:** we are able to identify the locations where people's mobility significantly differs from its origin patterns.
- **Service providing:** related to the system and the aim of each project the data is sent to the related drivers.

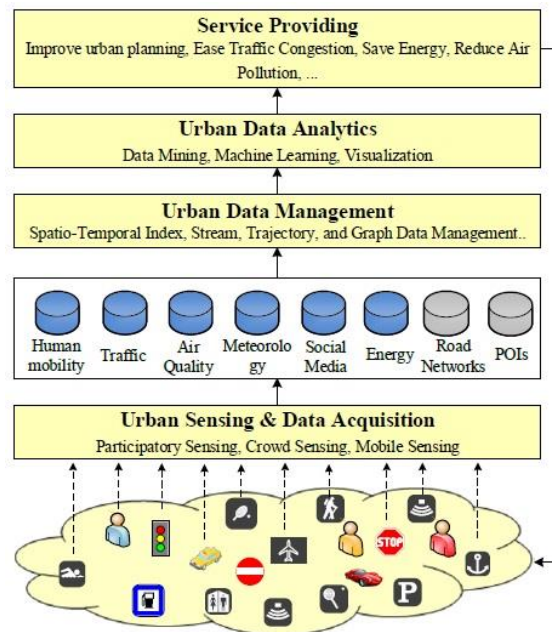


Figure 5 General framework of urban computing [3]

Compared with other systems, e.g., web search engines which are based on a single (modal)-data-single-task framework (i.e., information retrieval from web pages), urban computing holds a multi (modal)-data-multi-task framework. The tasks of urban computing include improving urban planning, easing traffic congestion, saving energy consumption, and reducing air pollution, etc.

2.1.2 Data sources in urban computing

The frequently data sources in urban computing are Geographical Data, Traffic Data, Mobile Phone Signals (Call Detail Record), Commuting Data, Environmental Monitoring Data, Social Network Data, Economy, Energy, Health Care, etc.

Land use data describes the function of a region, such as residential areas, suburban, and forests, originally planned by urban planners and roughly measured by satellite images in practice. For example, United States Geological Survey categorizes each 30m x 30m square of the U.S. into 21 types of ground cover [US Ground Cover], such as grass land, water, and commercial. In many developing countries where cities change over time with many new infrastructures built and old buildings removed, the reality of a city may be different from its original planning. As the satellite image cannot differentiate between fine-grained land use categories, such as educational, commercial, and residential areas, obtaining the current land use data of a big city is not easy [4]. Identification of these land uses could be a challenge that in the following we describe the worked done in this area.

2.2 Land use Identification

The classification of urban is essential for urban Computing. Urban land use, defined as the recognized human use of land in a city, can be differentiated either by its physical characteristics (such as reflectivity and texture) or social functions (i.e., residential areas are for living whereas industrial areas are for working). Among urban land use classification methods, remote sensing techniques are recognized as a vital method because of their ability to capture the physical characteristics of land use. Conventional land-use remote sensing methods classify land use based on spectral and textual characteristics [5] [6] [7] [8].

The daily activities of residents in various regions can be easily captured and used to **indicate the social function of the land use type**. In other words, within different land use areas, people may demonstrate different routine activities (for example, in residential areas, people usually leave home work in the morning and return in the evening, whereas in business areas the opposite pattern can be found). This may allow us to derive the activities of residents, and then the social functions of different land use types, from mobile phone data. As a result, mobile phone data may provide a new insight into traditional urban land use from the perspective of social function. While traditional approaches are based on questionnaires, which implies cost and time limitations, an automated method could help a lot in this stage.

2.2.1 Recent Land use identification methods from mobile phones

The retrieval of land use from mobile phone data can be divided into two stages. The first is to retrieve the residents' activities based on mobile phone data. The second is to infer land use from residents' activities. Regarding the first stage, recent research can be grouped into two categories. The first aims to reveal individual mobility patterns using cell detail record data, which consist of the different base transceiver station (BTS) locations from which users have made calls [9] [10] [11]. The second is based on the aggregation of the total calling time (or numbers) at each BTS in a certain temporal interval.

The spatiotemporal variation regarding BTS has been extensively studied to retrieve various residents' activities. Recent approaches include the description of urban landscapes (i.e., the space-time structure of residents' activities in a city), population estimates, the identification of specific social groups, and the detection of social events. The inference of land use types in this context is dependent on their social functions which can be derived from the residents' activities (namely, the overall characteristics of human communication in the urban area). This contains two main aspects:

The relative weekly **calling pattern** and the total **calling volume**. The pattern is defined as the share of hourly calling volume in a certain period. The calling volume of a BTS is defined as the total time or number of calls managed by that BTS in its area of coverage over a given period of time. Unlike the static residential population density, the volume is the overall characteristic of how many people actually use mobile phones, indicating the activeness of

their communicational interactions. To identify and extract recurring patterns of mobile phone usage and relate them to some land use types, [12] proposed the Eigen-decomposition method, a process similar to factoring but suitable for complex datasets. [13] Used an Eigen-decomposition analysis to reveal the relationship between mobile data and the residential and business areas. [14] Used a new tessellation technique to differentiate parks from residential areas by detecting changes in human density retrieved from mobile phone data. Although these studies have addressed the relationship between land use and mobile phone data, they have only focused on the identification of specific land use classification.

Computing the aggregated calling patterns of the antennas of the network and, after that, finds the optimum cluster distribution to automatically identify how the citizens use the different geographic regions within a city is one of the proposed method which is not only focused on the identification of specific land use classification [15]. This approach is comparing the planned use of a city with the actual use that citizens give to the different areas of the city without the need of on-site data collection. Among others [16] used aggregated cell-phone data to analyze urban planning in Milan, [17] identified behavioral patterns from the information captured by phones carrying logging software.

After making the all signature by using some clustering methods such as k-means clusters have been made and then validity of clusters have done. Then by checking the shape and behavior of the profiles and different analysis methods they try to categorize the city land use, the main different of recent works and us are the type of land use identification, what are going to do is to predict the land use based on the existed data and our prediction is going to deal with different categories of the land use types. In the next chapter we discuss shortly about other works solutions and you figure out that their methods are not working with this amount of data and idea.

2.2.2 Problems in land use identification

In land use identification there might be some issues such as the type of the time series and their model and also uncertainties. In the following we discuss about these two issues:

- 1) The time series model that represents land use type at the BTS level should be improved to enhance urban land use classification. The model should be more sophisticated and incorporate

more characteristics (say, the differences between weekdays and between weekends, new indices derived from aggregated mobile phone data) in order to better differentiate between different land use types. This is because the land use is not only dynamically changing, but is often also heterogeneous in some areas. Thus, either the pattern or the volume may not fully interpret the social functions of different land use types.

2) Because mobile phone data is a new data source in terms of urban planning, it is important to evaluate the uncertainties and influential factors behind land use classification. These include three aspects. one is related to the model, and especially the different characteristics in the time series. The second concerns the data, particularly the BTS density. The third considers the ground truth, and specifically the heterogeneity of land use.

2.3 Data Analysis of large data

As we discussed in the last parts about the type of the data in urban computing and different data sources, the importance of data analysis is vivid. In this section we discuss about data analysis concept.

Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data. According to [18] various analytic procedures provide a way of drawing inductive inferences from data and distinguishing the signal (the phenomenon of interest) from the noise (statistical fluctuations) present in the data [19].

While data analysis in qualitative research can include statistical procedures, many times analysis becomes an ongoing **iterative process** where data is continuously collected and analyzed almost simultaneously. Indeed, researchers generally analyze for patterns in observations through the entire data collection phase [20]. The form of the analysis is determined by the specific qualitative approach taken (field study, ethnography content analysis, oral history, biography, unobtrusive research) and the form of the data (field notes, documents, audiotape, videotape).

An essential component of ensuring data integrity is the accurate and appropriate analysis of research findings. Improper statistical analyses distort scientific findings, mislead casual

readers [21], and may negatively influence the public perception of research. Integrity issues are just as relevant to analysis of non-statistical data as well.

The key point of any data analysis is Data Exploration that is an informative search used by data consumers to form true analysis from the information gathered. Often, data is gathered in a non-rigid or controlled manner in large bulks. For true analysis, this unorganized bulk of data needs to be narrowed down. This is where data exploration is used to analyze the data and information from the data to form further analysis. Data often converges in a central warehouse called a data warehouse. This data can come from various sources using various formats. Relevant data is needed for tasks such as statistical reporting, trend spotting and pattern spotting. Data exploration is the process of gathering such relevant data. When you have done data exploration or you are in the middle of that you need to do some feature selections. In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. All these activities done in order to clean the data, in the next chapter you see how we clean the data sets and select features over them. In the next section we discuss about different data clustering which is one of the main activity in this kind of the project.

2.3.1 Data Clustering Algorithms

Definition

A loose definition of clustering could be the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are similar to each other and are dissimilar to the objects belonging to other clusters (Figure 6). Cluster analysis is also used to form descriptive statistics to ascertain whether or not the data consists of a set distinct subgroups, each group representing objects with substantially different properties. The latter goal requires an assessment of the degree of difference between the objects assigned to the respective clusters [22].

Central to clustering is to decide what constitutes a good clustering. This can only come from subject matter considerations and there is no absolute “best” criterion which would be independent of the final aim of the clustering.

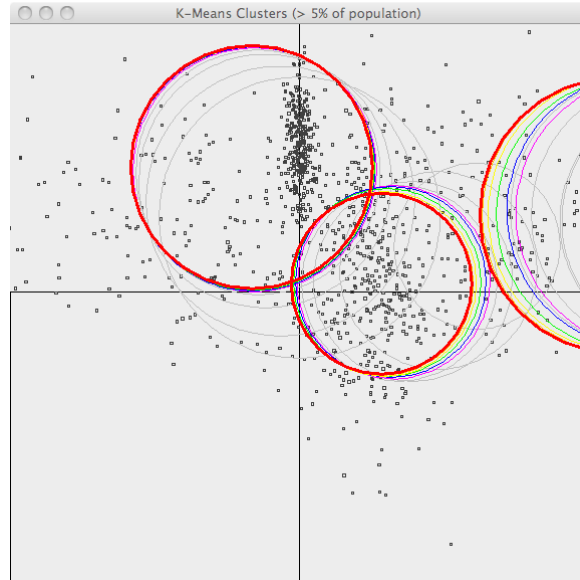


Figure 6 Clustering example by K-means [23]

For example, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection).

Two important components of cluster analysis are clustering algorithm and the similarity (distance) measure between two data samples. Which in the following we described them.

Clustering algorithm categories

Clustering algorithms can be categorized in four main below categories: [24]

1) Exclusive Clustering (e.g. K-means): In exclusive clustering data are grouped in an exclusive way, so that a certain datum belongs to only one definite cluster. K-means clustering is one example of the exclusive clustering algorithms.

2) Overlapping Clustering (e.g. Fuzzy C-means): The overlapping clustering uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership.

3) Hierarchical Clustering: Hierarchical clustering algorithm has two versions:

- Agglomerative clustering: It is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted. Basically, this is a bottom-up version
- Divisive clustering: It starts from one cluster containing all data items. At each step, clusters are successively split into smaller clusters according to some dissimilarity. Basically this is a top-down version.

4) Probabilistic Clustering (e.g. Mixture of Gaussians): It uses a completely probabilistic approach.

The main clustering algorithm that we used in this thesis is Exclusive Clustering algorithm, which we are going to explain more about that in the following section.

2.3.2 Exclusive Clustering or Centroid-based clustering (e.g. K-means)

K-means [25] is one of the simplest unsupervised learning algorithms that solves the clustering problem. The objective is to classify a given data set $S = \{s_1, s_2, \dots, s_N\}$ into a certain number of clusters (assume initial clusters) fixed a priori.

The idea is to define initial centroids, one for each cluster $c_i (i=1, \dots, k)$.

The procedure is:

- 1) Initial clusters: $\ell^0 = \{c_1^0, c_2^0, \dots, c_k^0\}$; the initial centroids should be placed as far as possible from each other.
- 2) Calculate the centroids of the clusters: $u_j^i = \frac{1}{|c_j^i|} \sum_{x \in c_j^i} x$ where $j = 1, \dots, k$ and i denotes the i^{th} iteration.
- 3) Take each point belonging to a given data set and associate it to the nearest centroid:
 - a. $c_j^{i+1} = \{x | d(x, u_j^i) \leq d(x, u_{j'}^i) \quad 1 \leq j' \leq k\}$
 - b. $\ell^{i+1} = \{c_j^{i+1} | 1 \leq j \leq k\}$
- 4) Repeat steps two and three until no more changes can be made to the clusters, that is, $\ell^{i+1} = \ell^i$. In other words centroids do not move any more.

Each iteration takes $O(Nk)$ time, but we don't know in how many iterations it will take to converge. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function:

$$J = \sum_{j=1}^k \sum_{x \in C_j} \|x - u_j\|^2$$

Although it can be proved that the k-means algorithm will always terminate, the algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. It might get stuck in a local minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centers. To get out of local minimum, the k-means algorithm can be run multiple times from different initial clustering or simulated annealing technique could be used.

K-means algorithm is provided by different tools such as python and R. We have used both tools to do our evaluation, in the following we discuss about this algorithm in Python.

K-means in python is defined as the following: [26]

```
class sklearn.cluster.KMeans (n_clusters=8, init='k-means++', n_init=10, max_iter=300,
                              tol=0.0001, precompute_distances=True, verbose=0,
                              random_state=None, copy_x=True, n_jobs=1)
```

The arguments are the following:

- **n_clusters** : int, optional, default: 8; The number of clusters to form as well as the number of centroids to generate.
- **max_iter** : int, default: 300; Maximum number of iterations of the k-means algorithm for a single run.
- **n_init** : int, default: 10; Number of time the k-means algorithm will be run with different centroid seeds. The final results will be the best output of n_init consecutive runs in terms of inertia.
- **init** : {'k-means++', 'random' or an ndarray}; Method for initialization, defaults to 'k-means++';
 - k-means++ : selects initial cluster centers for k-mean clustering in a smart way to speed up convergence.

- **random**: choose k observations (rows) at random from data for the initial centroids.

If an ndarray is passed, it should be of shape (n_clusters, n_features) and gives the initial centers.

- **precompute_distances** : boolean, default: True; Precompute distances (faster but takes more memory).
- **tol** : float, default: 1e-4; Relative tolerance with regards to inertia to declare convergence
- **n_jobs** : int, default: 1; The number of jobs to use for the computation. This works by breaking down the pairwise matrix into n_jobs even slices and computing them in parallel. If -1 all CPUs are used. If 1 is given, no parallel computing code is used at all, which is useful for debugging. For n jobs below -1, (n cpus + 1 + n jobs) are used. Thus for n jobs = -2, all CPUs but one are used.
- **random_state** : integer or numpy.RandomState, optional; The generator used to initialize the centers. If an integer is given, it fixes the seed. Defaults to the global numpy random number generator.

Kmeans VS K-means++

The k-means problem is to find cluster centers that minimize the intra-class variance, i.e. the sum of squared distances from each data point being clustered to its cluster center (the center that is closest to it). Although finding an exact solution to the k-means problem for arbitrary input is NP-hard, the standard approach to finding an approximate solution (often called Lloyd's algorithm or the k-means algorithm) is used widely and frequently finds reasonable solutions quickly.

However, the k-means algorithm has at least two major theoretic shortcomings:

- **First**, it has been shown that the worst case running time of the algorithm is super polynomial in the input size.
- **Second**, the approximation found can be arbitrarily bad with respect to the objective function compared to the optimal clustering.

The k-means++ algorithm addresses the second of these obstacles by specifying a procedure to initialize the cluster centers before proceeding with the standard k-means optimization iterations. With the K-means++ initialization, the algorithm is guaranteed to find a solution that is $O(\log k)$ competitive to the optimal k-means solution. The k-means++ approach has been applied since its initial proposal. In a review by [27] which includes many types of clustering algorithms, the method is said to successfully overcome some of the problems associated with other ways of defining initial cluster-centers for k-means clustering.

2.3.3 Similarity Measurements

Similarity is a quantity or measure of how much two objects are similar. This quantity is usually having range between ranges either -1 to +1 or normalized into 0 to 1. Similarity in data mining context usually described as a distance with dimensions representing features of the each objects. A large distance indicates a low degree of similarity and a small distance indicating a high degree of similarity. In the following we are going to explain measures which are used in this thesis:

1. Cosine Similarity measure
2. Pearson Correlation measure
3. Euclidian distance measure
4. Mahalanobis distance measure

1) Cosine similarity measure

Cosine similarity measure or CSM is a measure of similarity of two vectors of a DOT product that calculate the cosine of the angel between them. This metric is a measurement of orientation and not magnitude. So two vectors with a same orientation have cosine similarity of 1 which means highest degree of similarity and two vector diametrically opposed have a similarity of -1 which is lowest degree of similarity independent of magnitude (Figure 7). Usually cosine similarity used in positive space where the outcome in in range of 0 to 1 which are lowest similarity and highest similarity respectively. The similarity, $\cos(\theta)$, of two vectors of attributes, A and B, is represented as follow:

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

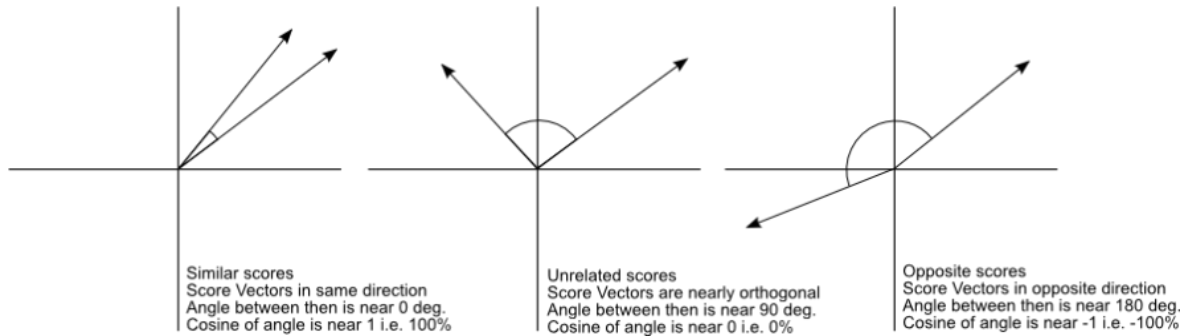


Figure 7 Cosine similarity values in different orientation²

2) Pearson Correlation measure

Pearson Correlation or Pearson Product Moment Correlation PPMC is a common measure of correlation between two variables that shows the linear relationships between two sets of data by giving value between +1 and -1 inclusive, where 1 is total positive correlation, 0 is no correlation and -1 is total negative correlation (Figure 8). This correlation is defined as covariance of the two variable divided by the product of their deviation. The formula for Pearson's correlation, r , of X and Y is:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{S_X}\right) \left(\frac{Y_i - \bar{Y}}{S_Y}\right)$$

Where $\left(\frac{X_i - \bar{X}}{S_X}\right)$ is standard score, \bar{X} is mean of X , and S_X is standard deviation of X

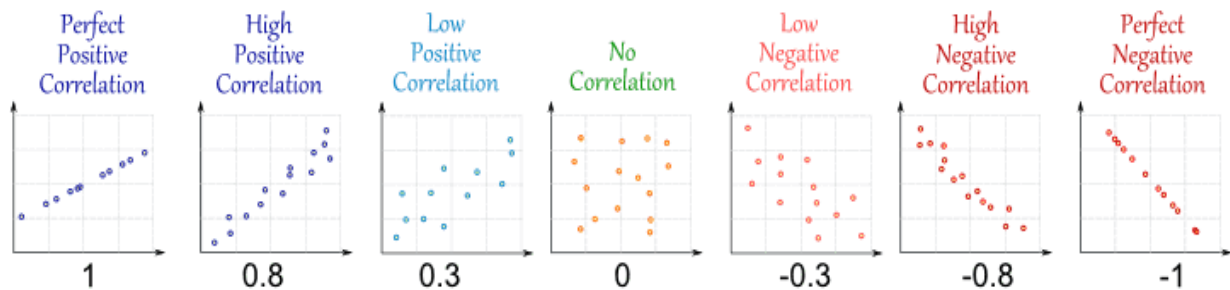


Figure 8 Level of correlation³

² <http://www.tuicool.com/articles/IR7jym>

³ <http://www.cqeacademy.com/>

3) Euclidian distance measure:

Euclidean distance is the most common use of distance. The Euclidian distance is the length of the straight line between two points or in other word it is the distance between two points in any dimension of space. The Euclidean can be computed as follow:

$$d(p, q) = d(q, p) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

4) Mahalanobis distance measure:

Mahalanobis distance is a distance between a point and a distribution. In other word Mahalanobis distance is the Euclidian distance with takes into account the covariance among the variable in calculating distance. So it will be equal to Euclidian distance when the covariance matrix is unit matrix. [28]

The Mahalanobis distance between an observation $x = (x_1, x_2, \dots, x_N)^T$ from a group observation with mean $\bar{Y} = (\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_N)^T$ and covariance of S is defined as:

$$D_m = \sqrt{(x - \bar{Y})^T \frac{1}{S} (x - \bar{Y})}$$

Unlike most other distance measure, this method is not depending on scale of variable. So in this method scale and correlation in Euclidean distance are no longer as issue. As Figure 9 shown Mahalanobis distance takes into account the distribution of the data points, whereas the Euclidean distance would treat the data as though it has a spherical distribution.

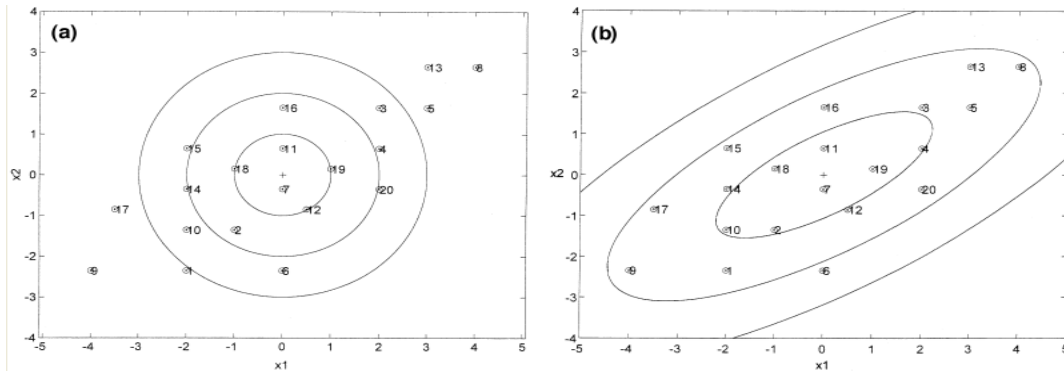


Figure 9 Illustration of Euclidean distance (a) and Mahalanobis distance (b) where the contours represent equidistant points from the center using each distance metric. [29]

2.4 Data visualization with web application technology

Analysing in the domain of urban computing and planning raises the need of visualization methods. If the application could provide a good presentation of the zones then the reviewer could interact with the application easier. Generally the applications that provided in this domain need to be observed and viewed by the users who are typically public administrations.

Data visualization is the presentation of data in a pictorial or graphical format [30]. It is an essential tool to understanding information. Users for understanding information and making value for them in quickest and easiest way are depended on visualizing representation techniques such as charts and maps. These techniques enable them to find relevance among the millions of variable, compare them, uncovering insight hidden in data, find trend and behavior, predict the future, and make hypothesis to others and so many other things. All of these are because of human brain or visual processing systems; it is faster for users to analyze data over graphics, charts or maps in compare of data tables or reports. Interactive data visualization is one step further; user can use computers or even mobile devices to drill down in charts or maps to find more details and even change immediately what data they need to see, add or remove some other dimensions or change visualizing technique immediately to see result in other view. One of the newer widely use interactive visualization technique is **Geo-visualization**. Geo-visualization, short for Geographical visualization, is a tool or technique of interactive visualization to support geospatial data analysis. It visualize geospatial information in such a way that user can explore on them in interactive form. Nowadays there are so many visualizing libraries or smart applications in commercial or open-source form to make interactive visual information just by give pure data to them in different area in easiest or simplest form. In the next section we mostly discuss about technologies in web which are useful to visualize data.

Web Technologies

In the last few years, web technology has a rapid growth and so many changes. We went form table or frames layout to responsive layout. From flash to HTML5, HTML4 to XHTML, SQL to NoSQL and big data, static map to responsive one and so on. In this part we will describe

about some state of the art tools, pattern and technology in web applications. In this section we discuss about:

- Web Application Frameworks
- Web Technologies to visualize data
- Web Application architecture and design pattern

Web Application Frameworks

Web application frameworks or WAF is a foundation, which is specially designed to help web developer to build their application. These types of framework provide core or common functionality of most web applications such as data persistence, template systems, and session management. In short we can say with frameworks developer can save significant amount of time to build a web application. In the following we are discussing two main important web application frameworks which we used during the thesis:

1. CodeIgniter
2. Twitter Bootstrap

1) CodeIgniter:

It is an Application Development Framework - a toolkit - for people who build web sites using PHP. Its goal is to enable you to develop projects much faster than you could if you were writing code from scratch, by providing a rich set of libraries for commonly needed tasks, as well as a simple interface and logical structure to access these libraries. CodeIgniter lets you creatively focus on your project by minimizing the amount of code needed for a given task [31].

From a technical and architectural standpoint, CodeIgniter was created with the following objectives:

- **Dynamic Instantiation:** In CodeIgniter, components are loaded and routines executed only when requested, rather than globally. No assumptions are made by the system regarding what may be needed beyond the minimal core resources, so the system is very light-weight by default. The events, as triggered by the HTTP request, and the controllers and views you design will determine what is invoked.

- **Loose Coupling:** Coupling is the degree to which components of a system rely on each other. The less components depend on each other the more reusable and flexible the system becomes. Our goal was a very loosely coupled system.
- **Component Singularity:** Singularity is the degree to which components have a narrowly focused purpose. In CodeIgniter, each class and its functions are highly autonomous in order to allow maximum usefulness.

2) Twitter Bootstrap:

Mobile clients have gained a substantial amount in web consumer ship, and cannot be ignored anymore. Therefore web application must be GUI responsive to be scalable for any smaller screen size or resolutions. Wise companies have even started designing user interfaces for the smallest resolution before even thinking of a desktop resolution. This makes scaling more easily, since the other way round is harder and tends to force painful workarounds.

Bootstrap [32] is the one of most popular open-source HTML, CSS, and JavaScript frameworks (front-end-framework) for developing responsive, mobile first projects on the web. It has been developed by Twitter and it is a combination of HTML, CSS and java script code (Figure 10). In order to use bootstrap you need only to add following code at the beginning of your HTML pages:

```
<link rel="stylesheet" href="/twitter-bootstrap/twitter-bootstrapv2/docs/assets/css/bootstrap.css">
```

There are many benefits of using Twitter's Bootstrap and one of the benefits is:

- *The User Interface CSS Styles:* Typography elements like quotes, headings, lists are actual included styles, flexible styles for tables and also very attractive form designs and buttons.
- *Extend Bootstrap with LESS:* One of the additional benefits to use Bootstrap is because it's enhanced through LESS preprocessor, so more efficient and faster coding in CSS.
- *Integrated JavaScript Plugins:* While there are many plugins out there to choose from, styling them and making them look integrated can take a long time. With Bootstrap it comes with everything from modal windows, flexible alert messages, carousels, tooltips and other fun JavaScript effects that are now a must have and provides a basic interactive layer on top of the included styles.



Figure 10 Responsive web application over mobile phone and PCs, the picture illustrated the proposed application usability on both mobile and pc devices

Web Application Visualization tools

Two well-known visualization tools which are used during this thesis work are the following:

1. LeafletJs interactive map
2. AMCharts data visualization

1) LeafletJs interactive map

Leaflet is a modern open-source JavaScript library for mobile-friendly interactive maps. It is developed by Vladimir Agafonkin with a team of dedicated contributors. Weighing just about 33 KB of JS, it has all the features most developers ever need for online maps. Leaflet is designed with simplicity, performance and usability in mind. It works efficiently across all major desktop and mobile platforms out of the box, taking advantage of HTML5 and CSS3 on modern browsers while still being accessible on older ones. It can be extended with a huge amount of plugins, has a beautiful, easy to use and well-documented API and a simple, readable source code that is a joy to contribute to. A sample of this framework is provided in Figure 11.

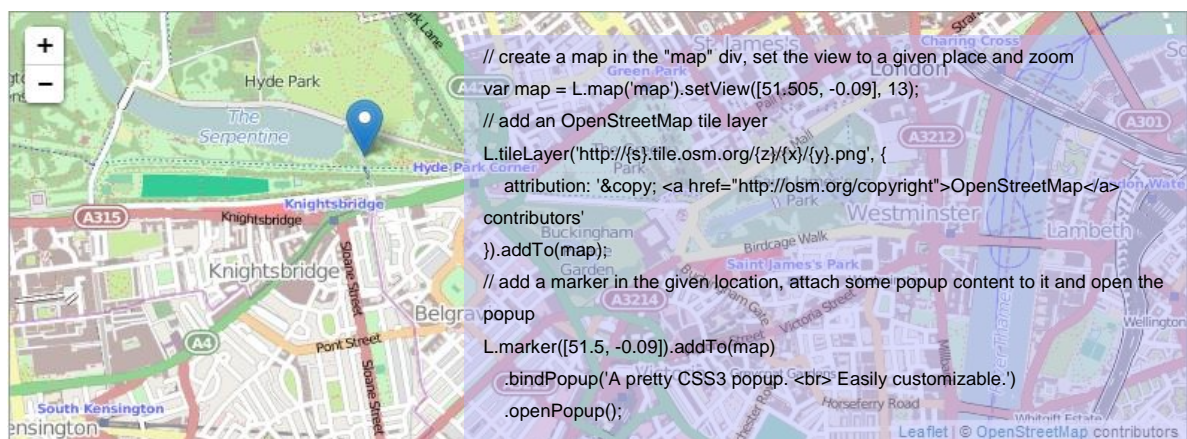


Figure 11 Basic example for leafletjs map which is used on mobile phone

2) AMCharts data visualization

amCharts is an advanced charting library that will suit any data visualization need. the charting solution include Column, Bar, Line, Area, Step, Step without risers, Smoothed line, Candlestick, OHLC, Pie/Donut, Radar/ Polar, XY/Scatter/Bubble, Bullet, Funnel/Pyramid charts as well as Gauges. Figure 12 displays different data charts.

The benefits of amCharts are the following:

- Editing charts visually
- Supports all modern browsers
- Free even for commercial use
- Super-powerful serial chart
- Scrollable and zoomable
- Exporting as an image or PDF
- A great set of themes
- Charts that look Hand-Drawn
- Motion-Charts

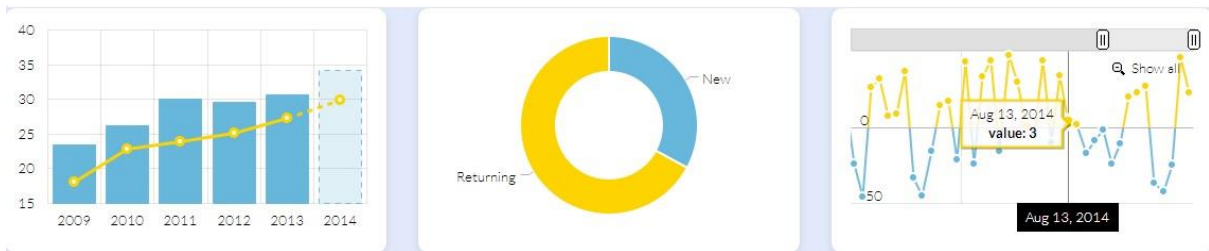


Figure 12 Various data visualization methods by amCharts

Web application architectures or design patterns

Design pattern is a solution to help developer to solve some similar design problem during development. From the developer point of view, using a good design pattern can produce a maintainable design and improve productivity of design is shortest time spending. From user`s prospective, it is also a solution to enhance the usability of application. Normally developers use their proper framework, which use their special or common design pattern. At below we list two most widely used web design pattern in these days:

1. Model-View-Controller architecture (MVC)
2. Service Oriented Front End Architecture (SOFEA)

1) Model-View-Controller architecture

Model-View-Controller or MVC [33] pattern has three component, model, view, and controller. **Model** represents the data and nothing else. Model does not depend on controller or view. **View** displays the model data to clients on the screens and sends user actions to the controller. **Controller** controls the interaction between view and model and initiates the creation of the application's new view. Figure 13 displays the architecture of the MVC.

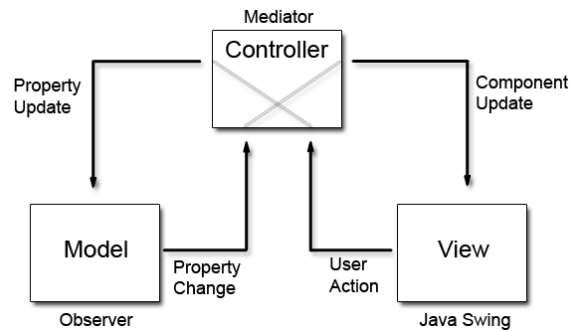


Figure 13 MVC architecture schema [33]

2) Service Oriented Front End Architecture

Service Oriented Front End Architecture or SOFEA is a well-known and widely used web application architecture in Thin Server Architecture (TSA). The main feature of SOFEA is no aspect of presentation logic runs on the server. As opposed to MVC architecture, all of the model, view and controller are client side component. The benefits of this architecture are scalability, a higher return of investment for each line of code, better use response, offline applications, interoperability, and an organized programming model. Figure 14 displays the architecture of the SOFEA.

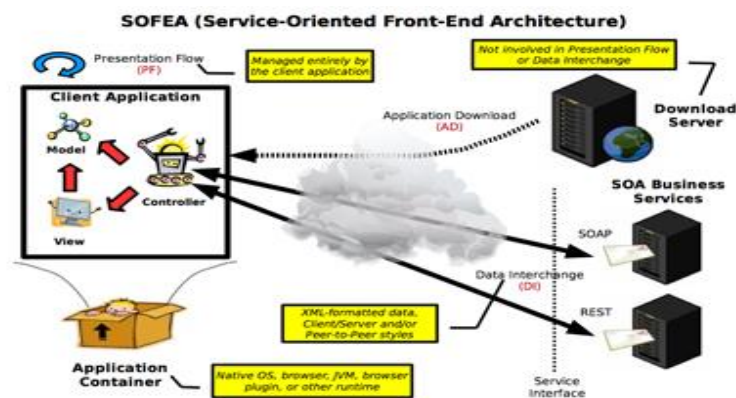


Figure 14 SOFEA architecture

Chapter 3

Problem Setting and Data Sources Exploration

In this chapter we are going to discuss more about the details of the thesis and the problem, and mainly about the type of the data sources we are dealing with. In this chapter we are going to represent the different profiles over these data set that are used in our proposed solution. This chapter is categorized in the two main section. In section 4.1: project proposal is discussed, and in section 4.2 the data sets from Telecom big data challenge and CORINE land use data set are discussed.

3.1 Project proposal

Generally as we mentioned in introduction we are going to provide:

- Innovative methods to understand actual land use based on activity data
- Experimental ways to automatically update land use classifications
- Concrete tools for urban planners to perform as-is and to-be analysis

The provided project is deriving **land use footprints** of Milano by analyzing the **activity data** provided by the telecom big data challenge 2013 and **comparing** the elicited land use footprints with the **land use classification** provided by CORINE in 2009. And finally **identifying relevant deviation** in land use between 2009 and 2013. Figure 15 explains the problem setting better. The Milano map is divided into 10000 cells and we have two different information about every cell; the first one is the type of the land use for each cell which could be from 1 to 24 different land uses, and the second one is the telecommunication data about every cell which includes the amount of activity in each cell. Fig1. Represents that cell id 6060

was mainly **construction site** in 2009, and the telecommunication data in 2013 could see that the land use may be changed.

But the question or problem which raises here is that, could we apply this assumption for all the cells and by using the mobile phone data try to predict the land use? If yes how it is possible? And how accurate it is? And what methods we need to apply to be successful. By analyzing the data sources we can find out the feasibility of the performing this idea. In the next part, we first explains about the data sources details and try to clarify the data type that we are going to work on. Then the limitations and type of these data are discussed.

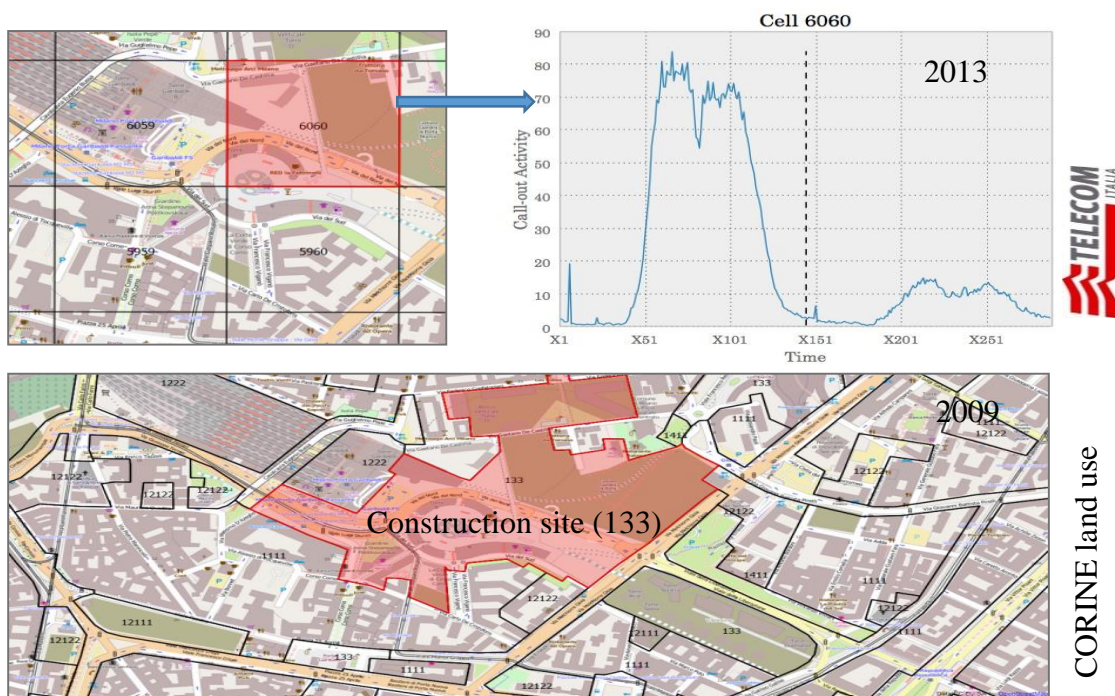


Figure 15 Milano grid and in-calls footprint for weekdays/weekends in cell id 6060

3.2 Raw Data Sources

In this part we are going to discuss more about the details of the data. As we hinted in the previous section we have two different data sets; which one of them provides the communication activity details by telecom Italia and the other one provides details about each cell land use in 2009. In the following we are analyzing these data to see what kind of solutions we can provide for them.

3.2.1 Telecommunication activity Data source

This dataset provides information about the telecommunication activities [34] in the city of Milano in the last two month of 2013. This dataset is the result of a computation over the Call Detail Records (CDRs)⁴ generated by the Telecom Italia cellular network over the city of Milano. CDRs log the user activity for billing purposes and network management.

Square id	Time interval	Country code	SMS-in activity	SMS-out activity	Call-in activity	Call-out activity	Internet activity
1	13835196	39	0.02613742	0.03087508	0.026137424	0.0552251	9.2601138
1	13835208	39	0.02792473	0.02792473	0.001787310	0.0546009	8.6690075
...							
40	13835196	39	0.04044804	0.01862799	0.001596025	0.0186279	4.5503902
40	13835196	49					0.0015960
...							
10000	13835196	39	0.02613742	0.03087508	0.026137424	0.0552251	9.2601138

Table 1 Telecommunication activity Data set

Table 1 displays in the first column the cell ids which are totally 10000 cells (the dimension of each cell is 250 * 250 meters), in the second column there is the time interval which is in Unix Epoch format and for instance the first cell means *11/4/2013 12:00:00 AM GMT+1*, then the country code that the connection is happened, and finally in other columns the activities. Notice that the Time interval is stored every 10 minutes. It means that for each cell id and country

⁴ Call Detail Record (CDR) databases are populated whenever a mobile phone makes/receives a call or uses a service (e.g. SMS, MMS). Hence, there is an entry for each interaction with the network, with its associated timestamp and the BTS that handled it, which gives an indication of the geographical location of the mobile phone at a given moment in time. Note that no information about the position of a user within a cell is known. The set of fields typically contained in a CDR include: (a) originating encrypted phone number; (b) destination encrypted phone number; (c) identifier of the BTS that handled the originating phone number (if available); (d) identifier of the BTS that handled the destination phone number (if available); (e) date and time of the call; and (f) duration of the call.

code there is maximum 144 time intervals. In the following we are explaining the details of this data source.

There are many types of CDRs, for the generation of this dataset; telecom Italia considered those related to the following activities:

- **Received SMS:** a CDR is generated each time a user receives an SMS
- **Sent SMS:** a CDR is generated each time a user sends an SMS
- **Incoming Calls:** a CDR is generated each time a user receives a call
- **Outgoing Calls:** CDR is generated each time a user issues a call
- **Internet:** a CDR is generate each time
 - a user starts an internet connection
 - a user ends an internet connection
 - during the same connection one of the following limits is reached:
 - 15 minutes from the last generated CDR
 - 5 MB from the last generated CDR

By aggregating the mentioned records it was created the dataset that provides SMSs, calls and Internet traffic activity. It measures the level of interaction of the users with the mobile phone network; for example the higher is the number of SMS sent by the users, the higher is the activity of the sent SMS. Measurements of call and SMS activity have the same scale (therefore are comparable); those referring to Internet traffic do not.

Spatial aggregation: different activity measurements are provided for each square of the Milano GRID.

Temporal aggregation: activity measurements are obtained by temporally aggregating CDRs in timeslots of ten minutes.

The data schema:

- **Square id:** the id of the square that is part of the Milano GRID; TYPE: numeric
- **Time interval:** the beginning of the time interval expressed as the number of millisecond elapsed from the Unix Epoch on January 1st, 1970 at UTC. The end of the time interval can be obtained by adding 600000 milliseconds (10 minutes) to this value. TYPE: numeric
- **Country code:** the phone country code of a nation. Depending on the measured activity this value assumes different meanings that are explained later. TYPE: numeric

- **SMS-in activity:** the activity in terms of received SMS inside the Square id, during the Time interval and sent from the nation identified by the Country code. TYPE: numeric
- **SMS-out activity:** the activity in terms of sent SMS inside the Square id, during the Time interval and received by the nation identified by the Country code. TYPE: numeric
- **Call-in activity:** the activity in terms of received calls inside the Square id, during the Time interval and issued from the nation identified by the Country code. TYPE: numeric
- **Call-out activity:** the activity in terms of issued calls inside the Square id, during the Time interval and received by the nation identified by the Country code. TYPE: numeric
- **Internet traffic activity:** the activity in terms of performed internet traffic inside the Square id, during the Time interval and by the nation of the users performing the connection identified by the Country code. TYPE: numeric

Files are in tsv format. If no activity was recorded for a field specified in the schema above then the corresponding value is missing from the file. For example, if for a given combination of the Square ids, the Time interval *i* and the Country code *c* no SMS was sent the corresponding record looks as follows:

```
s \t i \t c \t \t SMSout \t Callin \t Callout \t Internettraffic
```

Where `\t` corresponds to the tab character, `SMSout` is the value corresponding to the SMS-out activity, `Callin` is the value corresponding to the Call-in activity, and `Callout` is the value corresponding to the Call-out activity and `internet traffic` is the value corresponding to the Internet traffic activity.

Moreover, if for a given combination of the Square ids, the Time interval *i* and the Country code *c* no activity is recorded the corresponding record is missing from the dataset. This means that records of the following type

```
s \t i \t c \t \t \t \t \t
```

are not stored in the dataset. The data that we are working on that is almost huge, although it is only the collection of information for two months of November 2013, December 2013, and first day of January 2014. It includes 62 files which is totally about 5 GB text size.

Before displaying a sample of the data it is better to define the Unix Epoch which is used in the data. The Unix epoch (or Unix time or POSIX time or Unix timestamp) is the number of

seconds that have elapsed since January 1, 1970 (midnight UTC/GMT), not counting leap seconds (in ISO 8601: 1970-01-01T00:00:00Z). Literally speaking the epoch is Unix time 0 (midnight 1/1/1970), but 'epoch' is often used as a synonym for 'Unix time'. Many systems store epoch dates as a signed 32-bit integer, which might cause problems on January 19, 2038 (known as the Year 2038 problem or Y2038). [35] In the proposed application we convert this time unit to human readable time.

3.2.1.1 Mobile phone activities behaviors

In Figure 16 we are displaying the profile of all activities (Call-in, Call-out, Sms-in, Sms-out) for two weeks in one cell. The dark blue is showing the average of the two weeks. What we can understand is the zone of 6060 is one of the crowded zones maybe full of offices and commercial units because it has a pick time around 12 p.m. to 2 p.m. Therefore by observing these profiles we can say that the behavior of the signals are almost the same but the amount of the usage is different. In the main contribution of the project we tried to work on one of the data sets and apply the same solutions over other data sets and in most of the cases we have the similar result.

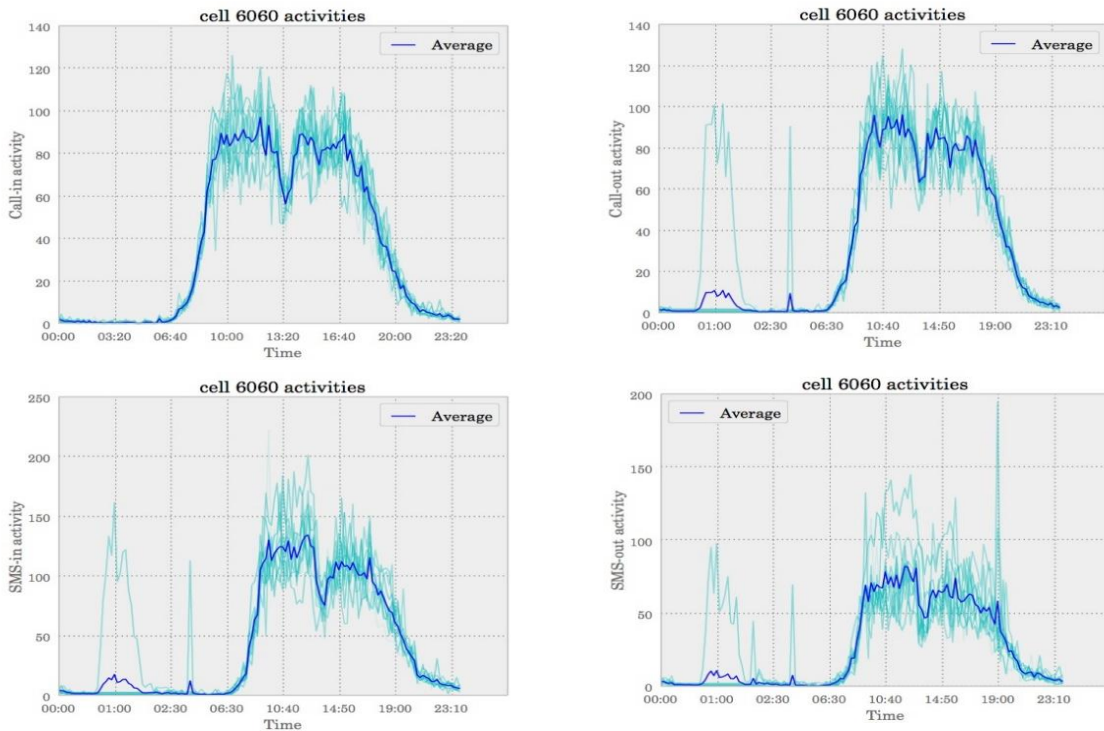


Figure 16 Cell 6060 activities during two weeks with the mean value

3.2.1.2 The Heterogeneity in communication data

As discussed before the communication data from telecom Italia are just for two last months of 2013, November and December. So the behavior of the last month of the year might be unusual and it causes some confusions in our analysis. We have analyzed that the behavior of mobile usage in the first 14 days of November is significantly different from the data of the last 14 days of the year. In Figure 17 we plotted this difference for one of the very crowded zone (6060).

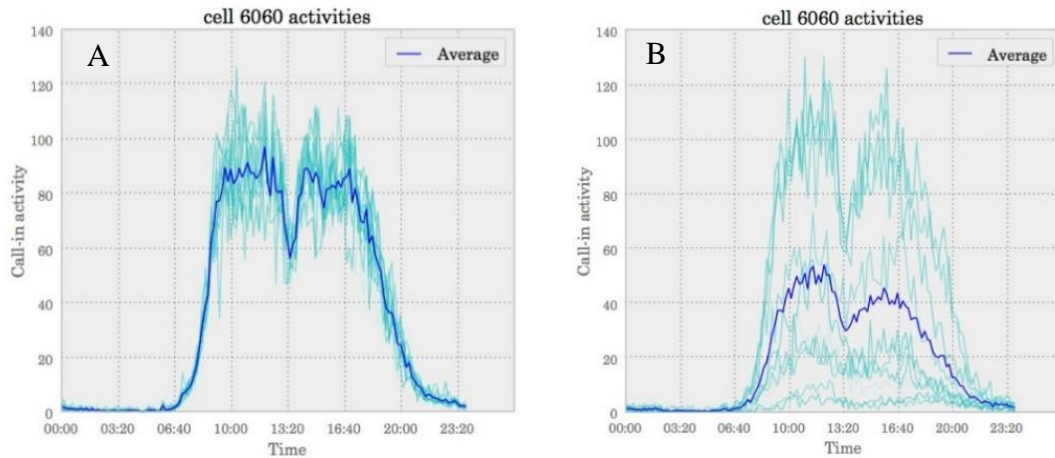


Figure 17 A) The call-in activity profile for the 4th to 17th of November
 B) The call-in activity profile between 18th to 31th December

3.2.2 CORINE data source of 2009

In this part we are going to explain about the data of 2009 which defines the land use of each cell. Every cell has different numbers and percentages of land uses that identify the type of that zone in the map; it is clear by observing the place and using GIS methods. These details have been collected in 2009 for Milano and by CORINE and you can see the in the Table 2.

cell	111	112	121	124	131	134	...	211	213	231	324	411	512
1	0	0	0	0	0	0		24.312	75.688	0	0	0	0
2	0	0	0	0	0	0			93.783	6.217	0	0	0
...													
9999	0	11.282	32.159	0	0	0		36.703	0	16.85	0	0	0
10000	0	11.077	0	0	0	33.514		0.297	0	4.752	0	0	0

10000 rows × 24 columns

Table 2 Milano Land use percentage data set information of 2009 from CORINE

The details of each columns are the following:

- **cell.id**: This is the id of the cell, i.e. the "square" in the Milano grid as in telecommunication activity Data set.
- **clc.codes**: these are the land use classification codes (European Commission, CORINE Land Cover initiative [36]) that indicates the type of land use of that portion of the cell. This number is the percentage of the cell area occupied by the portion classified with that land use, if you sum the percentages for the same cell.id you should get 100%. European Commission standardized a classification of land use which we discuss later. There are datasets available about most of Europe, but, specifically to us, Lombardy Region released as open data a dataset named DUSAF [37] which uses the CORINE land use classification to describe a large area including the city of Milano.
- **area**: this number is the surface area of the portion of the cell [38]; if you sum the areas for the same cell.id you should get around 55.225 are (235m x 235m)

In Table 3 we listed the summary of the CLC code with their title. In Appendix A, The full description of all the CLC codes are discussed.

CLC CODE	Title
111	Continuous urban fabric
112	Discontinuous urban fabric
121	Industrial or commercial units
122	Road and rail networks and associated land
124	Airports
131	Mineral extraction sites
132	Dump sites
133	Construction sites
134	Other unused areas
141	Green urban areas
142	Sport and leisure facilities
211	Non-irrigated arable land
213	Rice fields
221	Vineyards
222	Fruit trees and berry plantations
224	Other permanent crops

231	Pastures
311	Broad-leaved forest
314	Other forests
322	Moors and heathland
324	Transitional woodland-shrub
411	Inland marshes
511	Water courses
512	Water bodies

Table 3 CLC code labels for each land use in 2009

3.2.2.1 Distribution of land uses in each cell

In this section we are going to discuss about land use distribution base on CORINE land use data of 2009. The first thing that we figured out form this data source was in most of the cells (250 meter * 250 meter) we have different type of land use. As we can see in Figure 18, distribution of land uses in southwest and southeast of Milan is one (orange) or two (yellow) types in most of cells, probably the reason is the most of the cells in these area are agricultural area. But in the center of Milan, in most of cells we have three (light green) or four (dark green) type of land uses. And also we can figure out in northwest we have more than four type of land use. Number of land uses in each cell limits the quality of our analysis. In other words, in our raw data, most of the cells have about four to five type of land use as showed in the Figure 18, for instance 2677 of the cells have 3 different land uses. Then because of this high degree of differences in land use types we decide to categorize them based on their taxonomies to see that how the land uses distributed over the city based on their taxonomy. We discuss about this categorization in the next section.

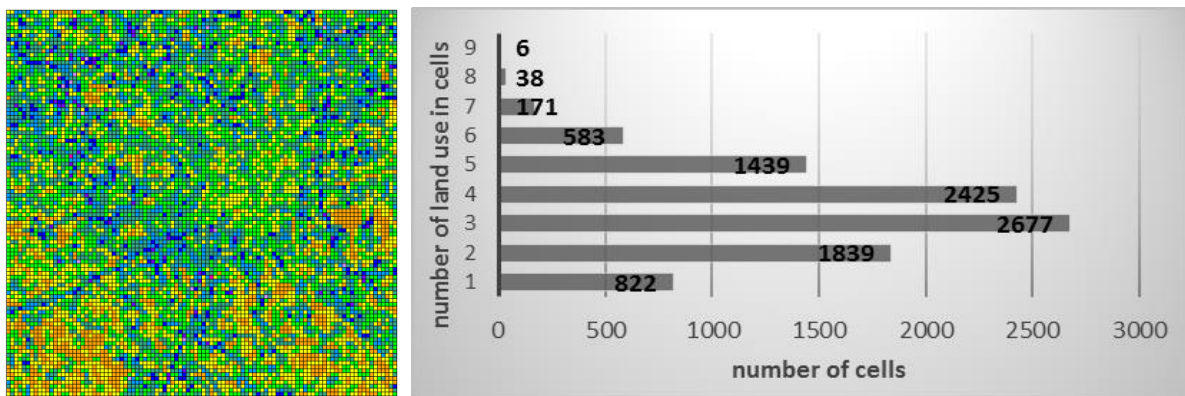


Figure 18 Distribution of land uses

3.2.2.2 Distribution of land uses base on taxonomy

Base on the taxonomy of the land uses, we can categorize the cells in five main category as below:

1. Artificial surfaces
2. Agricultural areas
3. Forests and semi-natural areas
4. Wetlands
5. Water bodies

As we can see in Figure 19 most of the cells are artificial surfaces and agricultural areas, which are in the center and corner of Milan. In more details we can say that 61.74% of our analyzing area (61.3% of cell`s prevalent) is artificial surfaces where most of them are in center, and 34.1% of the rest (36.5% of cell`s prevalent) is Agricultural surfaces where most of them are in corners of Milano. It means that about 96% of our analyzing area are in these two first category.

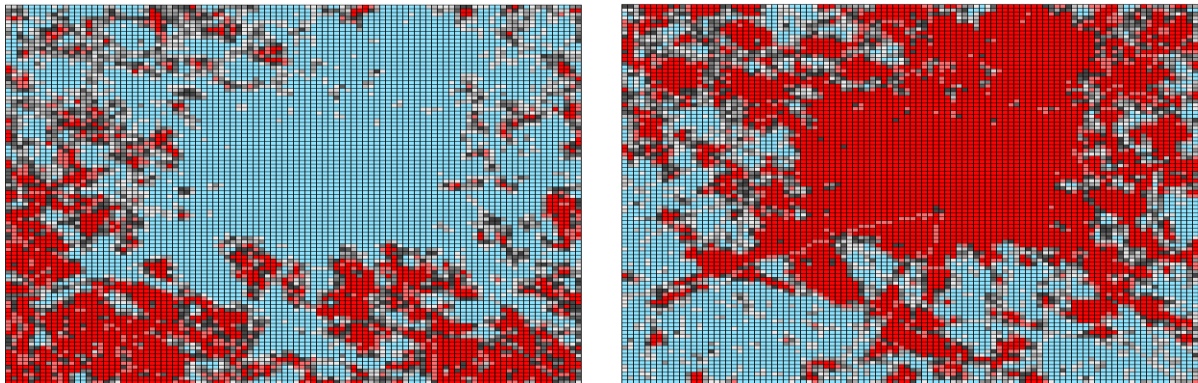


Figure 19 Distribution of land uses by taxonomy

3.2.2.3 Cell`s prevalent land use

Cell`s land use prevalent is a land use type which has the most percentage in the cell. As we described above, most of the cells have more than one land uses. In the Figure 20 we can see all land uses which their prevalent land use is at least 75% of the cell`s area. They are more than 38% of all cells which is around 3842 cells. More than 60% of these cells are Industrial/commercial units (code: 121), Non-irrigated arable land (code: 211), and Rice fields (code: 213), and in these 60%, more than 50% are just in non-irrigated arable land (code: 211). Moreover, the number of land uses reduced to 17.

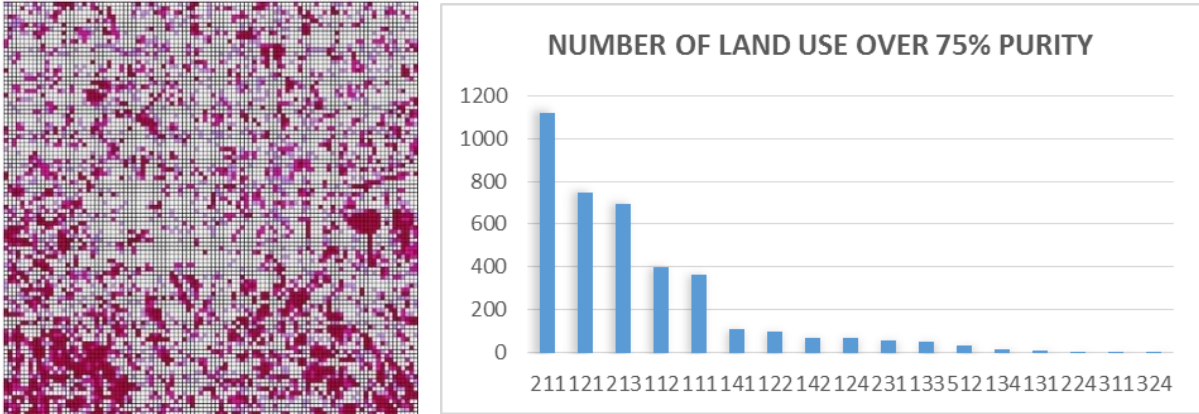


Figure 20 Prevalent land uses

3.3 Preprocessing of data

For the last two month of the year 2013 (November and December) we have 60 CDR files in total. We clean data in two main parts, at the first part we clean CDR data in terms of integration and separation for each day, and then in the second part, we merged all of them together to create the main data source for analyzing. In following the cleaning steps are discussed:

Step 1 (integration and separation for each day):

After visualizing some data and analysis over data we figure out that some attribute could be neglected. Table 1 displays the raw data extracted from Telecom Italia, we decided to aggregate and make summation of the activity of different country code for each time and cell, which do not make any difference for our computation. Because we are considered all the activities which took place in each cell, therefore it does not make any sense that the communication is from/to which country. So as an output of this phase, we had one row for each cell in each specific epoch time in all days.

Then, we decided to separate the data of weekdays and weekend in all the data sources. The reason is obvious since the human dynamics and behavior are well differentiated between weekdays and weekends [39]. So these behaviors of each of them has different meaning for analyzing. For example, in Rho fieria area (Cell-id: 7624), an area for exhibitions and EXPO, usage of mobile phone in weekends are more than weekdays and in opposite, in some places such as Citta Studi (Cell-id: 5773) area which is academic area (Politecnico di milano and Degli studi), usage of mobile in weekends is less than weekdays. You can see the difference

in Figure 21, the blue signal is displaying the signal for weekdays and the green one the signal for weekends.

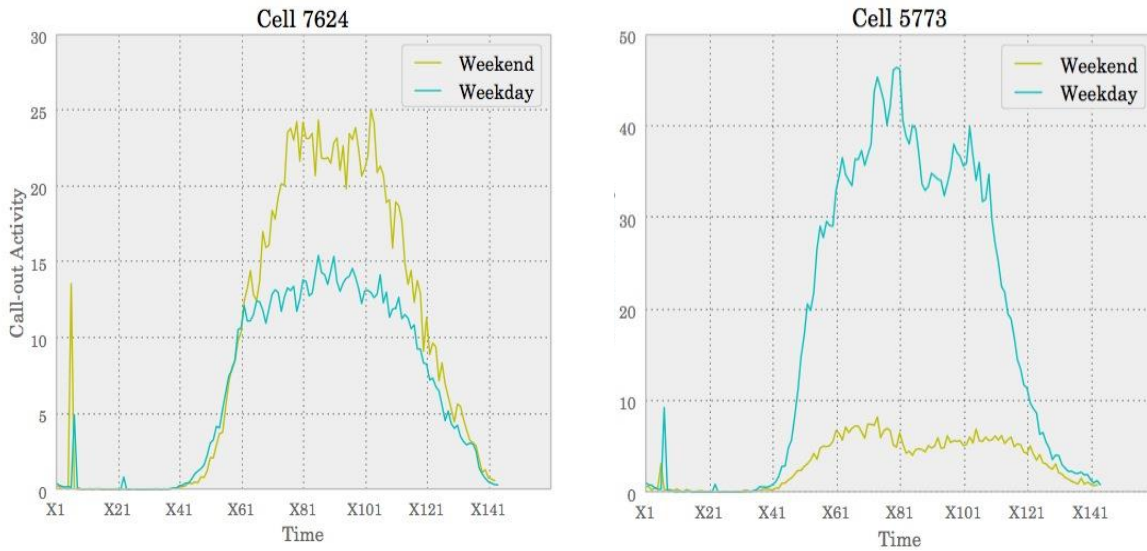


Figure 21 Usage of mobile phone during weekday and weekends for two different zones

CDR data collected in each 10 minutes, so for each day and for each cell we have 144 rows (144 * 10 = 1440 min = 24 h). We convert each those epoch times from X1 to X144 for weekdays and X145 to X288 for weekends. In 10000 rows × 144 columns

Table 4 you can see the output of this step which is a matrix of 10000 in 144 for each day.

	X1	X2	X3	X4	...	X141	X142	X143	X144
1	0.026137	0.0273	0.001787	0.0	...	0.108039	0.053438	0.217241	0.053438
2	0.027356	0.0273	0.000922	0.0	...	0.109257	0.054656	0.218459	0.054656
...
9999	0.316471	0.006571	0.17199	0.0	...	0.173896	0.237047	0.092566	0.237047
10000	0.176764	0.000000	0.17199	0.0	...	0.113237	0.090769	0.085995	0.090769

10000 rows × 144 columns

Table 4 Matrix of CDR data for the 4th of November (weekday)

Step 2 (total average):

In this step we made the average over previous outputs for weekdays (43 days) and weekends (17 days) respectively. The reason to do so is that the average of all the days could be nominate of that cell as we plotted in Figure 22 for cell id 6060.

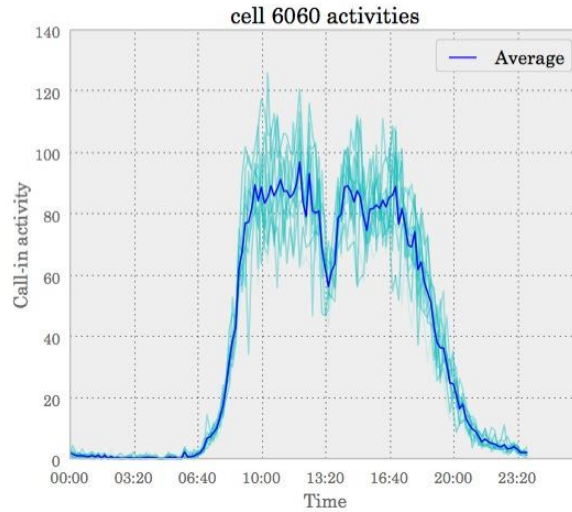


Figure 22 Average call in activity data for cell 6060 for the whole two months

So finally, we have two data frames, one for weekdays' average and one for weekends' average which are *average activity for each cell in each time unit for each type of activity*. After that we merged these two data frames for all the cells because we wanted to have only one data source for analyzing and clustering. That is our final and main data source for analyzing. Table 5 is showing the values of the merged data.

	X1	X2	X3	X4	...	X285	X286	X287	X288
1	0.123378	0.094360	0.064053	0.047685	...	0.094415	0.129664	0.105530	0.085839
2	0.124925	0.095664	0.065192	0.048246	...	0.095771	0.131311	0.107150	0.087264
3	0.126573	0.097053	0.066404	0.048844	...	0.097215	0.133066	0.108874	0.088781
...
9998	0.582413	0.364944	0.380144	0.282690	...	0.478331	0.484937	0.540031	0.441365
9999	0.359854	0.292194	0.234900	0.195788	...	0.340490	0.340622	0.357901	0.226425
10000	0.330093	0.232021	0.201159	0.187899	...	0.322601	0.376215	0.266842	0.170460

10000 rows × 288 columns

Table 5 Main data structure for computation

3.4 Naïve solution

In this section we discuss about the common approach of land use identification to see weather this approach is enough to predict land uses in Milano or not. This section includes:

- Clustering over the CDR data based on three different approaches
- Problem setting and challenges

This method is based on clustering, what we mean is to try to do some unsupervised clustering and assigning each land use to one of the clusters as the indicator of that cluster. And then by clustering try to identify the land use, this is the work that recent researches try to do. In other words, they are obtaining land use signature from CDR data. In the following, we demonstrate the same methods that could not solve our problem and causes us to define better solutions which discussed in the next chapter.

Clustering over CDR data to identify the land use

Here we did our first clustering experience and then tried to analysis the output. We used two different clustering algorithms one Kmenas and the other one K-means++. The problem here is defining the initial groups' centroids. The common way to do this is to assign random values for the centroids of all groups. This random behavior makes some inconsistency and we need to run the K-means algorithm several times to have the best result, more over in our experiment we found out that K-means++ have the better and accurate results for our research as we discussed in the methods. Our clustering samples are based on the meaningful data from CORINE and try to choose the K value in the following order:

1. Clustering based on the all types of land use
2. Clustering based on the taxonomy
3. Clustering over two majority land use taxonomy

1) Clustering based on the all types of land use

The first idea for the K (the number of clusters) is to choose 24, which is the number of all land uses, and hope to have different clusters with a meaningful output. The result of this clustering is depicted in Figure 23. We can say that a cluster is representing one type of land use when at least 75% of the cells on that cluster are in one type of that land uses category. But as we can see in figure below, about 80% percent of cells are in first three cluster, and it is impossible to find corresponding land use cluster.

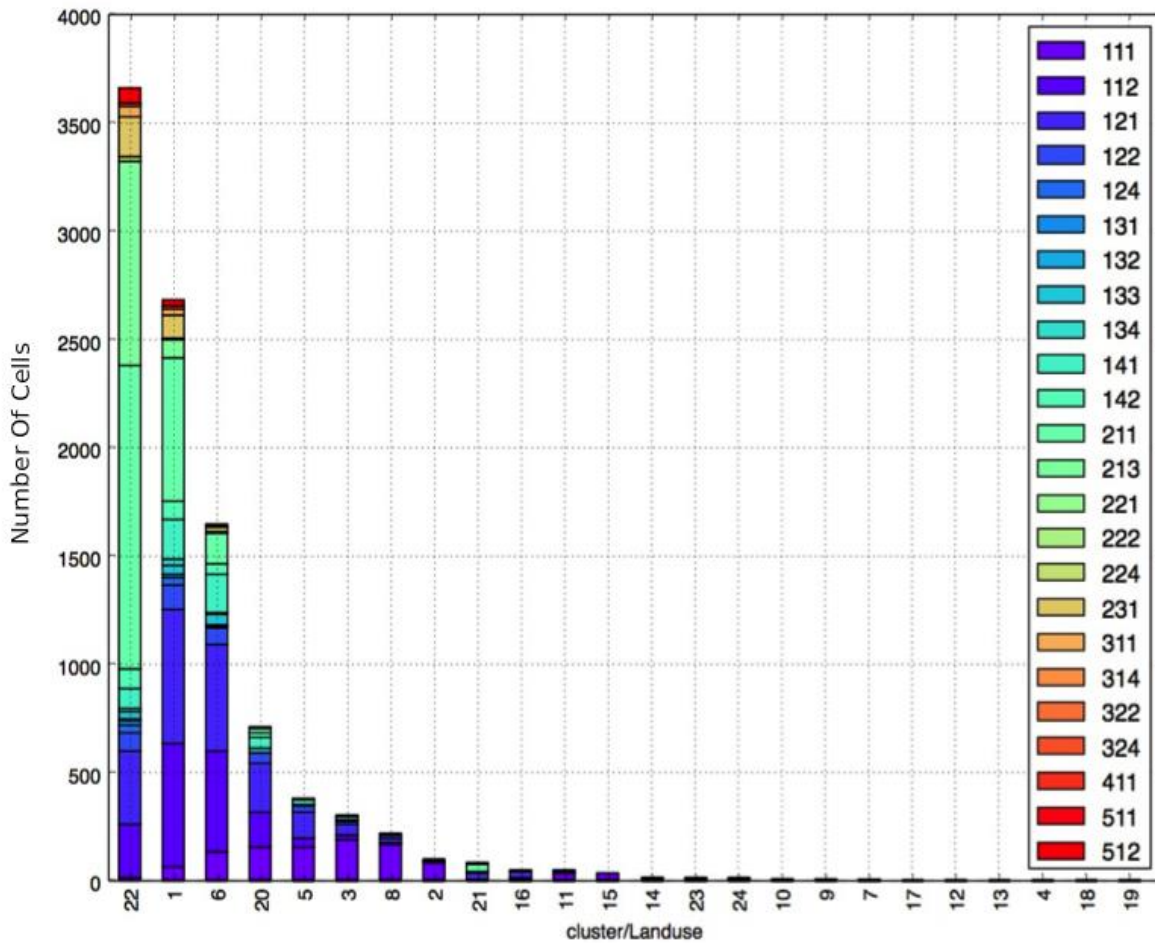


Figure 23 K-means++ clustering result over call-out activity by k=24 (which is all the land use type from CORINE)

2) Clustering based on the taxonomy

To do the clustering, trying different cluster numbers to find the best output is common. To improve our clustering we decide to choose smaller numbers and tried 18, 16, and so on. In this step we reached to 5, the reason of this number is the taxonomy of the land use' categories which are described in Appendix A. The categories are the following:

1. Artificial surfaces
2. Agricultural areas
3. Forests and semi-natural areas
4. Wetlands
5. Water bodies

The result is illustrated in Figure 24. Although we have the same problem as the one in 24 clustering and most of the cells are assigned by only two of the clusters, but this cluster number

have a better sense and could be analyzed more precisely. Then we tried to improve this clustering by using different parameters. The first step for improving our clustering method is changing distance measure.

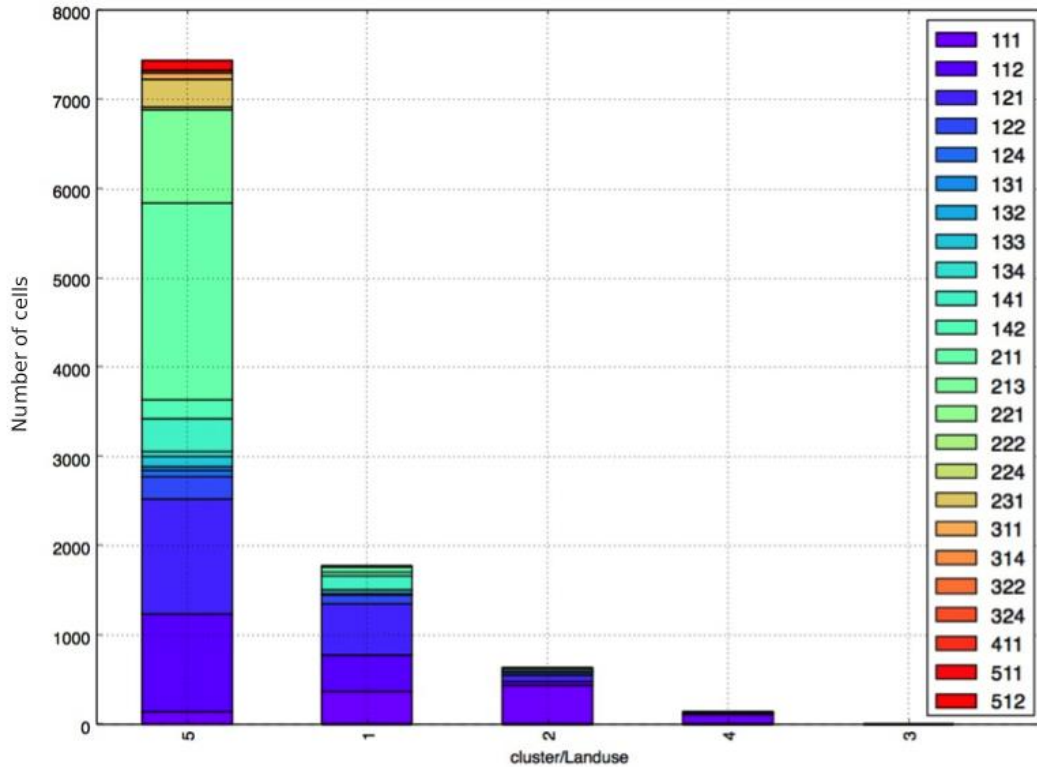


Figure 24 K-means++ clustering over call-out data by k = 5 (the number of taxamomy)

As we know K-means clustering was implemented using two distance measures: Euclidean and a Dynamic Time Warping (DTW) based metric. DTW is especially useful when dealing with signals that can have slight time shifts, but experimental evidence showed that better index values were obtained with the Euclidean distance [15]. Also the computational cost was considerably higher when using DTW. The second effort is using spherical k-means clustering, the problem of S-K-means is zero values in rows of the matrix which there is for our case study. We fixed it by setting a very low values near zero, but the results have no quality because S-kemans used cosine similarity for measurement. Since the shape of the profiles are similar to each other working with S-K-means has no benefits for us in this stage.

3) Clustering over two majority land use taxonomy

Base on section 3.2.2.3 (Prevalent land use), we removed all the cells that have prevalent less than 75% base on CORINE data source. By this cleaning we could say that each cells profile has behavior of one type of land use. Most of the land uses are in first two categories (Artificial surfaces and Agricultural area). We decided to choose $k=2$ to see whether we could distinguish this two main type of land uses by clustering or not? We do this over all cells with prevalent more than 75% in just first two category (code: 1xx and 2xx). As we can figured out in Figure 25 about 94 % of cell were in just one cluster and the other assigned to the second cluster, so it cannot separate in two cluster. But in the fact they are half and half but clustering could not separate them.

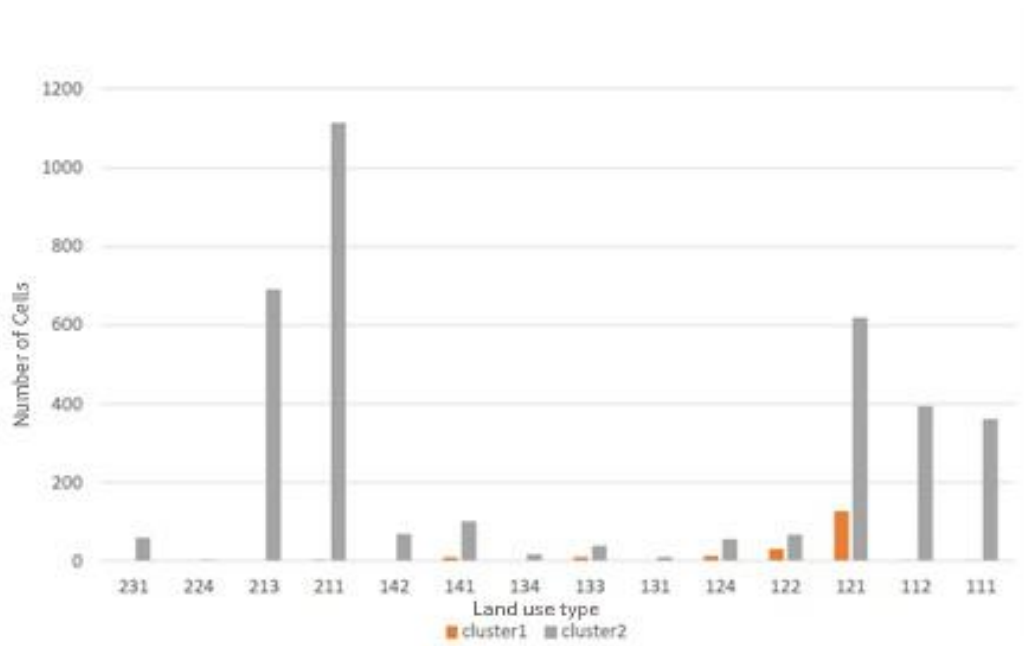


Figure 25 K-means++ clustering output over 2xx and 1xx

3.4.1 Problem setting (Challenges)

At this stage, we could say that the unsupervised learning or clustering solutions over the CDR data is not the suitable approach for our case study and it could not identify land uses signatures, the reason could be the following:

- **Size of the CDR data:** The size of the CDR data are not enough for this learning. As mentioned before the data are only for two months.
- **Heterogeneity in communication:** The heterogeneity in communication data (3.2.1.2) could be another reason that decrease the quality of this learning.
- **Diversity of Milano land uses:** The second reason could be diversity of Milano land uses, by analyzing over the Milano land uses we see most types of land uses are in two categories which are artificial and agricultural.
- **Characteristic of the city and mobile phone data**

The main challenge that could be interesting here is to research if there are any other solutions to solve the problem of automated land use identification or not. If yes, how they are going to solve this? And how accurate they are? In the next chapter we purpose our solutions.

Chapter 4

Solution Space

In this chapter we are going to discuss about the solutions to solve the problem described in chapter 3. We introduce the main structure of the proposed solutions, and in the two following sections we will go into details of each solution. Then in section 4.4 we compare these two solutions and finally in section 4.5 we explain the proposed application for visualizing and representing of prediction.

4.1 General Structure of proposed Solutions

As we described in problem setting the output of these solutions is prediction of the most land usages of each cell that can be more than just one type. In this thesis we proposed two solutions to solve the described problem in pervious chapter. Both of them have the same general idea or structure to solve the problem. Both of them at first detect profile of each land use type with using final set data and CORINE and then compare call-out signal of each cell with those profiles to predict land use of it. This idea is depicted in Figure 26.

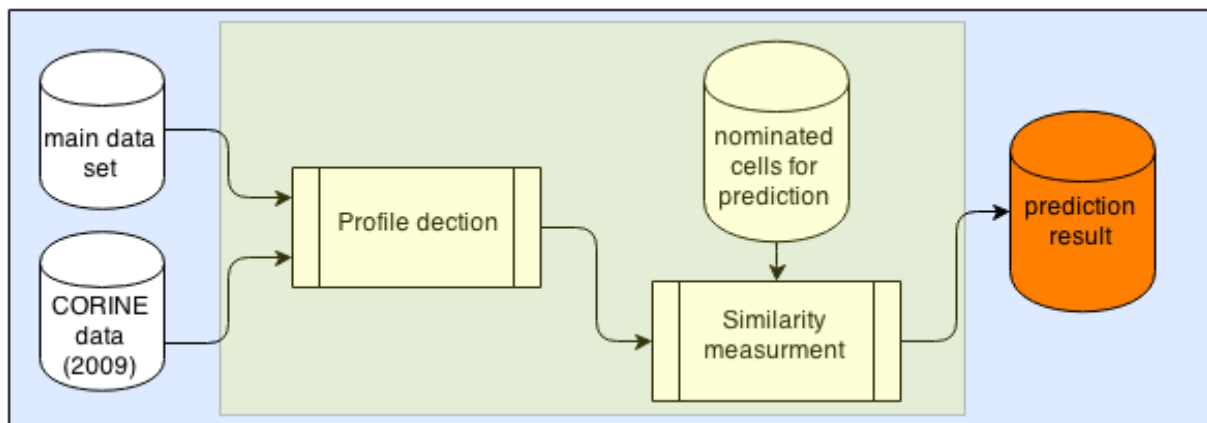


Figure 26 General structures of proposed approaches

Our solutions are based on some assumptions, the first and main assumption which is true in all the solution is:

Assumption-1: most of the cells and land uses have not been changed from 2009 to 2013.

This assumption can be one of the main differences between our solutions and previous works in this area, which most of them used unsupervised learning by clustering. Assumption 1 is derived from the nature of Milan because changing land use in this city is not too much. Obviously this assumption can be wrong for such cities with too many changes in their land usages for special periods. In order to apply assumption 1 to solutions we use CORINE data set as properties of each cell for profile detection phase. In the following two sections we explain detail of each solution and after that in section 4.4 we compare them.

4.2 Solution 1: Comparison to Weighted Profile

In this part we are going to explain the first solution that we used it as the data analyzing for backend of the Telecom Big Data Challenge application in April 2014.

This method apart from *assumption-1* has another assumption which is:

Assumption-2: percentage of each land use in each cell is percentage of mobile usage of that land use in that cell.

For example, in cell 6060, we have 80 % of construction site and 20% Continuous urban fabric base on CORINE data. We assume that 80 % of mobile activity for cell 6060 is for usage in construction site and 20 % is for Continuous urban fabric. This example is depicted in Figure 28 A.

Assumption 2 is exactly the way of how CORINE data can be used in profile detection phase. Therefore it is exactly the way of applying assumption 1 to profiling. This solution based on assumption 2 detects profile of each land use and with adjusted cosine similarity measure calculates similarity of each profiles and signal of each cells. Two most similar profiles can be prediction result. The conceptual model for this approach depicted in Figure 27.

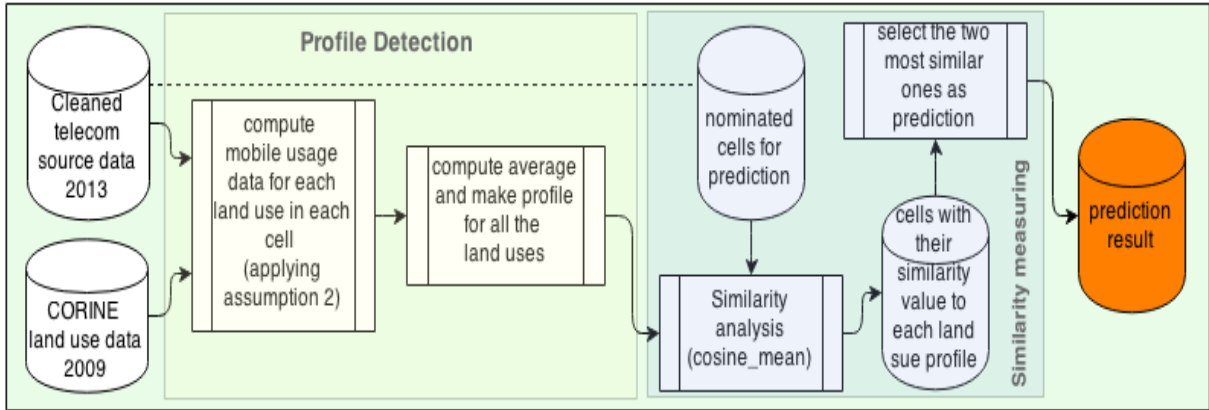


Figure 27 Conceptual model of solution 1

As we can see in figure shown above, this solution`s process has two main phase like general structure:

1. Profile Detection
2. Similarity measurement

In the following we explain these phases in detail.

1) Profile Detection

In this phase we prepare profile of each land uses. In order to do this at first base on assumption-2 we calculate signal of each land use in each cells (Figure 28 A). Land uses for each cell are the land uses with percentage more than zero in CORINE data for that cell. After that for make profile we take average of signals of all same land use type of all cells. Output of this phase is depicted in Figure 28 B.

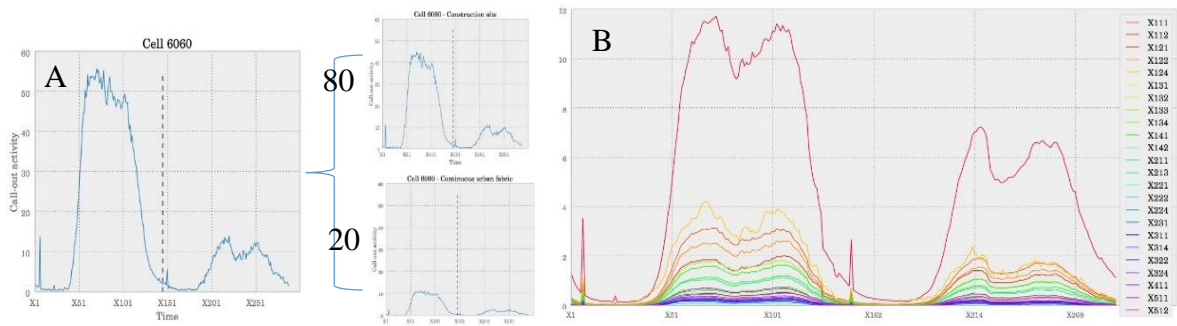


Figure 28 A) Calculation each land use signal in each cells based on assumption-2. It calculate these signal by weight of each land use which are in CORINE data B) profiles of all land uses, these profile are averages of same land uses in all cells

2) Similarity Measurement

After making profile in previous phase, this phase calculates similarity between each nominated cells (cell's call out signal) and all land use profiles. And selects two most similar land uses as a prediction.

To measure similarity we use the following method '*sim*' between the generated profile '*p*' of each land use type and a specific cell's signals '*c*':

$$Sim = cosine(p, c) * \left(\frac{\min(p_{avg}, c_{avg})}{\max(p_{avg}, c_{avg})} \right)$$

Where $p_{avg} = mean(p)$ And $c_{avg} = mean(c)$

This similarity measure has a domain of [0,1]: 0 indicating no similarity and 1 indicating high similarity. In fact, if '*p*' and '*c*' are similar in their dynamics (measured by cosine similarity) and similar in their amplitudes (roughly measured by the ratio of their means), the above formula will return a value close to 1. In contrast, if '*p*' and '*c*' are different in dynamic or have highly difference in their average amplitude, the above measure will return a value close to 0.

We have added the ratio-of-averages term inspired by the fact that the data of a large number of the cells are highly correlated. This is because of the expected rise and fall in the 'data' at certain day times. Thus the average of the cell's data seems to be relatively more distinctive.

Finally, two most similar land use of each nominated cell are predicted land use for it. The output of this solution is shown in Table 6

id	111	112	121	122	124	...	324	411	511	512	mx_sim	mx2_sim	diff_sim	prediction
1592	0.215405	0.154721	0.325676	0.216377	0.371211	...	0.005354	0.004422	0.011309	0.019826	0.371211	0.325676	0.045536	122,121
2447	0.056984	0.318826	0.789325	0.776136	0.666034	...	0.018948	0.015589	0.040488	0.069875	0.789325	0.776136	0.013189	121,133
2548	0.007132	0.190417	0.100234	0.149936	0.088045	...	0.145345	0.118896	0.302731	0.540146	0.540146	0.5333	0.006846	134,512
...
9713	0.002133	0.060273	0.029822	0.044892	0.02586	...	0.540395	0.446017	0.815582	0.492125	0.863709	0.815582	0.048126	231,511
9714	0.001124	0.031833	0.015654	0.023615	0.013484	...	0.843739	0.850273	0.45033	0.258561	0.850273	0.843739	0.006535	411,324
9814	0.006462	0.181836	0.090313	0.135867	0.078335	...	0.179488	0.148041	0.378368	0.664915	0.664915	0.637835	0.027081	512,134
9815	0.002981	0.084501	0.041506	0.062618	0.035764	...	0.387407	0.320279	0.823127	0.685836	0.903156	0.823127	0.080029	231,511

Table 6 Output of solution 1

4.3 Solution 2: Comparison to denoised clustered profile

This section introduces the second solution. Profiles in previous solution might be affected from outliers in process of profiling. The reason of that can be:

- Assumption-2 is probably wrong
- Those cells with radical change from 2009 to 2014 and nominated for prediction can make our profiles noisy.

So for this reason in this solution we proposed two following item for profile detection phase:

- Item 1: selection of cells which their prevalent land use has more than 75% of the cell's area.
- Item 2: find and exclude outlier cells in each group of cells with same prevalent land use type.

Item 1 and 2 are the main difference of profiling detection in compare of previous solution. Both of them are the way of using CORINE data in profiling. These two makes our classes more accurate than assumption-2 but they have their disadvantages that will explain in comparison section.

This solution based on these two items makes distribution of cells for signature of each land use and with Mahalanobis Distance calculates distance between of each nominated cell s and each land use class (which is distribution of inliers cells with same prevalent land use). Figure 29 shows the process of this solution. As we can see the process is consisting of three phases:

1. Preprocessing
2. Clustering
3. Distance Measure

The details of these three phases are discusses in following.

1) Preprocessing:

This phase prepares data set for clustering phase. In order to do this with information of CORINE data we select those cells that have prevalent with percentage more than 75% which are 3842 cells (item 1). The reason of this selection is to select those cells that show better or cleaned behavior of their prevalent land use type. After that data must be normalized for

clustering process. Therefore we ignore the shape of the distribution and transform the data to the center it by removing the mean value of each feature and scale it by dividing non-constant features by their standard derivation.

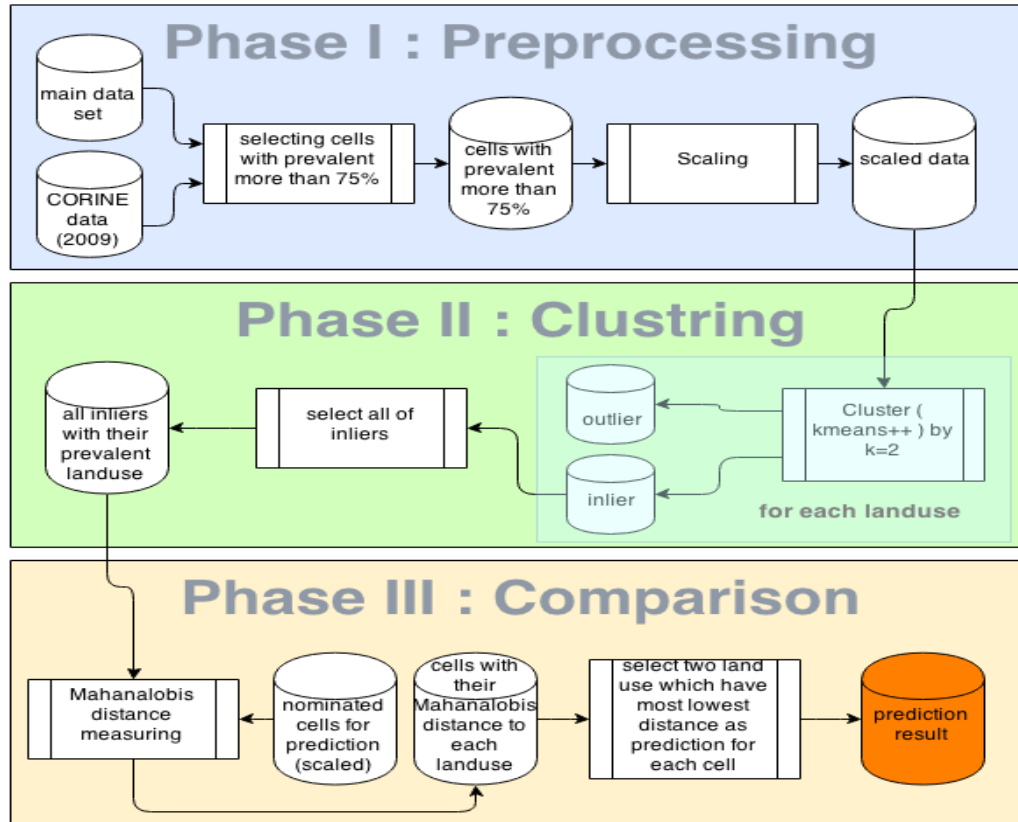


Figure 29 Conceptual model of solution 2

This operations performed by *scale* function of python. The output of this phase is shown in Table 7.

	X1	X2	X3	...	X286	X287	X288
1	-0.449487	-0.442598	-0.441358	...	-0.396145	-0.402296	-0.404846
2	-0.448380	-0.441508	-0.440191	...	-0.395028	-0.401084	-0.403710
3	-0.447201	-0.440348	-0.438949	...	-0.393838	-0.399795	-0.402500
...
9988	-0.264941	-0.304849	-0.280775	...	-0.283471	-0.248711	-0.207627
9989	-0.171385	-0.248015	-0.206077	...	-0.239161	-0.181987	-0.100697
9996	-0.171530	-0.212196	-0.154860	...	-0.198351	-0.103371	-0.170427

Table 7 Output of phase 1 (normalized data)

2) Clustering

In this phase we cluster all cells with same prevalent land use. To perform this we used k-means++ with $k=2$. The reason of choosing $k=2$ is to find outlier and inlier cells of each land use (item 2). So, for each land use we have to have two clusters that are cluster of outliers and inlier (Figure 30 A). In average outliers cells include 20% of all cells with the same prevalent land uses (Figure 30 B).

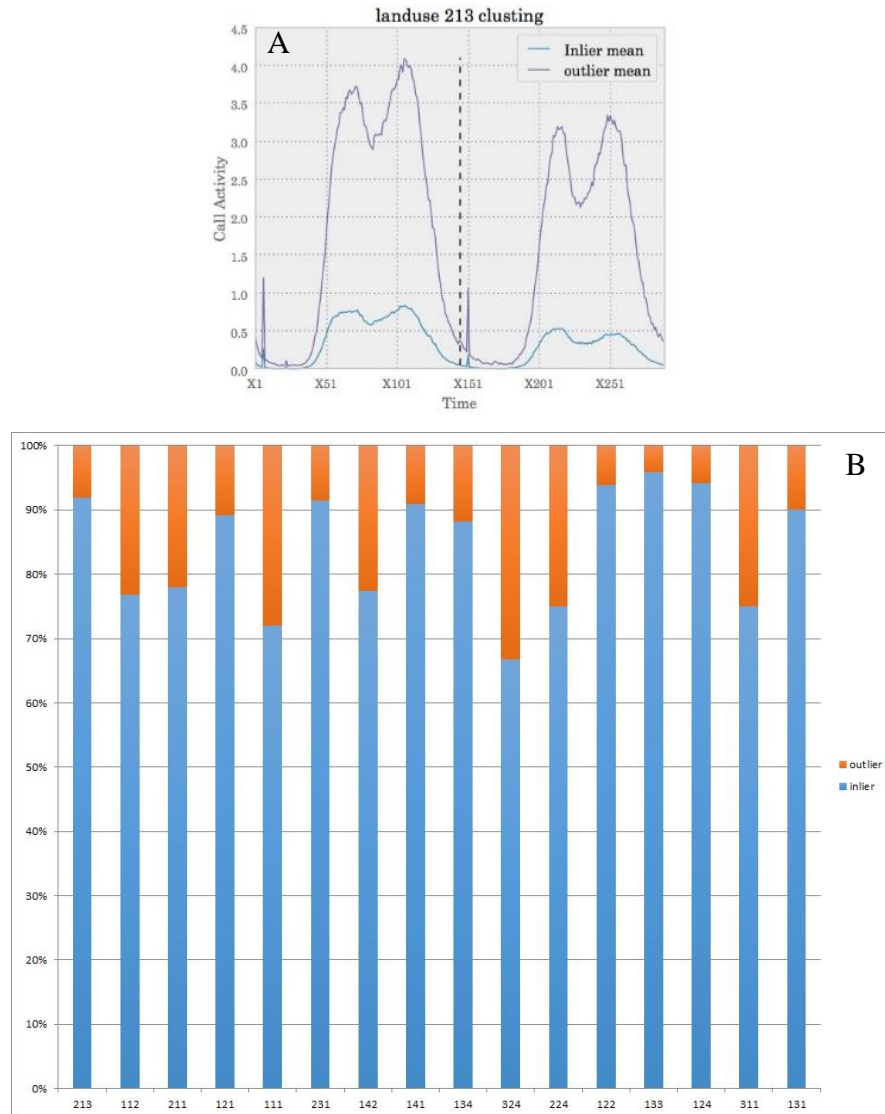


Figure 30 The difference percentage between outlier and inlier in each class of land use

After this clustering, we denoised data set with exclude all of outliers from previous phase's data and prepare new dataset for distance measuring phase.

3) Distance measure

This phase calculate distance between each cells and class of each land use. Again here we have to mention that the meaning of class of each land use here is each distribution of inliers cells with same prevalent land use. In order to calculate this distance we use Mahalanobis distance measure that explained in section 3.1.2. Instead of formula of Mahalanobis that we described in section 3.1.2 here we show the python code of this calculation:

```

...
list_of_landuse = inlier.prevaent.unique()
for x in list_of_landuse:
    #select all cells with same landuse in inliers cell
    class_of_x = inliers[inliers.landuse==x]
    #imake covariance inverce
    Covariance_inverce = np.linalg.pinv(class_of_x.cov())
    #mean of class of x
    mu =class_of_x.mean()
    #calculate mahalanobis distance between a nominatted cell with all class of landuse x
    output[name] = nominated_class.apply(lambda x: dist.mahalanobis(x,mu,C_inv),axis=1)
...

```

Figure 31 Snipped code for Mahalanobis distance calculation in python. In this snipped code some of the basic calculation are removed.

The output of this calculation is distance of each nominated cell with class of land uses. Finally by choosing two classes with least distance to the nominated cell we can predict land use of that cell. The output of this solution is shown in Table 8.

cell id	213	112	211	121	111	231	142	141	134	224	122	124	RESULT
1592	46384.2	70874.6	27.4	44.6	19.4	261.8	1527.1	255.1	167	282.6	149.6	11726.8	111,211
2447	8693.3	82999.6	111	33.7	577.2	2170.2	1810.7	363.3	162	136.5	394.5	4138.6	121,211
2548	27.2	41746.2	85.7	29.4	326.8	568.4	120	272.8	19.5	73.2	41.1	2645.9	213,134
...
9713	6081.2	3821.4	12.7	5.5	387.9	61.3	50.8	19.6	60.3	15.8	31.6	5017.2	121,211
9714	2396.7	3874.8	8.9	4	128.5	75.2	137.7	32.7	63.5	7.4	105.1	9453.5	121,224
9814	9571.7	9283.6	24.2	6.4	309.1	129.9	131.6	25.3	54.4	38.4	73.3	2255.8	121,211
9815	2860.3	3992.2	13.1	4.5	80.8	100.9	204.6	34.5	58.5	22	131	10595.3	121,211

Table 8 Output of solution 2

4.4 Comparison and discussion

As we discussed in 5.1 both proposed solutions have the same general structure and one basic assumption to solve the problem. So both of them have profile detection phase and similarity measure phase. But the methods for do these two phases are different as shown in Table 9. In general we can say that both of them are applicable to problem in this area, but they are depended on character of city in field of mobile usage and land use diversity.

	Profile detection phase	Similarity measure phase
Solution 1	Weighted Profiling	Adjusted Cosine Similarity
Solution 2	Denoised Cluster Profiling	Mahalanobis Distance

Table 9 Differences between solution 1 and 2

Solution-1 by using weighted profiling in one hand includes all of land uses of all cells to profiling procedure but in other hand create noisy profile. In opposed to solution-1, solution-2 makes filtered, denoised and pure profile class but in other hand loose about 60% of data. The type of profiles also is difference, in solution-1 we have just one signal of each profile but for second solution, we have a distribution of cell`s call-out signal as a class of profile.

At this stage, we could not easily conclude that which solution works better. Each of them has their advantages and disadvantages. Each of them can get better result in compare of another one in different situation. In the next chapter we evaluate these two solutions over our case study.

4.5 The proposed application

The proposed application, LivingLand use⁵, is an application for visualizing the results of our solutions. In the following we are going to explain the architecture, features and some implementation details of this application. The main interface of this application is a mobile-friendly interactive map that is coordinated to center of Milan ([45.4667, 9.1833]). Over this map there is different layers that show the gridded area of Milan with different aspect. By clicking over each cell in this grid the information of that cell will be shown in a popup over main interface such as predicted land use (result of our previous sections), signal of cell and etc. some main features of this application is listed as below:

1. Mobile friendly and responsive design
2. Model-View-Controller architecture
3. Real time charting
4. Safe and secure

⁵ <http://livingland use.cefruel.com>

Architecture and data flow of the application is depicted in Figure 32

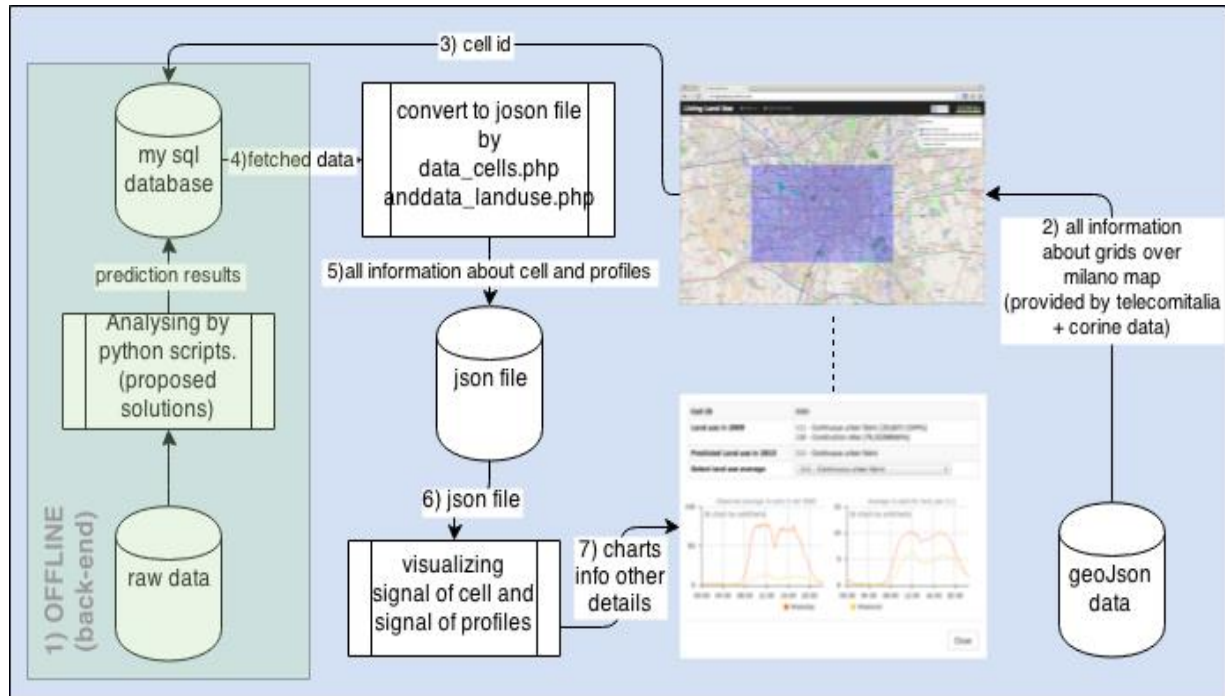


Figure 32 Livinglanduse application architecture and data flow

As we can see in figure above this application has seven step:

1. Offline data providing
2. Map initialization
3. Selection of a cell
4. Query over data base and fetch all information
5. Convert information to json file
6. Plotting and visualizing information
7. Show information as a pop up over map

In the following we are going to explain each step of this procedure.

1) Offline data providing

The main database of our application is filled by the analyzing activities that done in the previous sections. As we described, the raw data are converted and cleaned to the usable data sets and then analyzed by the different methods, solution 1 or solution 2, to create the final result that we represented in Table 6 and Table 8. These final results are stored into the database and then used by the web application. Notice that all the analyzing and data exploration executed by the professional related languages to data analysis such as Python and R in offline.

2) Map Initializing

This application benefits from Leaflet JavaScript library for visualizing interactive map. It used open source tile layer for map and GeoJson files for layers to initialize the main interface of this application. This interface has four layer of gridding which fetched from GeoJson files as below:

- Cells in Center of Milan
- Cells in Side of Milan
- Cells were Construction sites (CLC: 133) with prevalent land use more than 75% in 2009 which are colored by green
- Cells were Construction sites (CLC: 133) with prevalent land use more than 50% in 2009 which are colored by green

The conceptual model of this phase is depicted in Figure 33

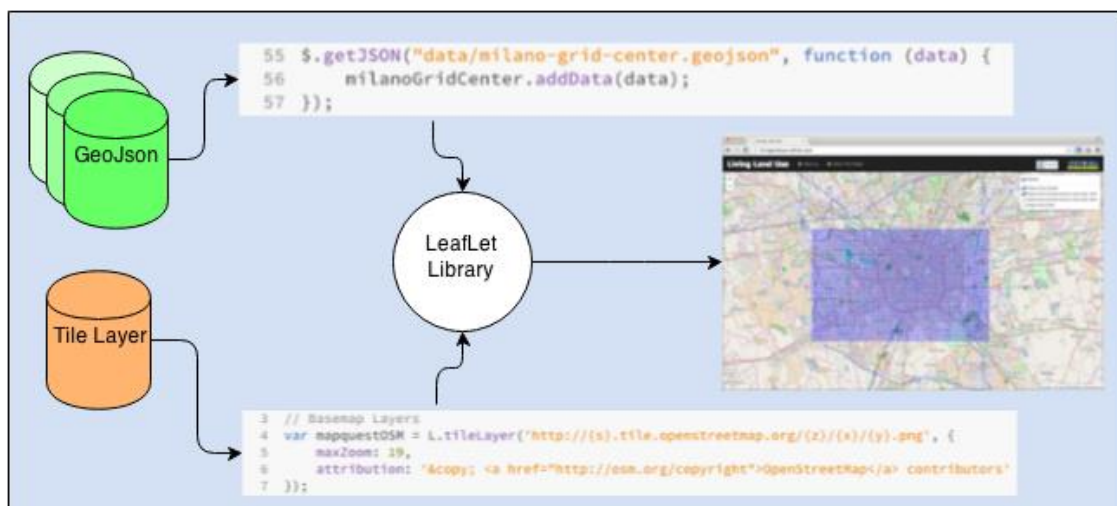


Figure 33 Conceptual model for initializing phase

3) Selection of a cell

As we described in section 3.2 analyzed area of Milano has 10000 cells that are 250 * 250 meters. By clicking over reach cell on map, cell id sends to the database to fetching all information about that.

4) Query over data base and fetch all information

All information about cells and prediction results are in database. For developing this application we used CodeIgniter framework. It uses a modified version of Active Record

Database Pattern (ARDP), this pattern allows information to be retrieved, insert and update on database with minimal scripting. Beyond of this simplicity, a major benefit to using Active Record features is that allows us to create database independent application that can make a safer query.

5) Convert information to json file

After fetching all information about a cell from database, they must be converted to the json file for visualizing by plotting library. In order to do this we implement two php scripts to convert this fetched information about cell and profiles. Figure shows php script for converting data about signal of a cell.

```

1  |<?php
2  $prefix = '';
3  echo "[\n";
4  foreach ( $cells as $row) {
5  echo $prefix . " {\n";
6  echo '  "cid": ' . $row->cellid . ', ' . "\n";
7  echo '  "time": ' . $row->d_time . ', ' . "\n";
8  echo '  "wd": ' . $row->wd_value . ', ' . "\n";
9  echo '  "we": ' . $row->we_value . ' ' . "\n";
10 echo " }";
11 $prefix = ",\n";
12 }
13 echo "\n]";
14 ?>

```

Figure 34 Translate retrieved data to json style. It retrieves weekday and weekend value of a cell in one day (10 min by 10 min)

```

[ { "cid": 6060, "time": "00:00", "wd": "2.19977808084602", "we": "2.82388995792008" }, { "cid": 6060, "time": "00:10",
"wd": "1.7962209665515", "we": "2.77648393093802" }, { "cid": 6060, "time": "00:20", "wd": "1.40564821483693", "we":
"2.47732509624602" }, { "cid": 6060, "time": "00:30", "wd": "1.28655066978581", "we": "2.10069385307516" }, { "cid": 6060,

```

Figure 35 Output of json file. It shows just weekday and week end value of first 30 min of the day for cell 6060

6) Plotting and visualizing information

This application uses AmCharts library. This library is one of the well-known open sources JavaScript charting libraries. It can plot and visualize information just by load them as a json file which is provided in previous step. Figure 36 and Figure 37 show some sniped code for loading and plotting data with AmCharts, respectively.

```

242 // load the data
243 var chartData = AmCharts.loadJSON(window.baseUrl+'index.php/data/index/'+id);
244 // SERIAL CHART
245 var chart = new AmCharts.AmSerialChart();
246 chart.dataProvider = chartData;
247 chart.categoryField = "time";

```

Figure 36 Load data which is provided from previous step to amchart

```

279 var graph1 = new AmCharts.AmGraph();
280 graph1.valueField = "wd";
281 graph1.title="Weekday";
282 chart.addGraph(graph1);
283
284 var graph2 = new AmCharts.AmGraph();
285 graph2.valueField = "we";
286 graph2.title="Weekend";
287 chart.addGraph(graph2);

```

Figure 37 Plot the loaded data for a cell

7) Show information as a pop up over map

All retrieved information and charts show in a pop up over map. As we can see in Figure 38, this pop up shows us:

- Cell id
- CORINE information of cell
- Predicted land use
- Profile of each land use. (For outputs of solution 2, instead of show distribution of land uses, it shows average of inlier of that land use)
- Cell's call out signal for weekdays and weekend

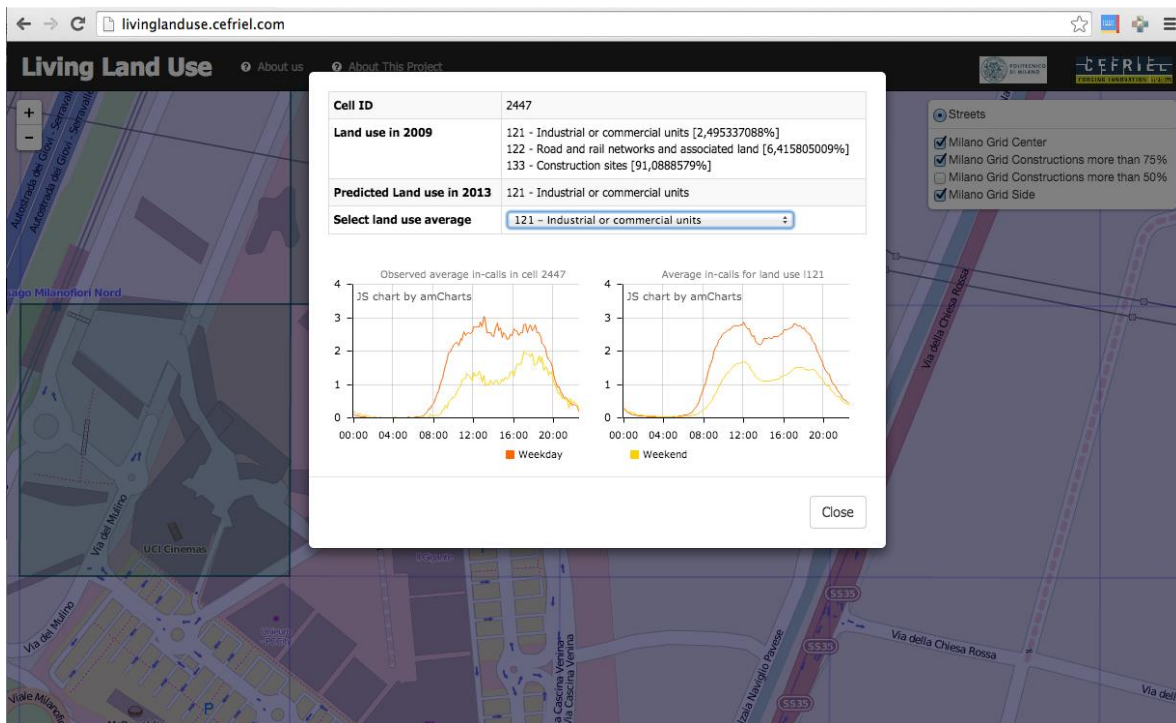


Figure 38 Visualized all information about a cell. Such as cell id, CORINE information, prediction result, signal of a cell, profile signal of land uses

Chapter 5

Evaluation

Ideally, any validation should be done using a ground truth to find out and check the accuracy of the training set's classification especially for supervised learning techniques. We do the process of gathering the objective data for our solution based on the Google service. Having information about the zones from urban planning departments could be more accurate than our information, but it is not available for us at the stage that we are writing this thesis. In this chapter, we evaluate and compare the described solution with some other basic methods (based lines) and check the prediction correctness of each method.

Ground Truth

In order to perform our evaluation, we make a selection of only 49 cells over 10000 which are all construction sites in 2009 with more than 75% prevalent (pure constructions sites), the reason to choose this number is logical since the probability of the change in construction sites are high and then we decided to considers only these cells as the case study.

In this stage we used the Google Street and Google Earth to create our ground truth by comparing the land usage of 2009 with the one provided by Google. Google recently has turned its Google Maps Street View into a time machine to let users travel back in time and see how places have changed in Figure 39 you can see the schema of this new technology. The new feature will let users track changes in landscape, buildings, roads and entire neighborhoods from around the world since the Street View mapping program began in 2007. That is a good feature for us to create the ground truth, the only problem is that for some zones the history data are not available. We try to use the data of the Google earth in some positions. The picture of the areas are available in the appendix B.

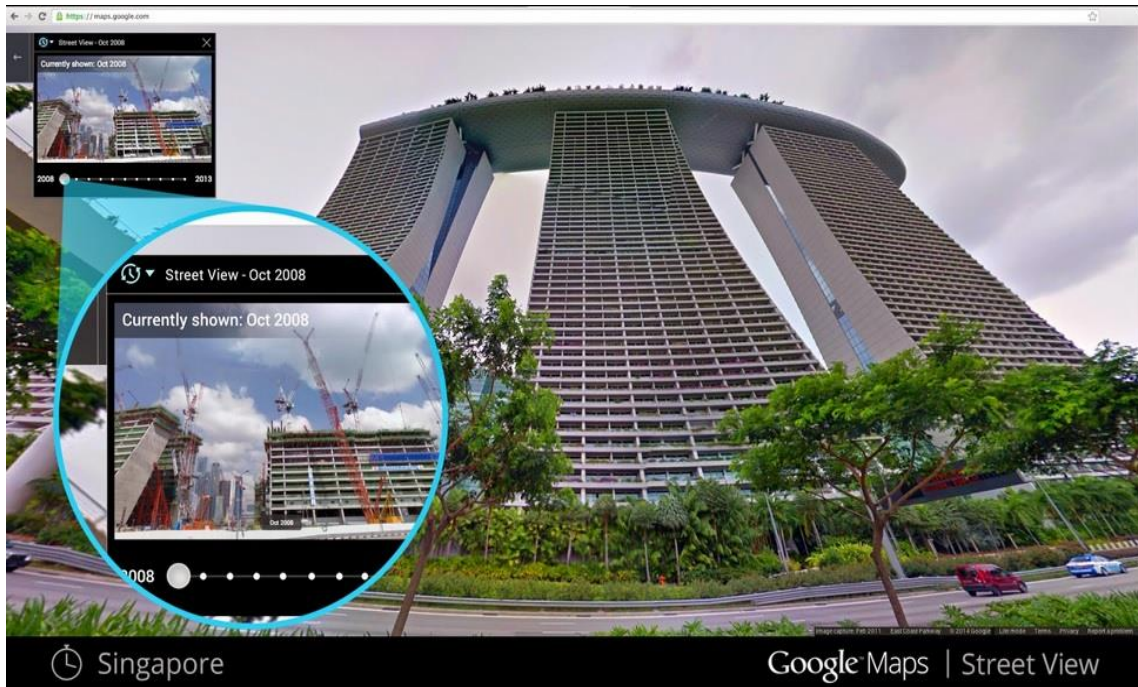


Figure 39 Google Maps Street View into a time machine⁶

In the Table 10 we are listing all the 49 selected current land uses in 2014. Then we use this information in order to compare with the result of different methods. In this table *Cell id* refers to the zone or cell which was a construction site in 2009, and *zone in Milano* column refers to the street which is in the middle of that cell, *current land uses* is defining the current land uses of that zone recently (2014) which are observed by Google map street time machine and the *CLC code* is chosen based on the CORRINE land use data set by our observation.

Notice that in this table the group of the colors means that the constructions sites are closed to each other. For example the cells 3777 to 4081 are all construction sites and are closed to LINATE airport.

⁶ <http://googleblog.blogspot.it/2014/04/go-back-in-time-with-street-view.html>

cell id	zone in milano	current landuses	CLC codes
1592	<i>via per Locate</i>	residential/Agricultural	111,211
2447	<i>Via del mulino</i>	Commercial/UCI	121
2548	<i>Via del mulino</i>	Commercial/Unused Aread/Rice field	121,134,213
3351	<i>Via Boffalora</i>	commercial/agricultural	121,211
3352	<i>Via Boffalora</i>	Commercial/Roads	121,122
3353	<i>Via Jan Palach</i>	commercial/agricultural/Roads/Residential	121,211,122,111
3452	<i>Via Don Ferrante</i>	Roads/Agricultural	122,211
3530	<i>Elsa morante</i>	commercial/agricultural/Roads/Residential/Unused	121,211,122,111,134
3691	<i>near Linate airport</i>	Airport/Commercial	124,121
3777	<i>Via del Futurismo</i>	Roads/Agricultural/Residential	122,211,111
3778	<i>Via del Futurismo</i>	Roads/Agricultural/Residential	122,211,111
3877	<i>Via Gino Severini</i>	Commercial/residential/agricultural	121,111,211
3878	<i>Via Alberto Savinio</i>	Roads/Agricultural	122,211
3879	<i>Via del Futurismo</i>	Roads	122
3881	<i>Via del Futurismo</i>	unused area	134
3978	<i>va luigi sordello</i>	unused area	134
3979	<i>va luigi sordello</i>	unused area	134
3980	<i>va luigi sordello</i>	unused area	134
3981	<i>va luigi sordello</i>	Aggricultural/Unused Aread	211,134
4079	<i>va luigi sordello</i>	commercial/Unused Aread	121,134
4080	<i>va luigi sordello</i>	Aggricultural/Unused Aread	211,134
4081	<i>va luigi sordello</i>	Aggricultural/Unused Aread	211,134
4999	<i>Microsoft, Longhignana</i>	commercial/agricultural	121,211
5648	<i>City life</i>	Commercial/unused	121,134
5649	<i>City life</i>	Commercial/unused	121,134
5749	<i>City life</i>	Commercial/residential/unused	121,111,134
6060	<i>via Gaetano de castillia</i>	Commercial	121
6146	<i>parco portello</i>	commercial/agricultural	121,211
6858	<i>Via Carlo Imbonati</i>	Commercial	121
6898	<i>Via caboto</i>	commercial/Aggricultural/Unused Aread	121,211,134
6980	<i>carlo cazzaniga</i>	Commercial	121
7187	<i>Via Olgettina</i>	Commercial	121
7246	<i>Filippo de pisis</i>	roads/unused area	122,134
7476	<i>Via Vittorio Gassman</i>	residential/unused	111,134
7477	<i>Via Ugo Tognazzi</i>	Aggricultural/Unused Aread	211,134
7478	<i>via publio Elio Adriano</i>	Aggricultural/Unused Aread	211,134
7575	<i>Via Roberto Tremelloni</i>	residential/unused	111,134
7576	<i>Via Vittorio Gassman</i>	residential/Agricultural	111,211
7577	<i>Via Vittorio Gassman</i>	Aggricultural/Unused Aread	211,134
7676	<i>Via Ugo Tognazzi</i>	Commercial	121
7677	<i>Via Ugo Tognazzi</i>	Commercial/unused	121,134
7975	<i>Tommas Edison</i>	Commercial	121
9081	<i>Via Vulcano</i>	Commercial	121
9281	<i>Via Vulcano</i>	Commercial	121
9477	<i>Via Panfilo Castaldi</i>	Commercial	121
9713	<i>Viale Alfa Romeo</i>	green area/commercial/Other permanent crops	211,141,121,224
9714	<i>Viale Alfa Romeo</i>	green area/commercial/Other permanent crops	211,141,121,224
9814	<i>Viale Alfa Romeo</i>	green area/commercial/Other permanent crops	211,141,121,224
9815	<i>Viale Alfa Romeo</i>	green area/commercial/Other permanent crops	211,141,121,224



Table 10 Milano Ground Truth details in 2013

Evaluation steps are done methods by methods and then we warp up all the methods together. Before going to the details of each method we are displaying the general structure of the evaluation in Figure 40. As you can see in the figure below the green cell is selected, in the proposed web application you can select and observe the 49 cells which are construction sites and have purity of 75% and see the result of the prediction.



Figure 40 Evaluation mechanism to validate proposed approach accuracy

After that by using the methods which we proposed in the solution chapters we compute the precision measurement for all the methods to compare them with some base lines.

The Table 11 represents the summary of prediction results which are described in details in the following. In this table we used some abbreviations: weighted for weighed random prediction, 1KM for top-1 in 1km, K for the number of samples in K-means++ clustering, CWP for solution 1, and CDCP for solution 2. The signs  and  are for correct and incorrect predictions respectively, and **PC** stands for prediction correctness which is the percentage of correct prediction.

Cell id	Random	weighted Random	1KM	cluster 5	cluster 16	cluster 24	CWP	CDCP1	CDCP2
1592	✗	✓	✓	✓	✗	✓	✗	✓	✓
2447	✗	✗	✗	✗	✗	✗	✓	✓	✓
2548	✗	✗	✗	✗	✗	✗	✓	✗	✓
3351	✗	✓	✗	✓	✓	✓	✓	✓	✓
3352	✗	✗	✗	✓	✓	✓	✓	✓	✓
3353	✗	✗	✗	✓	✓	✓	✓	✓	✓
3452	✗	✗	✗	✗	✗	✗	✗	✗	✓
3530	✗	✗	✓	✓	✓	✓	✓	✓	✓
3691	✗	✗	✓	✗	✗	✗	✓	✓	✓
3777	✗	✓	✗	✓	✗	✗	✓	✓	✓
3778	✗	✗	✓	✓	✗	✗	✓	✓	✓
3877	✓	✓	✓	✓	✓	✓	✓	✗	✓
3878	✓	✗	✓	✓	✗	✗	✗	✓	✓
3879	✗	✗	✓	✗	✗	✗	✗	✓	✓
3881	✗	✗	✗	✗	✗	✗	✗	✗	✗
3978	✗	✗	✗	✗	✗	✗	✗	✗	✓
3979	✗	✗	✗	✗	✗	✗	✗	✗	✓
3980	✗	✗	✗	✗	✗	✗	✗	✗	✓
3981	✗	✗	✗	✓	✗	✗	✗	✗	✓
4079	✗	✗	✓	✗	✓	✓	✓	✓	✓
4080	✗	✓	✗	✓	✗	✗	✗	✗	✓
4081	✗	✓	✗	✓	✗	✗	✗	✗	✓
4999	✗	✗	✓	✓	✓	✓	✗	✓	✓
5648	✗	✗	✗	✓	✓	✓	✓	✓	✓
5649	✗	✗	✗	✓	✓	✓	✓	✓	✓
5749	✗	✓	✓	✓	✓	✓	✓	✓	✓
6060	✗	✗	✗	✗	✗	✗	✓	✓	✓
6146	✓	✗	✗	✗	✓	✗	✓	✓	✓
6858	✗	✗	✗	✓	✗	✗	✓	✓	✓
6898	✗	✗	✓	✓	✓	✓	✓	✓	✓
6980	✗	✗	✗	✓	✓	✓	✓	✓	✓
7187	✗	✗	✗	✓	✗	✗	✓	✓	✓
7246	✓	✗	✗	✗	✗	✗	✓	✗	✓
7476	✗	✗	✓	✗	✗	✗	✗	✗	✗
7477	✗	✗	✗	✗	✗	✗	✗	✗	✓
7478	✗	✗	✗	✗	✗	✗	✗	✗	✓
7575	✗	✗	✗	✗	✗	✗	✓	✗	✗
7576	✗	✓	✗	✗	✗	✗	✓	✗	✗
7577	✓	✗	✗	✗	✗	✗	✗	✗	✓
7676	✗	✓	✓	✗	✓	✓	✓	✓	✓
7677	✗	✗	✓	✗	✓	✓	✓	✓	✓
7975	✗	✓	✓	✗	✓	✓	✓	✓	✓
9081	✗	✗	✗	✓	✓	✓	✗	✓	✓
9281	✗	✗	✗	✗	✓	✓	✗	✓	✓
9477	✗	✓	✓	✓	✓	✓	✗	✓	✓
9713	✗	✗	✓	✓	✓	✓	✗	✓	✓
9714	✗	✗	✓	✓	✓	✓	✗	✓	✓
9814	✗	✗	✓	✓	✓	✓	✗	✓	✓
9815	✗	✓	✓	✓	✓	✓	✗	✓	✓
PC	10.20%	24.50%	40.80%	51%	47%	45%	53%	65%	92%

Table 11 Summary result of all the methods with respect to the base lines

Baselines and proposed solution

We perform 4 different base lines which described below for the all 49 cells that we have in order to see how accurate our methods are.

- **Random:** In this method, we predict new land usage of each cell randomly by running 100 times.
- **Weighted random:** In this method we give the weighted value to each of the 24 land uses by percentage of them in Milan. As we know from data exploration the top three most land use are 211, 111, and 121. You can see the analyzed land use for all the

Milano in Figure 41. Then based on the weighted values, we predict each new land usage randomly but by including frequency properties of each land use type for selection.

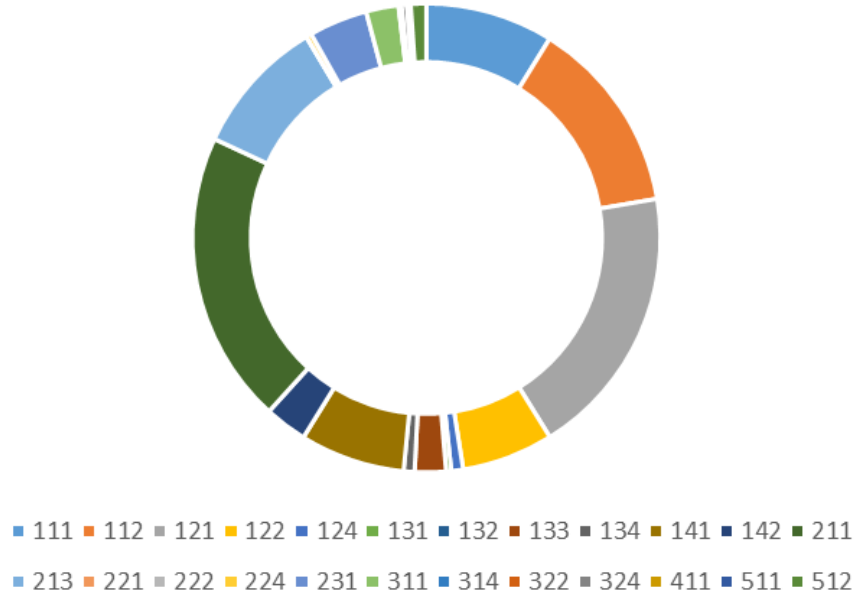




Figure 41 Analyzed land use for all the Milano

- **Top 1 in 1 KM:** In this method, we predict new land usage for selected cell by finding most used land use in radius of 1 KM from it by information of 2009.
- **Top-1 in K cluster:** In this method we cluster all the cells by k-means++ method with k equals to 5, 16, and 24. The same as pervious method we assign most frequented land use in that cluster to it. Then by checking the cluster of the selected cell we predict new land usage of that.

By having these base lines and our solutions we are going to list and discuss the result of each method one by one. In order to short the document we used four different tables to show all the results: (The signs  and  are for correct and incorrect predictions respectively, and **PC** stands for prediction correctness which is the percentage of correct prediction).

Random, weighted random and top 1 in 1 km

Table 12 displays the prediction result of these three base lines, as you could see in the table the prediction result is too low for random and weighted random ones, and better for Top 1 in 1 KM which are the cells near by the selected cell id.

cell id	Random	Predict	Weighted Random	Predict	Top 1 in 1 KM	Predict	Current
1592	311	✗	211	✓	211	✓	112,211
2447	322	✗	213	✗	213	✗	121
2548	222	✗	122	✗	213	✗	121,134,123
3351	322	✗	121	✓	112	✓	121,211
3352	131	✗	231	✗	112	✓	121,122
3353	324	✗	311	✗	112	✓	121,211,122,111
3452	311	✗	213	✗	112	✗	122,211
3530	112	✗	112	✗	121	✓	121,211,122,111,134
3691	322	✗	211	✗	121	✗	124,121
3777	134	✗	122	✓	121	✓	122,211,111
3778	133	✗	512	✗	122	✓	122,211,111
3877	111	✓	121	✓	121	✓	121,111,211
3878	122	✓	121	✗	122	✓	122,211
3879	131	✗	121	✗	122	✗	122
3881	132	✗	213	✗	133	✗	134
3978	322	✗	111	✗	121	✗	134
3979	111	✗	231	✗	133	✗	134
3980	322	✗	213	✗	133	✗	134
3981	511	✗	122	✗	133	✓	211,134
4079	112	✗	141	✗	121	✗	121,134
4080	324	✗	211	✓	133	✓	211,134
4081	322	✗	211	✓	133	✓	211,134
4999	314	✗	213	✗	211	✓	121,211
5648	222	✗	213	✗	111	✓	121,134
5649	411	✗	112	✗	111	✓	121,134
5749	311	✗	111	✓	111	✓	121,111,134
6060	112	✗	122	✗	111	✗	121
6146	121	✓	111	✗	111	✓	121,211
6858	511	✗	213	✗	111	✗	121
6898	314	✗	141	✗	211	✓	121,211,134
6980	142	✗	213	✗	111	✓	121
7187	222	✗	211	✗	211	✗	121
7246	134	✓	112	✗	121	✗	122,134
7476	231	✗	211	✗	111	✗	111,134
7477	314	✗	142	✗	111	✗	211,134
7478	231	✗	231	✗	111	✗	211,134
7575	221	✗	311	✗	121	✗	111,134
7576	131	✗	211	✓	121	✗	111,211
7577	134	✓	142	✗	121	✗	211,134
7676	133	✗	121	✓	121	✗	121
7677	222	✗	311	✗	121	✗	121,134
7975	133	✗	121	✓	121	✗	121
9081	511	✗	211	✗	134	✓	121
9281	224	✗	112	✗	112	✗	121
9477	141	✗	121	✓	121	✓	121
9713	112	✗	112	✗	211	✓	211,141,121,224
9714	222	✗	213	✗	121	✓	211,141,121,224
9814	231	✗	231	✗	121	✓	211,141,121,224
9815	311	✗	211	✓	121	✓	211,141,121,224
PC		10.20%		24.50%		40.80%	

Table 12 Random, Weighted random and Top 1 in 1 km result

Top 1 in k cluster (5, 16, and 24)

Table 13 displays the prediction results based on the Naïve solution that we discuss in 3.4 as you could see the results are not good enough.

cell id	cluster 5	predict	cluster 16	predict	cluster 24	predict	Current
1592	211	✓	121	✗	211	✓	112,211
2447	211	✗	211	✗	211	✗	121
2548	211	✗	211	✗	211	✗	121,134,123
3351	211	✓	121	✓	121	✓	121,211
3352	121	✓	121	✓	121	✓	121,122
3353	121	✓	121	✓	121	✓	121,211,122,111
3452	121	✗	121	✗	121	✗	122,211
3530	211	✓	121	✓	121	✓	121,211,122,111,134
3691	211	✗	211	✗	211	✗	124,121
3777	211	✓	121	✗	121	✗	122,211,111
3778	211	✓	121	✗	121	✗	122,211,111
3877	121	✓	121	✓	121	✓	121,111,211
3878	211	✓	121	✗	121	✗	122,211
3879	211	✗	121	✗	121	✗	122
3881	211	✗	121	✗	121	✗	134
3978	211	✗	121	✗	121	✗	134
3979	211	✗	121	✗	121	✗	134
3980	211	✗	121	✗	121	✗	134
3981	211	✓	121	✗	121	✗	211,134
4079	211	✗	121	✓	121	✓	121,134
4080	211	✓	121	✗	121	✗	211,134
4081	211	✓	121	✗	121	✗	211,134
4999	211	✓	211	✓	211	✓	121,211
5648	121	✓	121	✓	121	✓	121,134
5649	121	✓	121	✓	121	✓	121,134
5749	121	✓	121	✓	121	✓	121,111,134
6060	111	✗	111	✗	111	✗	121
6146	121	✓	121	✓	111	✗	121,211
6858	111	✗	111	✗	111	✗	121
6898	211	✓	211	✓	211	✓	121,211,134
6980	121	✓	121	✓	121	✓	121
7187	211	✗	211	✗	211	✗	121
7246	211	✗	121	✗	211	✗	122,134
7476	211	✗	121	✗	121	✗	111,134
7477	121	✗	121	✗	121	✗	211,134
7478	121	✗	121	✗	121	✗	211,134
7575	121	✗	121	✗	121	✗	111,134
7576	121	✗	121	✗	121	✗	111,211
7577	121	✗	121	✗	121	✗	211,134
7676	211	✗	121	✓	121	✓	121
7677	211	✗	121	✓	121	✓	121,134
7975	211	✗	121	✓	211	✗	121
9081	121	✓	121	✓	121	✓	121
9281	211	✗	121	✓	121	✓	121
9477	121	✓	121	✓	121	✓	121
9713	211	✓	211	✓	211	✓	211,141,121,224
9714	211	✓	211	✓	211	✓	211,141,121,224
9814	211	✓	211	✓	211	✓	211,141,121,224
9815	211	✓	211	✓	211	✓	211,141,121,224
PC		51%		47%		45%	

Table 13 Top 1 in 5 cluster and 16 cluster and 24 cluster

Solution 1: Comparison to weighted profile (CWP)

Table 14 displays the result of the first solution that we introduced in 4.2. The result of this evaluation is better than the above ones.

<i>cell id</i>	<i>CWP</i>	<i>Predict</i>	<i>Current</i>
1592	122,121	✗	112,211
2447	121,133	✓	121
2548	134,512	✓	121,134,123
3351	111,121	✓	121,211
3352	111,121	✓	121,122
3353	111,121	✓	121,211,122,111
3452	111,121	✗	122,211
3530	111,124	✓	121,211,122,111,134
3691	121,133	✓	124,121
3777	111,121	✓	122,211,111
3778	111,121	✓	122,211,111
3877	111,121	✓	121,111,211
3878	111,121	✗	122,211
3879	111,121	✗	122
3881	111,121	✗	134
3978	111,121	✗	134
3979	111,121	✗	134
3980	111,121	✗	134
3981	111,121	✗	211,134
4079	111,121	✓	121,134
4080	111,121	✗	211,134
4081	111,121	✗	211,134
4999	141,134	✗	121,211
5648	111,121	✓	121,134
5649	111,121	✓	121,134
5749	111,121	✓	121,111,134
6060	111,121	✓	121
6146	111,121	✓	121,211
6858	111,121	✓	121
6898	121,122	✓	121,211,134
6980	111,121	✓	121
7187	121,122	✓	121
7246	121,122	✓	122,134
7476	121,122	✗	111,134
7477	111,121	✗	211,134
7478	111,121	✗	211,134
7575	111,121	✓	111,134
7576	111,121	✓	111,211
7577	111,121	✗	211,134
7676	121,122	✓	121
7677	121,122	✓	121,134
7975	121,122	✓	121
9081	111,122	✗	121
9281	111,122	✗	121
9477	111,122	✗	121
9713	231,511	✗	211,141,121,224
9714	411,324	✗	211,141,121,224
9814	512,134	✗	211,141,121,224
9815	231,511	✗	211,141,121,224
PC		53%	

Table 14 The result of the CWP

Solution 2: Comparison to denoised clustered profile (CDCP)

Table 15 displays the results of the second solutions, we choose first and second nearest distance to evaluate the solution more precise.

<i>cell id</i>	<i>CDCP1</i>	<i>predict</i>	<i>CDCP2</i>	<i>predict</i>	<i>Current</i>
1592	111	✓	111,211	✓	112,211
2447	121	✓	121,211	✓	121
2548	213	✗	213,134	✓	121,134,123
3351	121	✓	121,211	✓	121,211
3352	121	✓	121,122	✓	121,122
3353	121	✓	121,122	✓	121,211,122,111
3452	121	✗	121,211	✓	122,211
3530	121	✓	121,211	✓	121,211,122,111,134
3691	124	✓	124,121	✓	124,121
3777	122	✓	122,121	✓	122,211,111
3778	122	✓	122,121	✓	122,211,111
3877	122	✗	122,121	✓	121,111,211
3878	122	✓	122,121	✓	122,211
3879	122	✓	122,121	✓	122
3881	121	✗	121,122	✗	134
3978	121	✗	121,134	✓	134
3979	121	✗	121,134	✓	134
3980	121	✗	121,134	✓	134
3981	121	✗	121,211	✓	211,134
4079	121	✓	121,134	✓	121,134
4080	121	✗	121,211	✓	211,134
4081	121	✗	121,211	✓	211,134
4999	121	✓	121,211	✓	121,211
5648	121	✓	121,134	✓	121,134
5649	121	✓	121,134	✓	121,134
5749	121	✓	121,111	✓	121,111,134
6060	121	✓	121	✓	121
6146	121	✓	121,211	✓	121,211
6858	121	✓	121	✓	121
6898	121	✓	121,134	✓	121,211,134
6980	121	✓	121,134	✓	121
7187	121	✓	121,134	✓	121
7246	121	✗	121,134	✓	122,134
7476	121	✗	121,122	✗	111,134
7477	121	✗	121,211	✓	211,134
7478	121	✗	121,211	✓	211,134
7575	121	✗	121,122	✗	111,134
7576	121	✗	121,122	✗	111,211
7577	121	✗	121,211	✓	211,134
7676	121	✓	121,122	✓	121
7677	121	✓	121,134	✓	121,134
7975	121	✓	121	✓	121
9081	121	✓	121	✓	121
9281	121	✓	121	✓	121
9477	121	✓	121	✓	121
9713	121	✓	121,211	✓	211,141,121,224
9714	121	✓	121,224	✓	211,141,121,224
9814	121	✓	121,211	✓	211,141,121,224
9815	121	✓	121,211	✓	211,141,121,224
PC		65%		92%	

Table 15 The result of CDCP 1 which displays the first lowest distance and the CDCP 2 which displays the second lowest distance

In the Figure 42 you can see the final result of each method as the matter of their prediction correctness, as it displays the last proposed method has the highest percentage of the prediction.

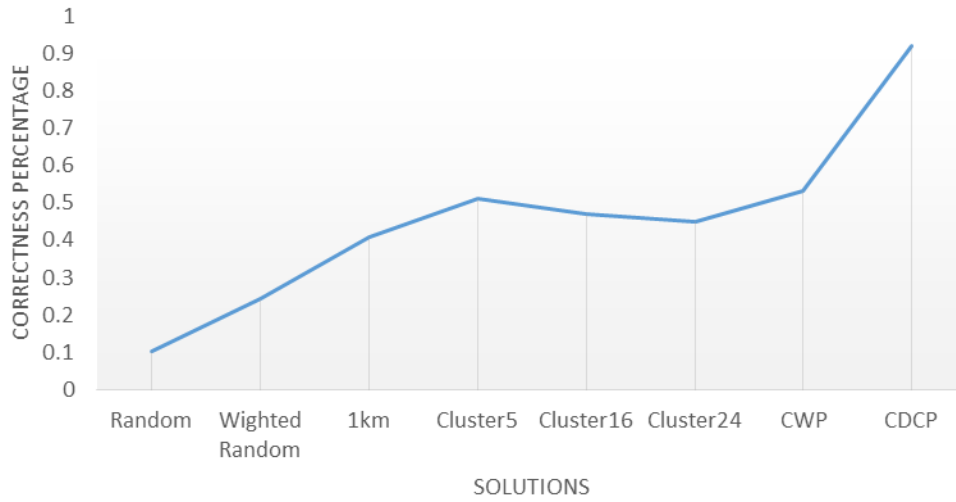


Figure 42 The improvement of the result base on the approaches

By considering this evaluation and results that we have, we can say that the method is working for the defined case study well. In the next chapter we conclude our solutions and we see the barriers we faced during this thesis work.

Chapter 6

Conclusion and Future work

In this thesis we proposed two solutions to predict land use of a specific area by using of CDR data and old CORINE data. While traditional system are based on questionnaires, and costs too much in both money and time, our solutions try to solve these problems and bring new advantages like the capability of tracing land use evolution over time or the ability to focus the study in a particular social background (elders, tourists, socio-economic levels, etc.). In order to improve and complement the classification of land uses, new methods based on the data mining approaches are introduced which mostly are based on the telecommunication data. Recent methods in this field are based on clustering, what we mean is to try to do some unsupervised clustering and assigning each land use to one of the clusters as the indicator of that cluster. And then by clustering try to identify the land use, this is the work that recent researches try to do. In other words, they are obtaining land use signature from CDR data. These method's outputs do not have enough quality for our case study because of the size of the CDR data, Heterogeneity in communication, Diversity of Milano land uses, and Characteristic of the city and mobile phone data. Tuning the clustering method is one of the solution that can be come up at the first sight. Instead of the pure clustering method our method is based on the CORINE properties that is information about land uses of Milan area in 2009, and using supervised learning.

Our proposed solutions have same general structure to solve described problem. But the difference between these solutions are their profile detection mechanism and similarity measurement which is summarized in the Table 16.

		Solution1	Solution2
Profile detection	Date entry	All cells	Cells with prevalent more than 75%
	Mechanism	Weighted profiling	Denoised by clustering
	output	A single signal as a profile for each land use	Distribution of cell's signal as a profile for each land use
Similarity measure		Adjusted cosine	Mahalanobis distance

Table 16 Comparison of solutions

One of the impressive benefits of our approaches is to predict the changes in land uses. In order to evaluate our claim, we define a set of cells as a ground truth which were mostly construction sites in 2009. The evaluation results shows us that first and second solution predict 54% and 92% of the land uses correctly. Generally we could say that our approaches is not intended to substitute traditional urban analysis approaches but they could be a useful tools to complement and improve them.

6.1 Future works

Future works that could be defined in the area of the land use identification is the following:

- **Adding support for multiple comparisons of the elicited land use:** by repeating the analysis over different time frames to understand seasonal (winter vs. summer) or longer-term variability of behavior and results.
- **Enhancing land use footprints representation and computation:** Including additional data sources (e.g. public transport, bike sharing, other location-based social media, hotel/museum/restaurant bookings, event-specific information); Combining and harmonizing data from different sources into a coherent land use footprint.
- **Introducing a paradigm shift from batch processing to continuous analysis:** Incrementally updating land use footprint computation on specific interval batches;

continuously analyzing streaming data in quasi-real time (feasible only on specific real-time data sources that do not require pre-processing)

Bibliography

- [1] Frias-Martinez, Vanessa & Soto, Victor & Hohwald, Heath & Frias-Martinez, Enrique, "Characterizing urban landscapes using geolocated tweets," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, 2012.
- [2] "Coordination of Information on the Environment (CORINE) of the European Environment Agency (EEA)," 10 August 2014. [Online]. Available: <http://en.wikipedia.org/wiki/CORINE>.
- [3] Zheng, Yu & Capra, Licia & Wolfson, Ouri & Yang, Hai, "Urban Computing: concepts, methodologies, and applications," *ACM Transaction on Intelligent Systems and Technology (ACM TIST)*, 2014.
- [4] Yuan, J. & Zheng, Y. & Xie, X. & Sun, G. , "T-Drive: Enhancing driving directions with taxi drivers' intelligence.," *Knowledge and Data Engineering*, vol. 25, no. IEEE, pp. 220-232, 2013.
- [5] Gong, Peng & HOWARTH, PHILIPJ & etc., "The use ostructural information for improving land-cover classification accuracies at the rural-urban fringe," *Photogrammetric engineering and remote sensing*, vol. 56, pp. 67-73, 1990.
- [6] P. Fisher, "The pixel: a snare and a delusion," *International Journal of Remote Sensing*, vol. 18, no. Taylor Francis, pp. 679-685, 1997.

- [7] Shaban, MA & Dikshit, o, "Improvement of classification in urban areas by the use of textural features: the case study of Lucknow city, Uttar Pradesh," *International Journal of Remote Sensing*, no. Taylor Francis, pp. 565-593, 2001.
- [8] Lu, Dengsheng & Weng, Qihao, "Use of impervious surface in urban land-use classification," *Remote Sensing of Environment*, vol. 102, no. Elsevier, pp. 146-160, 2006.
- [9] Gonzalez, Marta C & Hidalgo, Cesar A & Barabasi, Albert-Laszlo, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. Nature Publishing Group, pp. 779-782, 2008.
- [10] Song, Chaoming & Qu, Zehui & Blumm, Nicholas & Barabasi, Albert-Laszlo, "Limits of predictability in human mobility," *Science*, no. American Association for the Advancement of Science, pp. 1018-1021, 2010.
- [11] Calabrese, Francesco & Di Lorenzo, Giusy & Liu, Liang & Ratti, Carlo, "Estimating origin-destination flows using mobile phone location data," *Pervasive Computing, IEEE*, 2011.
- [12] Reades, Jonathan & Calabrese, Francesco & Ratti, Carlo, "Eigenplaces: analysing cities using the space-time structure of the mobile phone network," *Environment and Planning B: Planning and Design*, no. Pion Ltd, London, pp. 824-836, 2009.
- [13] Calabrese, Francesco & Reades, Jonathan & Ratti, Carlo, "Eigenplaces: segmenting space through digital signatures," *Pervasive Computing, IEEE*, vol. 9, no. IEEE, pp. 78-84, 2010.
- [14] Caceres, Ramon & Rowland, James & Small, Christopher & Urbanek, Simon, "Exploring the use of urban greenspace through cellular network activity," *Proc. of 2nd Workshop on Pervasive Urban Applications (PURBA)*, 2012.

- [15] Soto, Victor & Fias-Marinez, Enrique, "Automated land use identification using cell-phone records," *Proceedings of the 3rd ACM international workshop on MobiArch*, pp. 17-22, 2011.
- [16] Ratti, Carlo & Williams, S & Frenchman, D & Pulselli, RM, "Mobile landscapes: using location data from cell phones for urban analysis," *Environment and Planning B Planning and Design*, vol. 33, no. PION LTD, p. 727, 2006.
- [17] Eagle, Nathan & Pentland, Alex, "Reality mining: sensing complex social systems, volume 10(4), 2006," *Personal and ubiquitous computing*, pp. 255-268, 2006.
- [18] Shamoo, Adil E & Resnik, David B, *Responsible conduct of research*, Oxford University Press, 2003.
- [19] L. A. Gottschalk, "Content analysis of verbal behavior: New findings and clinical applications," *Routledge*, 2014.
- [20] Savenye, Wilhelmina C & Robinson, Rhonda S, "Using qualitative research methods in higher education," *Journal of computing in Higher education*, no. Springer, pp. 65-95, 2005.
- [21] Shephard, Roy J, "Ethics in exercise science research," *Sports Medicine*, no. Springer, pp. 169-183, 2002.
- [22] Hastie, Trevor & Tibshirani, Robert & Friedman, Jerome & Hastie, T & Friedman, J and Tibshirani, R, *The Elements of Statistical Learning*, Springer, 2009.
- [23] "k-means-clustering," 1 september 2014. [Online]. Available: <https://mahout.apache.org/users/clustering/k-means-clustering.html>.
- [24] "Clustering," 10 August 2014. [Online]. Available: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/.

- [25] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281-297, 1967.
- [26] "KMEANS CLUSTERING," 2 september 2014. [Online]. Available: <http://scikitlearn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans>.
- [27] Arthur, David & Vassilvitskii, Sergei, "How Slow is the k-means Method?," *Proceedings of the twenty-second annual symposium on Computational geometry*, pp. 144-153, 2006.
- [28] Durak, Bahadir, "A classification Algorithm Using mahalanobis Distance Clustering of Data With Applications on Biomedical Data Sets," 2011.
- [29] De Maesschalck, Roy & Jouan-Rimbaud, Delphine & Massart, "The mahalanobis distance," *Chemometrics and intelligent laboratory systems*, vol. 50, pp. 1-18, 2000.
- [30] Lu, Yun & Zhang, Mingjin & Li, Tao & Guang, Yudong & Rishe, Naphtali, "Online spatial data analysis and visualization system," *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics*, pp. 71-78, 2013.
- [31] "CodeIgniter," 10 August 2014. [Online]. Available:<http://www.phpframeworks.com/php-frameworks/index.php?id=9+CodeIgniter>.
- [32] "twitters-bootstrap," 1 september 2014. [Online]. Available: <http://www.tectale.com/an-introduction-to-twitters-bootstrap/>.
- [33] Thung, Phek Lan & Ng, Chu Jian & Thung, Swee Jing & Sulaiman, Shahida, "Improving a web application using design patterns: A case study," *Information Technology (ITSim), 2010 International Symposium in*, vol. 1, no. IEEE, pp. 1-6, 2010.
- [34] "TELECOM big data challenge," 1 september 2014. [Online]. Available: <http://www.telecomitalia.com/tit/en/bigdatachallenge.html>.

- [35] "epoch converter," 15 August 2014. [Online]. Available: <http://www.epochconverter.com/>.
- [36] "CLC2006 technical guidelines," 10 August 2014. [Online]. Available: http://www.eea.europa.eu/publications/technical_report_2007_17. [Accessed August 2014].
- [37] "Uso-Suolo-Dusaf," 1 september 2014. [Online]. Available: <https://www.dati.lombardia.it/Territorio/Uso-Suolo-Dusaf-2009/y6xb-wpka>.
- [38] "Hectare," 1 september 2014. [Online]. Available: <http://en.wikipedia.org/wiki/Hectare#Are>. [Accessed August 2014].
- [39] Candia, J. & Gonzalez, M. & Wans, P. & Schoenharl, T. & Barabasi, A.L., "Uncovering individual and collective human dynamics from mobile phone records. In: J. Phys. A: Math. Theor. Volume 41. (2008)," vol. 41, 2008.
- [40] Pei, Tao & Sobolevsky, Stanislav & Ratti, Carlo & Shaw, Shih-Lung & Li, Ting & Zhou, Chenghu, "A new insight into land use classification based on aggregated mobile phone data," *International Journal of Geographical Information Science*, no. Taylor Francis, pp. 1-20, 2014.
- [41] Zhong, Chen & Arisona, Stefan Muller & Huang, Xianfeng & Schmitt, Gerhard, "Identifying spatial structure of urban functional centers using travel survey data: a case study of Singapore," in *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place*, 2013, pp. 28-33.
- [42] Toole, Jameson L & Ulm, Michael & Gonzalez, Marta C & Bauer, Dietmar, "Inferring land use from mobile phone activity," *Proceedings of the ACM SIGKDD international workshop on urban computing*, pp. 1-8, 2012.
- [43] Rhind, David & Hudson, Ray & others, Land use, Methuen., 1980.
- [44] Tranos, Emmanouil & Steenbruggen, John & Nijkamp, Peter, "Mobile Phone Data and Urban Analysis: An Exploratory Space--Time Study," 2013.

- [45] Soto, Victor & Frias-Martinez, Enrique, "Robust land use characterization of urban landscapes using cell phone data," in *1st Workshop on Pervasive Urban Applications, in conjunction with 9th Int. Conf. Pervasive Computing*, 2011.
- [46] Marshall, Elizabeth & Shortle, J, "Urban development impacts on ecosystems," *Land use problems and conflicts: causes, consequences and solutions*. Routledge, New York, pp. 79-93, 1005.

Appendix A

Full description of each land use.

1. Artificial surfaces

1.1. Urban fabric

1. 1. 1. Continuous urban fabric

Most of the land is covered by Buildings, roads and artificially surfaced area cover almost all the ground.

Non-linear areas of vegetation and bare soil are exceptional.

1.1.2. Discontinuous urban fabric

Most of the land is covered by structures. Buildings, roads and artificially surfaced areas associated with vegetated areas and bare soil, which occupy discontinuous but significant surfaces.

1.2. Industrial, commercial and transport

1.2.1. Industrial or commercial units

Artificially surfaced areas (with concrete, asphalt, tarmacadam, or stabilized, e.g. beaten earth) devoid of vegetation, occupy most of the area in question, which also contains buildings and/or vegetated areas.

1.2.2. Road and rail networks and associated land

Motorways, railways, including associated installations (stations, platforms, embankments).
Minimum width to include: 100 m.

1.2.3. Port areas

Infrastructure of port areas, including quays, dockyards and marinas.

1.2.4. Airports

Airport installations: runways, buildings and associated land.

1.3. Mine, dump and construction sites

1.3.1. Mineral extraction sites

Areas with open-pit extraction of industrial minerals (sandpits, quarries) or other minerals (opencast mines). Includes flooded gravel pits, except for river-bed extraction.

1.3.2. Dump sites

Landfill or mine dump sites, industrial or public.

1.3.3. Construction sites

Spaces under construction development, soil or bedrock excavations, earthworks.

1.4. Artificial, non-agricultural vegetated areas

1.4.1. Green urban areas

Areas with vegetation within urban fabric. Includes parks and cemeteries with vegetation.

1.4.2. Sport and leisure facilities

Camping grounds, sports grounds, leisure parks, golf courses, racecourses, etc. Includes formal parks not surrounded by urban zones.

2. Agricultural areas

2.1. Arable land

Cultivated areas regularly ploughed and generally under a rotation system.

2.1.1. Non-irrigated arable land

Cereals, legumes, fodder crops, root crops and fallow land. Includes flower and tree (nurseries) cultivation and vegetables, whether open field, under plastic or glass (includes market gardening). Includes aromatic, medicinal and culinary plants. Excludes permanent pastures.

2.1.2. Permanently irrigated land

Crops irrigated permanently and periodically, using a permanent infrastructure (irrigation channels, drainage network). Most of these crops could not be cultivated without an artificial water supply. Does not include sporadically irrigated land.

2.1.3. Rice fields

Land developed for rice cultivation. Flat surfaces with irrigation channels. Surfaces regularly flooded.

2.2. Permanent crops

Crops not under a rotation system which provide repeated harvests and occupy the land for a long period before it is ploughed and replanted: mainly plantations of woody crops. Excludes pastures, grazing lands and forests.

2.2.1. Vineyards

Areas planted with vines.

2.2.2. Fruit trees and berry plantations

Parcels planted with fruit trees or shrubs: single or mixed fruit species, fruit trees associated with permanently grassed surfaces. Includes chestnut and walnut groves.

2.2.3. Olive groves

Areas planted with olive trees, including mixed occurrence of olive trees and vines on the same parcel.

2.3. Pastures

2.3.1. Pastures

Dense, predominantly graminoid grass cover, of floral composition, not under a rotation system. Mainly used for grazing, but the fodder may be harvested mechanically. Includes areas with hedges (bocage).

2.4. Heterogeneous agricultural areas

2.4.1. Annual crops associated with permanent crops

Non-permanent crops (arable lands or pasture) associated with permanent crops on the same parcel.

2.4.2. Complex cultivation

Juxtaposition of small parcels of diverse annual crops, pasture and/or permanent crops.

2.4.3. Land principally occupied by agriculture, with significant areas of natural vegetation

Areas principally occupied by agriculture, interspersed with significant natural areas.

2.4.4. Agro-forestry areas

Annual crops or grazing land under the wooded cover of forestry species.

3. Forests and semi-natural areas

3.1. Forests

3.1.1. Broad-leaved forest

Vegetation formation composed principally of trees, including shrub and bush understories, where broadleaved species predominate.

3.1.2. Coniferous forest

Vegetation formation composed principally of trees, including shrub and bush understories, where coniferous species predominate.

3.1.3. Mixed forest

Vegetation formation composed principally of trees, including shrub and bush understories, where broadleaved and coniferous species co-dominate.

3.2. Shrub and/or herbaceous vegetation associations

3.2.1. Natural grassland

Low productivity grassland. Often situated in areas of rough uneven ground. Frequently includes rocky areas, briars, and heathland.

3.2.2. Moors and heathland

Vegetation with low and closed cover, dominated by bushes, shrubs and herbaceous plants (heath, briars, broom, gorse, laburnum, etc.).

3.2.3. Sclerophyllous vegetation

Bushy sclerophyllous vegetation. Includes *maquis and garrigue*. *Maquis*: a dense vegetation association composed of numerous shrubs associated with siliceous soils in the Mediterranean environment.

Garrigue: discontinuous bushy associations of Mediterranean calcareous plateaus. Generally composed of kermes oak, arbutus, lavender, thyme, cistus, etc. May include a few isolated trees.

3.2.4. Transitional woodland/shrub

Bushy or herbaceous vegetation with scattered trees. Can represent either woodland degradation or forest regeneration/colonization.

3.3. Open spaces with little or no vegetation

3.3.1. Beaches, dunes, and sand plains

Beaches, dunes and expanses of sand or pebbles in coastal or continental, including beds of stream channels with torrential regime.

3.3.2. Bare rock

Scree, cliffs, rocks and outcrops.

3.3.3. Sparsely vegetated areas

Includes steppes, tundra and badlands. Scattered high-attitude vegetation.

3.3.4. Burnt areas

Areas affected by recent fires, still mainly black.

3.3.5. Glaciers and perpetual snow

Land covered by glaciers or permanent snowfields.

4. Wetlands

4.1. Inland wetlands

Non-forested areas either partially, seasonally or permanently waterlogged. The water may be stagnant or circulating.

4.1. 1. Inland marshes

Low-lying land usually flooded in winter, and more or less saturated by water all year round.

4.1.2. Peatbogs

Peatland consisting mainly of decomposed moss and vegetable matter. May or may not be exploited.

4.2. Coastal wetlands

Non-wooded areas either tidally, seasonally or permanently waterlogged with brackish or saline water.

4.2.1. Salt marshes

Vegetated low-lying areas, above the high-tide line, susceptible to flooding by sea water. Often in the process of filling in, gradually being colonised by halophilic plants.

4.2.2. Salines

Salt-pans, active or in process of. Sections of salt marsh exploited for the production of salt by evaporation. They are clearly distinguishable from the rest of the marsh by their segmentation and embankment systems.

4.2.3. Intertidal flats

Generally unvegetated expanses of mud, sand or rock lying between high and low water-marks. On contour on maps.

5. Water bodies

5.1. Inland waters

5.1. 1. Water courses

Natural or artificial water-courses serving as water drainage channels. Includes canals. Minimum width to include: 100 m.

5.1.2. Water bodies

Natural or artificial stretches of water.

5.2. Marine waters

5.2.1. Coastal lagoons

Unvegetated stretches of salt or brackish waters separated from the sea by a tongue of land or other similar topography. These water bodies can be connected with the sea at limited points, either permanently or for parts of the year only.

5.2.2. Estuaries

The mouth of a river within which the tide ebbs and flows.

5.2.3. Sea and ocean

Zone seaward of the lowest tide limit.

Appendix B

The last updated pics of Google for the cells with the land use of 133 (construction site) more than 75% in 2009 and current land use.

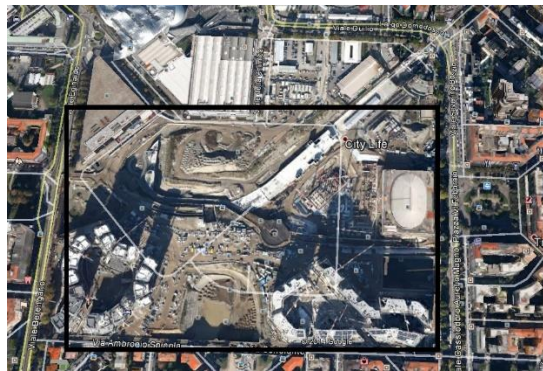
Cell id: 6146 near parco portello – construction site to commercial and green areas



Cell id: 6858 Via Carlo Imbonati – construction site to commercial and residential



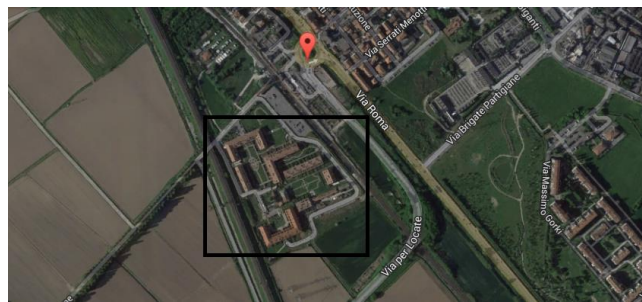
Cell id: 5648, 5649, 5749 City Life project – construction site to commercial and residential



Cell id: 4999 Microsoft, Longhignana – commercial and agricultural



Cell id: 1592 near Via per Locate – commercial and agricultural



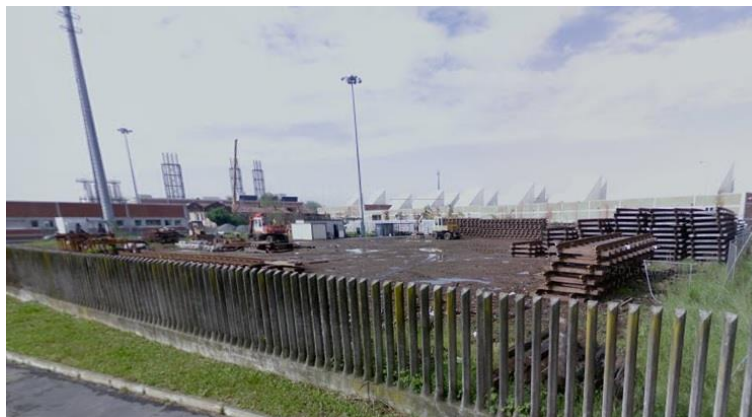
Cell id: 2447, 2548 near via del mulino – construction site to commercial



Cell id: 3452 near Via Don Ferrante – construction site to Residential and green areas



Cell id: 3352 near Via Boffalora – construction site



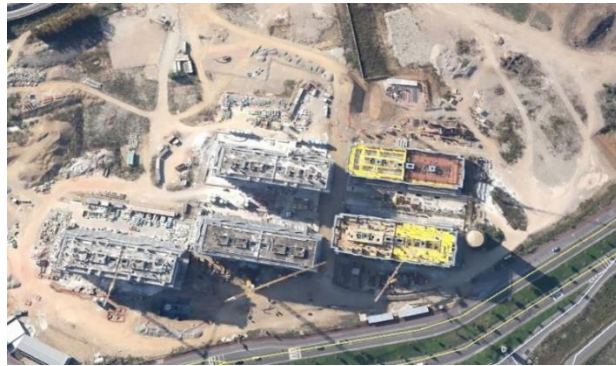
Cell id: 3530 near via elsa morante – construction site



Cell id: 9477 near Via Panfilo Castaldi – construction site to residential and green areas



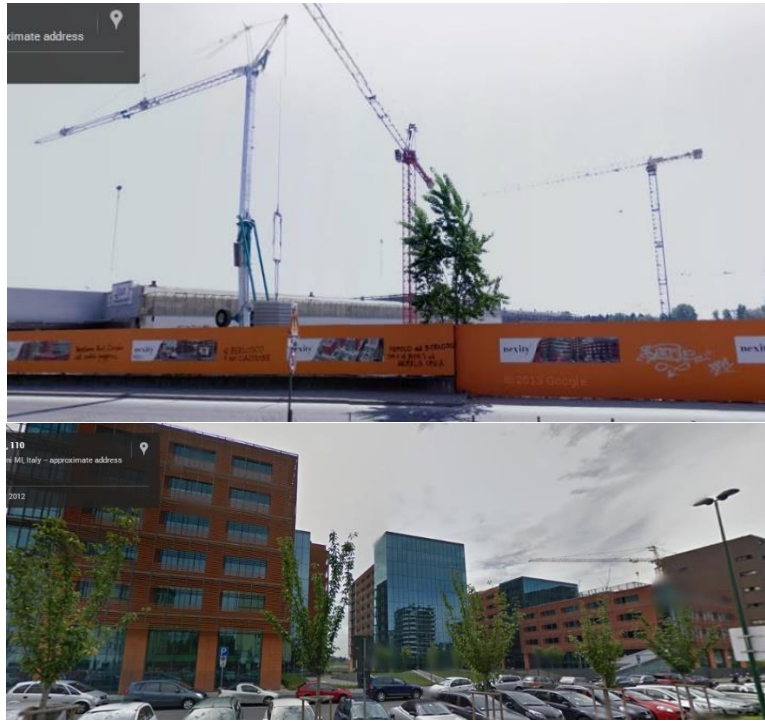
Cell id: 9081 near Via Vulcano, Sesto San Giovanni – construction site to construction site and residential



Cell id: 9281 near Via Vulcano, Sesto San Giovanni - construction site to residential and green areas



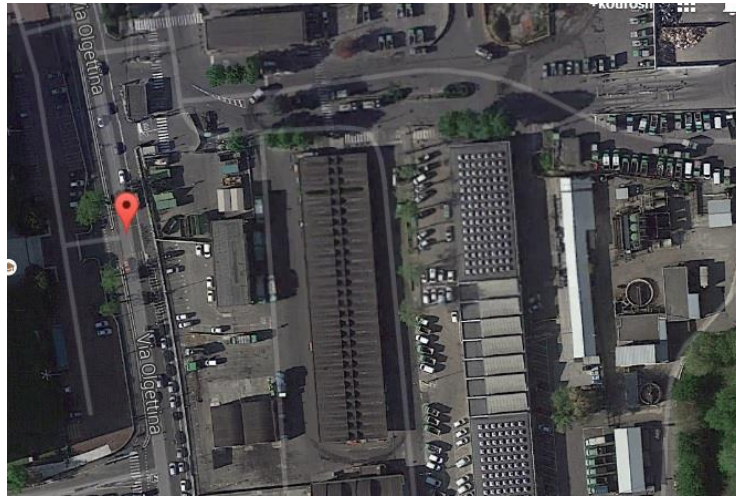
Cell id: 7975 – construction site to commercial



Cell ids: 9814, 9815, 9713, 9714 near Viale Alfa Romeo. It seems that it remain construction site.



Cell: 7187 now is parking lot ehtemale ziad



Cell id: 6898 the image is for 2014 and now it is a residential or commercial unit



Cell id: 6980 near Carlo Cazzaniga 2011 it seems to have a change to residential and sports

