POLITECNICO DI MILANO

Faculty of Environmental and Civil Engineering

Master of Science in Environmental and Geomatic Engineering

OPEN GEODATA: DEFINITIONS AND QUALITY ASSESSMENT OF SOME DATASETS IN ITALY

Supervisor: Prof. Maria Antonia BROVELLI

Co-Supervisor: Dr. Marco MINGHINI

Master Graduation Thesis by:
Miriam MOLTENI
Student Id. 820519

Academic year 2015/2016

Dedication

I would like to thank my supervisor Professor Maria Antonia Brovelli and my co-supervisor Dr. Marco Minghini for their expert advice and their support in overcoming numerous obstacles I have been facing through my research.

I am also grateful to all the members of Geomatics and Earth Observation Laboratory, especially to Dr. Monia Elisa Molinari who helped me with very valuable suggestions on this thesis.

I also want to show my recognition to all the people who, directly or indirectly, have lent their hand in this work.

I must express my very profound gratitude to my family, for providing me with unfailing support and finally, last but by not least, I must thank my boyfriend Edoardo, who lovingly encouraged me since the very beginning of this work.
This accomplishment would not have been possible without them.

Introduction

The diffusion and the availability of Open Data has considerably increased over the last decade. Thanks to their open license, they have become fundamental tools for spreading different kinds of information which are easily accessible to anyone. Among all the available types of Open Data, this thesis focuses on the evaluation of open geospatial products (named open geodata), i.e. datasets characterized by geographical information. The importance of this Open Data was demonstrated in many studies, which showed e.g. that spatial data products have the highest commercial value of re-use as they can be involved in many different sectors (Carrara et al. 2015), and that a business growth of 15% was found in countries where spatial products are open (Koski 2011). Hence, over time open geodata have become more and more important in many disparate fields and even in the daily life. The work investigates whether the current situation of open geodata in the area of the Metropolitan City of Milan follows the worldwide tendency of general increase of open data. Another important information investigated in this work is the spatial data quality, the most relevant aspect after the open data availability. Low quality datasets are harmful, as they may spread wrong information. For this reason, experts and producers have been studying spatial data quality since 1970 (Goodchild 2010) and the subject is still relevant. Thus, the objective of the thesis is to solve two different issues for the Metropolitan City of Milan: the analysis of the availability of open geodata and the evaluation of their quality. For the latter aspect, beside the general methodology, two relevant examples of quality assessment are provided. The thesis is divided in 4 Chapters. Chapter 1 is an overview on Open Data, with a particular focus on open geodata and their quality. Chapter 2 explores and catalogues the available open data for the Metropolitan City of Milan, adding detailed information on geodata. Chapters 3 and 4 propose the quality assessment of two open geodata, following two different procedures. Chapter 3 assesses the positional accuracy of an orthophoto following specific international guidelines, while Chapter 4 assesses the positional and semantic accuracy of a Land Use map through comparison against a chosen reference dataset.

Index

List of figures

List of tables

CHAPTER 1
OPEN DATA

The term Open Data (OD) has gained an increased popularity over the last years. The Open Definition delineates it as data that "can be freely used, modified, and shared by anyone for any purpose" (The open definition 2016). Various authors gave their own definition, as Halonen who described them as "usually referred to non-personal data that is accessible to all and can be freely used, re-used and distributed by anyone" (Halonen 2012). Braunschweig defined them as "freely available” data which “can be used as well as republished by everyone without restrictions from copyright or patents" (Braunschweig et al. 2012). They can be designed as datasets, with human or machine-readable formats, that can be freely accessed on the web. Similarly to the concept of open source (which is applied to software), the licenses for Open Data must ensure permission for use, modification, separation, redistribution, compilation, non-discrimination, propagation, application to any purpose, and no charge; furthermore some other conditions may be required such as attribution, integrity, share-alike, notice, source, technical restriction prohibition, and non-aggression (The open definition 2016). In this chapter, various aspects of Open Data will be analysed and explained, paying particular attention to the Italian situation.

The Chapter will be divided in 5 parts. The first one will be an introduction on the doctrine of openness, analysing its influence on Government, citizens and science. In the second section, the importance of licenses is explained. In the third part, an analysis of the geodata and its main typologies (authoritative data and Volunteered Geographic Information or VGI) will be illustrated. After that, the evolution of data quality fundamentals is explained, with a particular focus on GIS data. Finally, a revisit on the possible risks that Open Data can cause will be made.

1.1 Doctrine of Openness

1.1.1 Open Government

The doctrine of openness is the movement that promotes the spreading of information as open. It was introduced by Perritt (1997), who stated "freedom of information issues are centrally important in countries around the world, and the Internet's World Wide Web offers the potential to provide freedom of information at low cost". It has a strong impact on multiple contexts and disciplines: one of these is the access to Public Sector Information (PSI) consented by the Public Administration (PA). The doctrine of openness focuses on the improvement of government’s efficiency and effectiveness, enhances the economic growth and guarantees an overall transparency (Davies 2013). Innovating the society and making the public more participating are among its purposes (Braunschweig et al. 2012). Lately, more and more public administrations have approached this movement, promoting the diffusion of clearer and more transparent information. Openness ensures stability, provides services to the citizens and, as Harrison and Falvey (2001) affirmed:

> "The idea of using new technologies to support, enhance, expand, or re-invigorate democratic practices is not novel. The history of 20th century media has demonstrated that the introduction of new communication technologies routinely gives rise to intense speculation about their impact on the processes and practices of democracy"

Lakomaa and Kallberg (2013) defined Open Data as a "catalyst for innovation": indeed, the lack of availability of Open Data slows the innovation processes or prevents the inceptions of entrepreneurial inventive. Despite that, the amount of Open Data varies from nation to nation: it is strictly connected to the government type. Generally, **centralized countries have a higher availability of data** and, in particular, the Anglo-Saxon countries are more open to the voluntarily dissemination of Open Data. This can be explained underlying a peculiar aspect of this government type: non-Anglo-Saxon nations are characterized by a weaker bureaucratic culture compared to other centralized administrations (Lakomaa and Kallberg 2013).

Despite all the differences and problems that might arise by publishing this type of data (e.g. citizen privacy), the global will is to follow this trend. In 2009, United States’ president Barack Obama mentioned that openness of government would have been a pillar of his administration. In 2011 at the U.N. General Assembly meeting, US and other seven countries stipulated the first act of **Open Government Partnership** (OGP, http://www.opengovpartnership.org), "a multilateral initiative that aims to secure concrete commitments from governments to promote transparency, empower citizens, fight corruption and harness new technologies to strengthen governance" (Open Government Partnership 2016). With the purposes of collaborating and improving the civil society, it was built according to five "grand challenges":

- Improving Public Services;
- Increasing Public Integrity;
- More Effectively Managing Public Resources;
- Creating Safer Communities;
- Increasing Corporate Accountability.

To be included in this partnership, each country should demonstrate its interest and should prove its dedication. In particular, participating countries must endorse Open Government Declaration, deliver a country action plan developed with public consultation and commit to independent reporting on their progresses.

In 2012 at the first high-level meeting in Brasilia, OGP passed from eight to forty-two nations involved (Italy included). It delineated there the first biennial program that each participant had to respect, called the “first action plan cycle”. The successive one was planned in 2014, named the “second action plan cycle”, and lastly in 2016 the “third action plan cycle” was drafted. Each participant country had to follow this path: once the requirements of a plan are fulfilled, it is possible to proceed to the successive one.

Nowadays OGP is counting a total amount of seventy nations from all over the world (Open Government Partnership 2016).

Figure 1.1 shows the nations involved in the OGP. It is easy to see that the distribution of the countries is not restricted to the high-developed area. There are many countries from the Third World and some EU countries are missing (e.g. Germany, Poland and Austria). In the figure, the colours of the countries represent their situation: the red countries are applying the “third action plan cycle”, the yellow ones are implementing the “second action plan cycle” and the blue ones are handling the first. The countries that want to become partners are brown. The status, on a global scale, appears heterogeneous.

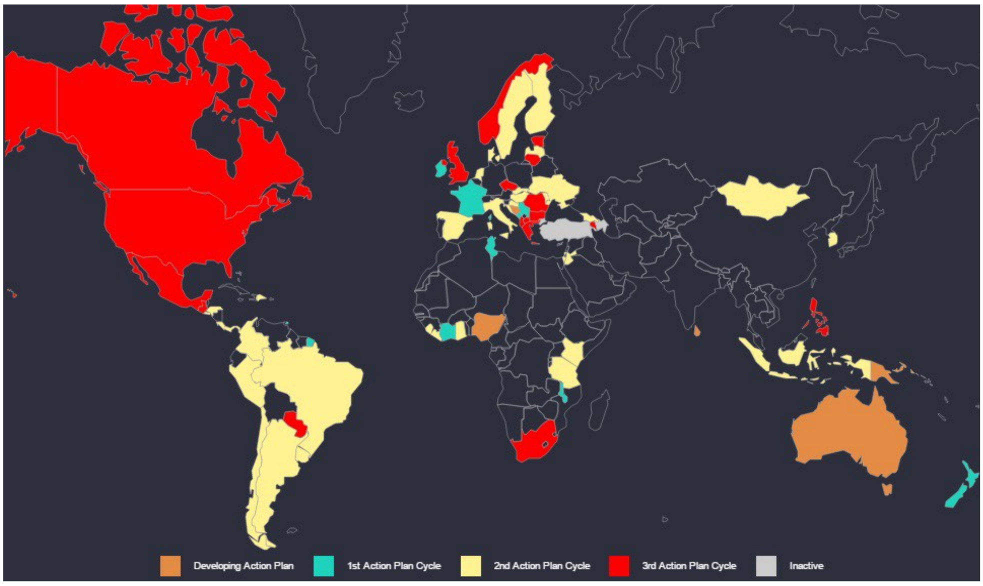


Figure 1.1 – Open Government Partnership Countries (source: http://www.opengovpartnership.org/countries, last accessed December 2016)

Italy is currently applying the second action plan cycle. In 2011, it joined OGP and conferred the first Action Plan in 2012 in Brasilia. In 2014 a collaboration of Department for Public Administration (DPA), Agency for Digital Italy (*Agenzia per l'Italia Digitale*, AgID) and Anticorruption National Authority (*Autorità Nazionale Anticorruzione*, ANAC) published the second OGP Action Plan. This has as fundamentals:

- Access to Information;
- Anti-Corruption;
- E-Government;
- Open Data;
- Public Service Delivery.

On June 6th, 2016, the Minister Marianna Madia (*Ministro senza portafoglio per la semplificazione e la pubblica amministrazione*) launched the first Open Gov Forum, which started the process of designing and drafting the third Italian National Action Plan (Open Government Partnership 2016). This was published on September 20, 2016 and it is currently available online (Il Terzo Piano Nazionale d'Azione 2016)

At this point the question can be: "is this partnership worthy?"
WorldJusticeProject.org can give an answer. In 2015, it made an analysis that delineated an index based on Open Government Partnership data (World Justice Project 2016). It measured government openness based on the worldwide experiences and perceptions of the general public. Figures 1.2 and 1.3 show the investigation results. Figure 1.2 demonstrates

that the OGP members have an Open Government Index higher than the countries that are not involved. Specifically, the countries that are applying the "second action plan cycle" have a higher value.

Figure 1.3 shows the index distribution according to the monetary earnings of the countries. The trend is similar to the general one, except for the higher income countries. In this case, countries that are implementing the first plan cycle have a lower value compared to the ones that are not involved. Even though, OGP members in the second action plan cycle have the greatest index.

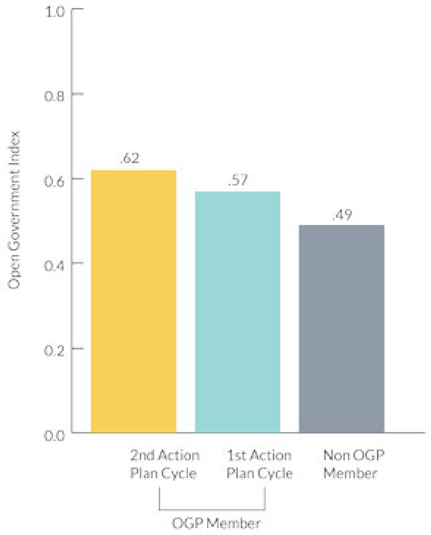


Figure 1.1.2 – WJP Global Open Government Index (source: http://www.worldjusticeproject.org)

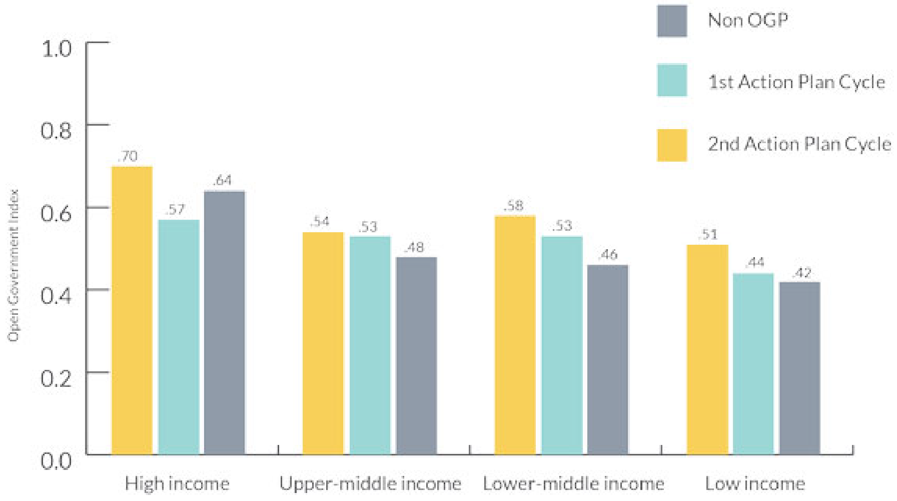


Figure 1.3 – WJP Open Government Index by income (source: http://www.worldjusticeproject.org)

Another important deal in the field of OD is the **Open Data Charter** (http://opendatacharter.net), proposed by the former G8 government in 2013. It is a plan for transparency and development, which relies on the open release of high-value governance datasets in national geoportals (G8 leaders 2013). Defined also as "a Global Multi-Stakeholder Action Network", it involves not only countries but also organizations (The international Open Data Charter 2016).

The principles taken as fundamentals are:

- Open by Default;
- Timely and Comprehensive;
- Accessible and Usable;
- Comparable and Interoperable;
- For Improved Governance and Citizen Engagement;
- For Inclusive Development and Innovation.

Open Data Charter consists of two different groups: Stewards and Lead Stewards. At the time of writing (December 2016), Italy belongs to the Stewards category, together with French government, city of Buenos Aires and others. In order to develop a sustainable expansion, each member is involved in different working groups, according to its interests (The international Open Data Charter 2016). Figure 1.4 shows the countries and the governments engaged at the time of writing (December 2016).

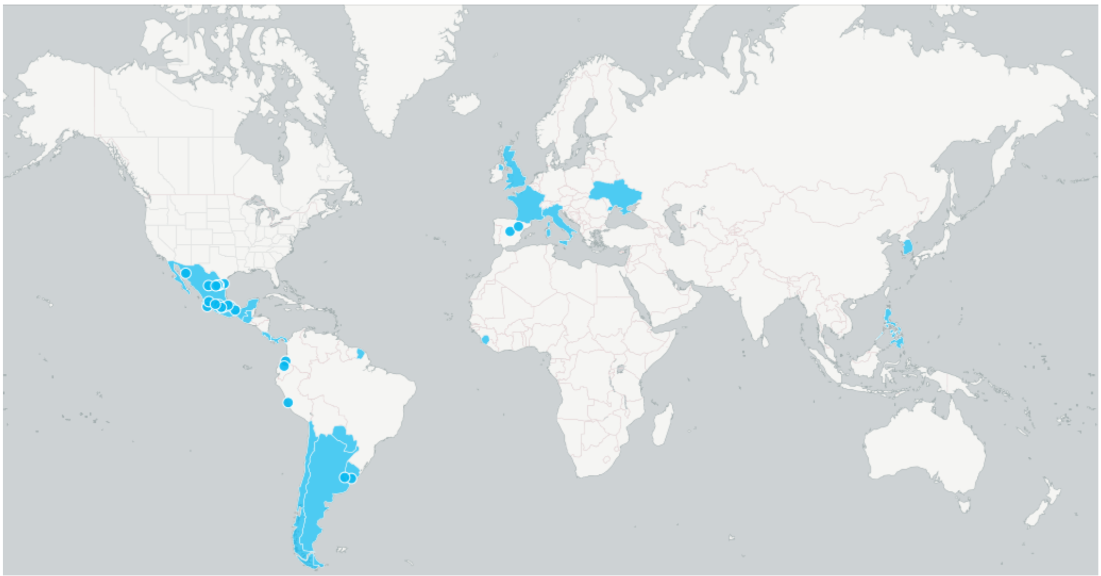


Figure 1.4 – Countries and governments that adopt Open Data Charter (source: http://opendatacharter.net)

The light blue colour represents the involved entities: countries (as polygons) and administration (as points).

From a European point of view, a good source of information is the European Data Portal (http://www.europeandataportal.eu). This project harvests Open Government metadata from public data portals of European countries, with the overall purposes of improving Open Data accessibility and increase their value. Carrara et al. (2015) show that the direct market size of Open Data for the 28+ EU Member States in 2016 is 55.3 billion EUR, with an expected increase of 36.9% to a value of 75.7 billion EUR in 2020. Using indicators such as number of jobs created, cost savings, and efficiency gains, Carrara et al. (2015) quantified additional benefits of the re-use of Open Data in Europe. Therefore, the cumulative total market size of Open Data between 2016 and 2020 is expected to be around 1200 billion EUR. Estimates are based on the so-called **Open Data maturity**, a parameter attributed to each EU Member State by looking at the progress made so far in terms of Open Data. Countries are ranked in three groups: trend setters, followers and beginners (Fisher et al. 2015) (fig. 1.5).

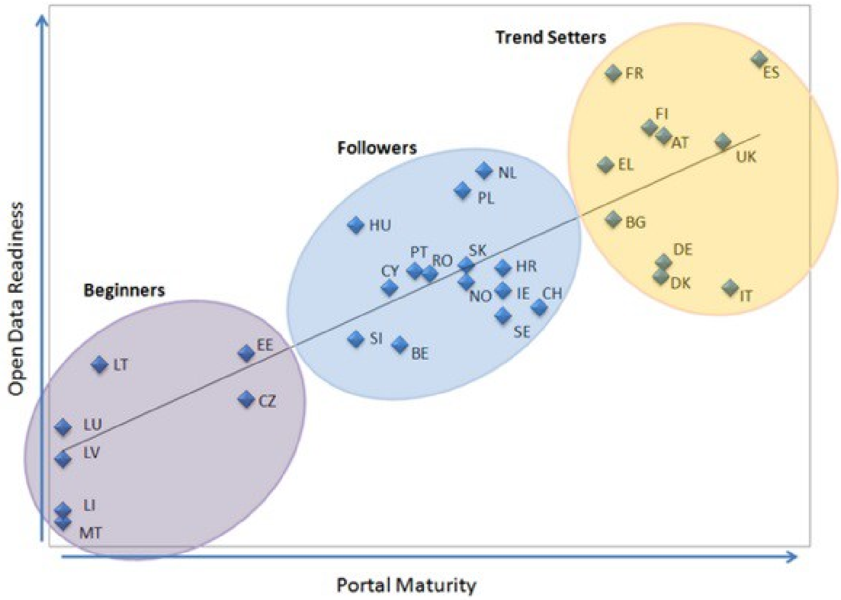


Figure 1.5 – European Open Data Maturity (source: https://www.europeandataportal.eu)

Italy, whose portal is available at http://www.dati.gov.it, belongs to the group of trend setters, that also includes Spain, the first EU Member State with a national Open Data portal. It is expected that by 2020 almost all EU 28+ Member States will have fully operating portals and thus they also will become trend setters.

Another indicator to evaluate the situation of Open Government is the **Global Open Data Index** (http://index.okfn.org).

Since 2013, it has provided the most comprehensive snapshot of the global state of Open Data, by monitoring the increase of Open Government datasets on an annual basis. The number of entries has increased from 597 to 1'586 and the number of countries has passed from 60 to 122. The percentage of Open Data has decreased from 15% to 9% due to the introduction of new countries datasets (fig. 1.6). This is probably a consequence of the fact that the countries with the greatest number of open datasets have been already involved in the analysis. Hence, the increase of the number of countries has produced an increase of the total number of datasets, but a decrease of the available Open Data.

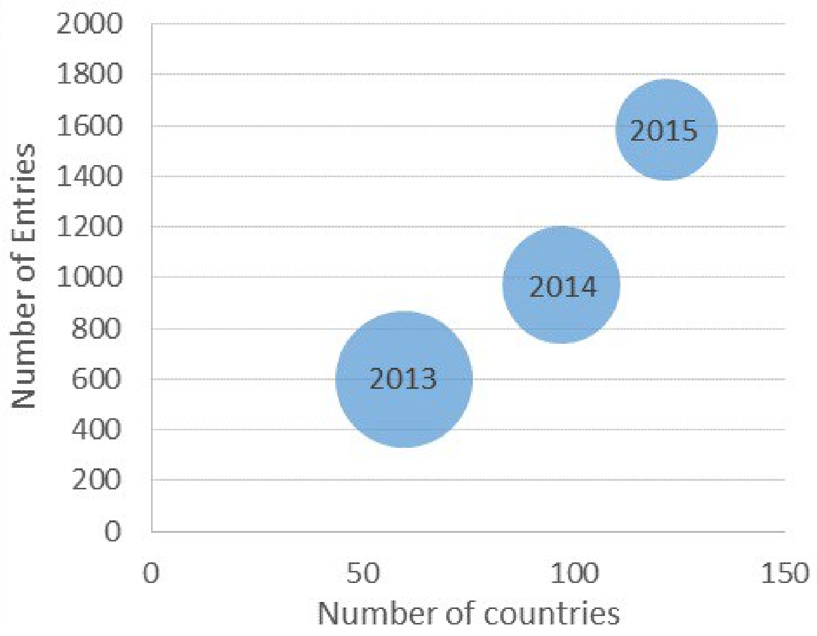


Figure 1.6 – Global Open Data Index

In 2016, the country with the highest percentage (78%) of released Open Data is Taiwan, while Italy ranks 17th with 55% of its datasets being open.

To conclude, in order to deeply understanding the current situation, table 1.1 summarizes the previous statements. It lists the first ten countries with the highest OD index plus Italy. It also indicates the kind of government and shows whether the country is involved in Open Data Partnership, is part of Open Data Charter and (if it is in the EU) the maturity of its portals.

Table 1.1 – List of the first 10 nations with the highest OD indexes plus Italy

	OD index	**Kind of government**	**Open Data Partnership**	**Open Data charter**	**OD portal maturity**
Taiwan	76%	Centralized	No	Yes	\
United Kingdom	70%	Centralized	3rd Action Plan Cycle	No	Trend Setters
Denmark	68%	Centralized	2nd Action Plan Cycle	Yes	Trend Setters
Colombia	67%	Centralized	2nd Action Plan Cycle	No	\
Finland	67%	Centralized	No	No	Trend Setters
Australia	66%	Federation	Developing a Plan	No	\
Uruguay	64%	Centralized	2nd Action Plan Cycle	Yes	\
United States	64%	Federation	3rd Action Plan Cycle	No	\
Netherlands	63%	Centralized	2nd Action Plan Cycle	Yes	\
Norway	63%	Centralized	3rd Action Plan Cycle	No	Followers
Italy	55%	Centralized	2nd Action Plan Cycle	Yes	Trend Setters

Most of the countries with the highest OD indexes have a centralized government and they are involved at least in the “Second Action Plan Cycle” of the Open Data Partnership. Being part of Open Data Charter is not so common, hence being or not involved in it does not affect much the Open Data index. Lastly, the classification of European portal maturity is mostly Trend Setters, with only Norway falling into the Followers.

1.1.1.1 Examples of benefits taken by an Open Government

Many examples of advantages taken from Open Government could be listed. In 2010, UK saved £4 million using Open Data. Liam Maxwell, the UK Government Chief Technology Officer (CTO), realized that the same expensive market research reports were bought many times by different UK offices. Hence, he decided to open the total government spending. Instead of having a specific file for each government department, he decided to list the total transactional spending data, allowing the employees to check whether a document was already purchased. This allowed to save not only time in the government work, but also money (Pollock and Rogers 2010).

Another interesting example is related to the Danish government, that in 2002 released the official Danish address database openly by affirming that (Rogers 2010):

> “Free and unrestricted access to addresses of high quality is beneficial to the public and forms the basis for reaping substantial benefits in public administration and in industry and commerce”

In 2010, 1’236 entities (being 70% private companies, 20% from the central government and 10% municipalities) used the data. In the period 2005-2009, an indipendent study demonstrated that the direct benefit of the address data was 62 million EUR, expecting to rise to 14 million EUR in 2010. Moreover, the indirect advantages have been the improving of the public service coordination, the elimination of duplicate collection and a standardization with higher quality data (Rogers 2010).

Also in the not-developed countries the Open Government can be very effective (Open Data Insitute 2015). In 2013-2014, in Nepal 22% of the national budget was foreign aid. In 2013 the Nepal’s Ministry of Finance launched a trasparency program that allows the monitoring of the aid flow, the Aid Management Platform. It allows non-Governmental Organizations (NGOs), journalists and civil society to control where the money is directed. In this scenario, Open Data can be used to produce analyses for policy reform, or by NGOs which use Open Government data for programme planning, monitoring and evaluation. The Nepal’s Ministry of Finance uses them to formulate the entire government’s budget, helping to trace gaps between spending and output (Open Data Insitute 2015). Similar cases occur in Uganda, Kenya, Tanzania and countires from Western and Southern Africa. The details are different but the aim of monitoring aid while improving the governement is comparable.

1.1.2 Open Data for Citizens

The publication of Open Data can be provided on the web in two main formats: human-readable or machine-readable. The first one has an interface that allows the user to understand them directly with graphs and maps, while the second is made up by raw data readable only by specific software (Braunschweig et al. 2012). A citizen (a generic, not expert person) can interface only with the amount of available data that are human-readable, as they are the easier to interface. Salvemini (2015) suggested that, nowadays, Open Data are most used by technicians, who develop applications and services that only later end in the hands of citizens. This statement adds something that should be a foundation of the Open Government movement: Open Data should be easily and clearly accessible by anyone. They should be published in a format and in a website that best allow robust and diverse third party use. They should be freely available on the web in a structured, human-readable format so that anyone can easily access them at minimal cost (Robinson et al. 2009). Not considering those standards would introduce a new barrier between government and citizens and the fundament of transparency would fail.

1.1.3 Open Data in science

The Word Wide Web has radically modified the way data can be shared. In 2007, Tim O'Reilly (2007) introduced the definition of Web 2.0: an evolution of the old web that allows a high potentiality of sharing and interacting with online materials. The spreading of the Open Data can be seen as one consequence of this new state.

In a research environment this can create positive benefits. Sharing data makes analysis more robust, trustable and detects its vulnerabilities. It can speed up discoveries and identify large scale trends (Gewin 2016). Hence, the sentence "the scientists must share online their own data" can be considered as a mantra, since researchers can easily distribute their codes and data with a mouse click (Boulton et al. 2011).

However, in practice this is not so simple (Murray-Rust 2008). Nowadays many scientists are unfavourable or worried about making their researches open. The reasons of these behaviours are numerous. Some of these are the fear of being scooped on future works, the impossibility of sharing because of signed agreements (Van Noorden 2014) or the fright of being scientifically wrong (Gewin 2016). As an interesting example, Nature reported stance of Steve Simpson of the University of Exeter (UK), a biologist who studied a rare tropical fish. He is available to distribute his raw data privately with potential collaborators, but not to put them online. Even if he gave guidelines to replicate the collection of the data, he did not openly supply them because of the difficulty of obtaining the specific licenses needed to operate in the field (Van Noorden 2014).

Despite all, junior researchers appear the ones that suffer more from the lack of Open Data availability. An inquiry in 2014 declared that in 217 works between 2000 and 2013 only 40% of them achieved the requested data, based on the seniority of the applicant. The difficulty in accessing the data, problematic if someone else could not ask for them, has limited and reshaped their works (Magee et al. 2014).

Open science should be more common because it can also deter fraud. If the research could be verified by third parties, unlawful acts could be avoided. This may prevent scandals such as the anesthesiologist Yoshitaka Fujii, who has published 172 articles containing counterfeit data (Boulton et al. 2011). That situation has created an outrage in the medical industry, but in general has weakened the reliability of science itself.

Sharing data promotes the reproducibility and reliability of an analysis. Generally, the peer-review publications provide summaries of the available data, but not an effective usable amount (Boulton et al. 2011). In order to avoid that situation, the Open Knowledge

Foundation founded the Public Library of Science (PLoS), a journal that follows the Open Access doctrine, in which authors must provide online all the data needed to reproduce their researches (Pubblic Library of Science 2016). A study demonstrated that, even with this statement, not every author published all the data needed to replicate their investigations: in 2011-2012 only the 12% have fully supplied the requests, while in 2014 the trend increased at 40% (Van Noorden 2014). The trend is positive, but in order to increase it the education of the authors is necessary. Once scientists will truly understand the benefits that everyone could get with an open science, this closing condition could finally change.

Concluding, as Boulton et al. (2011) affirmed:

> “Science has become woven into the fabric of modern civilisation and should be, and be seen to be a public enterprise, not a private enterprise done behind closed laboratory doors”

1.2 Open Licenses

The main aim of publishing Open Data is to promote the interoperability, but the respect of the producer’s work must not be taken for granted. If data are shared in a tiny group, their exchange would easily respect the producer’s will, but if they are published on the web and thus are accessible by anyone, some rules need to be fixed. For this reason, it is important to underline the permissions that authors attribute to their products, in view of third party usage. In this situation, without an appropriate protection of the information, only two scenarios could be reached: the locking of information or the spreading of information in a public-domain form (Miller et al. 2008).

Nevertheless, the reality is different thanks to the usage of licenses. They represent the tool to solve that problem based on the intellectual property rights explained in the national jurisdictions. Licenses explicit the terms in which data can be used, stating for example if they can be reused and redistributed (Open Data Handbook Documentation 2012).

From an international point of view, the most important licenses for Open Data are the Creative Commons (CC) licenses, an extension of copyright law. They recognize the holder of the rights the ability to extend particular permissions for third party use (Creative Commons 2016). Different CC licenses can be used, based on the features that the data producer wants to include. As an example, if an adaptation of the work is permitted a CC-BY license is suggested (CC-BY Creative Commons 2016).

The Italian Open Data License version 2.0 (IODL 2.0) is the most significant and common license in Italy. It was developed considering the copyright law, the Italian law n. 248/2000 [1] and personal data protection (Italian Open Data License v2.0 2016).

1.3 Open geodata

Geospatial data are a "model of reality", a reasoned and simplified representation of the world (Goodchild 2010). They are representations of geographical data (visible or invisible) located in the territory with their geometrical or photographical georeferenced images and annexed information (Figure 1.7) (Biallo 2012).

The elements have as attributes (features) some geographical information: they can be numbers (coordinates) or names (as addresses or specific locations). In both cases elements can be georeferenced using a specific reference system, local or global.

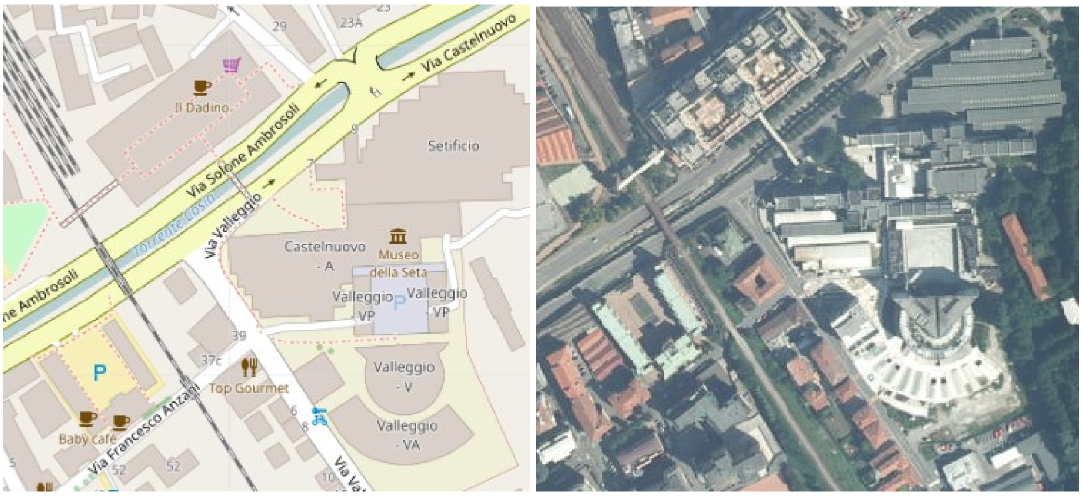


Figure 1.7 – Geometrical and photographical georeferenced images (ortophoto) (source: left © OpenStreetMap contributors, right AGEA 2012)

The uses of geodata, as suggested by Crăciunescu and Ilie (2013), are various:

- Management of natural resources (e.g. water management, forestry conservation);
- Government use (e.g. urban planning, public works);
- Mapping purpose (e.g. cartography, topography, navigational use);
- Transportation (e.g. public transit, infrastructure management);
- Communications purpose (e.g. definition of the location of pipelines or electrical lines);
- Locationing of services (e.g. position of restaurants, hotels or infrastructures);

[1] http://www.parlamento.it/parlam/leggi/00248l.htm (accessed March 22, 2017)

- Military scope (e.g. geospatial intelligence);
- Others.

Making these datasets open spreads their usage and allows the interoperability by different subjects. Even in a research environment the usage of geodata is wide, because they allow studying trends from a spatial point of view. Data are distributed not only a temporally but also geographically, allowing the creation of maps and visualizations. This is very important in topics such as, for example, archaeology, hydrology, geophysics and climatology.

The relevance of open geodata is underlined by some reports of the European community, which analysed the datasets available in the EU Open Data portal (https://data.europa.eu). Carrara, et al. (2015) declared that geospatial datasets have the highest commercial value of re-use between all the thirteen identified categories (see figure 1.8). The authors underlined the potential of this kind of data because they can be involved in many different sectors compared to the other typologies.

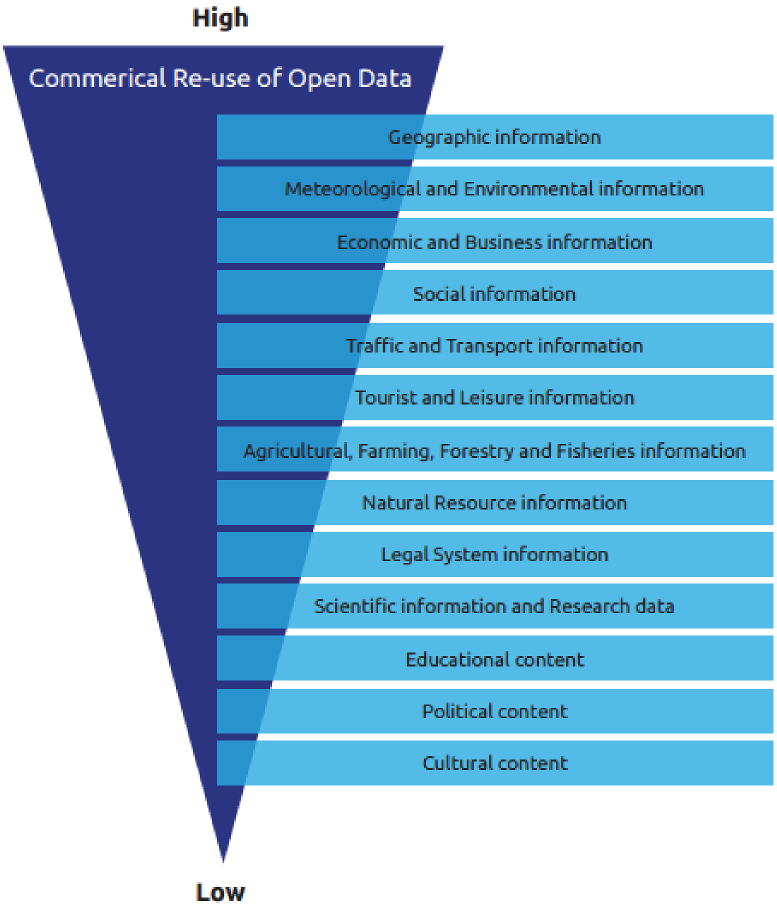

High
Commerical Re-use of Open Data
Geographic information
Meteorological and Environmental information
Economic and Business information
Social information
Traffic and Transport information
Tourist and Leisure information
Agricultural, Farming, Forestry and Fisheries information
Natural Resource information
Legal System information
Scientific information and Research data
Educational content
Political content
Cultural content
Low

Figure 1.8 – Commercial Re-use of Open Data (source: https://www.europeandataportal.eu)

Moreover, Koski (2011) demonstrated a business growth of 15% in countries where geodata are open (or at least sold at reduced prices). Hence, the importance of open geodata appears relevant.

Generally, geodata can be divided by the kind of producers. By tradition, they are supplied by public authorities (e.g. PA, governmental agencies and institutions) and nowadays more and more of these are available under open licenses. More recently as low-cost alternative, volunteers decide to join and create their own Open datasets. That movement, called Volunteered Geographic Information (VGI) is now an alternative to official ones. This concept, important in the current geodata situation, will be illustrated in section 1.2.2.

1.3.1 Authoritative geodata

Over the last decades, collecting information to build geodata was a procedure affordable mainly by public or commercial institutions (Auer and Zipf 2009) because of the pricey methods. Nowadays the evolution of new technologies brought a general decrease of the costs and a spreading of new possibilities. Some examples are the usage of satellite images instead of information taken from photogrammetric flights.

In a global scale, the measures taken to have Open Governments brought more and more geodata to being open too. Over the last years, new geoportals and new open datasets are available and usable by anyone worldwide.

Biallo (2012) described the situation in Italy. The producers of geodata can be listed as:

- PAC (*Pubblica Amministrazione Centrale*, Central Public Administration): ministries (Environmental, Culture, Agriculture, Tourism, etc.), agencies and institutes, e.g. Military Geographical Institute (*Istituto Geografico Militare,* IGM*),* Italian National Institute for Environmental Protection and Research (*Istituto Superiore per la Protezione e la Ricerca Ambientale*, ISPRA), Agency for the Agricultural Supplies (*Agenzia per le Erogazioni in Agricoltura*, AGEA), Italian National Research Council (*Consiglio nazionale delle ricerche,* CNR);
- PAL (*Pubblica Amministrazione Locale*, Local Public Administration): regions and their agencies, provinces and municipalities;
- Others.

Among the institutions belonging to the first category, of particular interest is the Military Geographical Institute (*Istituto Geografico Militare,* IGM*),* which is the Italian National Mapping Agency (NMA). It is defined as a national cartographic institution, which builds and updates the state's geodetic network.

1.3.2 Open Data created by citizens

The rise of the consumer-oriented Web 2.0, with its user-generated contents (Auer and Zipf 2009), created a new situation in which people voluntarily collect data.

This has created a deep innovation: geographic information system is now a discipline where the general public can interact directly (Goodchild 2007). Goodchild (2007) termed this new phenomenon Volunteered Geographic Information (VGI), a particular case of user-generated web contents. Citizens collect data and make them freely available on the web, using specific license conditions. People involved are both users and producers, or "produsers" to use a recent definition (Coleman et al. 2009).

Thanks to the spreading of this condition in a vast variety of domains, nowadays VGI can be considered as an alternative to the expensive and/or proprietary data. The amount of these datasets is increasing day by day. In fact, the more this freely available spatial data are available on the Internet, the higher is the demand for them (Zielstra and Zipf 2010).

The most popular VGI project is OpenStreetMap (OSM, http://www.openstreetmap.org). It is a collaborative project aiming to create a map of the whole world, which is freely available and openly-licensed. Started in 2004, OSM has seen the involvement of more and more people, as shown in Figure 1.9 (data are available at http://wiki.openstreetmap.org/wiki/Stats). The number of contributors has grown since then, reaching 3'537'047 in March 2017.

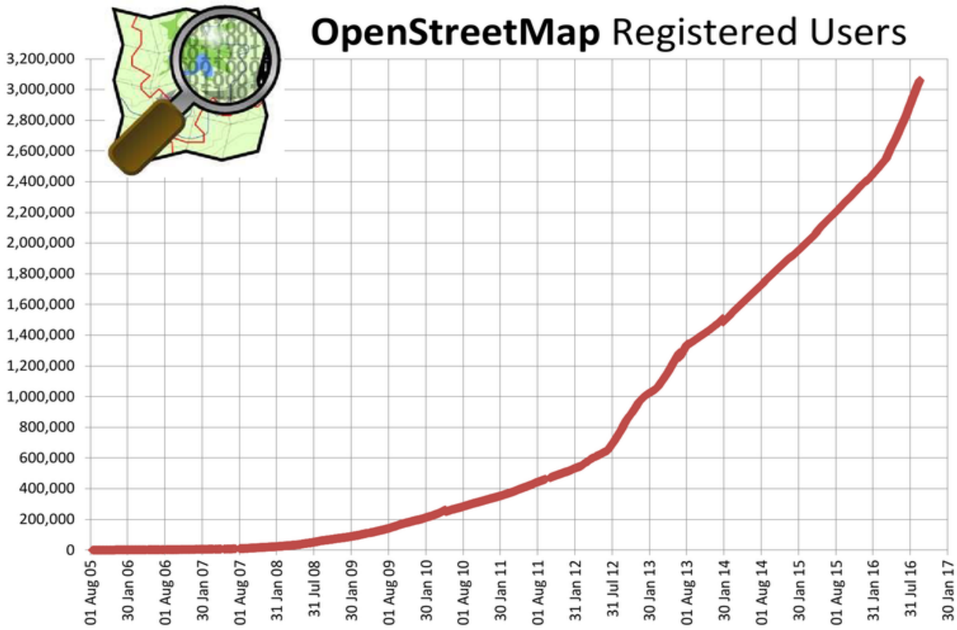


Figure 1.9 – OSM registered users (source: http://wiki.openstreetmap.org/wiki/Stats)

In OSM it is also possible to upload GPS tracks in the GPS eXchange (GPX) format, a particular XML format used to transfer GPS tracks in software and Web applications. Figure 1.10 shows (in pink) the amount of GPX track points uploaded in the OSM database together with the total amount of registered users (in blue). The track points show an almost linear increase in time. In March 2017, the total number of track point uploaded was higher than 600'000.

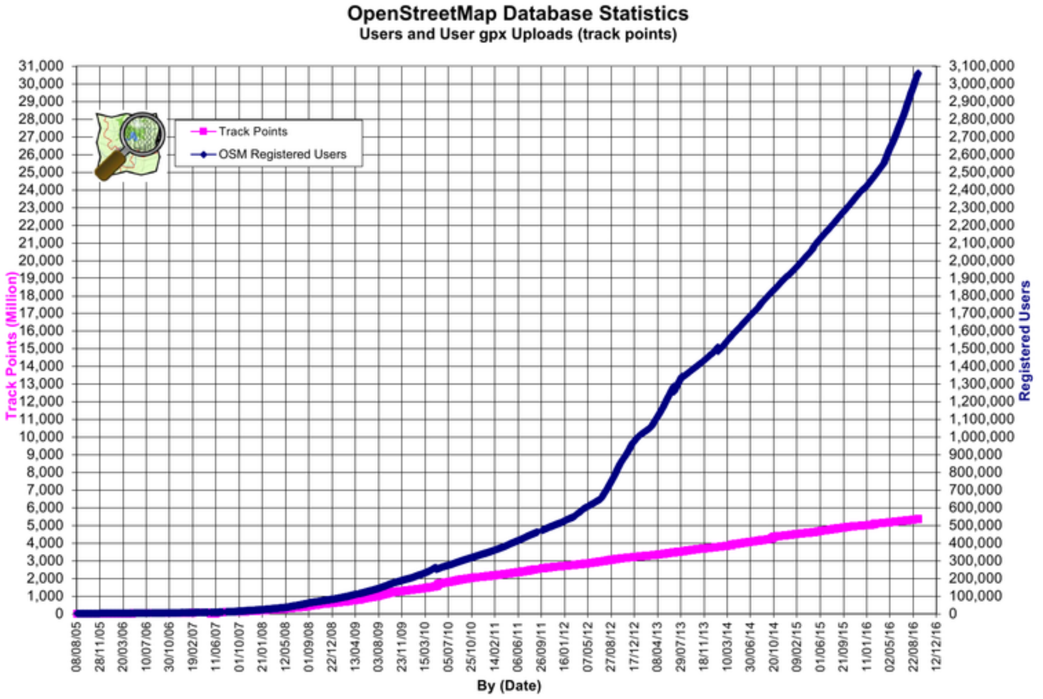


Figure 1.10 – OSM users and user GPX track points uploaded (source: http://wiki.openstreetmap.org/wiki/Stats)

Another interesting information is given by figure 1.11, which shows the total user contribution in the edits of OSM nodes. It is easy to see that the historical trend features a peak between 2007 and 2009, a moment in which smartphones (with integrated GPS receivers) witnessed a huge spread. This allowed more and more people to access Web 2.0 applications and enabled to register and share positions online without the need to buy specific (and typically expensive) devices.

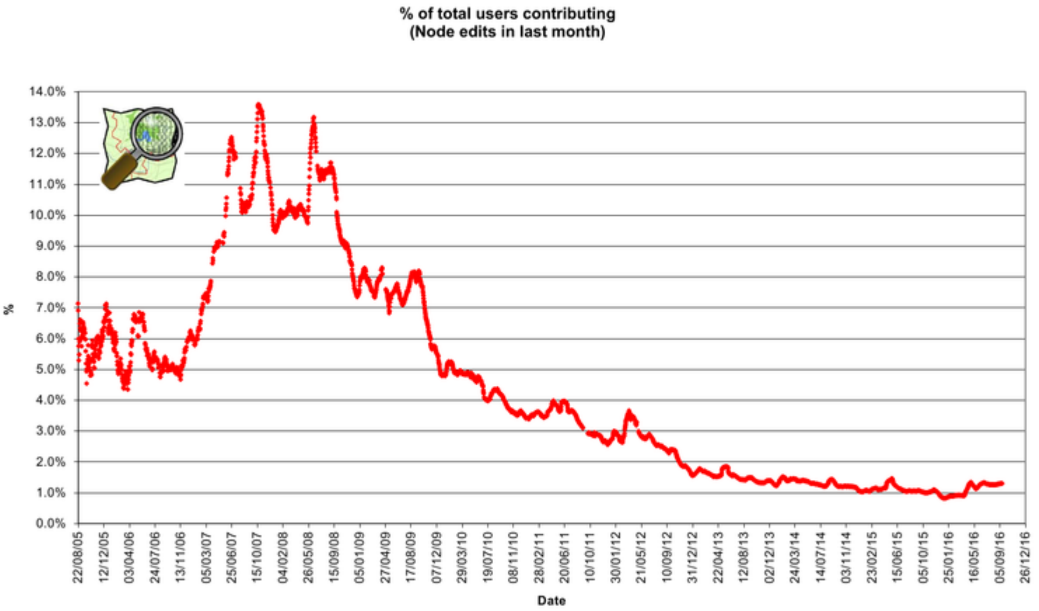


Figure 1.11 – OSM node edits (source: http://wiki.openstreetmap.org/wiki/Stats)

The potential of VGI lies also in the fact that datasets are typically available at a continental or even global level and are created and updated almost continuously. This is a very useful tool to collect information and manage decisions (Boney et al. 2009).

Lastly, the amount of VGI data has the potential to redistribute the right to define and judge the value of the corresponding authoritative geodata already on the market and the production of the new ones (Coleman et al. 2009).

1.4 Data quality

Goodchild (2010) stated that all geospatial data are imprecise, inaccurate, out-of-date and incomplete at different levels. For these reasons, a measure of their goodness is needed: their quality. Data quality is defined as the measure of the differences between the data and the reality that they represent. The more they diverge, the more their quality decreases. Hence, data quality is a fundamental component needed to truly undestand the data (Natale 2011). Geodata with low quality contain little information about the world in general and their value is small (Goodchild 2010). Low quality data are unsatisfying, if not even contradictory. They could result to be not usable and generate unhappy users (Natale 2011). Thus, having lots of data does not imply that they are of a high quality.

Experts and producers have been aware of this problem since 1970. Geodata quality has been discussed in many scientific conferences as the International Symposium on Spatial Accuracy Assessment in Natural Resources & Environmental Sciences (Accuracy) and the biennial International Symposium on Spatial Data Quality (ISSDQ). Data quality has been also analyzed in specific working groups such as “Quality of Spatio-Temporal Data and Models” in the International Society for Photogrammetry and Remote Sensing (ISPRS), the “WG on Spatial Data Usability” in the Association of Geographic Information Laboratories in Europe (AGILE) or the “WG on Spatial Data Uncertainty and Map Quality” in the International Cartographic Association (ICA) (Goodchild 2010).

Despite the awareness that experienced people have on the importance of data quality, the spreading of open geodata allows anyone to access and reuse them. In this status GIS approaches a new problem: communicate the quality so that users having different skills can understand it (Boin and Hunter 2007) and realize its importance.

This new situation is summarized by Tòth and Tomas (2011) in five statements:

- Increasing exchange and use of spatial data;
- Increasing of the number of users without knowledge about spatial data quality;
- GIS datasets usable in different applications, regardless the original purpose;
- Lack of spatial quality tools;
- Increasing distance between the end users and the producer (person who knows quality).

If, on one side, authoritative data must adhere to specific requirements for being validated and published, there are not simple tools available to estimate the quality of VGI data. For that reason, despite VGI brought a fundamental action as a geographic data enhancer, it created prompted concerns regarding their quality, reliability, and overall value (Flanagin and Metzger 2008). VGI quality has become a major research topic over the past few years. In particular most of the studies have focused on the OSM quality, already tested for a few European countries. As examples, Haklay (2010) evaluated the UK OSM road, Girres and Touya (2010) extended the same work in France and Zielstra and Zipf (2010) proposed a similar study for Germany.

1.4.1 Concepts of quality in GIS

Korzybski (1933) affirmed in one of his works that “the map is not the territory”. Each geographic information has a particular level of abstraction, as it is a model that represents the real world, and quality can give parameters that can assess the maps.

In cartography, data quality is traditionally associated to the positional accuracy. It is defined as the correctness of the position of a particular feature. It has to stay under a fixed threshold, which respects the current technological progress (Goodchild 2010). However, with the advent of the digital era, this element has been recognized as being not enough.

Hunter and Beard (1992) proposed a classification of errors in GIS data (see Figure 1.12) to which data quality tries to give answers. Errors are divided in three categories:

- Source errors: related to data collection, processing and usage;
- Forms errors: divided in primary (positional or attribute error) and secondary (logical consistency and completeness);
- Final product errors.

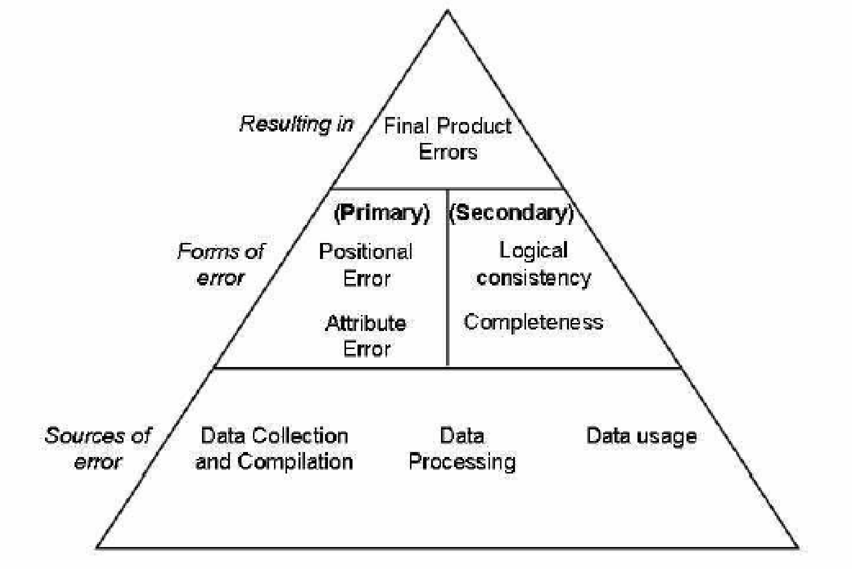


Figure 1.12 – GIS errors classification (source: Hunter and Beard 1992)

Starting with the delineation of these errors, many different quality elements are detected and set as standard nowadays.

1.4.2 Requirements and metadata

Measures of quality parameters must be somehow/somewhere expressed. To comprehend this, it is of fundamental importance to understand the differences between two key terms in the development of a geographic dataset: requirements and metadata. The creation process of geodata starts with the identification of current specific directives to arrange its quality, a priori requirements that fix a target result. The requirement is defined as “a document that prescribes the requirements to which the product must conform” (Benoît and Fasquel 1997). Once the directives are selected, the producer can proceed and create the geodata.

Metadata are defined as measures of the evaluation of the data itself and they are an a posteriori statement (Tòth and Tomas 2011) based on direct measurements and calculations processed in the final product by an analyst.

At the end of the creation process, there is the end user, who can interface with the geodata and its metadata (see Figure 1.13).

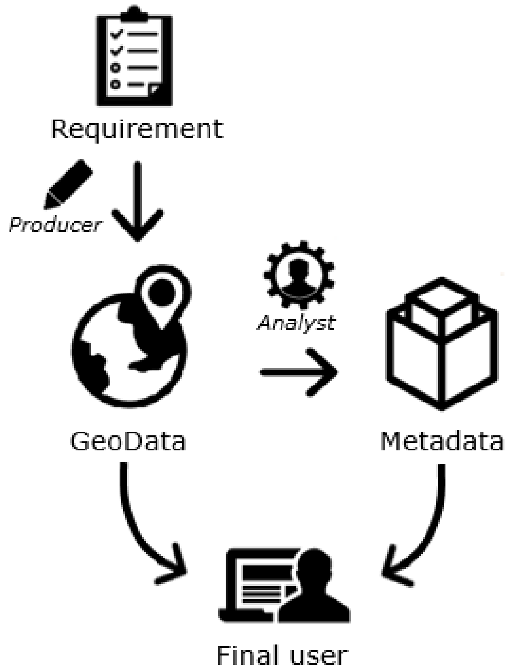


Figure 1.13 – Requirement VS Metadata

As the requirements formalize the specifications that the dataset will have, the connection between requirements and metadata is strict. Nevertheless, the quality of the final product could be different from the one guaranteed in the requirement, causing a potential confusion in the final (not expert) user.

1.4.3 Data quality: internal and external

Devillers and Jeansoulin (2006) described the previous statement in another way, by dividing the concept of quality into internal and external.
Figure 1.14 depicts these definitions. The internal quality is measured through a comparison between the requirements and the produced dataset, while the external quality is the one assessed through a comparison between the geodata and the end use.

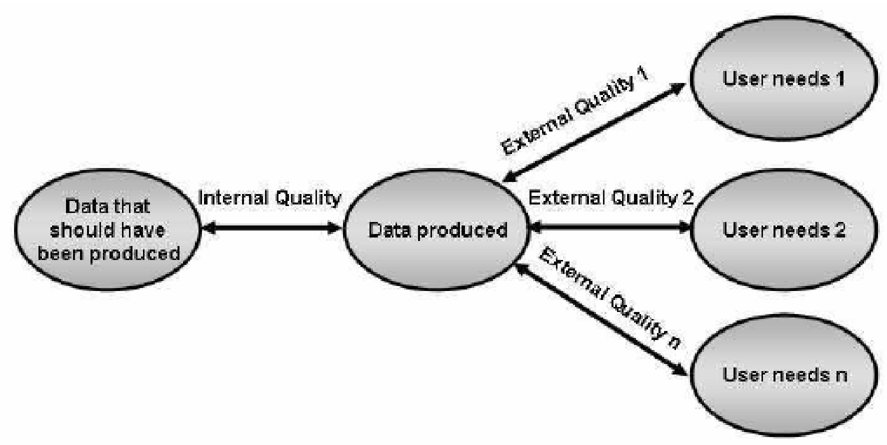


Figure 1.14 – Internal and external data quality (source: Goodchild 2010)

The quality measured between requirements and metadata is the internal; the one between metadata and end-use is the external.

1.4.3.1 Internal quality

Internal quality expresses the resemblance of the dataset to the required one. Data are compared to the "nominal ground" defined by David and Fasquel (1997) as "an image of the universe, at a given date, through the filter defined by the specification of the product". It represents the "perfect" data. In practice, it consists of small datasets with high accuracy, called also control data or reference data, which are compared with the final (wider) product.

Different criteria describe the internal quality. The International Organization for Standardization (ISO) has defined a list of quality elements since 2002 in "ISO 19113 – Geographic information, Quality principle", then updated in 2003 in "ISO 19114 - Geographic information, Quality evaluation procedures", later revised in 2006 with "ISO 19138 - Geographic information, Data quality measures" and finally adjusted in 2013 in "ISO 19157 - Geographic information, Data quality". ISO 19157 replaces the previous editions.

ISO provides a standard list of five elements to evaluate data quality (ISOquality 2013):

- Completeness: presence and absence of features, their attributes and their relationships;
- Logical consistency: degree of adherence to logical rules of data structure, attribution and relationships (data structure can be conceptual, logical or physical);
- Positional accuracy: accuracy of the position of features;
- Temporal quality: accuracy of the temporal attributes and temporal relationships of features;
- Thematic accuracy: accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships.

1.4.3.2 External quality

The external quality measures the concordance between the product and the user’s needs (e.g. spatial and temporal coverage, precision, completeness, and updates). It measures how much the dataset fulfills user’s expectations. For example, a hydrological dataset having high internal quality could be very useful for an environmental expert (high external quality) but useless for a land surveyor (poor external quality). This implies that external quality is not absolute but relative and, for this reason, it is also called “fitness for use”. Nevertheless, it is important to remember that external quality is strictly related to the internal one.

Satisfying the user’s need is a dynamic process, because he may change his requirements over time. Furthermore, external quality is highly important in decision-making processes to understand if the dataset is suitable or not.

Defining criteria to evaluate it is not an easy task. However, the general approach of producers is that external quality or the correct interpretation of a geographical dataset is under the user's responsibility (Goodchild 2010). Even the International Organization for Standardization (ISO) did not produce a list of parameters to evaluate external quality, giving it only a general consideration in “ISO 9000 - Quality management systems, Fundamentals and vocabulary” (ISO 2000).

Despite that, few authors suggested their own criteria. Bèdard and Vallière (1995) introduced six parameters to evaluate geodata quality:

- Definition: to evaluate whether the exact nature of a data and the object that it describes, that is, the “what”, corresponds to user needs (semantic, spatial and temporal definitions);
- Coverage: to evaluate whether the territory and the period for which the data exists, that is, the “where” and the “when”, meet user needs;
- Lineage: to find out where data come from, their acquisition objectives, the methods used to obtain them, that is, the “how” and the “why”, and to see whether the data meet user needs;
- Precision: to evaluate what data is worth and whether it is acceptable for an expressed need (semantic, temporal, and spatial precision of the object and its attributes);
- Legitimacy: to evaluate the official recognition and the legal scope of data and whether they meet the needs of de facto standards, respect recognized standards, have legal or administrative recognition by an official body, or legal guarantee by a supplier, etc.;

- Accessibility: to evaluate the ease with which the user can obtain the data analysed (cost, time frame, format, confidentiality, respect of recognized standards, copyright, etc.).

Hence, external quality is not as easy to evaluate as the internal one, but its importance cannot be neglected.

1.5 The dark side of the Open Data

The freedom to publish data online, which allows anyone to access and produce derived works from them, brings also some drawbacks.

Governments do not want to be too open because of that. In fact, each dataset published, even if it has a high internal quality, could have poor external quality for the citizens or, in the worst case, could be misinterpreted. This would cause a decrease in the support to the government itself by the mass (Lakomaa and Kallberg 2013). Hence, the path to the openness could also cause a political failure.

Even private companies can risk by publishing Open Data. An example is the backlash against Google in 2009 after it published some Japanese historical maps of the 18th and 19th century on Google Earth (Morozov 2012). In that period, the Japanese population was divided into classes. At the bottom of the hierarchy, there was a group called "burakumin", which was avoided and marginalized. The publication of these maps in Google Earth showed the location where the different casts lived in the past. Japanese population payed particular attention on the position of burakumin's villages, identifying the descendants of that class as the people who still lived in those areas. This inflamed new discriminatory activities. Furthermore, the Japanese Justice Ministry also faced an inquiry to Google (Alabaster 2009).

While the PAs and the private enterprises must be aware of what they publish, avoiding the possible problems that they can create, the situation is different for VGI. Despite the fact that data quality is still a fundamental topic that any producer considers, VGI does not have a specific guarantee on its internal quality, e.g. OSM heterogeneous quality results (Hacklay 2010). External quality, as it is strictly connected to the internal one, could be defined heterogeneous too.

Piotrowski (2008) described four ways for disseminating datasets: public meetings, leaks, voluntary disseminations and freedom of information requests. He introduced another category of datasets that were not analysed before, the leaks. The so-called Climatic Research Unit email controversy is an example. In 2009, a thousand of mails were hacked from the server of the Climatic Research Unit (CRU) at the University of East Anglia and thousands of documents of the Climatic Research Unit were published a week before the Copenhagen Summit about climate change (Closing the Climategate 2010). The hacker published online the raw data about global temperatures, making them available to download in various servers on the Internet. This created a huge scandal. Citizens thought there was a scientific conspiracy and a manipulation of global warming data, complicit the misunderstanding of some sentences in the mails (Lakomaa and Kallberg 2013). This situation enhanced the fact that there must be the will of sharing the data, while forcing them to be opened could be counterproductive.

However, in light of this, it has to be said that the conviction that digital technology brings only positive changes is not true (Morozov 2012). Web 2.0 is a powerful tool that must be used carefully in order to propagate information. Those who puts their data online must be aware of the possible consequences.

Publishing Open Data is in general a good way to improve the situation in various fields, but producers must be conscious of the potential negative consequences.

CHAPTER 2
CATALOGUING OF THE OPEN DATA AVAILABLE IN MILAN

The previous Chapter has stressed the importance of Open Data (OD). This part of the elaborate will investigate the real availability of OD regarding the Metropolitan city of Milan. The inquiry will first concentrate on how to access the authoritative portals that supply Open Data and then focus on a deeper study of geodata. The VGI (Volunteered Geographic Information) datasets will not be included in this analysis because we want to focus on the available authoritative data.

This research will underline the characteristics of the authoritative Open Data available in the Metropolitan city of Milan, in order to evaluate their status. The evaluation will be divided in five parts. The first one will define the study area, setting the spatial boundaries where the inquiry will be operated. The second part will refer to the methodology that the study will follow. It will list and explain the most important analysed characteristics of Open Data. Then the third section will focus on the available portals. Different databases will be illustrated and their different features will be underlined. Then, the fourth part will proceed with a deep analysis of the available geodata, classifying them according to some specific aspects (e.g. format, typology, and license). Lastly, the fifth section will show and explain a recap of the investigation, in order to comprehend the status of open geodata for the Metropolitan city of Milan.

2.1 Study area

This work focuses on the area of the Metropolitan city of Milan. Milan is the capital of the Lombardy region (in the North of Italy): it is the most populous metropolitan area and the second most populous municipality in Italy. The metropolitan area is 1'575 km^2 wide and includes 134 municipalities.

The datasets will be searched by restricting the analysis to that area or by simplifying it to a bounding box (from now on abbreviated as BBox) as shown in Figure 2.1. The selected area is quite wide due to the presence of San Colombano al Lambro, an exclave that is 22 kilometres apart from Milan in the South-East direction.

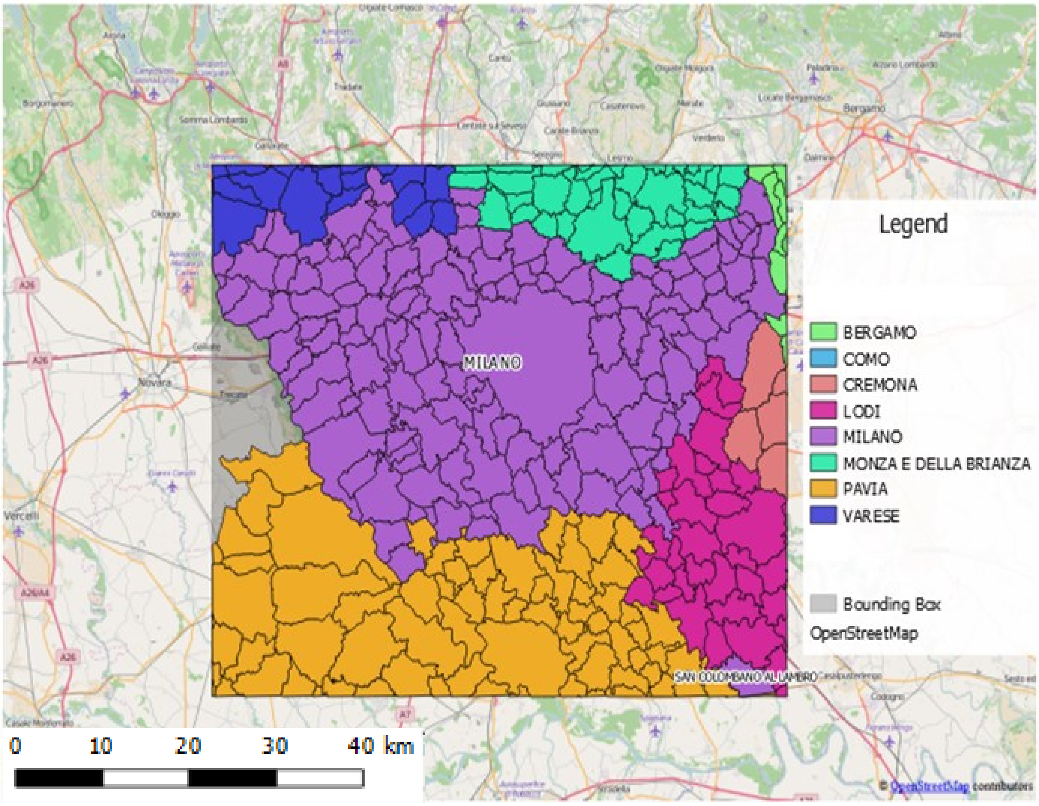


Figure 2.1 – Study area (base map: © OpenStreetMap contributors)

As Figure 2.1 shows, in the BBox some parts of other Italian provinces are considered. The image displays the provinces of Varese in blue, Monza e della Brianza in light green, Bergamo in green, Cremona in pink, Lodi in purple and Pavia in orange. On the left side is possible to see a lack of municipalities because that territory belongs to Piedmont, an Italian region alongside Lombardy.

2.2 Methodology

A good methodology is needed for underlining the relevant aspects and for operating a reasonable analysis. In this way, the cataloguing will interest information about the open datasets available in the study area. Therefore, in order to have a general idea of the available data, some parameters will be assessed. For all the available Open Data portals, we will first evaluate: the spatial coverage, the data level, the temporal coverage, the formats and the contents. In the second part of the assessment, where a deeper analysis on the datasets will be performed, even their contents and update time will be analysed. All those terms will be explained in the next subsections.

2.2.1 Spatial coverage

Generally, spatial coverage is defined as an area, specified with coordinates or with names, in which the data fit. Its edges can have different shapes: not regular (e.g. boundary of a municipality or a province) or regular.

In this case of study, the spatial coverage is classified into:

- Worldwide;
- National scale (i.e. Italy);
- Regional scale (i.e. Lombardy Region);
- Provincial scale (i.e. Metropolitan city of Milan);
- Municipal scale (i.e. Municipality of Milan);
- Local scale.

Figure 2.2 shows the different classes. In the left image it is possible to see the first two categories (worldwide and national), while the right one displays the others except the local scale. In particular, the yellow area represents Lombardy Region (regional scale), the orange shape delineates the Metropolitan city of Milan (provincial scale) and, lastly, the red polygon shows the Municipality of Milan (municipal scale).

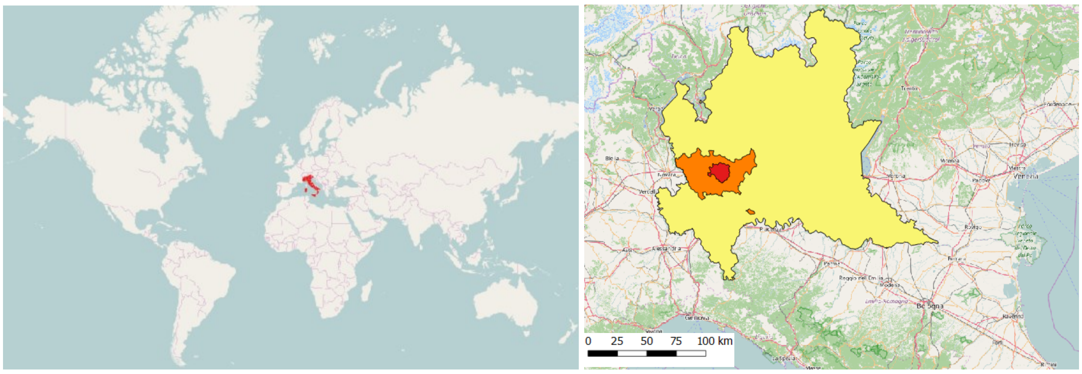


Figure 2.2 – Spatial coverage (sources: Italian boundary and Lombardy Region Geoportal for Lombardy region, Metropolitan city of Milan and Milan Municipality shapefiles; base map: © OpenStreetMap contributors)

Spatial coverage information allows to limit the research within a specific geographic location. Thus, different datasets regarding the study are can be collected, in order to compile a catalogue of the available territorial information.

2.2.2 Data level

Another geographic parameter analysed is the data level. It refers to how specific a dataset is. For example, a product can have a provincial spatial coverage (maximum extent of the data) with a provincial data level (see Figure 2.3 left) or a municipal data level (see Figure 2.3 right). In the former case, there is only one piece of information, while in the latter a multitude of them. Obviously, the latter allows to better understand the spatial distribution of the variable described by the dataset.

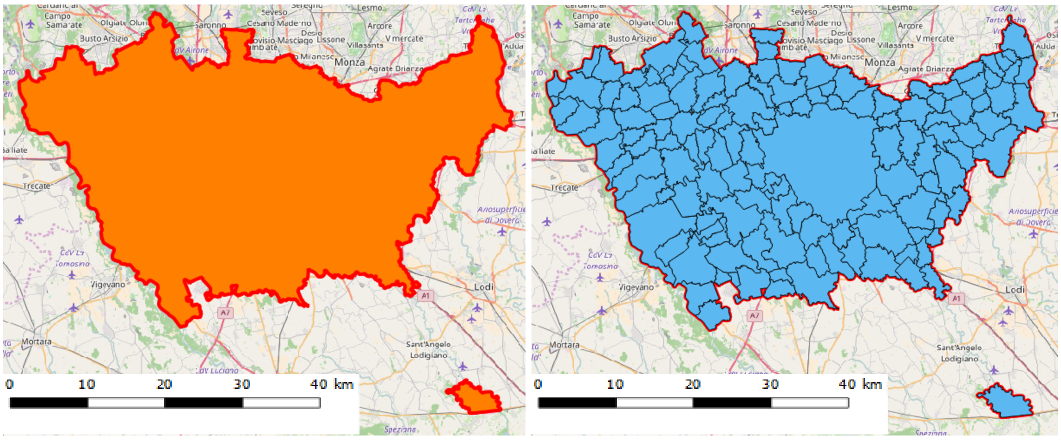


Figure 2.3 – Metropolitan city of Milan with different data levels (sources: shapefiles: Region Geoportal for Lombardy region; base map: © OpenStreetMap contributors)

The data level has always an equal or smaller dimension compared to the spatial coverage. Similarly to the spatial coverage, it can be national, regional, BBox, provincial, municipal or local.

2.2.3 Temporal coverage

Temporal coverage is defined as the time period during which the data was collected or the observations were made. In other cases, it represents the temporal length which an activity or collection is thematically linked to (e.g. from 2014 to 2016; in the 18th century). It is often set in years.

In this analysis, it represents the amount of time that the available information covers, from the oldest to the newest one. For example, if in a portal there are two datasets, one related to 1990 and the other to 2010, the temporal coverage will be 1990-2010.

Chapter 1 illustrated the huge spreading of available online data underlined by a positive trend of publications, because more and more entities have decided to embrace the policy of data openness. Hence, a further cataloguing operation is required to verify this statement for the Metropolitan city of Milan case study. Specifically, the analysis will be performed by indicating the year in which each evaluated datasets was published. This will provide information about the amount of data published online in the last decades and will allow to verify if the geodata related to the Metropolitan city of Milan follow or not the global trend towards openness.

2.2.4 Format

Formats are sets of standards used to encode the data for storing them in a computer. They can be mainly classified, as explained in Subsection 1.1.2, into human or machine-readable formats. In this case, only the second ones are taken into consideration because they can show all the information they contain, hence they are the most suitable for the chosen investigations.

In the Metropolitan city of Milan case study, datasets are classified into non-geodata and geodata (defined in Section 2.3). The first ones are presented in formats such as online visualizations (images or graphs) or tables (XLS, HTML, CSV, etc.).

On the other hand, geodata can be classified into:

- Vector datasets (e.g. Shapefiles, CSV and JSON, etc.);
- Raster datasets (e.g. GeoTIFF, ASCII grid, JPEG, etc.);
- GeoWeb Services (e.g. Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), etc.).

Of particular interest are some text-formatted Open Data such as CSV and JSON that may contain addresses. As they contain geographic information that allows them to be georeferenced with geocoding algorithms, they can be considered vector geodata too.

2.2.5 License

The licenses, as already stated in Section 1.2, are the tool used by data creators to define the copyright permission for their products. Licenses are a way to declare if data can be copied, edited, distributed, remixed and built upon, following the copyright law (Creative Common 2016).

Licenses are a fundamental characteristic of the open datasets. Table 2.1 summarizes the characteristics of the licenses found together with the permissions they allow on data. Each of these licenses is fully explained in the following paragraphs.

Table 2.1 – Table of licenses for the datasets of the Metropolitan city of Milan case study.

	Share	Adapt	Commercial use	Attribution	Indicate changes	Share A Like
CC0	✓	✓	✓	X	X	X
CC-BY	✓	✓	✓	✓	✓	X
CC-BY-SA 3.0 IT	✓	✓	✓	✓	✓	✓
CC-BY-NC-SA 3.0 IT	✓	✓	X	✓	✓	✓
IODL 2.0	✓	✓	✓	✓	✓	X

2.2.5.1 CC0, Public Domain Dedication

People who associate this license to a work have decided to dedicate the product to the public domain by waiving all of their rights worldwide under copyright law, including all the related and neighbouring rights, to the extent allowed by law. It is possible to copy, modify, distribute and perform the work, even for commercial purposes, without asking permission (CC0 Creative Commons 2016).

2.2.5.2 CC-BY, Attribution

It gives the freedom to (CC-BY Creative Commons 2016):

- Share: copy and redistribute the material in any medium or format;
- Adapt: remix, transform, and build upon the material for any purpose, even commercial.

Under the following conditions:

- Attribution: give appropriate credit, provide a link to the license, and indicate if changes were made. It is possible to do so in any reasonable manner, but not in any way that suggests the licensor endorses the use;
- No additional restrictions: not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

CC-BY-SA 3.0 IT, Attribution-ShareAlike 3.0 Italy

It gives the freedom to (CC-BY-SA 3.0 IT Creative Commons 2016):

- Share: copy and redistribute the material in any medium or format;
- Adapt: remix, transform, and build upon the material.

Under the following conditions:

- Attribution: give appropriate credit, provide a link to the license, and indicate if changes were made. It is possible to do so in any reasonable manner, but not in any way that suggests the licensor endorses the use;
- ShareAlike: if the data are remix, transform, or build upon the material, it has to be distributed under the same license as the original.

2.2.5.3 CC-BY-NC-SA 3.0 IT, Attribution-NonCommercial-ShareAlike 3.0 Italy

It gives the freedom to (CC-BY-NC-SA 3.0 IT Creative Commons 2016):

- Share: copy and redistribute the material in any medium or format;
- Adapt: remix, transform, and build upon the material.

Under the following conditions:

- Attribution: give appropriate credit, provide a link to the license, and indicate if changes were made. It is possible to do so in any reasonable manner, but not in any way that suggests the licensor endorses the use;
- NonCommercial: not use the material for commercial purposes;
- ShareAlike: if the data are remix, transform, or build upon the material, it has to be distributed under the same license as the original.

2.2.5.4 IODL 2.0

It gives the freedom to (Italian Open Data License v2.0 2016):

- Share: copy and redistribute the material in any medium or format;
- Adapt: remix, transform, and build upon the material.

Under the following conditions:

- Attribution: give appropriate credit, provide a link to the license, and indicate if changes were made. It is possible to do so in any reasonable manner, but not in any way that suggests the licensor endorses the use;
- Not publish he data in a way that they can be mistaken on purpose.

2.2.6 Content

The contents of geodata can be extremely various. They can contain extensive areal information of the territory (e.g. land use maps) or just few point-wise information (e.g. location of something). For this reason, the cataloguing will hold a classification based on the typologies of data. The classification will divide geodata products into five categories:

- Topographic data – "T" (e.g. roads, railways, buildings, Digital Terrain Models - DTMs, hydrographic data, etc.);
- Environmental data – "A" (e.g. land use maps, geological maps, forest maps, maps of landslides and earthquake susceptibility, etc.);
- Governance data – "G" (e.g. data related to population census, culture, tourism, education, services for citizens, etc.);
- Airborne observations – "O" (e.g. LiDAR data, orthophotos, aerial and satellite imagery, products derived from SAR data, etc.);
- Sensor observations – "S" (e.g. rain, temperature, river flow, etc.).

2.3 Databases consulted

This section illustrates the evaluation of the available OD databases. A first evaluation was developed in April 2016 (Brovelli et al. 2016) and this thesis will describe the same research updated in January 2017. A total amount of 22 portals providing OD for the Metropolitan city of Milan were detected. These portals will be divided into GIS databases, as the ones created by authorities working with geographic information systems (e.g. Environmental Agencies, Spatial Agencies and scientific research centres) and non-GIS databases, the more general ones (e.g. Open Government data and statistics authorities). The difference is that in the GIS databases each dataset is a geodata, while in non-GIS databases datasets in other (non-GIS) formats can be found.

2.3.1 GIS databases

7 geoportals providing OD for the Metropolitan city of Milan have been catalogued as GIS databases. 3 of them are Italian and the others are international. They will be described in the following subsections.

2.3.1.1 Lombardy Region Geoportal

It is the official geoportal of Lombardy Region (*Geoportale Regione Lombardia*), containing all its cartographic resources. It is available at http://www.geoportale.regione.lombardia.it. It gives access to a wide range of geospatial datasets, from the most used territorial information to the thematic ones. It provides a service for data visualization, which allows the composition of new maps by overlapping different levels, and a download section (Geoportale della Regione Lombardia 2016). The spatial coverage regards Lombardy Region in general. In some cases, the geoportal contains information of provinces and municipalities.

In January 2017, for the Metropolitan city of Milan 694 datasets have been found in the Lombardy Region Geoportal. They have different formats (vector, raster and services) and a temporal coverage that goes from 1975 to 2016. The licenses are various, based on the dataset involved. The most recurring are IODL 2.0 and CC-BY-NC-SA 3.0 Italy.

2.3.1.2 National Cartographic Portal

The Italian National Cartographic Portal (*Portale Geografico Nazionale)* is the official national geoportal of the Italian Environmental and Territorial and Maritime Safeguard Ministry. It is the national access point for the European Infrastructure for Spatial Information in Europe (INSPIRE) Directive, allowing third parts and citizens to access the available territorial information (Geoportale Nazionale 2016). It is available at http://www.pcn.minambiente.it/GN.

Products concern the entire Italian territory with different data levels and a temporal coverage from 2013 to 2015. Considering Milan area, in January 2017, 137 datasets have been found. All of them are web services (WMS, WFS, WCS) available under a CC-BY-SA 3.0 IT or CC-BY-NC-SA 3.0 IT licenses.

2.3.1.3 GlobeLand30

GlobeLand30 is a Chinese product that consists of a Global Land Cover (GLC) map. It covers the entire Earth with squared cells of 30 meters, computed with an automated classification on more than ten thousands of satellite images. It is available at http://www.globallandcover.com. At the time of writing, the available data correspond to 2000 and 2010; hence, only two maps are accessible. The copyright is an open customized license and it defines that products are free of charge for scientific researches or for public uses only.

2.3.1.4 EarthExplorer USGS

EarthExplorer is a huge catalogue of satellite images, aerial photographs and cartographic products of the U.S. Geological Survey, http://earthexplorer.usgs.gov. This database provides, as an example, all the Landsat archives, some Digital Elevation Models (e.g. ASTER, GMTED2010 and SRTM) and the NASA's Land Processes Distributed Active Archive Centre (LP DAAC). Even the temporal interval is quite wide. The oldest pieces of information are declassified data of the first generation U.S. photo-intelligence satellite of Earth surface (1962), while the newest are up-to-date ones. Data can be queried using a spatial range (i.e. bounding box) and/or a temporal range. Unfortunately, not all the datasets are open and some of them have to be purchased. The licenses vary based on the specific selected dataset.

For the Metropolitan city of Milan, more than a thousand datasets are available both for free of charge and for a fee.

2.3.1.5 GEOSS

GEOSS is the geoportal provided by GEO (Group on Earth Observations), an archive of Earth Observation implemented and operated by European Space Agency (ESA). It is available at http://www.geoportal.org. As EarthExplorer, a BBox is needed in order to query the archive.

For Milan area, at the time of writing, 13 "full and Open Datasets" have been found. The temporal coverage is 1990-2014. Most information about the datasets are missing in the metadata interface (e.g. the organization or the description of the product), hence they are not analysed further.

2.3.1.6 ESA Earth Online

It is an archive of the Earth observation (EO) data products available for each ESA and Third Party mission, https://earth.esa.int/web/guest/data-access. It contains a wide set of information about EO data.

In January 2017, in the "Browse Data Product" section, 216 datasets have been listed and described for the area of study. It is possible to sort them by topic (e.g. Earth Topic, Agriculture, Atmosphere and Land), by mission (e.g. GOCE, IKONOS, LANDSAT and SPOT), by instrument (e.g. DORIS, ETM, NAOMI and TM), by typology (e.g. Atmospheric data, Gravimetry, Optical Multi Spectral Radiometry and Radar Altimetry) or by processing level. The temporal coverage is quite wide as it goes from 1978 to 2016. The license is an open customized one, defined by the European Space Agency. It can be summarized as non-commercial use and attributing the credit.

2.3.1.7 RAPu Archives

RAPu Archive (*Rete Archivi Piani urbanistici, Archivio RAPu*) is a project developed by Politecnico di Milano and Triennale di Milano in 2006. The database is a virtual catalogue of urbanistic plans and related documents. It is constituted by many digitalized materials originally located in different places in Italy, such as internal archives, local authorities, libraries and National archives. It is available at http://www.rapu.it.

For Milan area, at the time of writing (January 2017) 16 datasets are available, with a temporal coverage ranging from 1888 to 1938, in JPG format. The license is an open customized one; data can be used for didactical and research purposes only.

2.3.2 Non-GIS databases

15 portals have been catalogued as non-GIS databases. As the presence of Open Data portals is more widespread, a comprehensive classification was not possible and thus only Italian databases were considered. Few of them are also part of some European projects, hence their datasets can be found even in other international databases.

2.3.2.1 Italian National Institute of Statistics (ISTAT)

It is the Italian National Institute of Statistics (*Istituto Nazionale di Statistica*, ISTAT). ISTAT is a public research institution, which conducts state censuses and surveys. Available data can be visualized and downloaded at http://www.istat.it. The available formats are XLS and CSV, with variable spatial coverages. Some of them are at provincial level, other at municipal one.

For the study area, in January 2017, 12 different topics have been found. The fields are disparate (e.g. environmental and energy, assistance and welfare, work etc.) corresponding to 2’203 datasets. The time interval goes from 1982 to 2016 and the license is CC-BY.

2.3.2.2 Lombardy Open Data Portal

Lombardy Open Data Portal (*Portale Open Data Regione Lombardia*) is an Italian application of the 2003/98/CE Directive of the European Parliament, which regards the re-use of the public sector information. This database is reachable at https://www.dati.lombardia.it. It contains more than 1’200 open datasets, with a temporal coverage ranging from 2005 to 2016, divided in twenty topics (agriculture, commerce, government, healthcare, security, tourism, etc.).

In January 2017, 355 products related to the Metropolitan city of Milan are available at the municipal scale. They are available in CSV, JSON, XLS, XLSX and XML formats and they are updated each year. A total of 132 datasets contain also coordinates or addresses, therefore they can be considered as geodata. The license for all of them is IODL 2.0.

2.3.2.3 *Italian Portal of the Open Data*

Italian Portal of the Open Data (*Portale Italiano dell'Open Data*) is one of the widest OD databases, which contains more than 2000 datasets with a temporal coverage that ranges from 1881 to 2016. The database is reachable at http://www.datiopen.it. The available data are taken from other portals such as "Lombardy Open Data Portal" and "Data Portal of Milan Municipality". Hence, they are not analysed further.

2.3.2.4 *Data Portal of Milan Municipality*

Data portal of Milan Municipality (*Portale dati Comune di Milano*), available at http://dati.comune.milano.it, is the database where the Public Administration of Milan Municipality published its datasets. Hence, the data level is municipal.

In January 2017, the database divided the information in 10 different topics (e.g. environmental and weather, geographic information, social, education and formation and municipal administration) and, between all the categories, 273 datasets are listed in CSV, TXT or shapefile formats. Among them, 35 can be considered as geodata. Datasets refer to a large amount of years, from 1881 to 2016. The used licenses are CC0 and CC-BY.

2.3.2.5 *Provincial Statistical Yearbook*

Provincial statistical yearbook (*Annuario Statistico Provinciale,* ASP) is a regional portal created with the collaboration of Lombardy Region, Unioncamere Lombardia and ISTAT. The available datasets are sets of information collected by these authorities. The section related to the Metropolitan city of Milan is available at http://www.asr-lombardia.it/ASP-Milano.

In January 2017, the total amount of datasets has been detected as 492, divided in 26 topics. The spatial coverage is provincial and the data level provincial or municipal. The format in which they are available is XLS and the license is CC-BY. The database contains some historical data; in fact, the temporal coverage goes from 1861 to 2016.

2.3.2.6 *Quality of life*

At the time of writing (January 2017), quality of life is a list of 12 indicators (*Qualità della Vita*) listed by one of the most important economical newspapers in Italy, Sole24Ore (http://www.ilsole24ore.com). This work is aimed at determining the quality of life in different Italian provinces. It compiles and yearly updates a list, changing both the chosen parameters and the global evaluation. This project started in 1990, but at the time of writing the available data date back only to 2014. The previous data are available in a fee-paying e-book. The spatial coverage is national, while the data level is provincial. Datasets are available in HTML format under a customized open license.

2.3.2.7 Open Data of the Public Administration

The Open Data of the Public Administration (*Dati aperti della Pubblica Amministrazione*) is the Open Data portal of the Italian PA. At the time of writing, it contains 10'338 products created by 76 authorities. It is available at http://www.dati.gov.it. The datasets are taken from "Lombardy Open Data Portal" and "Data Portal of Milan Municipality".

2.3.2.8 OpenCivitas

OpenCivitas (http://www.opencivitas.it) is the Italian portal that allows citizens to access local authorities' information. It is part of an initiative to promote transparency created by the Italian Ministry of Finance and an agency called Solutions for the Public and Private Economical System (*Soluzioni per il Sistema Economico Pubblico e Privato,* SOSE).

In January 2017, it includes 126 datasets available in CSV and RDF formats under a CC-BY license. The temporal coverage goes from 2009 to 2012.

2.3.2.9 Ministry of Finance

It is the Open Data section of the Italian Ministry of Finance database. It is part of the transparency operations of the Italian government, available at http://www.mef.gov.it/operazione-trasparenza.

In January 2017, it collects 20 different topics in which 63 datasets are sorted. They are available as PDF documents under different open licenses. The spatial coverage and data level are both national, and as it is too general, this database is not of interest for the purpose of this work.

2.3.2.10 Digital Agenda of Lombardy

The Digital Agenda of Lombardy (*Agenda Digitale Lombardia*) is an initiative promoted by Lombardy Region to address and sustain the technological innovation development for its territory. It is reachable at http://www.agendadigitale.regione.lombardia.it. The datasets contained are also presented in "Lombardy Region Open Data", hence this data source is not analysed further.

2.3.2.11 Regional Agency for Protection of the Environment of Lombardy

Regional Agency for Protection of the Environment of Lombardy (*Agenzia Regionale per la Protezione dell'Ambiente,* ARPA) is an institution that concerns protection and environmental monitoring, integrating the efforts of some regional and local agencies. The database allows citizens to download data that has been collected in monitoring stations located all over the territory. The portal is available at http://ita.arpalombardia.it/ITA/servizi/richiesta_dati/idro_pluvio_termo.asp.

For the selected study area, there are 9 stations in Milan Municipality and 17 in the Metropolitan city of Milan. Their locations are known in terms of either geographic coordinates or addresses. Hence, they can be considered as geodata. The spatial coverage is regional, while the data level is local.

At the time of writing, the reached datasets contain various kinds of information such as temperature, wind velocity and direction and relative humidity. The total amount of datasets is 108. The temporal coverage is quite wide: in some locations, it goes from 1989 to the present. They are available as row data, hourly or daily. The formats in which they are downloadable are CSV or PDF. The license is an open customized one: data can be used also for commercial purposes, but the credits must be explicated.

2.3.2.14 Italian National Institute for Environmental Protection and Research

The Italian National Institute for Environmental Protection and Research (*Istituto Superiore per la Protezione e la Ricerca Ambientale,* ISPRA) is the scientific centre in charge of implementing the INSPIRE Directive in Italy by publishing datasets regarding environmental information.

At the time of writing, 306 datasets are available which sorted in nine categories (http://www.isprambiente.gov.it/it/banche-dati). Furthermore, other environmental indicators are collected in the "Yearly indicators Database" (*Banca dati indicatori Annuario*). The 2016 edition is available online at http://annuario.isprambiente.it. The spatial coverage is national, while the data level is different from dataset to dataset (some are national, other are provincial or municipal) with a temporal coverage that ranges from 2007 to 2016. The information can be visualized in charts or in the online version of the reports. The license of the data is IODL 2.0.

2.3.2.12 Postmetropoli Atlas

Postmetropoli Atlas (*Atlante Postmetropoli*) is a project of national interest that derives information starting with public data. It has been financed by the Italian Ministry of Educational, University and Research and coordinated by Politecnico di Milano. It is available at http://www.postmetropoli.it/atlante.

In January 2017, it contains 231 elaborations that took as input some public data (except for some few cases), such as from ISTAT, ISPRA, OSM, etc., sorted, then, in 9 categories. The temporal coverage goes from 1971 to 2016. Data are related to the entire Italian territory and, in some cases, they contain a more specific analysis with regional or provincial information. The datasets can be visualized in a webpage, both with a map and some tables. The license is an open customized one; it allows to use them for non-commercial scope, by citing the source.

2.3.2.13 Statistical Information System of Local Authorities

The Statistical Information System of Local Authorities (*Sistema Informativo Statistico Enti Locali, PORTALE SIS.EL.*) is a regional governmental portal, reachable at http://dwh.servizirl.it/SASPortal/main.do. It allows to access two different sections: E-Health and E-Government. The available data derive from ISTAT and ASP Milano, hence, it was not analysed further.

2.3.2.15 Mobility Agency for Environment and Territory

The Mobility Agency for Environment and Territory (*Agenzia Mobilità Ambiente Territorio,* AMAT) is an authority that supplies an archive of information on local public transportation of the city of Milan. In January 2017, it lists the time at which a public transportation has passed by a specific station since 2016. It is reachable at https://www.amat-mi.it/it/mobilita/dati-strumenti-tecnologie/dati-gtfs. Datasets are available in the General Transit Feed Specification (GTFS) format, used for public transportation schedules and associated geographic information), under a CC-BY 4.0 license.

2.3.3 Summary of the databases

The characteristics of all the portals described above are summarized in two table (tables 2.2 and 2.3). Specifically, table 2.2 shows the GIS databases, while table 2.3 displays the non-GIS ones. They list the parameters analysed before (i.e. spatial coverage, data level, temporal coverage, format and licenses) and add the last access to the portal.

Table 2.2 shows the 7 sources catalogued as GIS databases. They have different spatial coverages; half of them are worldwide, while the others are related to Italy. The data level is heterogeneous. The ones related most to satellite information require a BBox, while the others are more specific (municipal and local).

Even the time interval is quite wide: nevertheless, most of the geoportals contain recent datasets. The only exceptions is "RAPu Archive" because it contains digitalized historical maps.

Table 2.2 – GIS databases

	Source	**Spatial coverage**	**Data level**	**Temporal coverage**	**Format**	**License**	**Last access**
1	Lombardy Region Geoportal	Regional	Various	1975-2016	Various	Various	10/01/2017
2	National Cartographic Portal	National	Various	2013-2015	Various	Various	24/01/2017
3	GlobeLand 30	Worldwide	30 x 30 m pixels	2000, 2010	Various	Custom	09/01/2017
4	EarthExplorer USGS	Worldwide	BBox	1962-2015	Various	Various	09/01/2017
5	GEOSS	Worldwide	BBox	2006	Services	Custom	16/01/2017
6	ESA Earth Online	Worldwide	BBox	1978-2015	Various	Custom	09/01/2017
7	RAPu Archives	Milano	Local	1888-1930	Rasters	Custom	09/01/2017

Table 2.3 shows the non-GIS databases, by listing all the 15 portals. It enumerates more than double sources with respect to the previous table; they are more common, due to their more general contents. In fact, they are mostly Italian application of the Open Government fundamentals, realized by national or local authorities (e.g. Open Data of the Public Administration, Data Portal of Milan Municipality and Ministry of Finance). In other cases, they are public statistical agencies that make their data open on the web (e.g. National Statistical Institution, Provincial Statistical Yearbook) or, lastly, they are portals that supply scientific data.

The spatial coverage is more expanded with respect to the GIS databases; in fact, it is generally national. On the other hand, the data level is different from portal to portal; quite often it varies based on the selected datasets. The temporal coverage is vast: some historical data regard even the 18th century. Nevertheless, all the databases contain recent datasets. Considering the licenses, the most common one is the CC-BY, while some others use a proprietary one. Under this parameter, the situation appears heterogeneous.

Table 2.3 – Non-GIS databases

	Source	**Spatial coverage**	**Data level**	**Temporal coverage**	**License**	**Last access**
1	National Statistical Institution	National	Various	1982-2016	CC-BY	09/01/2017
2	Lombardy Open Data Portal	Regional	Municipal	2005-2016	IODL 2.0	09/01/2017
3	Italian Portal of the Open Data	National	Municipal	1881-2016	Various	09/01/2017
4	Data Portal of Milan Municipality	Municipal	Municipal	1881-2016	Various	09/01/2017
5	Provincial Statistical Yearbook	Provincial	Various	1861-2016	CC-BY	09/01/2017
6	Quality of life	National	Provincial	2014-2016	Custom	09/01/2017
7	Open Data of the Public Administration	National	Various	1881-2016	Various	09/01/2017
8	OpenCivitas	National	Provincial	2009-2012	CC-BY	09/01/2017
9	Ministry of Finance	National	National	2014-2016	Various	09/01/2017
10	Digital Agenda of Lombardy	Regional	Various	2005-2016	IODL	09/01/2017
11	Regional Agency for Protection of the Environment of Lombardy (ARPA)	Regional	Local	1989-2016	Custom	09/01/2017
12	Superior Institute for the Environmental Defence and Research	National	Various	2014-2016	IODL 2.0	09/01/2017
13	Postmetropoli Atlas	National	Various	1971-2016	Custom	09/01/2017
14	Statistical Information System of Local Authorities	National	Municipal	1861-2016	CC-BY	09/01/2017
15	Mobility Agency for Environment and Territory	National	Local	2016	CC-BY4.0	09/01/2017

2.4 Datasets used for the cataloguing operation

Due to the vastness of the available datasets and the extreme difficulty in finding and cataloguing them all, the further analysis is only based on open geodata regarding the Metropolitan city of Milan, released by Italian institutions in the last few decades. The cataloguing operation has been made in January 2017 and it records a total of 1'099 open geodata. The Italian institutions' portals providing them are:

- Lombardy Region Geoportal;
- National Cartographic Portal;
- Regional Agency for Protection of the Environment of Lombardy (ARPA);
- Lombardy Open Data Portal;
- Data Portal of Milan Municipality.

2.4.1 Lombardy Region Geoportal

Globally 694 layers are available; 669 of them are vectors (shapefiles, abbreviate as SHP), 8 are raster and 17 are WMS. Table 2.4 lists the available vector layers, by sorting them in categories. It shows a column ("n. of available layers") in which the number of datasets included is specified. Then, Table 2.5 shows the raster data and Table 2.6 lists the services. The "Content" column includes "T" as topographic data, "G" as governance data, "A" as Environmental data and "O" as Airborne observation.

Table 2.4 – List of vector layers of Lombardy Region Geoportal

	Number of layers	**Update**	**Format**	**License**	**Content**
AGAPU analysis and agricultural administrations	14	2013	SHP	CC-BY-NC-SA 3.0 Italia	G
Health care agency	1	2016	SHP	IODL 2.0	G
Agritourisms	1	2012	SHP	CC-BY-NC-SA 3.0 Italia	G
Settings of PTRA	1	2015	SHP	None	T
Architectures of cultural interest	2	2011	SHP	CC-BY-NC-SA 3.0 Italia	G
Secured architectures of particular interest	1	2011	SHP	CC-BY-NC-SA 3.0 Italia	G
Agricultural areas of the state of art	1	2009	SHP	IODL 2.0	G
Area of wine quality	1	2013	SHP	IODL 2.0	G
Dismissed area	1	2009	SHP	CC-BY-NC-SA 3.0 Italia	T

Priority areas for biodiversity	5	2012	SHP	CC-BY-NC-SA 3.0 Italia	A
Protected areas	8	2016	SHP	CC-BY-NC-SA 3.0 Italia	A
Hydrogeological linked area (2013)	1	2013	SHP	CC-BY-NC-SA 3.0 Italia	T
Hydrographical basins	4	2006	SHP	CC-BY-NC-SA 3.0 Italia	T
Geological banks of the underground	2	2015	SHP	CC-BY-NC-SA 3.0 Italia	A
Informational bases of environmental cartography - Exploitation activity map	6	1987	SHP	IODL 2.0	A
Informational bases of environmental cartography - map of the hydrogeological disease and dangerousness	10	1987	SHP	IODL 2.0	A
Informational bases of environmental cartography - Land use map of vegetation orientation	11	1987	SHP	IODL 2.0	A
Informational bases of environmental cartography – Land use capacity map	1	1987	SHP	IODL 2.0	A
Informational bases of environmental cartography - map of the soil productive predisposition	1	1987	SHP	IODL 2.0	A
Informational bases of environmental cartography – Map of the naturalistic and landscaping relevance	16	1987	SHP	IODL 2.0	A
Informational bases of environmental cartography – Geomorphological map	12	1987	SHP	IODL 2.0	A
Informational bases of environmental cartography – Hydrological map with permeability indications	24	1987	SHP	IODL 2.0	A
Informational bases of environmental cartography – Lithological map	8	1987	SHP	IODL 2.0	A
Environmental bases of the plain – Exploitation soil activity	12	2007	SHP	CC-BY-NC-SA 3.0 Italia	A
Environmental bases of the plain – Geomorphology	5	2007	SHP	CC-BY-NC-SA 3.0 Italia	A
Environmental bases of the plain – Hydrology	3	2007	SHP	CC-BY-NC-SA 3.0 Italia	A
Environmental bases of the plain – Lithology	2	2003	SHP	CC-BY-NC-SA 3.0 Italia	A
Environmental bases of the plain – Naturalistic and landscape relevance	20	2007	SHP	CC-BY-NC-SA 3.0 Italia	A
Informatics bases of the ground	2	2013	SHP	IODL 2.0	A
Seed forests	1	2011	SHP	IODL 2.0	A
Maps of Lombardy forestall management plans	1	2016	SHP	IODL 2.0	A
Lombardy forestry real types map	1	2016	SHP	IODL 2.0	A
Avalanche probable location map (CLPV)	8	2016	SHP	IODL 2.0	A
Geological map 250.000	2	1990	SHP	CC-BY-NC-SA 3.0 Italia	A

Fishing regional map	11	2010	SHP	IODL 2.0	A
Technical Regional Map scale 1:10000	41	2006	SHP	IODL 2.0	T
Mine cadastre	2	2015	SHP	IODL 2.0	A
Georeferenced waste cadastre	2	2016	SHP	IODL 2.0	A
Municipalities seismic classification	1	2016	SHP	IODL 2.0	A
Corine 1990	1	1990	SHP	CC-BY-NC-SA 3.0 Italia	A
Corine 2000	1	2006	SHP	CC-BY-NC-SA 3.0 Italia	A
Topographic Regional Database - Access	2	2015	SHP	IODL 2.0	T
Topographic Regional Database - Altimetry	3	2015	SHP	IODL 2.0	T
Topographic Regional Database - Building	5	2015	SHP	IODL 2.0	T
Topographic Regional Database - Administrative extension, street toponyms	2	2016	SHP	IODL 2.0	T
Topographic Regional Database - Railways and streets	7	2015	SHP	IODL 2.0	T
Topographic Regional Database - terrain shapes	7	2015	SHP	IODL 2.0	T
Topographic Regional Database - Photogrammetry, cartography and geodesy	8	2015	SHP	IODL 2.0	T
Topographic Regional Database - Locality	2	2015	SHP	IODL 2.0	T
Topographic Regional Database - Manufactory	13	2015	SHP	IODL 2.0	T
Topographic Regional Database - Transport infrastructure manufactories	1	2015	SHP	IODL 2.0	T
Topographic Regional Database - Transport infrastructure works	2	2015	SHP	IODL 2.0	T
Topographic Regional Database - Soil defence works	1	2015	SHP	IODL 2.0	T
Topographic Regional Database - Hydraulically and regulation works	5	2015	SHP	IODL 2.0	T
Topographic Regional Database - Pertinence	7	2015	SHP	IODL 2.0	T
Topographic Regional Database - Technological network	5	2015	SHP	IODL 2.0	T
Topographic Regional Database - Hydrographic network	4	2015	SHP	IODL 2.0	T
Topographic Regional Database - Hydrographic surfaces	4	2015	SHP	IODL 2.0	T
Topographic Regional Database - Vegetation	8	2015	SHP	IODL 2.0	T
Topographic Regional Database - Viability, mobility and transport	13	2015	SHP	IODL 2.0	T

Statistical data	2	2014	SHP	IODL 2.0	G
Flooding directive 2007/60/CE	8	2015	SHP	IODL 2.0	A
Commercial districts	6	2016	SHP	IODL 2.0	G
DUSAF 1.0 - Landuse 1999-2000	2	2003	SHP	IODL 2.0	A
DUSAF 1.1 - Landuse 1999/00	2	2008	SHP	IODL 2.0	A
DUSAF 2.0 - Landuse 2005/7	3	2010	SHP	IODL 2.0	A
DUSAF 2.1 - Landuse 2007	2	2010	SHP	IODL 2.0	A
DUSAF 3.0 - Landuse 2009	2	2011	SHP	IODL 2.0	A
DUSAF 4.0 - Landuse 2012	2	2012	SHP	IODL 2.0	A
Resurgences of Lombardy	1	2013	SHP	IODL 2.0	A
Geology, aquifers - Group A	4	2007	SHP	IODL 2.0	A
Geology, aquifers - Group B	5	2007	SHP	IODL 2.0	A
Geology, aquifers - Group C	6	2007	SHP	IODL 2.0	A
Geology, aquifers - Group D	5	2007	SHP	IODL 2.0	A
Lombardy glaciers	14	2016	SHP	IODL 2.0	A
Big dams	1	2000	SHP	CC-BY-NC-SA 3.0 Italia	T
Big sales structures	1	2014	SHP	IODL 2.0	T
Grids	2	2000	SHP	IODL 2.0	T
Methane gas service stations	1	2014	SHP	IODL 2.0	G
Historical and traditional signs	1	2014	SHP	IODL 2.0	G
Catalogue of Lombardy's landslide phenomena	6	2014	SHP	CC-BY-NC-SA 3.0 Italia	A
Administrative limits 2012 with DBT/PGT update	6	2012	SHP	IODL 2.0	T
Administrative limits 2013	5	2014	SHP	IODL 2.0	T
Administrative limits 2014 with DBT/PGT update	5	2014	SHP	IODL 2.0	T
Administrative limits 2015 DBT/PGT updated	6	2015	SHP	IODL 2.0	T
Administrative limits 2016 DBT/PGT updated	5	2016	SHP	IODL 2.0	T
Municipal mosaic of urbanistic tools MISURC	24	1999	SHP	CC-BY-NC-SA 3.0 Italia	G

Local and historical shops	1	2014	SHP	IODL 2.0	G
Soil defence works (ODS)	8	2012	SHP	IODL 2.0	A
Landscape - Safeguard orientation	4	2012	SHP	IODL 2.0	A
PGT - Table of plan forecast	28	2016	SHP	IODL 2.0	T
PGT to be transmitted to the region	1	2015	SHP	CC-BY-NC-SA 3.0 Italia	T
Acoustic classification plans	1	2015	SHP	IODL 2.0	S
Regional landscape plan	21	2012	SHP	CC-BY-NC-SA 3.0 Italia	G
Lombardy region pilot book	2	2016	SHP	IODL 2.0	T
Protection and water usage program - PTUA	6	2013	SHP	CC-BY-NC-SA 3.0 Italia	A
Negotiated programming	1	2015	SHP	IODL 2.0	G
Integrated program of local development (PISL Montagna)	1	2014	SHP	IODL 2.0	T
Shipping of Lombardy	20	2015	SHP	IODL 2.0	A
Points refinement of the IGM network	3	2016	SHP	None	T
Regional Cycle Network	2	2014	SHP	CC-BY-NC-SA 3.0 Italia	T
Ecological Regional Network (RER)	6	2011	SHP	CC-BY-NC-SA 3.0 Italia	A
Unified regional hydrographic network	10	2016	SHP	IODL 2.0	T
Schools in Lombardy	1	2012	SHP	IODL 2.0	G
Overflow service, hydraulical and hydrogeological protection (d.g.r. 3723 of 19/06/2015)	3	2016	SHP	CC-BY-NC-SA 3.0 Italia	T
Trade show system	1	2011	SHP	IODL 2.0	A
Reclaim sites and contaminated sites	1	2016	SHP	CC-BY-NC-SA 3.0 Italia	A
Topographic database (DbT) development state	1	2016	SHP	IODL 2.0	T
Hydro meteorological survey stations 10000	1	2000	SHP	IODL 2.0	T
Agricultural, forest and pasture streets	1	2015	SHP	IODL 2.0	T
Streets, railways and metropolitans	10	2016	SHP	IODL 2.0	T
Health care structures	2	2013	SHP	IODL 2.0	G
Municipal geological studies	12	2016	SHP	CC-BY-NC-SA 3.0 Italia	A

Table of territorial plans coordinated by the provinces (PTCP)	30	2014	SHP	IODL 2.0	T
Forest cuts location	1	2012	SHP	IODL 2.0	A
Landuse 1980	1	2011	SHP	IODL 2.0	A
Historical landuse (1954)	2	2011	SHP	IODL 2.0	A
Landscape restrictions	11	2014	SHP	CC-BY-NC-SA 3.0 Italia	A
Zoning for air quality evaluation	2	2013	SHP	CC-BY-NC-SA 3.0 Italia	T

Table 2.5 – List of raster layers of Lombardy Region Geoportal

	Number of layers	**Update**	**Format**	**License**	**Content**
Map of the ecological coherent forest type of Lombardy	1	2012	raster	IODL 2.0	A
Lombardy forest real types map	1	2012	raster	IODL 2.0	A
Regional Technical Map scale 1:10.000 updated by Topographical Database	1	2016	raster	IODL 2.0	T
Technical regional map black/white CT50	1	2003	raster	IODL 2.0	T
Agricultural use map - SIARL data 2012	1	2013	raster	IODL 2.0	A
Digital Terrain Model DTM5x5	1	2015	raster	IODL 2.0	T
Diffused background of Lombardy	1	2008	raster	IODL 2.0	T
Agricultural value	1	2008	raster	IODL 2.0	A

Table 2.6 – List of services of Lombardy Region Geoportal

	Number of layers	**Update**	**Format**	**License**	**Content**
Physical map of Lombardy Region	1	2009	WMS	CC-BY-NC-SA 3.0 Italia	O
Political map of Lombardy Region	1	2009	WMS	CC-BY-NC-SA 3.0 Italia	T
Regional technical chart 1:10000 updated from the topographic database	1	2015	WMS	IODL 2.0	T
Regional technical chart 1:10000 ed. 1980-94	1	2008	WMS	IODL 2.0	T
Flood directive 2007/60/CE - 2015 Revision	1	2015	WMS	CC-BY-NC-SA 3.0 Italia	T
Digital Terrain Model DTM5x5	1	2015	WMS	CC-BY-NC-SA 3.0 Italia	T
DTM20 - Digital Terrain Model	1	2008	WMS	IODL 2.0	T
DUSAF 2.1 - Landuse 2007	1	2010	WMS	CC-BY-NC-SA 3.0 Italia	A

Historical flood between Ticino and Adda	1	2015	WMS	CC-BY-NC-SA 3.0 Italia	A
Sentinel 2 satellite image	1	2016	WMS	CC-BY-NC-SA 3.0 Italia	O
Cadastre map	1	2015	WMS	CC-BY-NC-SA 3.0 Italia	T
Orthophoto 1975	1	1975	WMS	CC-BY-NC-SA 3.0 Italia	O
Orthophoto 2007	1	2014	WMS	IODL 2.0	O
Orthophoto AGEA 2012	1	2013	WMS	IODL 2.0	O
PGT - Table of plan forecast	1	2016	WMS	IODL 2.0	T
Unified regional hydrographic network	1	2015	WMS	IODL 2.0	T
GAI fly 1954	1	2014	WMS	CC-BY-NC-SA 3.0 Italia	O

2.4.2 National Cartographic Portal

It has 137 different available layers. All of them are web services: 70 WMS (Table 2.7), 54 WFS (Table 2.8) and 13 WCS (see Table 2.9). The "Content" column includes: "T" as topographic data, "G" as governance data, "A" as Environmental data and "O" as Airborne observation.

Table 2.7 – WMS layers of National Cartographic Portal

	Update	License	Content
Interferometric products - ERS ENVISAT ascending comparison	2013	CC-BY-SA 3.0 IT	O
Interferometric products - ERS ENVISAT descending comparison	2013	CC-BY-SA 3.0 IT	O
IBA - Important Birds Areas	2013	CC-BY-SA 3.0 IT	A
National atlas of the desertification risk areas	\	CC-BY-SA 3.0 IT	A
Basin authorities	2013	CC-BY-SA 3.0 IT	G
Principal and secondary hydrological basins	2013	CC-BY-SA 3.0 IT	A
Italian ecopedological map	2013	CC-BY-SA 3.0 IT	A
Italian phytoclimatical map	2013	CC-BY-NC-SA 3.0 IT	A
Italian geolithological map	2013	CC-BY-SA 3.0 IT	A

Italian geological map	2013	CC-BY-NC-SA 3.0 IT	A
Forestry antifire cartography (AIB)	2014	CC-BY-SA 3.0 IT	A
Basic cartography - DeAgostini	2013	CC-BY-NC-SA 3.0 IT	T
Basic cartography - IGM 100.000	2013	CC-BY-NC-SA 3.0 IT	T
Basic cartography - IGM 25.000	2013	CC-BY-NC-SA 3.0 IT	T
Basic cartography - IGM 250.000	2013	CC-BY-NC-SA 3.0 IT	T
Landslide catalogue	2013	CC-BY-SA 3.0 IT	A
Seismic classification of the Italian municipalities up to 2012	2013	CC-BY-SA 3.0 IT	G
Images datafile SAR ENVISAT ascending	2013	CC-BY-SA 3.0 IT	O
Images datafile SAR ENVISAT descending	2013	CC-BY-SA 3.0 IT	O
Images datafile SAR ERS ascending	2013	CC-BY-SA 3.0 IT	O
Images datafile SAR ERS descending	2013	CC-BY-SA 3.0 IT	O
Directive 2007/60/CE - Units of Management (UoM)	2013	CC-BY-SA 3.0 IT	G
LiDAR product - Lombardy region	2013	CC-BY-SA 3.0 IT	O
Buildings in provincial administrative centres	2013	CC-BY-SA 3.0 IT	G
Railway infrastructures	2013	CC-BY-SA 3.0 IT	T
Road infrastructures	2013	CC-BY-SA 3.0 IT	T
Italian landuse inventory (IUTI)	2013	CC-BY-SA 3.0 IT	A
Digital Terrain Model - 20 meters	\	CC-BY-SA 3.0 IT	T
Digital Terrain Model - 40 meters	\	CC-BY-SA 3.0 IT	T
Digital Terrain Model - 75 meters	\	CC-BY-SA 3.0 IT	T
House numbers in provincial administrative centres	2013	CC-BY-SA 3.0 IT	T
House numbers - 2012 updated	2013	CC-BY-SA 3.0 IT	T
Coloured orthophoto year 2000 with related flight date	2013	CC-BY-NC-SA 3.0 IT	O
Colour 2006 orthophoto with related flight date	2013	CC-BY-NC-SA 3.0 IT	O
Colour 2012 orthophoto with related flight date	2014	CC-BY-SA 3.0 IT	O
Coloured orthophoto of province capitals with relative date of the flight	2014	CC-BY-SA 3.0 IT	O

Colour orthophoto of Metropolitan cities with related flight date			
Black and white orthophoto years 1988-1989 with related flight date	2013	CC-BY-NC-SA 3.0 IT	O
Black and white orthophoto years 1994-1998 with related flight date	2013	CC-BY-NC-SA 3.0 IT	O
PAI - Hydrogeological danger	2013	CC-BY-SA 3.0 IT	A
PAI - Hydrogeological risk	2013	CC-BY-SA 3.0 IT	A
References seismic dangerousness, step of 0.02 grades	2013	CC-BY-SA 3.0 IT	A
References seismic dangerousness, step of 0.05 grades	2013	CC-BY-SA 3.0 IT	A
Persistent Scattered Interferometry Project	\	CC-BY-SA 3.0 IT	O
COSMO SKY-MED ascending union framework	2013	CC-BY-SA 3.0 IT	O
COSMO SKY-MED descending union framework	2013	CC-BY-SA 3.0 IT	O
ENVISAT ascending union framework	2013	CC-BY-SA 3.0 IT	O
ENVISAT descending union framework	2013	CC-BY-SA 3.0 IT	O
ERS ascending union framework	2013	CC-BY-SA 3.0 IT	O
ERS descending union framework	2013	CC-BY-SA 3.0 IT	O
LiDAR tables union framework	2013	CC-BY-SA 3.0 IT	O
Hydrographic network	2013	CC-BY-SA 3.0 IT	T
Kindergarten and school on the national territory	2013	CC-BY-SA 3.0 IT	G
Census sections - ISTAT 1991	2013	CC-BY-NC-SA 3.0 IT	G
Census sections - ISTAT 2001	2013	CC-BY-SA 3.0 IT	G
Protected sites - VI Official list of the protected area (EUAP)	2013	CC-BY-SA 3.0 IT	G
Protected sites - Natural network 2000 - Communitarian importance site (SIC)	2014	CC-BY-SA 3.0 IT	A
Ecological protection zone (ZPE)	2014	CC-BY-SA 3.0 IT	A
Protected sites - Natural network 2000 - Special protection zone (ZPS)	2014	CC-BY-SA 3.0 IT	A
Lakes and other internal watersheds	2013	CC-BY-SA 3.0 IT	A
Italian toponyms IGM	2013	CC-BY-SA 3.0 IT	T

Regional, provincial and municipal administrative units	2013	CC-BY-SA 3.0 IT	T
Landuse - Corine Land Cover 1990	2013	CC-BY-SA 3.0 IT	A
Landuse - Corine Land Cover 2000	2013	CC-BY-SA 3.0 IT	A
Landuse - Corine Land Cover 2006	2013	CC-BY-SA 3.0 IT	A
Landuse - Corine Land Cover 2000 IV Level	2013	CC-BY-SA 3.0 IT	A
Landuse - Corine Land Cover 2006 IV Level	2013	CC-BY-SA 3.0 IT	A
Landuse - Corine Land Cover 2012 IV Level	2013	CC-BY-SA 3.0 IT	A
Protected sites - International importance wetland (Ramsar)	2011	CC-BY-SA 3.0 IT	A
Earthquake - prone areas ZS9	2013	CC-BY-SA 3.0 IT	A

Table 2.8 – WFS layers of National Cartographic Portal

	Update	**License**	**Content**
Derived acceleration ERS ENVISAT Ascending	2013	CC-BY-SA 3.0 IT	O
Derived acceleration ERS ENVISAT Descending	2013	CC-BY-SA 3.0 IT	O
National atlas of the desertification risk areas	2013	CC-BY-SA 3.0 IT	A
IBA - Important Birds Areas	2013	CC-BY-SA 3.0 IT	A
Principal and secondary hydrological basins	2013	CC-BY-SA 3.0 IT	A
Italian ecopedological map	2013	CC-BY-SA 3.0 IT	A
Italian phytoclimatical map	2013	CC-BY-NC-SA 3.0 IT	A
Italian geolithological map	2013	CC-BY-SA 3.0 IT	A
Italian geological map	2013	CC-BY-NC-SA 3.0 IT	A
Forestry antifire cartography (AIB)	2013	CC-BY-SA 3.0 IT	A
Landslide catalogue	2013	CC-BY-SA 3.0 IT	A
Seismic classification of the Italian municipalities up to 2010	2013	CC-BY-SA 3.0 IT	G
Seismic classification of the Italian municipalities up to 2012	2013	CC-BY-SA 3.0 IT	G
Datafile SAR COSMO SKY-MED Ascending – PST 2013	2013	CC-BY-SA 3.0 IT	O
Datafile SAR COSMO SKY-MED Ascending – PST 2013	2013	CC-BY-SA 3.0 IT	O

Datafile SAR ENVISAT Ascending	2013	CC-BY-SA 3.0 IT	O
Datafile SAR ENVISAT Descending	2013	CC-BY-SA 3.0 IT	O
Datafile SAR ERS Ascending	2013	CC-BY-SA 3.0 IT	O
Datafile SAR ERS Descending	2013	CC-BY-SA 3.0 IT	O
Directive 2007/60/CE - Units of Management (UoM)	2013	CC-BY-SA 3.0 IT	G
Buildings in provincial administrative centres	2013	CC-BY-SA 3.0 IT	T
Railway infrastructures	2013	CC-BY-SA 3.0 IT	T
Road infrastructures	2013	CC-BY-SA 3.0 IT	T
Italian landuse inventory (IUTI)	2013	CC-BY-SA 3.0 IT	A
House numbers in provincial administrative centres	2013	CC-BY-SA 3.0 IT	T
House numbers - 2012 updated	2013	CC-BY-SA 3.0 IT	T
PAI - Hydrogeological danger	2013	CC-BY-SA 3.0 IT	A
PAI - Hydrogeological risk	2013	CC-BY-SA 3.0 IT	A
References seismic dangerousness, step of 0.02 grades	2013	CC-BY-SA 3.0 IT	A
References seismic dangerousness, step of 0.05 grades	2013	CC-BY-SA 3.0 IT	A
COSMO SKY-MED ascending union framework	2013	CC-BY-SA 3.0 IT	O
COSMO SKY-MED descending union framework	2013	CC-BY-SA 3.0 IT	O
ENVISAT ascending union framework	2013	CC-BY-SA 3.0 IT	O
ENVISAT descending union framework	2013	CC-BY-SA 3.0 IT	O
ERS ascending union framework	2013	CC-BY-SA 3.0 IT	O
ERS descending union framework	2013	CC-BY-SA 3.0 IT	O
LiDAR tables union framework	2013	CC-BY-SA 3.0 IT	O
Hydrographic network	2013	CC-BY-SA 3.0 IT	T
Kindergarten and school on the national territory	2013	CC-BY-SA 3.0 IT	G
Protected sites - VI Official list of the protected area (EUAP)	2013	CC-BY-SA 3.0 IT	G
Protected sites - Natural network 2000 - Communitarian importance site (SIC)	2013	CC-BY-SA 3.0 IT	A
Ecological protection zones (ZPE)	2013	CC-BY-SA 3.0 IT	A

Protected sites - Natural network 2000 - Special protection zone (ZPS)	2013	CC-BY-SA 3.0 IT	A
Protected sites - International importance wetland (Ramsar)	2013	CC-BY-SA 3.0 IT	A
Internal watersheds	2013	CC-BY-SA 3.0 IT	A
Italian toponyms IGM	2013	CC-BY-SA 3.0 IT	T
Regional, provincial and municipal administrative units	2013	CC-BY-SA 3.0 IT	G
Landuse - Corine Land Cover 1990	2013	CC-BY-SA 3.0 IT	A
Landuse - Corine Land Cover 2000	2013	CC-BY-SA 3.0 IT	A
Landuse - Corine Land Cover 2006	2013	CC-BY-SA 3.0 IT	A
Landuse - Corine Land Cover 2000 IV Level	2013	CC-BY-SA 3.0 IT	A
Landuse - Corine Land Cover 2006 IV Level	2013	CC-BY-SA 3.0 IT	A
Landuse - Corine Land Cover 2012 IV Level	2013	CC-BY-SA 3.0 IT	A
Earthquake-prone areas ZS9	2013	CC-BY-SA 3.0 IT	A

Table 2.9 – WCS layers of National Cartographic Portal

	Update	**License**	**Content**
Desertification atlas - Potentially salted aquifers	2013	CC-BY-SA 3.0 IT	A
Desertification atlas - Protected areas	2013	CC-BY-SA 3.0 IT	A
Desertification atlas - Urban areas and principal infrastructures	2013	CC-BY-SA 3.0 IT	A
Desertification atlas - Normalized Difference Vegetation Index - NDVI	2013	CC-BY-SA 3.0 IT	A
Desertification atlas - Grazing intensity	2013	CC-BY-SA 3.0 IT	A
Desertification atlas - Arable land agri-environmental measures	2013	CC-BY-SA 3.0 IT	A
Desertification atlas - Yearly averages number of dry land days	2013	CC-BY-SA 3.0 IT	A
Desertification atlas - Presence of erosion phenomenon	2013	CC-BY-SA 3.0 IT	A
Desertification atlas - Potential droughts	2013	CC-BY-SA 3.0 IT	A
Desertification atlas - Thin soils on steep gradients	2013	CC-BY-SA 3.0 IT	A
Digital Terrain Model - 20 meters	\	CC-BY-SA 3.0 IT	T
Digital Terrain Model - 40 meters	\	CC-BY-SA 3.0 IT	T

Digital Terrain Model - 75 meters	\	CC-BY-SA 3.0 IT	T

2.4.3 Regional Agency for Protection of the Environment of Lombardy (ARPA)

Total amounts of 103 datasets have been found. For each dataset, the location, the kind of sensor, the time interval, the station, the data type and the content is explained in Table 2.10. This table provides also information about the water level of some rivers: Lambro, Lura and Seveso. For these datasets, the license is a customized open license drafted by ARPA. All the dataset contents are catalogued as Sensor observation, "S".

Table 2.10 – Datasets of Regional Agency for Protection of the Environment of Lombardy

	Station	Sensor	Time interval	Data types	Format	Content
Milano	Brera street	Temperature	19/03/1990 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity	25/03/2014-	raw/hourly/daily	CSV, PDF	S
		Wind velocity gust	16/06/2014-	raw/hourly/daily	CSV, PDF	S
		Wind direction	25/03/2014-	raw/hourly/daily	CSV, PDF	S
		Wind direction gust	19/06/2014-	raw/hourly/daily	CSV, PDF	S
		Relative humidity	5/12/2002-	raw/hourly/daily	CSV, PDF	S
		Precipitation	11/07/2016 -	raw/hourly/daily	CSV, PDF	S
		Global radiation	08/07/2016 -	raw/hourly/daily	CSV, PDF	S
	Lambrate (Lambro park - aqueduct)	Temperature	01/01/2004 -	raw/hourly/daily	CSV, PDF	S
		Precipitation	01/01/2004 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity	15/11/2012-	raw/hourly/daily	CSV, PDF	S
		Wind velocity gust	15/11/2012-	raw/hourly/daily	CSV, PDF	S
		Wind direction	15/11/2012-	raw/hourly/daily	CSV, PDF	S
		Wind direction gust	15/11/2012-	raw/hourly/daily	CSV, PDF	S
		Relative humidity	29/06/2001-	raw/hourly/daily	CSV, PDF	S
		Global radiation	12/09/2003-	raw/hourly/daily	CSV, PDF	S

	Zavattari square	Temperature	01/01/1989 -	raw/hourly/daily	CSV, PDF	S
		Precipitation	22/11/2004 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity gust	19/06/2014-	raw/hourly/daily	CSV, PDF	S
		Wind direction	26/09/2013-	raw/hourly/daily	CSV, PDF	S
		Relative humidity	22/04/2002-	raw/hourly/daily	CSV, PDF	S
		Wind velocity	29/09/2013-	raw/hourly/daily	CSV, PDF	S
	Confalonieri street (ARPA Lombardy office)	Precipitation	28/06/2006 - 02/04/2012	raw/hourly/daily	CSV, PDF	S
	Feltre street	Lambro Level	24/07/1998 -	raw/hourly/daily	CSV, PDF	S
		Temperature	05/07/2000 -	raw/hourly/daily	CSV, PDF	S
	Juvara street	Temperature	01/01/1989 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity	28/02/2014-	raw/hourly/daily	CSV, PDF	S
		Wind velocity gust	19/06/2014-	raw/hourly/daily	CSV, PDF	S
		Wind direction	28/02/2014-	raw/hourly/daily	CSV, PDF	S
		Wind direction gust	19/06/2014-	raw/hourly/daily	CSV, PDF	S
		Relative humidity	01/01/1999-	raw/hourly/daily	CSV, PDF	S
		Global radiation	01/01/1999-	raw/hourly/daily	CSV, PDF	S
		Precipitation	01/01/1989 -	raw/hourly/daily	CSV, PDF	S
	Niguarda	Seveso river hydrometric level	18/10/2013 -	raw/hourly/daily	CSV, PDF	S
	Rosellini street	Precipitation	12/04/2012 -	raw/hourly/daily	CSV, PDF	S
	Marche boulevard	Temperature	01/01/1989 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity	29/10/2013-	raw/hourly/daily	CSV, PDF	S
		Wind velocity gust	19/06/2014-	raw/hourly/daily	CSV, PDF	S
		Wind direction	29/10/2013-	raw/hourly/daily	CSV, PDF	S
		Relative humidity	21/03/2002-	raw/hourly/daily	CSV, PDF	S
Arconate	De Gasperi street, Dell'Usignolo street	Precipitation	02/07/2002 -	raw/hourly/daily	CSV, PDF	S
		Temperature	02/07/2002 -	raw/hourly/daily	CSV, PDF	S

	corner	Wind velocity	02/07/2002 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity gust	22/01/2009 -	raw/hourly/daily	CSV, PDF	S
		Wind direction	05/04/2012 -	raw/hourly/daily	CSV, PDF	S
		Wind direction gust	05-04-2012 -	raw/hourly/daily	CSV, PDF	S
		Relative humidity	02/07/2002 -	raw/hourly/daily	CSV, PDF	S
		Global radiation	02/07/2002 -	raw/hourly/daily	CSV, PDF	S
Cinisello Balsamo	Parco Nord	Precipitation	17/10/2008 -	raw/hourly/daily	CSV, PDF	S
		Temperature	30/05/2002 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity	30/05/2002 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity gust	03/02/2009 -	raw/hourly/daily	CSV, PDF	S
		Wind direction	13/03/2012 -	raw/hourly/daily	CSV, PDF	S
		Wind direction gust	13/03/2012 -	raw/hourly/daily	CSV, PDF	S
		Relative humidity	30/05/2002 -	raw/hourly/daily	CSV, PDF	S
		Global radiation	30/05/2002 -	raw/hourly/daily	CSV, PDF	S
Corsico	Italia boulevard	Precipitation	29/03/2002 -	raw/hourly/daily	CSV, PDF	S
		Temperature	12/02/1997 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity	30/10/2013 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity gust	19/06/2014 -	raw/hourly/daily	CSV, PDF	S
		Wind direction	30/10/2013 -	raw/hourly/daily	CSV, PDF	S
		Relative humidity	01/01/1999 -	raw/hourly/daily	CSV, PDF	S
Lacchiarella	Molise street	Temperature	15/09/1998 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity	30/10/2013 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity gust	19/06/2014 -	raw/hourly/daily	CSV, PDF	S
		Wind direction	30/10/2013 -	raw/hourly/daily	CSV, PDF	S
		Relative humidity	01/01/1999 -	raw/hourly/daily	CSV, PDF	S
Lainate	XXV Aprile street	Level Lura a Lainate	16/04/2014 -	raw/hourly/daily	CSV, PDF	S
Locate Triulzi	Staffora street	Level Lambro meridionale a Locate Triulzi	15/10/2012 -	raw/hourly/daily	CSV, PDF	S

Motta Visconti	A. de Gasperi street	Precipitation	06/11/2008 -	raw/hourly/daily	CSV, PDF	S
		Temperature	15/05/2003 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity	15/05/2003 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity gust	28/10/2008 -	raw/hourly/daily	CSV, PDF	S
		Wind direction	21/05/2012 -	raw/hourly/daily	CSV, PDF	S
		Wind direction gust	21/05/2012 -	raw/hourly/daily	CSV, PDF	S
		Relative humidity	15/05/2003 -	raw/hourly/daily	CSV, PDF	S
		Global radiation	15/05/2003 -	raw/hourly/daily	CSV, PDF	S
Paderno Dugnano	Palazzolo Borghetto park	Precipitation	01/01/2000 -	raw/hourly/daily	CSV, PDF	S
		Precipitation	01/01/2000 -	raw/hourly/daily	CSV, PDF	S
	Lampugnani square - Palazzolo	Level Seveso a Palazzolo	24/07/1998 -	raw/hourly/daily	CSV, PDF	S
		Precipitazione a Palazzolo	07/02/2006 – 28/11/2014	raw/hourly/daily	CSV, PDF	S
Pogliano Milanese	Cesare Battisti street	Precipitation	16/04/2014 -	raw/hourly/daily	CSV, PDF	S
Rescaldina	Marco Polo Ovest boulevard	Precipitation	16/04/2014 -	raw/hourly/daily	CSV, PDF	S
Rho	Fiorenza station - Grass	Precipitation	24/02/2015 -	raw/hourly/daily	CSV, PDF	S
		Temperature	24/02/2015 -	raw/hourly/daily	CSV, PDF	S
		Relative humidity	24/02/2015 -	raw/hourly/daily	CSV, PDF	S
	Fiorenza station - Roof	Wind velocity	24/02/2015 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity gust	24/02/2015 -	raw/hourly/daily	CSV, PDF	S
		Wind direction	24/02/2015 -	raw/hourly/daily	CSV, PDF	S
		Wind direction gust	24/02/2015 -	raw/hourly/daily	CSV, PDF	S
Rodano	F. Turati street	Precipitation	01/01/1999 -	raw/hourly/daily	CSV, PDF	S
		Temperature	01/01/1999 -	raw/hourly/daily	CSV, PDF	S
		Relative humidity	01/01/1999 -	raw/hourly/daily	CSV, PDF	S
S. Colombano al Lambro	Serafina street	Precipitation	01/01/2004 -	raw/hourly/daily	CSV, PDF	S
		Temperature	01/01/2004 -	raw/hourly/daily	CSV, PDF	S

		Wind velocity	27/06/2001 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity gust	19/06/2012 -	raw/hourly/daily	CSV, PDF	S
		Wind direction	19/06/2012 -	raw/hourly/daily	CSV, PDF	S
		Wind direction gust	19/06/2012 -	raw/hourly/daily	CSV, PDF	S
		Relative humidity	27/06/2001 -	raw/hourly/daily	CSV, PDF	S
		Global radiation	12/09/2003 -	raw/hourly/daily	CSV, PDF	S
Segrate - Milano 2	Milano 2	Precipitation	23/09/1994 - 16/06/2004	raw/hourly/daily	CSV, PDF	S
		Temperature	21/06/1995 - 16/06/2004	raw/hourly/daily	CSV, PDF	S
		Relative humidity	21/06/1995 - 16/06/2004	raw/hourly/daily	CSV, PDF	S
Trezzo d'Adda	Nenni street	Precipitation	21/02/2001 -	raw/hourly/daily	CSV, PDF	S
		Wind direction	22/02/2001 -	raw/hourly/daily	CSV, PDF	S
		Wind velocity	22/02/2001 -	raw/hourly/daily	CSV, PDF	S
		Relative humidity	22/02/2001 -	raw/hourly/daily	CSV, PDF	S

2.4.4 Lombardy Open Data Portal

The Lombardy Open Data Portal provides 115 vector georeferenced datasets. They are classified in different categories that are catalogued in Table 2.11. The contents of the dataset are governance data “G” and Environmental data “A”.

Table 2.11 – Lombardy Open Data Portal layers

	Number of layers	**Update**	**Format**	**License**	**Format**
Agriculture	9	2016-2017	CSV, JSON, XLS, XLSX, XML	IODL 2.0	A
Environmental	12	2015-2016	CSV, JSON, XLS, XLSX, XML	IODL 2.0	A
Productive activities	4	2015-2016	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G
Commerce	12	2016-2017	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G
Culture	13	2015-2017	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G
Energy	4	2015-2017	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G

Family	31	2015-2017	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G
Government	12	2016-2017	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G
Instruction	11	2015-2016	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G
Civil protection	2	2015	CSV, JSON, XLS, XLSX, XML	IODL 2.0	A
Sanity	5	2016-2017	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G
Security	4	2015-2016	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G
Solidarity	3	2015-2016	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G
Sport	2	2016	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G
Territory	1	2016	CSV, JSON, XLS, XLSX, XML	IODL 2.0	A
Transparency	1	2017	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G
Tributes	1	2016	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G
Tourism	2	2016-2017	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G
Transport and mobility	3	2015-2016	CSV, JSON, XLS, XLSX, XML	IODL 2.0	G

2.4.5 Data Portal of Milan Municipality

The Data Portal of Milan Municipality, catalogued as a non-GIS database, has a geographical information section that contains 33 geodata. The contents are governance "G" and topographic data "T". They are listed in Table 2.12.

Table 2.12 – Geospatial layers of Data Portal of Milan Municipality

	Update	**Format**	**License**	**Format**
Culture: locations of cultural associations	2010	SHP	CC0	G
Territory: locations of dog areas	2012	SHP	CC0	G
Sport: locations of swimming pools	2011	SHP	CC0	G
Sport: locations of leisure centres	2011	SHP	CC0	G
Instruction: locations of kindergartners	2013	SHP	CC0	G
Social: locations of nursery schools	2012	SHP	CC0	G
Instruction: locations of primary	2013	SHP	CC0	G

schools				
Culture: locations of libraries and archives	2013	SHP	CC0	G
Culture: locations of conference centres	2011	SHP	CC0	G
Mobility: locations of interchange parking	2012	SHP	CC0	G
Instruction: locations of secondary school	2013	SHP	CC0	G
Commercial activities: locations of kiosks	2014	SHP	CC0	G
Mobility: locations of Bike Sharing (BikeMI) parking areas	2014	SHP	CC0	G
Public transport: locations of subway stops - ATM	2015	SHP	CC0	T
Public transport: locations of subway lines	2015	SHP	CC0	T
Territory: locations of municipal cemetery structures	2012	SHP	CC0	G
Instruction: locations of high schools	2010	SHP	CC0	G
Mobility: locations of Car Sharing (GuidaMI) parking areas	2012	SHP	CC-BY	G
Public transport: locations of railway stops	2011	SHP	CC0	T
Public transport: locations of railway networks	2011	SHP	CC0	T
Mobility: locations of electronic crosses	2012	SHP	CC0	G
Mobility: locations of parking	2013	SHP	CC0	T
Instruction: locations of Milan's universities	2011	SHP	CC0	G
Sanity: locations of pharmacies	2011	SHP	CC0	G
Territory: locations of cycling lanes	2013	SHP	CC0	T
Territory: locations of city's neighbourhood (Local Identity Nucleus)	2011	SHP	CC0	G
Territory: locations of parks and gardens	2012	SHP	CC-BY	G

Territory: locations of census areas	2011	SHP	CC-BY	G
Environmental: green areas	2011	SHP	CC-BY	T

2.5 Results

The analysis was firstly been performed in April 2016 and the result were published in Brovelli, et al. (2016). This work updates the results for January 2017. The results show the charateristics of the 1'099 datasets and summarize them in tables and graphs. They are catalougued taking into account the classification based on provider, content, format, spatial coverage, license, year of publication and lastly a crossed investigation between content and provider.

2.5.1 Providers

A first analysis can be done considering the provider of the data. The amount of information supplied by each portal is shown in Table 2.13.

Table 2.13 – Cataloguing according to the providers

	Datasets	**Percentage**
Lombardy Region Geoportal	694	63%
Italian National Cartographic Geoportal	137	12%
Regional Agency for Protection of the Environment of Lombardy (ARPA)	108	10%
Lombardy Open Data Portal	132	12%
Municipality of Milan	28	3%

Most of the datasets are provided by Lombardy Region Geoportal (see Figure 2.4): a total amount of 694 (63%). The second wider supplier is the National Cartographic Portal of the Italian Environmental Ministry, which has 137 geodata (12%). Then there are Lombardy Open Data Portal and Regional Agency for Protection of the Environment of Lombardy (ARPA) that release together around 10% of the total, and lastly there is Milan Municipality with a 3%.

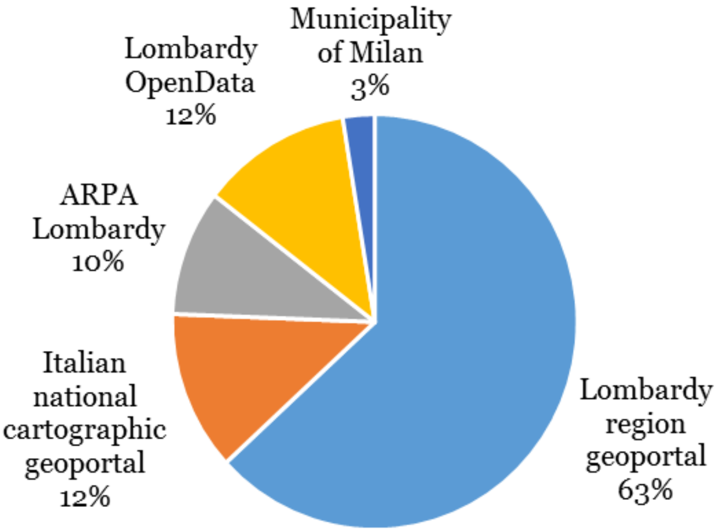


Figure 2.4 – Distribution of available open geodata for the Metropolitan city of Milan according to their provider

2.5.2 Contents

Based on a content classification, the results are shown in Table 2.14.

Table 2.14 – Cataloguing according to the content

	Datasets	Percentage
Sensor Observation	109	10%
Topography	318	29%
Environmental	408	37%
Governance	224	20%
Airborne Observation	40	4%

The greatest number of datasets (37%) falls in the environmental category (Figure 2.5), followed by the topographic one (29%) and the governance one (20%). Airborne and sensor observations correspond to smaller percentages of the available information.

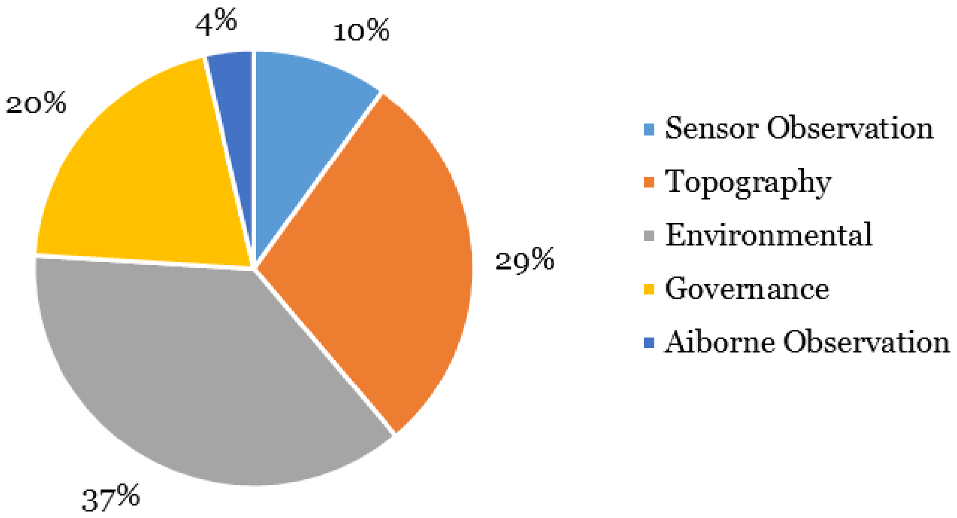


Figure 2.5 – Distribution of available open geodata for the Metropolitan city of Milan according to their content

2.5.3 Formats

Table 2.15 shows the list of the types and percentages of data formats available.

Table 2.15 – Cataloguing according to the format

	Datasets	**Percentage**
Vector	937	85%
Raster	8	1%
Web service	154	14%

Most of them are vector (85%); the 14% are available as GeoWeb services, while less than 1% are raster (Figure 2.6).

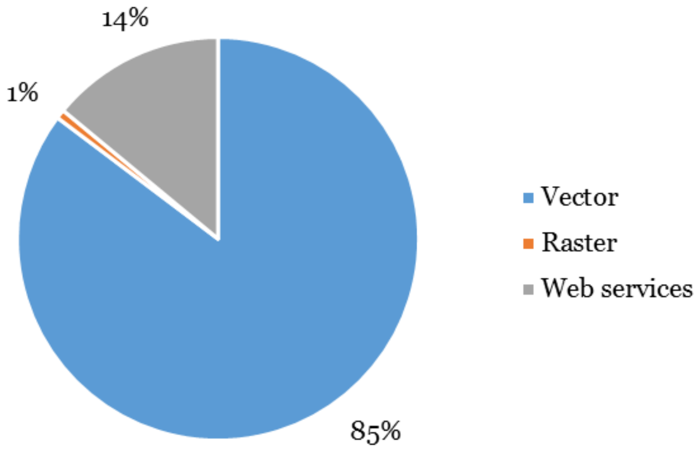


Figure 2.6 – Distribution of available open geodata for the Metropolitan city of Milan according to their format

2.5.4 Spatial coverages

Considering the spatial coverages of the datasets, Table 2.16 summarizes the characteristics of all the products.

Table 2.16 – Cataloguing according to their spatial coverage

	Datasets	**Percentage**
National scale	137	12%
Regional scale	854	78%
Local scale	108	10%

Figure 2.7 shows that the great majority of open geodata have a regional scale (78%). These mainly correspond to the datasets provided by the Lombardy Region Geoportal and Lombardy Open Data portal (which together have 854 products).

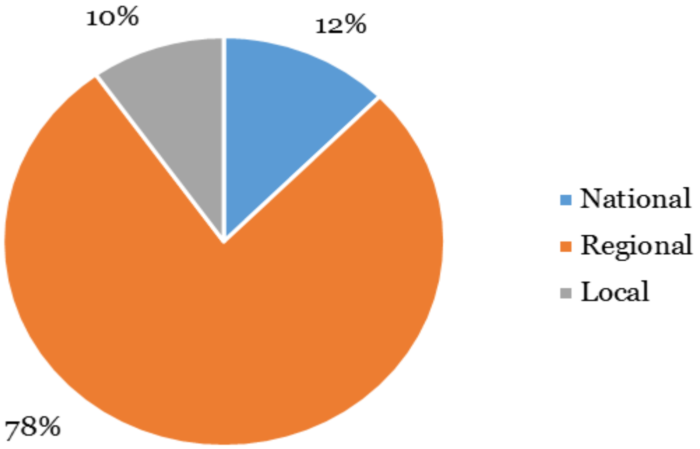


Figure 2.7 – Distribution of available open geodata for the Metropolitan city of Milan according to their spatial coverage

2.5.5 Licenses

Table 2.17 lists all the types of license used.

Table 2.17 – Cataloguing according to the licenses

	Datasets	Percentage
CC0	24	2%
CC-BY	4	0%
CC-BY-SA 3.0 IT	124	11%
CC-BY-NC-SA 3.0 IT	204	19%
IODL 2.0	631	57%
No license	4	0%
Custom open license	108	10%

More than half of the datasets are available under the IODL v2.0 license (see Figure 2.8). Among the remaining datasets, a high percentage is released under a CC-BY-NC-SA 3.0 IT license, followed by CC-BY-SA 3.0 IT. CC0 and CC-BY licenses are quite infrequent, as they represent together 2% of the total. All the products from ARPA (the 10 % of the total) are available under a custom open license specified by the provider, while the rest of the data (less than 1 % of the total) have no license at all.

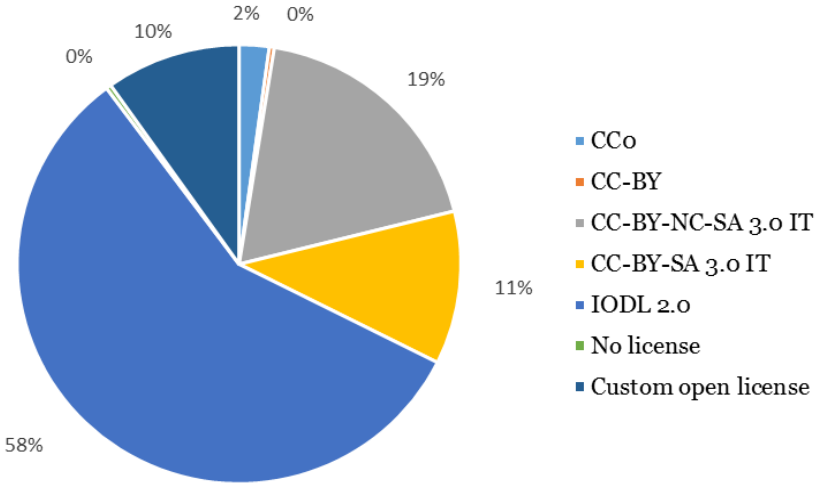


Figure 2.8 – Distribution of available open geodata for the Metropolitan city of Milan according to their license

2.5.6 Years of publication

A final classification of the open geodata available for the Metropolitan city of Milan is made according to the year of publication. As shown in Figure 2.9, the oldest datasets were published in 1975 and 1987. The Lombardy Region Geoportal published these old datasets.

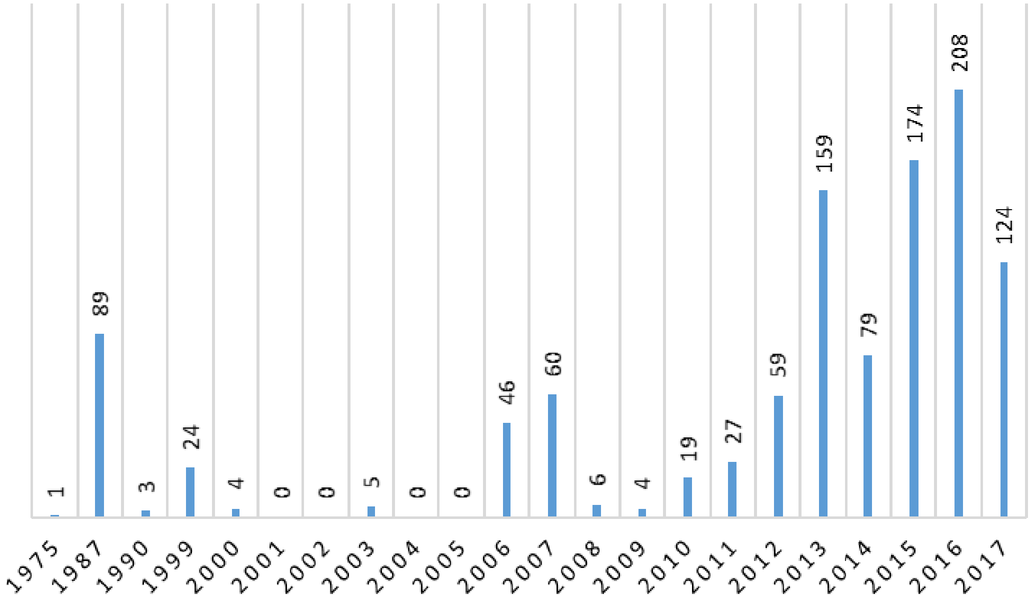


Figure 2.9 – Classification of the open geodata available for the Metropolitan city of Milan according to the years of publication

The number of datasets released in 2017 is expected to further increase, as the time of writing was January 2017. Finally, it is worth mentioning that 8 out of the 1099 datasets (around 2%) have no indication of the year of publication.

2.5.7 Categories vs contents

Finally, Table 2.18 gives a better idea of the categories of datasets published by each provider.

Table 2.18 – Cataloguing according to the categories and contents

	Sensor	**Topography**	**Environmental**	**Governance**	**Airborne Observation**
Lombardy Region Geoportal	1	286	320	81	6
Italian National Cartographic Geoportal	0	25	64	14	34
Regional Agency for Protection of the Environment of Lombardy (ARPA)	108	0	0	0	0
Lombardy Open Data Portal	0	0	24	108	0
Municipality of Milan	0	7	0	21	0

The main source of geodata, i.e. the Lombardy Region Geoportal, is focused on topographic, environmental and governance information (see Figure 2.10).

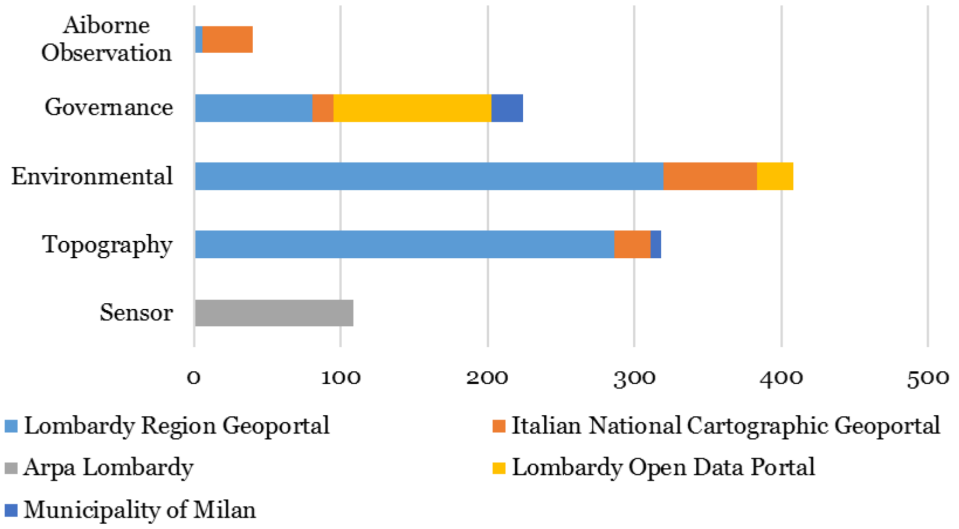


Figure 2.10 – Proportion of open geodata available for the Metropolitan city of Milan according to their category and provider

Governance data receive the most significant contributions from the Lombardy Region Geoportal and the Lombardy Open Data portal. Open geodata from the Italian National Cartographic Geoportal are mainly environmental data and airborne observations, while almost all the available sensor observations are provided by the Lombardy section of ARPA.

2.6 Conclusions

From this analysis, a total amount of 22 authoritative databases containing information about the Metropolitan city of Milan have been detected. 7 databases have been classified as GIS portals and 15 as non-GIS portals. Most of them were Italian databases that supply original information; the remaining were collections of datasets that are already published in other databases, like in the case of Italian Portal of the Open Data and Open Data of the Public Administration. These kinds of databases mainly regard the non-GIS portals, because of the spreading of Open Data datasets that can be found on the web. This can be considered as an application of the Open Government transparency plan, which makes the search and access to open datasets easier. In fact, having a huge collection to browse is simpler than looking up for datasets in many smaller portals. The European Data Portal also follows this strategy. It collects OD datasets from the portals of European countries and makes them available together. This improves their accessibility and their value (European Data Portal, 2016). On the other hand, users must be aware that this approach mixes many disparate datasets coming from different portals and, consequently, proposes a catalogue with heterogeneous characteristics.

Even if VGI and non authoritative databases were not included in the analysis, they are accessible on the web for the study area. They include international projects such as OSM and other Italian databases such as Spaghetti Open Data (http://www.spaghettiopendata.org).

This study has listed an amount of more than 6'000 datasets contained in the considered databases. This investigation has showed that the Italian availability of OD is quite wide and demonstrated that the trend of data publication is positive. Hence, it can be affirmed that the Metropolitan city of Milan reflects the global spreading of OD.

The analysis has performed a deep evaluation of 1'099 geodata that were included in 5 portals. Among all of them, Lombardy Region Geoportal published the majority of the data. The most common contents of these geodata were environmental and topography. The classification based on the format was quite interesting, too: almost all the datasets were vectors. The other categories contained very few data. The spatial coverage was mainly regional, because regional geoportals supplied most of the datasets. The most used license was IODL 2.0. Of particular significance is the fact that the second most used license is CC-BY-SA 3.0 IT, which is the Italian versions of the international definitions. These licenses, in their explanation, refer to the Italian copyright law.

Taking into account the classification based on the year of publication, a significant increase can be outlined in the number of open geodata released over the last five years. This is mainly a consequence of the European openness strategy applied also in Italy. In fact, even if position of Italy in the Open Data Index is not so high (see Subsection 1.1.1), the trend of geodata publication is positive.

The cross evaluation providers vs contents has shown that Lombardy Region Geoportal supplied most of the topography and environmental data; Italian National Cartographic Geoportal produced most of the airborne observation; ARPA supplied sensor data; and Lombardy Open Data Portal contributed to the governance data. The analysis underlined that the production of data is sectorized, i.e. the portals tend to produce mainly a specific category of datasets.

Concluding, the evaluation has shown that more than a quarter of the total amount of the datasets were geodata, even if the number of non-GIS databases were more than twice the GIS ones. This accentuates the importance of the geodata between the other Open Data categories and highlights the attention many authorities actually pay, which is bringing them to an unceasing creation of new data.

CHAPTER 3
ORTHOPHOTO QUALITY ASSESSMENT

In the previous chapter, the availability of open geodata for the Metropolitan city of Milan has been discussed. Between such datasets, it is important to understand the goodness of these information. In order to reach that, this second part of this work aims to evaluate the quality of an exemplifying number of open Geodata already catalogued in Chapter 2, i.e. a web service and a vector datasets. Specifically, they will assess the planimetric accuracy of building's roofs in an orthophoto (Chapter 3) and the resemblance between a land use map and a official cartography vector (Chapter 4).

This chapter will illustrate a quality assessment for the planimetric accuracy of the Italian orthophoto called AGEA 2012, a service (WMS) previously listed as Open Data, by comparing it against the official vector cartography of Milan Municipality, namely the Regional Topographic Database (*Database Topografico Regionale,* DBTR) of Lombardy Region, which will be considered as the ground truth.

As the orthophoto is an image of the territory taken by flight or satellite, before being published it needs to minimize the distortions with geometrically and stereoscopically corrections. This case study will focus on the assessment of the roof of the building in Municipality of Milan represented in AGEA 2012 orthophoto respect to their ground true position described by the corresponding DBTR layers. In fact, with low distortions in the orthophoto, the position of the roofs would theoretically be over the basement of their buildings. To evaluate it, this study will compute the ISO parameter "positional accuracy" in order to compute a quality assessment index for the roof of the buildings represented in AGEA 2012 orthophoto respect to the ones of DBTR, by obtaining the distance between the positions of the two different information.

Generally, information on the quality of a geographic product allows its producer to evaluate how well the metadata of that dataset meet its specifications and assist users in evaluating the product's ability to satisfy their requirements. In this case study, an analysis will be made to check whether AGEA 2012 metadata meet the specifications of the Italian law, considering the planimetric accuracy of the building roofs in Milan through a methodology suggested by ISO.

This Chapter will be divided in four parts. The first part will focus on the definition of the case study. AGEA 2012 and a specific building layer of the DBTR used in the comparison will be fully described together with their metadata. Then an overview on the ISO methodology exploited in the analysis will be provided and finally a brief description of the software used in the evaluation will be illustrated.

The second part will be the methodology. It will describe the procedure suggested by the ISO directive and it will conclude with a detailed description of the technique to determine the digitalizing error.

Then, the third section will focus on the orthophoto quality assessment. , This part will describe how the methodology illustrated in the second section has been practically implemented. This part will systematically review all the passages allowing obtaining some results useful to evaluate the quality of the orthophoto.

A conclusion section will finally recap and comment the whole procedure. A summarized version of this study can be found in Brovelli et al. (2016).

3.1 Case study

This part of the work will introduce the datasets involved in the quality assessment and their specific characteristics. This will be followed by an overview on the ISO quality evaluation method and the software used.

3.1.1 Agency for the Agricultural Supplies (AGEA) and AGEA 2012 orthophoto

The Agency for the Agricultural Supplies (*Agenzia per le Erogazioni in Agricoltura,* AGEA) is a governmental agency established in 1999 by the Italian Legislative Decree (LD) 165/1999 (II). It replaced the agency called National Company for the interventions in the agricultural market (*Azienda di Stato per gli interventi nel mercato agricolo,* AIMA) and integrated new tasks defined by the 18th article of the EU Commission Regulation n.

[II] http://www.parlamento.it/parlam/leggi/deleghe/99165dl.htm (accessed March 22, 2017)

885/2006 (III). The European Union, in fact, supports the agricultural production of its countries by distributing helps, contributions and awards financed by the European Agricultural Guarantee Fund (EAGF) and the European Agricultural Fund for Rural Development (EAFRD). AGEA is the Italian application of this international regulation.

One of the most important tasks of AGEA, according to the Italian LD 165/1999, is to act as the Italian Coordination Organism and Paying Organism. As Coordination Organism, AGEA tasks are (AGEA 2013):

- To watch and coordinate the Paying Organisms;
- To verify the coherence of their activities according to European guidelines;
- To promote the application of the European normative and its relative procedures.

On the other hand, as Paying Organism, AGEA functions are (AGEA 2013):

- To authorize the payments;
- To execute the payments;
- To count the payments.

Beside these functions, AGEA produces orthophotos of the Italian territory in the management and control activity of the National Agriculture Informative System (*Sistema Informativo Agricolo Nazionale,* SIAN). Following an agreement taken in 2014, the orthphotos produced by AGEA were published on the Italian National Cartographic Geoportal, managed by the Italian Ministry of Environment, Land and Sea, took an agreement with AGEA. These orthophotos have a resolution of 50 cm/pixel and they provide an indication of the flight date and related metadata for the temporal interval 2009-2012. The orthophotos are available in the Geoportal as WMS layers (Italian National Cartographic Geoportal 2014).

AGEA 2012 orthophoto is a national colour orthophoto, acquired by photogrammetric flight by AGEA in 2012. It is available in the Italian National Cartographic Geoportal along with the information on the date when the different aerial photos that compose it were taken (Italian National Cartographic Geoportal 2014).

It is available as service (WMS) in the National Cartographic Portal with a scale of 1:10'000 (Italian National Cartographic Geoportal 2014). The orthophoto WMS layer in the Italian National Cartographic Portal is named "Colour 2012 orthophoto with related flight date" and it is available under a CC-BY-SA 3.0 IT license (CC-BY-SA 3.0 IT Creative Commons 2016).

[III] http://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:32006R0885 (accessed March 22, 2017)

3.1.1.1 Requirements

Being an Italian orthophoto, AGEA 2012 orthophoto has to respect some requirements, precisely the "technical specifications for the formation, documentation and sharing of digital orthophotos with the nominal scale 1:10'000" (*Regole tecniche per la formazione, la documentazione e lo scambio di ortofoto digitali alla scala nominale 1:10000*). This document was established by the National Committee for the technical rules on the geographic datasets of public administrations (*Comitato nazionale per le regole tecniche sui dati territoriali delle pubbliche amministrazioni*) in the Italian Decree of November 10, 2011 (IV). The document also set the adoption of the national geodetic system, the definition of the specific contents of the geo-topographical databases and the delineation of the compositions and the contents of the National Inventory of the Geographic Data (*Repertorio Nazionale dei Dati Territoriali,* RNDT). The Italian Decree of November 10, 2011 gives different guidelines based on the final application of the orthophotos: the cartographic one and the thematic one.

AGEA orthophotos fall into the thematic one. The technical specifications divide the instructions for thematic orthophotos into:

- Product characteristics;
- Image acquisition;
- Pre-processing;
- Processing;
- Trial.

Of particular interest in this evaluation is the definition of the aerial images orientation through Aerial Triangulation (AT), the first step of the processing phase. The specifications state that the adjustment on a block are accepted when the residual errors are compliant with the following:

- Residual errors on Ground Control Points (GCP):
    - 2.0 meters in planimetry;
    - 1.8 meters in altimetry.
- Residual errors on Control Points (CP):
    - 4.0 meters in planimetry;
    - 3.6 meters in altimetry

[IV] http://www.gazzettaufficiale.it/eli/id/2012/02/27/12A01799/sg (accessed March 22, 2017)

The technical specifications also define a strategy to evaluate the quality of the orthophoto in the trial section. They suggest taking the 5% of the total geometric points and verifying that they respect the above values.

3.1.1.2 Metadata

AGEA set some stricter requirements compared to the Italian legislation in order to increase the final quality of the orthophoto (AGEA 2012). In fact, AGEA decided to accept the block adjustment on a block when the residual errors comply with the following:

- Residual errors on Ground Control Points (GCP):
  - 1.5 meters in planimetry;
  - 1.8 meters in altimetry.
- Residual errors on Control Points (CP):
  - 3.0 meters in planimetry;
  - 3.6 meters in altimetry

This information can be considered as metadata, as it represents a measure of the evaluation in an a posteriori statement (see Subsection 1.4.2).

3.1.2 Topographic Database of Lombardy Region

The Topographic Database of Lombardy Region (*DataBase Topografico Regionale,* DBTR) is a database defined by the Italian Decree of November 10, 2011 containing regional geographic information organized in Layers, Themes and Classes and developed by an experienced working group formed by Italian Regions, IGM, Civil Protection, AGEA, Ministry of the Environmental, Land and Sea, etc. It was establish following the article 59 of the LD 82/2005 (V), which applies the INSPIRE European Directive for the geographic information.

The Topographic Database of Lombardy Region contains 11 different Layers:

- Layer 00: Geodetic and photogrammetric information;
- Layer 01: Viability, mobility and transport;
- Layer 02: Properties and anthropization;
- Layer 03: Viability and address management;
- Layer 04: Hydrography;
- Layer 05: Orography;
- Layer 06: Vegetation;

v http://www.camera.it/parlam/leggi/deleghe/05082dl.htm (accessed March 22, 2017)

- Layer 07: Underservice networks;
- Layer 08: Significant places and cartographic tags;
- Layer 09: Administrative areas;
- Layer 10: Relevance areas.

For this case study, the DBTR is used as the ground truth to evaluate the goodness of AGEA 2012 orthophoto. In particular, the evaluation proposed in this chapter will use Layer 02 **Proprieties and anthropization**, as it contains information on buildings. This Layer includes the whole information about objects derived by the anthropic activities that are not related to transport. It includes five different Themes:

- Theme 0201: Buildings;
- Theme 0202: Artefacts;
- Theme 0203: Transport works;
- Theme 0204: Soil defence works;
- Theme 0205: Hydraulic defence and water regulation works.

In this case study, Theme 0201 **Buildings** will be used to understand which among its Classes are the most suitable to be as ground truth for comparison to the AGEA 2012 orthophoto.

3.1.2.1 Theme 0201: Buildings

It corresponds to the definition of buildings, intended as stable constructions covered by a roof which are mainly used for living, working or leisure purposes. The buildings are classified according to their volumetric and architectural characteristics (Italian Decree of November 10, 2011).

The Classes of this Theme are:

- Class 020101: Volumetric units;
- Class 020102: Building (maximum extension of the building bodies);
- Class 020103: Building caissons (soil encumbrance of the building bodies);
- Class 020104: Coverage elements (e.g. canopy, gallery platform roof);
- Class 020105: Architectural particular (e.g. staircase, external stair, skylight);
- Class 020106: Minor building (e.g. shack, garage, leisure building).

All the datasets are available as polygon vector layers with the exception of the dataset Coverage elements represented by a polyline vector layer. Figure 3.1 shows all the datasets belonging to Theme 0201 **Buildings** in a map representing Milan city centre. It is easy to see that Class Buildings covers almost everything. The only exceptions are some Coverage elements (lines) and some Architectural particular (orange polygons).

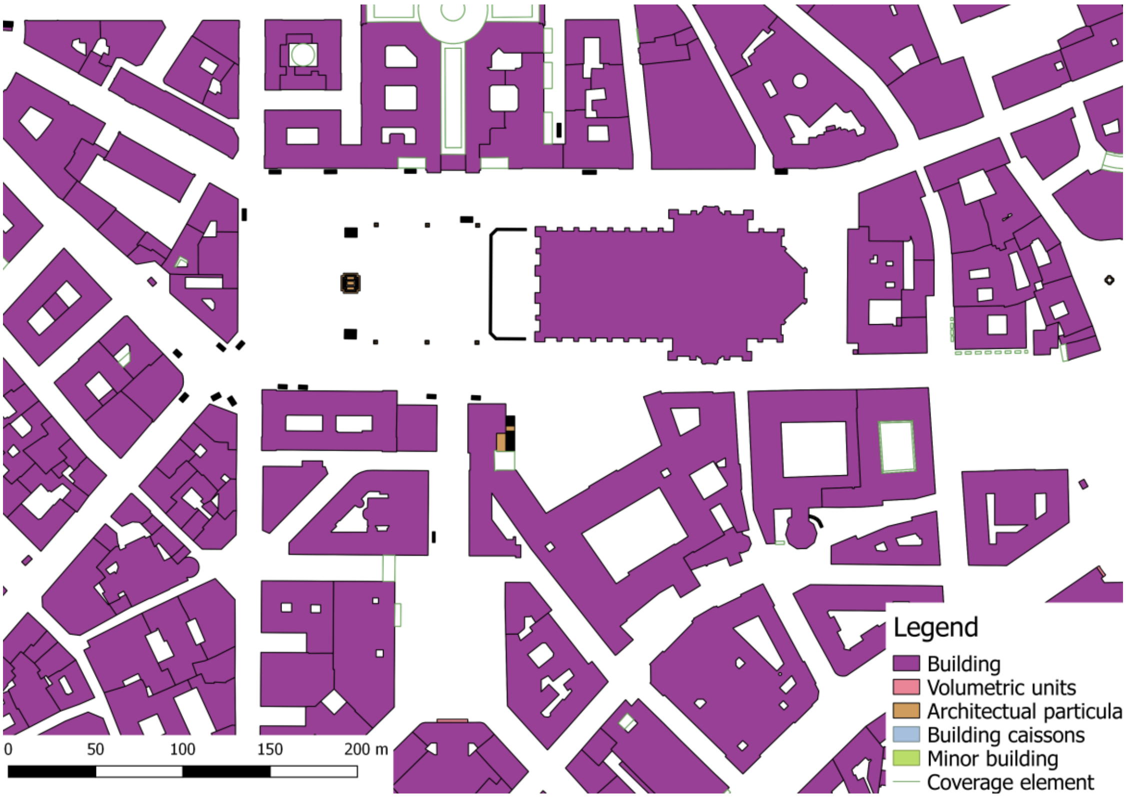


Figure 3.1 – DBTR, Theme 0201 building layers

Therefore, the dataset corresponding to **Building** Class fully includes the datasets of the Classes Volumetric units, Building caissons and Minor building. It also includes the eaves. Of particular interest is the comparison between the polygon layers corresponding to the Building and Volumetric units Classes represented in figure 3.2. It is possible to see that the figure on the left (Volumetric units) describe the building structure in a much more specific way than the one on the right (Buildings).

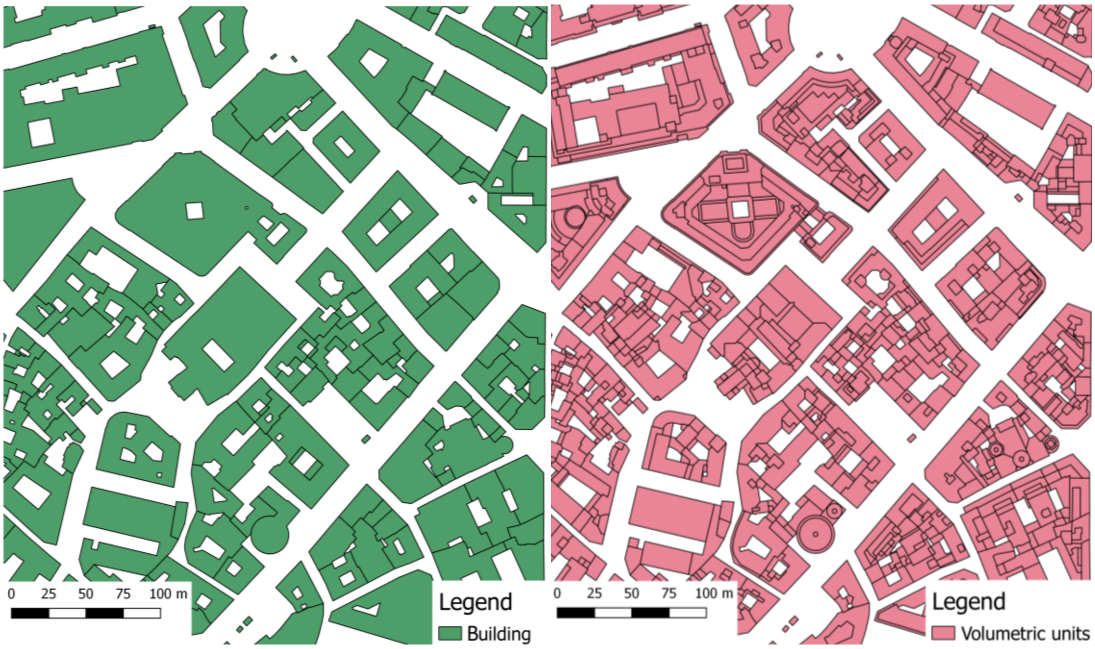


Figure 3.2 – Polygon layers of Building vs Volumetric units Classes

Both represent the ground projection of the maximum extension of buildings, however the Volumetric units Class includes further information. In fact, the polygons that compose this layer have vertices at the same height; hence the shapes have the same altitude. On the other hand, the Building layer is composed by polygons that can have vertexes with different height according to the shape of the building, as they simply indicate the maximum extension of each construction. The further information that the Volumetric units Class includes is not relevant in this case study, indeed it can introduce some misunderstanding problems. In fact, as this work involve the identification of correspondent homologous roof corners of buildings between DBTR and AGEA 2012 orthophoto, it is more suitable to select perimeter points respect to chose internal ones. This concept is easy to detect in figure 3.3. The image shows AGEA 2012 orthophoto overlapped with the two polygon layers (Building and Volumetric units). It is quite difficult to understand the correspondence of the Volumetric units layer corners (yellow line) and these points on the orthophoto, while the correspondence is more recognisable whereas the Building layer is considered (green line).

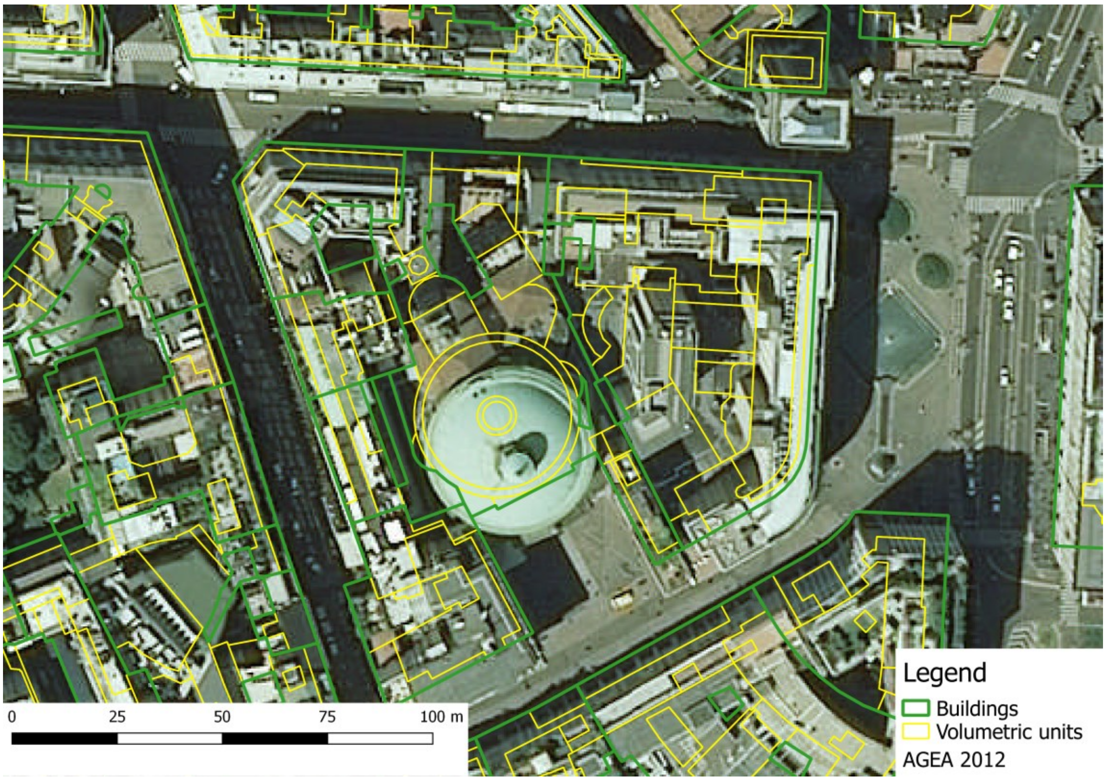


Figure 3.3 – AGEA 2012 orthophoto, Building and Volumteric units

For this reason, the DBTR dataset used as ground truth for comparison with the AGEA 2012 orthophoto is the one corresponding to Layer 02: Proprieties and anthropizations, Theme 0201: Buildings, Class 020102: Building. From now on, this dataset will be simply referred to as the building layer.

3.1.2.2 Metadata

All the datasets of the DBTR are available in the Lombardy Region Geoportal under the IODL 2.0 license. The different Layers of DBTR have different metadata based on the municipality. In the case of Milan the scale is 1:1'000, the Layers have been updated in 2012 and the declared planimetric accuracy is 0.20 m (Specifiche Tecniche aerofotogrammetriche per la realizzazione del Data base topografico alle scale 1:1.000 e 1:2.000 2007). This is the most up-to-date, large-scale and accurate dataset available for the area of interest.

3.1.3 ISO 19157

ISO 19157 is the selected set of guidelines used to perform the assessment. They provide principles for describing the quality of geographic data and a consistent and standard manner to determine and report quality information of a dataset. It also aims at providing guidelines for quantitatively evaluating quality information of geographic data.

In order to describe the quality of geographic data, different quality elements are considered. A data quality unit is the combination of a scope and data quality elements, as showed in the scheme of figure 3.4. Specifically, the data quality units are completeness, logical consistency, usability element, positional accuracy, thematic accuracy and temporal accuracy, as stated in Subsection 1.3.3. On the other hand, data quality elements are components that describe a certain aspect of the quality of geographic data. For instance, the data quality elements of completeness are commission and omission.

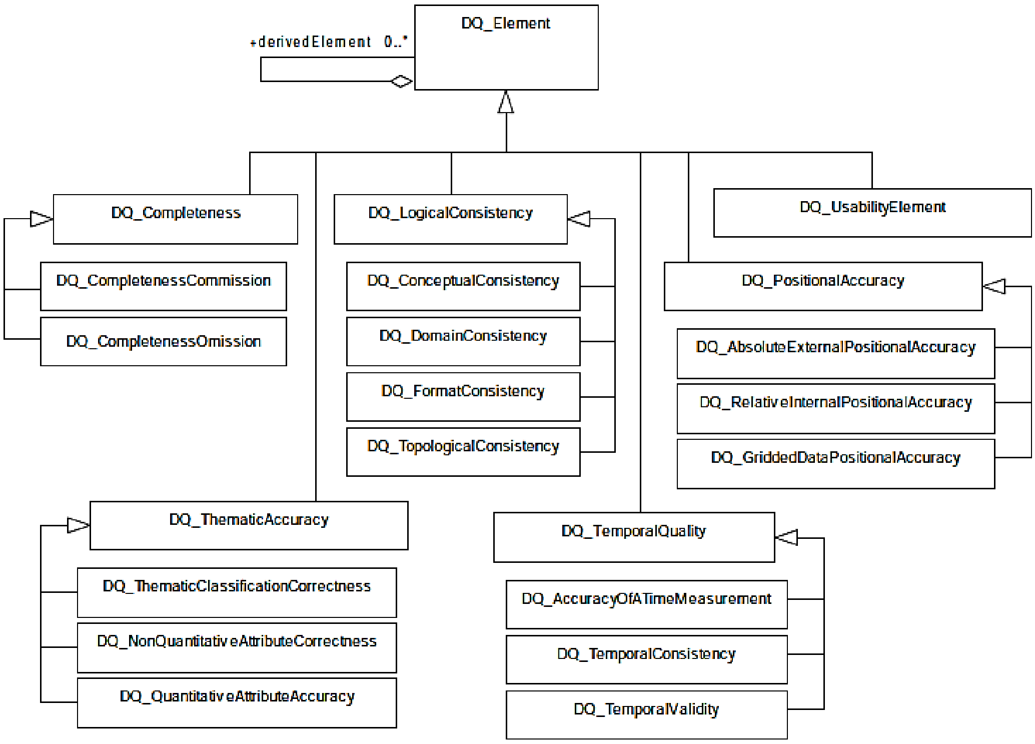

DQ_Element
+derivedElement 0..*
DQ_Completeness
DQ_CompletenessCommission
DQ_CompletenessOmission
DQ_LogicalConsistency
DQ_ConceptualConsistency
DQ_DomainConsistency
DQ_FormatConsistency
DQ_TopologicalConsistency
DQ_UsabilityElement
DQ_PositionalAccuracy
DQ_AbsoluteExternalPositionalAccuracy
DQ_RelativeInternalPositionalAccuracy
DQ_GriddedDataPositionalAccuracy
DQ_ThematicAccuracy
DQ_ThematicClassificationCorrectness
DQ_NonQuantitativeAttributeCorrectness
DQ_QuantitativeAttributeAccuracy
DQ_TemporalQuality
DQ_AccuracyOfATimeMeasurement
DQ_TemporalConsistency
DQ_TemporalValidity

Figure 3.4 – Overview of the components of data quality (source: ISO 2013)

To enable the comparison between different datasets, it is necessary that the results in the data quality reports are expressed in a standardized, comparable way. For this reason, Annex D of ISO 19157 describes the indexes useful to define a specific data quality element. The choice of the index to use depends on the type of data and its intended purpose (ISO 2013).

3.1.3.1 Data quality evaluation methods

A data quality evaluation procedure includes one or more data quality evaluation methods. The latter can be divided into two main classes: direct and indirect. Direct evaluation methods determine data quality through the comparison of the data with internal and/or external reference information. Indirect evaluation methods infer or estimate data quality using information on the data themselves, such as the lineage. Direct evaluation methods should be used in preference to indirect evaluations. The direct evaluation methods are further sub-classified according to source of information required to perform the evaluation, which can be internal or external (ISO 2013).

3.1.3.2 The data quality process flow

ISO 19157 suggests a sequence of steps to produce a quality assessment result. Figure 3.5 shows an example of such a quality assessment process.

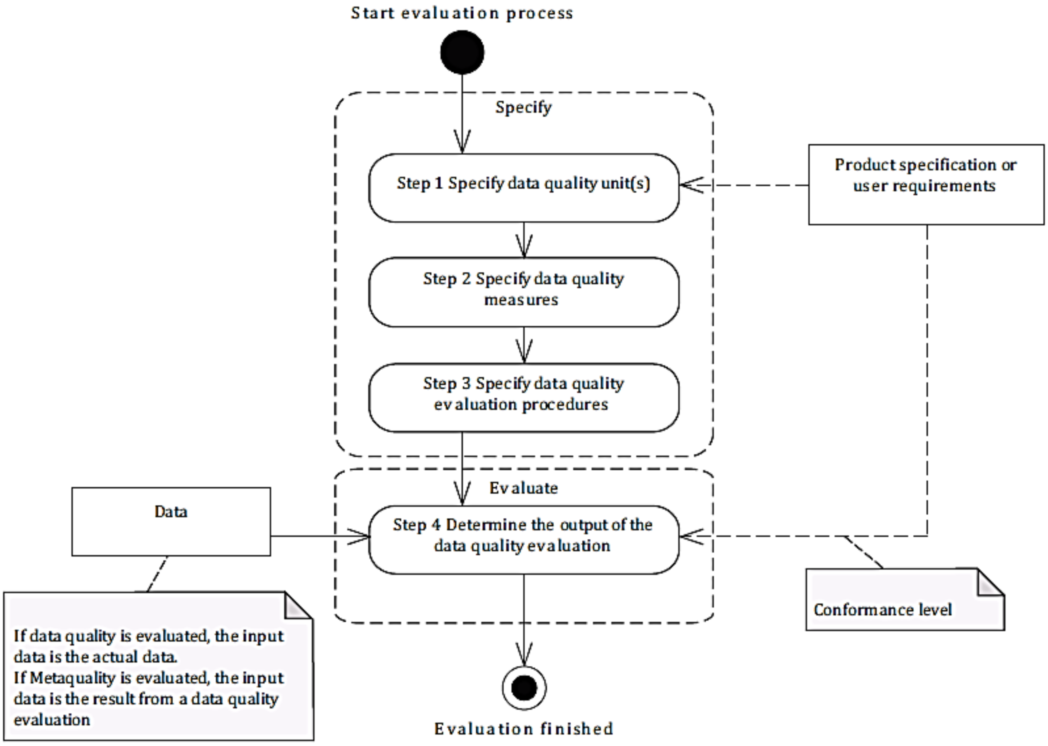

Start evaluation process
Specify
Step 1 Specify data quality unit(s)
Product specification or user requirements
Step 2 Specify data quality measures
Step 3 Specify data quality evaluation procedures
Evaluate
Data
Step 4 Determine the output of the data quality evaluation
Conformance level
If data quality is evaluated, the input data is the actual data.
If Metaquality is evaluated, the input data is the result from a data quality evaluation
Evaluation finished

Figure 3.5 – ISO 19157 process flow (source: ISO 2013)

The process is composed of four steps which are fully explained in table 3.1. The first three steps focus on the specification of data quality units, measures and procedures, while the fourth one involves the data and finally performs the evaluation. In order to evaluate if the product specifications meet the user requirements, the conformance level must be checked.

This procedure can be applied on both data and metadata.

Table 3.1 – Process flow steps (source: ISO 2013)

Step	Action	Description
1	Specify data quality unit(s)	Data quality elements relevant to the data for which quality is to be described should be used.
2	Specify data quality measures	Specify each data quality elements involved.
3	Specify data quality evaluation procedures	Define the applied evaluation methods.
4	Determine the output of the data quality evaluation	A result is the output of applying the evaluation

3.1.4 Software used in the evaluation

The software used for the evaluation of the quality of the orthophoto is QGIS version 2.14 Essen, the last long-term release available at the time of writing (available at https://www.qgis.org/en/site/). It is one of the most popular free and open source software for geospatial applications and it is available under the GNU General Public License (GPL, https://www.gnu.org/licenses/gpl-3.0.en.html). QGIS offers many typical GIS functions (QGIS 2014) including:

- View and explore data;
- Compose and print maps;
- Create, edit, manage and export data;
- Analyse data;
- Publish maps on the Internet;
- Extend functionalities through plugins (core plugins or external Python plugins).

In this evaluation, the most used functions will be the ones connected to the manipulation of vector datasets and the digitization tools.

The possibility to create and openly share plugins is one of the core features of QGIS. Even in this study, to achieve specific tasks a plug-in is needed named MMQGIS (see Subsection 3.3.2.1).

3.2 Methodology

The process flow followed to assess the quality of the AGEA 2012 orthophoto is showed in figure 3.6. It reflects the example suggested in ISO 19157 (see Subsection 3.1.3.2) applied to the specific case study. The first step regards the specification of the data quality unit, which is the positional accuracy. The second step specifies the measures taken into account, which in this case are represented by the relative or internal accuracy (see Subsection 3.2.1). The evaluation procedure is performed in the third step and it is a direct external evaluation (see Subsection 3.2.2). These steps make it possible to determine the quality (meant here as positional accuracy) of the orthophoto and to compare the result of the assessment with the planimetric accuracy declared in the metadata (3 m).

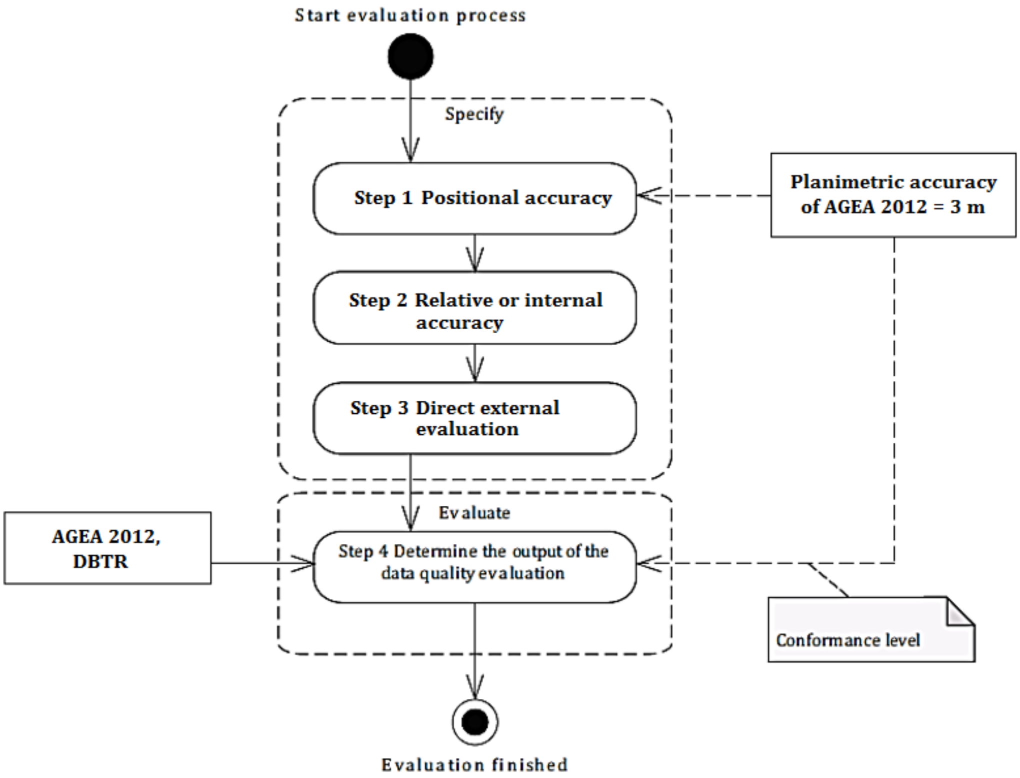

Start evaluation process
Specify
Step 1 Positional accuracy
Step 2 Relative or internal accuracy
Step 3 Direct external evaluation
Planimetric accuracy of AGEA 2012 = 3 m
Evaluate
AGEA 2012, DBTR
Step 4 Determine the output of the data quality evaluation
Conformance level
Evaluation finished

Figure 3.6 – Process flow to assess the positional accuracy of the AGEA 2012 orthophoto (Source: ISO 2013)

Each of the four steps is fully explained in the following subsections.

3.2.1 Positional accuracy

ISO 19157 defines positional accuracy as the accuracy of the position of features within a spatial reference system (ISO 2013). It consists of three data quality elements:

- Absolute or external accuracy: closeness of reported coordinate values to values accepted as or being true;
- Relative or internal accuracy: closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true;
- Gridded data positional accuracy: closeness of gridded data spatial position values to values accepted as or being true.

In this case, the selected data quality element is the relative or internal accuracy. In fact, this study will estimate the distance between the points represented in the orthophoto and the positions of their homologous points, determined using the DBTR building layer (see Subsection 3.3.4) which is considered as the ground truth.

3.2.2 Direct evaluation

A direct evaluation method is a procedure for assessing the quality of a dataset based on inspection of the items within it.

Direct evaluation methods can be classified as internal or external. Internal direct data quality assessment makes only use of the same dataset being evaluated. Conversely, external direct quality assessment requires reference data external to the dataset being evaluated. For both external and internal evaluation methods, one of the following inspection methods may be used:

- Full inspection;
- Sampling.

Full inspection tests every item in the population specified by the data quality scope. It is most appropriate for small populations or for tests that can be accomplished by automated means. On the other hand, sampling means that tests are performed on subsets of the geographic data defined by the data quality scope. Some guidelines are given in Annex F of ISO 19157 (ISO 2013).

In this case study, a direct external evaluation is performed because a reference external dataset (extracted from DBTR, see Subsection 3.1.2.1) will be used in the assessment. As the dimension of the dataset is quite large (the orthophoto covers the whole Italy and is here assessed on Milan Municipality), the sampling method will be chosen as inspection process.

3.2.3 Sampling procedure

The sampling procedure is described in ISO 2859 series "Sampling procedures for inspection by attributes" (1989) and in ISO 3951 "Sampling procedures for inspection by variables" (2013) and it is originally developed for a more general (non-spatial) use. Nevertheless, ISO 19157 focuses on its application and introduces sampling techniques specific to geographic data.

To understand the sampling procedure the difference between "lot" and "item" has to be introduced. A "lot" is the minimum unit for which quality may be evaluated. An "item" is the minimum unit to be inspected and should be defined by the data producer in accordance with the data product specification.

The size of a population, and consequently the size of samples, is defined according to the selected items. The definition of a sample size requires an explicit indication of the items, which can be features, area covered, curves or vertices (ISO 2013). In this case study, the chosen items are the features, specifically the shapes of the buildings. Knowing their total number, it is possible to select an appropriate percentage to evaluate them (see Subsection 3.3.2).

3.2.3.1 Sampling strategies

The sampling strategy is shown in figure 3.7. Two aspects compose it: the items to be sampled (areas or features) and the manner by which they are selected (probability or judgement).

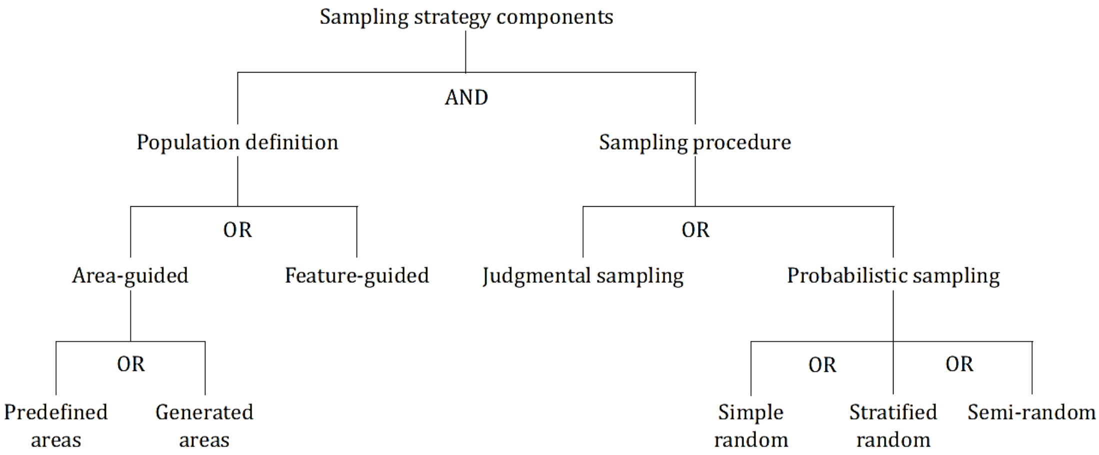


Figure 3.7 – Sampling strategy (source: ISO 2013)

According to the type of procedure, the sampling can be either probabilistic or judgmental. The first one applies the sampling theory and is based on a random selection of the sample items. The essential characteristic of probabilistic sampling is that each member of the population from which the sample is extracted has a known probability of selection. Three kinds of probabilistic sampling are possible:

- Simple random sampling: it extracts a random sample from the whole population. It is useful if the dataset is homogeneous in the sampled characteristic;
- Stratified random sampling: it divides the population into non-overlapping strata or subpopulations that are more homogeneous in the items to be sampled, and extracts a random set from each of them;
- Semi-random sampling: it randomly extracts the items of the initial sample and then applies a rule for selecting the remaining items.

Conversely, the design of a judgemental sampling involves the selection of samples based on experts' knowledge or professional judgement.

In this case study, the probabilistic sampling with a stratified random sampling technique best fits the intended evaluation. In fact, the distribution of buildings in Milan Municipality is not homogenous, but it depends on the city zoning as shown in figure 3.8. Here, the city is divided using a squared grid of 1 km x 1 km size and the colours of the grid cells show the density of buildings: the darker it is, the higher the number of building contained. It can be noticed that there is a ring around the city centre where the building density is higher and that the density generally decreases towards the city outskirts. For this reason, the stratified random sampling technique best fits the evaluation as it allows to create an homogeneous sampling on the area of interest. Nevertheless, the stratified random sampling has the potential for a greater precision in the estimates of mean and variance compared to a non-stratified sampling technique applied on the same population (ISO 2013).

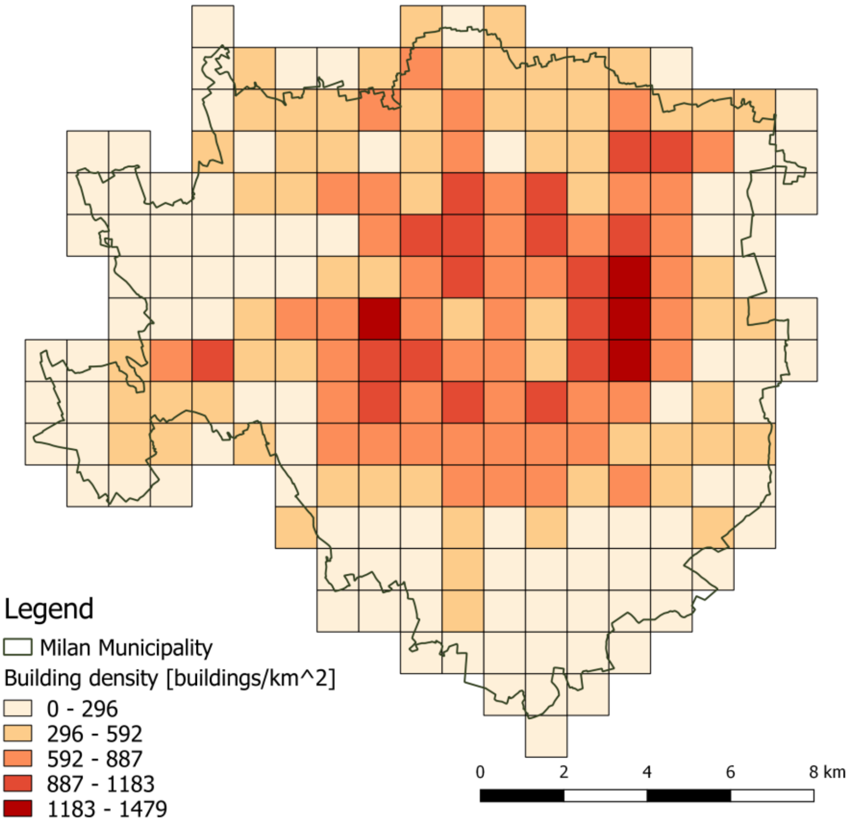


Figure 3.8 – Building density in Milan Municipality

In addition to the sampling procedure, the population must be also defined (see figure 3.8). According to ISO (2013), population definition can be area-guided or feature-guided. The second one does not imply any spatial sampling, hence it is not feasible in this analysis. On the contrary, the area-guided approach selects sampling units as geographic areas (e.g. administrative or statistical areas) or some other forms of partitioning of the population of interest. In fact ISO recommends to use an area-guided approach if the coverage of the entire area is important, in order to sample according to a regular or semi-regular pattern (ISO 2013), which best fit this work.

Therefore, in this study the total area of interest (Milan Municipality) will be subdivided into cells.

Summarizing, the sampling procedure applied in this case study will have a generated area as the population definition and a stratified random sampling as the probabilistic sampling technique.

3.2.3.2 Probability-based sampling

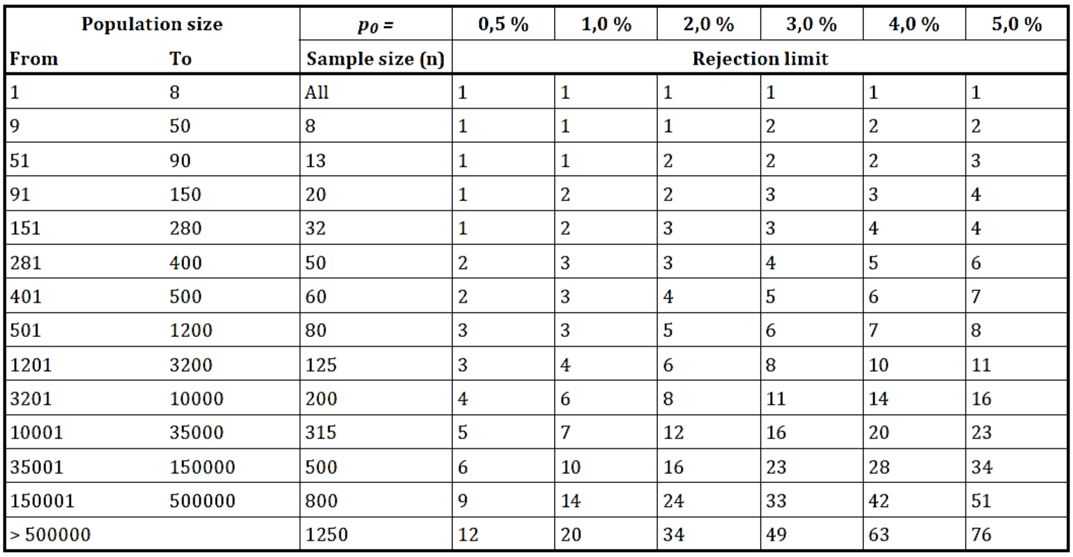
Before starting the sampling, ISO suggests to pay particular attention to the following:

- The areas covered by a geographic dataset have to form a continuous space. When splitting the dataset into lots, special attention should be paid to the omission or commission of items crossing over the lot boundaries;
- A variety of factors, including the quality of source data and skill of operators, may affect the quality of geographic data. The data producer should be careful to define lots to achieve homogeneity in terms of quality.

The quality assessment of a sample is defined as AQL (Acceptance Quality Limit) and it is explained in ISO 3534-2:2006 (2006). It is a requirement limit to evaluate the conformity of each item, based on the data product specifications.

ISO specifications are included in table 3.2, which defines the recommended sample size according to the population size, as well as the associated rejection limit (p_o) of the AQL based on the hypergeometric distribution (ISO 2013). It is assumed that the deviations fit this distribution.

Table 3.2 – Statistical values to test the number of conforming/non-conforming items with a significance level of 95% (source: ISO 2013)



Population size		p_0 =	0,5 %	1,0 %	2,0 %	3,0 %	4,0 %	5,0 %
From	**To**	**Sample size (n)**	**Rejection limit**					
1	8	All	1	1	1	1	1	1
9	50	8	1	1	1	2	2	2
51	90	13	1	1	2	2	2	3
91	150	20	1	2	2	3	3	4
151	280	32	1	2	3	3	4	4
281	400	50	2	3	3	4	5	6
401	500	60	2	3	4	5	6	7
501	1200	80	3	3	5	6	7	8
1201	3200	125	3	4	6	8	10	11
3201	10000	200	4	6	8	11	14	16
10001	35000	315	5	7	12	16	20	23
35001	150000	500	6	10	16	23	28	34
150001	500000	800	9	14	24	33	42	51
> 500000		1250	12	20	34	49	63	76

The table must be used as follows:

- Decide the population size of the items to be checked;
- Select the corresponding sample size (*n*) from the table;
- Perform the evaluation, and count the number of “failed items”;
- The sample is rejected if the number of failed items is equal or higher than the rejection limit for the chosen values of *n* and p_o (AQL).

In this case study, the AQL is set as p_o=5,0%.

3.2.4 Positional uncertainty

The output of the data quality assessment consists of the values of the positional uncertainty. It is defined in ISO as the distance between the measured position of a point (x_{mi}, y_{mi}) and what is considered to be its corresponding true position (x_{ti}, y_{ti}).

In this case study, positional uncertainty is evaluated on the *x* axis (see Equation 3.1), on the *y* axis (see Equation 3.2) and on the bi-dimensional plane, computing an Euclidean distance based on the two previous measures (see Equation 3.3).

$$e_{xi} = \sqrt{(x_{mi} - x_{ti})^2} \qquad \text{Equation 3.1}$$

$$e_{yi} = \sqrt{(y_{mi} - y_{ti})^2} \qquad \text{Equation 3.2}$$

$$e_i = \sqrt{(x_{mi} - x_{ti})^2 + (y_{mi} - y_{ti})^2} \qquad \text{Equation 3.3}$$

ISO suggests that a criterion to establish correspondence between the measured and the true position should also be stated (e.g. allowing for correspondence to the closest position, on vertices or along lines). The criterion/criteria used to find the corresponding points shall be reported as well together with the data quality assessment result (ISO 2013).

In this case study, the measured positions are the digitized positions of building corners on the AGEA 2012 orthophoto, while the corresponding true positions are the positions of building corners on the DBTR (see figure 3.9).

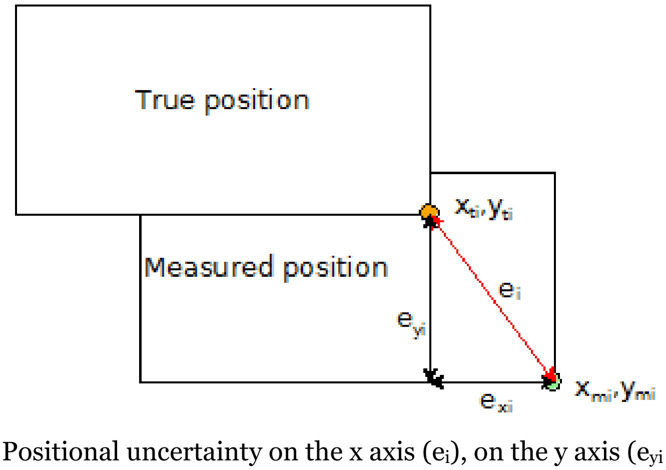


Figure 3.9 – Positional uncertainty on the x axis (e_i), on the y axis (e_{yi}) and on a bi-dimensional plane (e_{xi}) measured from the orange dot (x_{ti}, y_{ti}) representing the positions of building corners on the DBTR (True position) respect to green dot (x_{mi}, y_{mi}), which represents the position of the homologous point on the AGEA 2012 orthophoto (Measured position)

3.2.5 Statistical measures

As the sampling procedure is probabilistic, many statistical measures can be computed on the positional uncertainties. In particular, 7 different indexes will be evaluated: the mean, the median, the standard deviation, the minimum, the maximum, the number of positional uncertainties which are above a given threshold, and the confidence area. The last two measures are specifically indicated by ISO (2013) for the internal positional accuracy in its Annex D.

3.2.5.1 Mean

The mean μ is the measure of the central tendency either of a probability distribution or of the random variable characterized by that distribution. It is the average of all the values.

3.2.5.2 Median

The median *Me* is the value separating the higher half of a data sample, a population, or a probability distribution, from the lower half. In simple terms, it may be thought as the "middle" value of a data set.

3.2.5.3 Standard deviation

The standard deviation σ is the measure used to quantify the amount of variation or dispersion of a set of data values. For this case study, the standard deviation is calculated on the positional uncertainties (see Equation 3.4).

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i - \bar{e})^2}$$ Equation 3.4

3.2.5.4 Minimum

The minimum *min* is the minimum value in the total set of values.

3.2.5.5 Maximum

The maximum *max* is the maximum value in the total set of values.

3.2.5.6 *Number of positional uncertainties above a given threshold*

It is the number of computed positional uncertainties (see Equations 3.1, 3.2 and 3.3) which are above a given threshold e_{max} for a set of positions.

3.2.5.7 *Confidence area*

It is described as a circle around the best estimation for the true value. The probability for the true value to lie in this area is calculated by area integration over the two-dimensional density function of the normal distribution (see Equation 3.5). Its radius *R* characterizes a circular area and it is used as a measure for the accuracy of two-dimensional random variables.

$$P\left(R, \sigma_x, \sigma_y\right) = \frac{1}{2\pi\sigma_x\sigma_y} \iint_{(x-x_t)^2+(y-y_t)^2=R^2} e^{-\frac{1}{2}\left(\frac{(x-x_t)^2}{e_x^2}+\frac{(y-y_t)^2}{e_y^2}\right)} dx\, dy \qquad \text{Equation 3.5}$$

Table 3.3 shows some estimations of the radius *R* for some particular probability values. For the case study of AGEA 2012 orthophoto, these indexes of confidence area will be evaluated in Subsection 3.3.6.

Table 3.3 – Relationship between the probability *P* and the corresponding radius of the circular area (source: ISO 19157)

Probability P	**Radius R**	**Name**
P = 39.4%	$\frac{1}{\sqrt{2}}\sqrt{\sigma_x^2+\sigma_y^2}$	CE39.4
P = 50%	$\frac{1.1774}{\sqrt{2}}\sqrt{\sigma_x^2+\sigma_y^2}$	CE50
P = 90%	$\frac{2.146}{\sqrt{2}}\sqrt{\sigma_x^2+\sigma_y^2}$	CE90
P = 95%	$\frac{2.4477}{\sqrt{2}}\sqrt{\sigma_x^2+\sigma_y^2}$	CE95
P = 99.8%	$\frac{3.5}{\sqrt{2}}\sqrt{\sigma_x^2+\sigma_y^2}$	CE99.8

3.2.6 Digitization error

In an experimental process, e.g. when a map is digitized, different kinds of error can occur. Traditionally, when an analogic map is transformed into a digital one, the global instrumental error must be taken into consideration. This error can be seen as the combination of different factors (Spalla and Galletto 2000):

- Intrinsic error (i.e. connected to the precision of the tool);
- Collimation error (i.e. due to the depth and the shape of the tracking tool);
- Parallax error (i.e. if the glass is not correctly laid out on the collimation point);
- Local imperfection of the digitization table;
- Different response of the digitizer based on the position of the cursor (i.e. solenoid eccentricity with respect to the centre of the collimation tool).

In this case study the situation is slightly different, as the digitization process is performed on an orthophoto that is already in a digital form, hence not all the errors listed above can occur. Nevertheless, the intrinsic error and the collimation error must be taken into consideration.

Spalla and Galletto (2000) defined an evaluation method used to define the global instrumental error. This approach considers some hundreds of points (*NP*) that are digitized many times by different operators. For each point digitizated NM times, the average values of its measured positions on the two axes (X_i, Y_i) are assumed as true (see Equation 3.6 and Equation 3.7)

$$X_i = \frac{\sum_{j=1}^{NM} x_{ij}}{NM}$$ Equation 3.6

$$Y_i = \frac{\sum_{j=1}^{NM} y_{ij}}{NM}$$ Equation 3.7

Then, for each point, the residual s_{ij} with respect to the mean value is computed (see Equation 3.8).

$$s_{ij} = \sqrt{(x_{ij} - X_i)^2 + (y_{ij} - Y_i)^2}$$

for $i = 1 \div NP, j = 1 \div NM$ Equation 3.8

After that, the mean square error m_i can be computed (see Equation 3.9).

$$m_i = \sqrt{\frac{\sum_{j=1}^{NM} s_{ij}^2}{NM}} \qquad \text{Equation 3.9}$$

Finally, the digitization mean square error σ_D is computed as the mean of the mean square errors (see Equation 3.10).

$$\sigma_D = \frac{\sum_{i=1}^{NP} m_i}{NP} \qquad \text{Equation 3.10}$$

In general, it is expected that the loss of precision introduced in the digitization process does not affect the final product too much, i.e. it is at least one order of magnitude lower than the positional uncertainty.

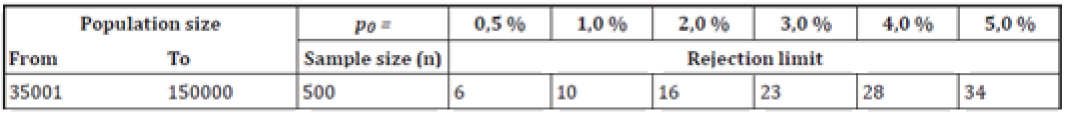
3.3 Results

In this Section, each step of the AGEA 2012 orthophoto quality assessment is explained from a practical point of view. The procedure consists of the application of the ISO guidelines to this case study, in order to carry out the assessment process flow (see Section 3.2). Results give an indication on whether the positional accuracy of the AGEA 2012 orthophoto for Milan Municipality is compliant to the declared one.

3.3.1 Definition of the sample

The DBTR of Milan Municipality contains 89’000 building elements. Following ISO recommendations (see table 3.2), the corresponding sample size is 500 items and the rejection limit connected to an AQL of 5,0% is 34 (see the detail of table 3.4). Hence, a minimum of 466 buildings must be detected, otherwise the sample will be rejected.

Table 3.4 – Population, sample size and rejection limit in the case of Milan Municipality buildings (source: ISO 2013)



Population size		p_0 =	0,5 %	1,0 %	2,0 %	3,0 %	4,0 %	5,0 %
From	To	Sample size (n)	Rejection limit					
35001	150000	500	6	10	16	23	28	34

3.3.2 Sampling strategy

As described in Subsection 3.2.3.1, the sampling procedure is divided in two parts. The first one is the population definition, that is area-guided and based on the generation of a grid, while the sampling procedure consists of a stratified random sampling (see figure 3.10).

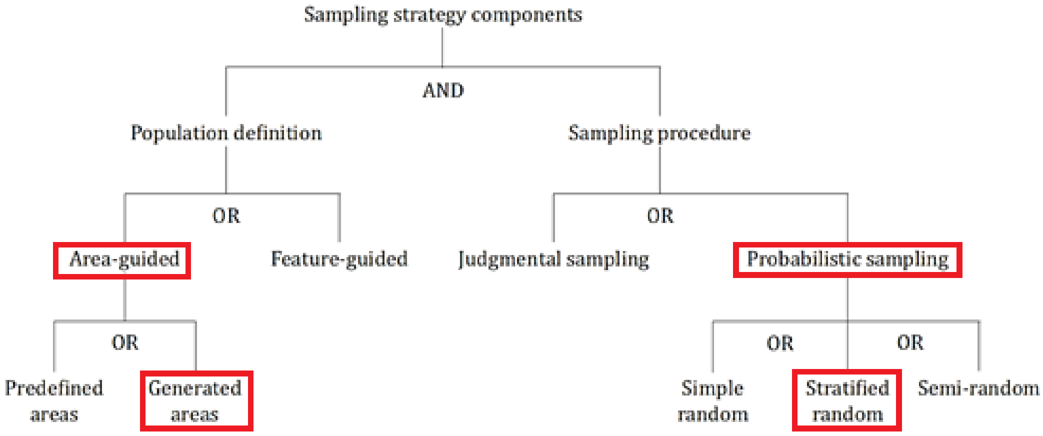


Figure 3.10 – Sampling procedure (source: ISO 19157)

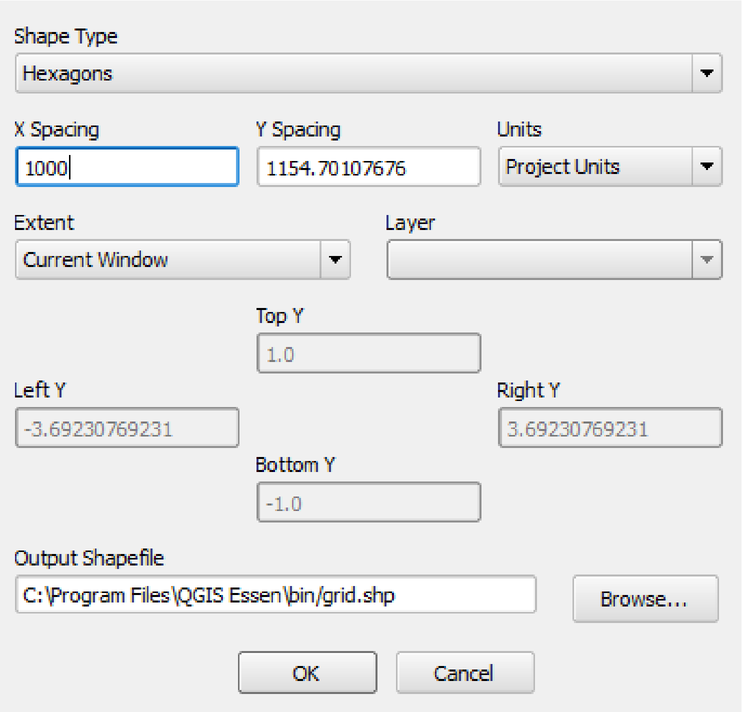
3.3.2.1 Population definition

In order to cover the dataset with a continuous space, the study area is divided by means of a hexagonal grid. Hecht et al. (2013) suggest this choice by stating the following:

> “Hexagonal raster offers the advantage of more closely approximating the circle while providing the same complete coverage of the study area”.

To obtain a sample of 500 buildings, the developed strategy consists of the creation of a grid composed of 250 cells and the selection of 2 buildings in each cell.

To achieve that a specific QGIS is used: MMQGIS. It is a set of Python scripts that allow to integrate specialized tools to manipulate vectors developed by Michael Minn (http://michaelminn.com/linux/mmqgis/). The MMQGIS plugin allows the creation of the grid using the command *Create > Create grid layer*. Then a GUI tool allows the user to define some parameters (see figure 3.11).



Shape Type: Hexagons

X Spacing: 1000

Y Spacing: 1154.70107676

Units: Project Units

Extent: Current Window

Layer:

Top Y: 1.0

Left Y: -3.69230769231

Right Y: 3.69230769231

Bottom Y: -1.0

Output Shapefile: C:\Program Files\QGIS Essen\bin/grid.shp

Figure 3.11 – MMQGIS menu to create the grid

It is possible to define the shape of the grid cellas (choices include, among the others, rectangles, diamonds and hexagons), the dimension of the grid cells, the extension of the grid and the location of the vector grid file to be saved.

Figure 3.12 shows the meaning of the parameters *X Spacing* and *Y Spacing*. The value of *Y Spacing* is double than the radius of the circle in which the hexagon is circumscribed, while *Y Spacing* depends on *X Spacing* according to Equation 3.11.

$$Y\ Spacing = \frac{X\ Spacing}{\sqrt{3}}\ 2 \qquad \text{Equation 3.11}$$

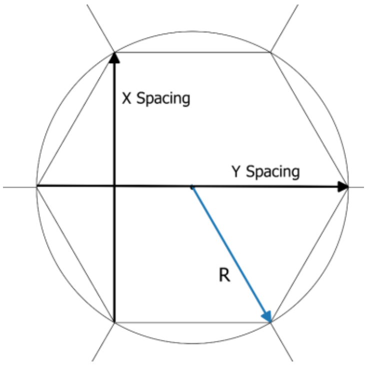
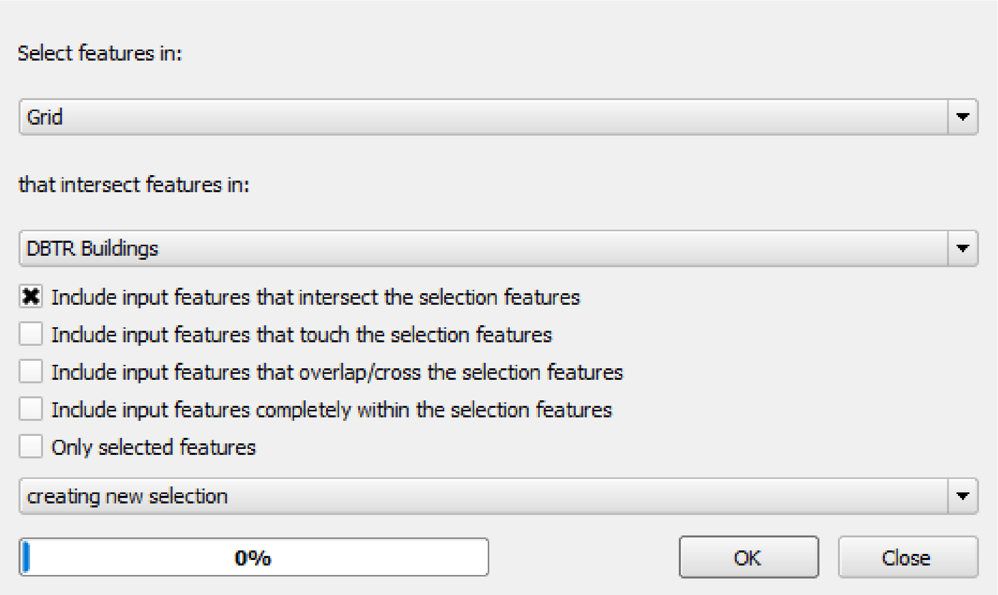


Figure 3.12 – Geometrical meaning of X Spacing and Y Spacing

The extension of the grid was set as the “Current Window”, in order to create a grid that is wider than the extension of the DBTR layer (provided of course that the QGIS window shows the entire DBTR layer). Then, using the QGIS command *Vector > Research Tools > Select by location*, only the grid cells that contain at least one building are selected. The input files are clearly the grid and the building vector layers (see figure 3.13).



Select features in:

Grid

that intersect features in:

DBTR Buildings

☒ Include input features that intersect the selection features
□ Include input features that touch the selection features
□ Include input features that overlap/cross the selection features
□ Include input features completely within the selection features
□ Only selected features

creating new selection

Figure 3.13 – QGIS *Select by location* menu to select only the grid cells containing buildings

After that, the grid cells which do not contain any building are selected and deleted. The final number of grid cells containing at least a building is found. . Table 3.5 lists the different tests performed to find the grid that best fits the requirements (i.e. that includes at least 500 buildings with a maximum of 2 buildings for each hexagonal cell). At the end of this process, a grid with a radius of 550 m is chosen. This grid contains 269 hexagons, which is slightly more than the required number (250). In fact, while in principle a maximum number of 538 buildings can be found, there are some cells (mainly towards the outskirts of Milan Municipality) containing only one building.

Table 3.5 – List of tested grids (values are rounded to the nearest integer number)

Y Spacing (m)	X Spacing (m)	Radius (m)	Number of hexagons
866	750	433	365
1’000	866	500	324

1'010	950	505	318
1'039	900	520	299
1'100	**953**	**550**	**269**
1'150	996	575	251
1'154	1000	577	250
1'200	1039	600	231

Figure 3.14 shows the chosen hexagonal grid. It can be noticed that there is one cell without buildings in the bottom-right parts this is mainly an agricultural/rural area.

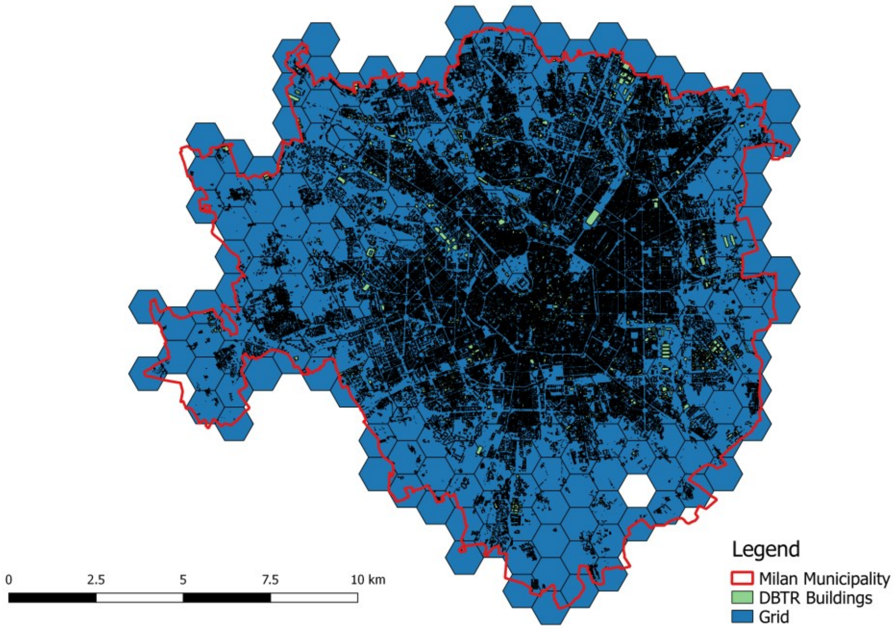
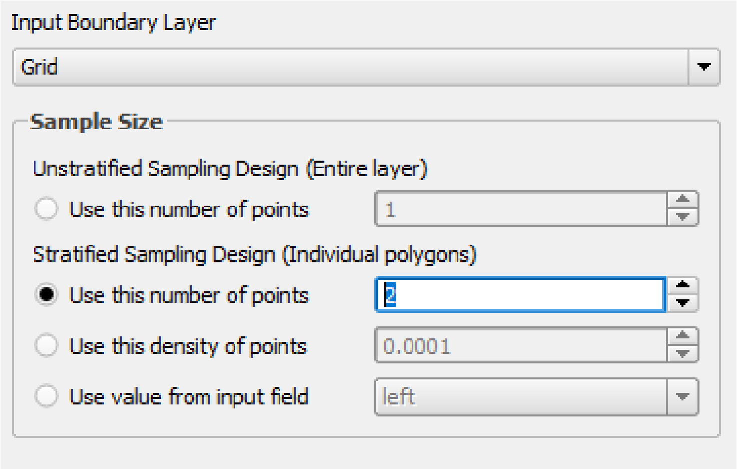


Figure 3.14 – Hexagonal grid superimposed to the boundary of Milan Municipality and the DBTR building layer.

It can be noticed as well that some portions included in the boundary of Milan Municipality do not again belong to any grid cell, as they do not contain buildings.

3.3.2.2 Sampling procedure

The sampling procedure consists of randomly picking up two points in each hexagonal cell. This is performed in QGIS using the command *Vector* > *Research tool* > *Random points*. The number of points was set to 2 for each cell and the result of this operation is the creation of a new point vector layer (see figure 3.15). Figure 3.16 shows the hexagonal grid with the random points.



Input Boundary Layer

Grid

Sample Size

Unstratified Sampling Design (Entire layer)

○ Use this number of points: 1

Stratified Sampling Design (Individual polygons)

● Use this number of points: 2

○ Use this density of points: 0.0001

○ Use value from input field: left

Figure 3.15 – QGIS *Random points* tool for the random extraction of 2 random points in each grid cell

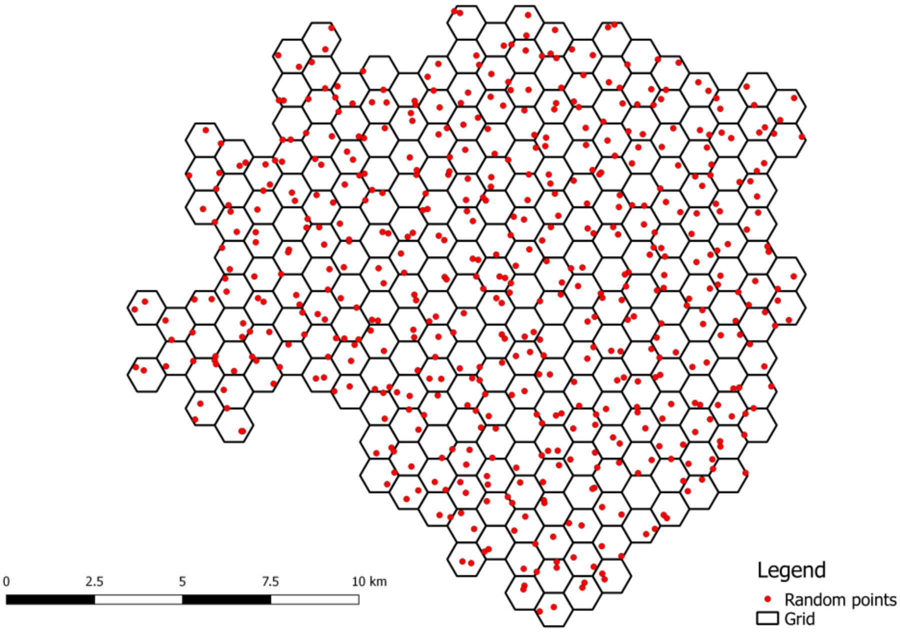


Figure 3.16 – Grid with 2 random points in each cell

The buildings which have the perimeter closest to the random points constitute the sample (see figure 3.17). Starting with 538 points (corresponding to the 269 grid cells), a total amout of 502 buildings are selected. This happens because some hexagonal cells, mainly located in the outskirts of Milan Municipality, containe only one building.

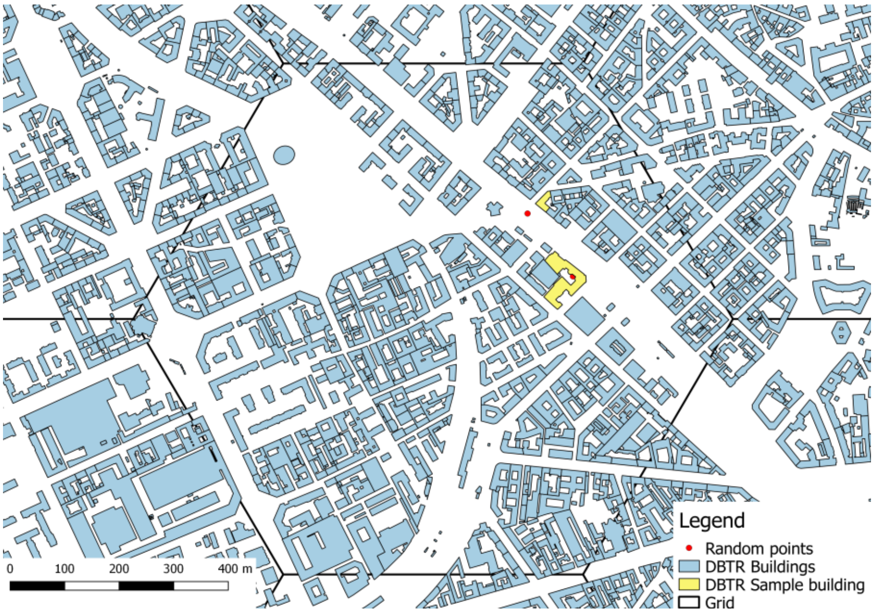


Figure 3.17 – Selection of the nearest building (yellow shapes) from the DBTR building layer respect to the random points (red points)

In this way, a new polygon vector layer composed of 502 buildings is created.

3.3.3 Digitization of the sample

Once the DBTR building sample is extracted, the corresponding buildings on the AGEA 2012 orthophoto must be digitized. This operation is performed in QGIS using the *Digitizing toolbar* tool. First, a new empty polygon layer is created. Then, as shown in figure 3.18, the *Toggle editing* button (pencil button) is activated to enable the editing of the layer. With the *Add feature* function, items (i.e. the buildings digitized on the AGEA 2012 orthophoto) are added to the new layer.

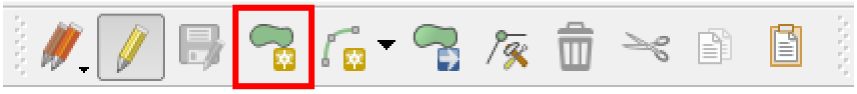


Figure 3.18 – QGIS Digitizing toolbar with the *Add Feature* button highlighted

Figure 3.19 represents the DBTR buildings sample, the grid layer and the AGEA 2012 orthophoto before the digitization of buildings on the AGEA 2012 orthophoto. In each cell, at least one building (ochre polygons) was selected as DBTR sample layer. The blue line represents the grid.

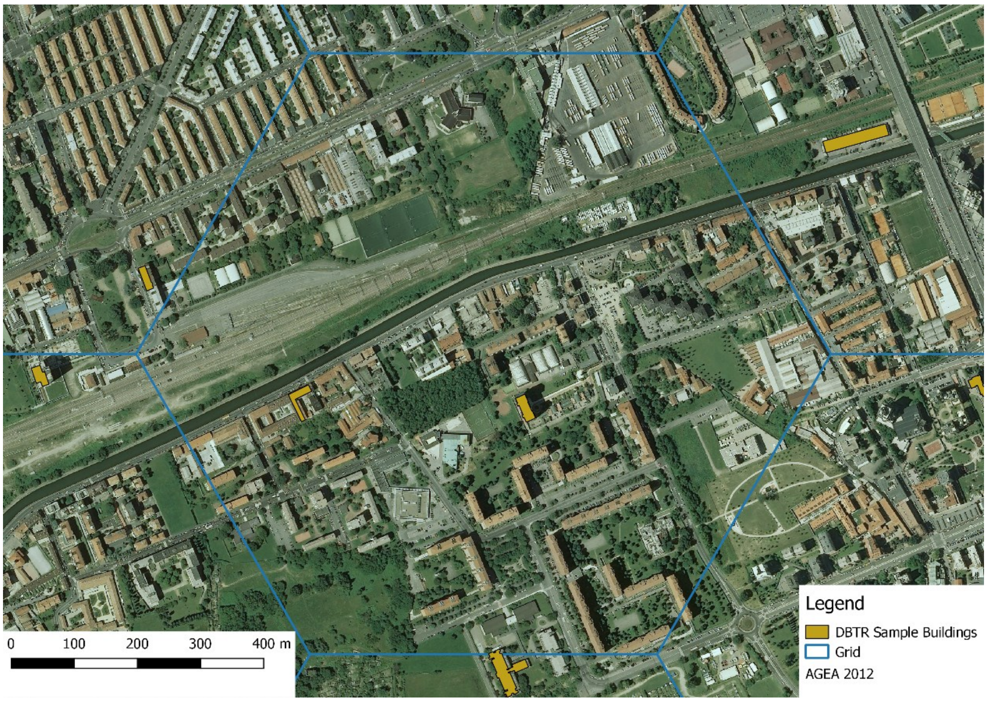


Figure 3.19 – DBTR buildings sample, grid and AGEA 2012 orthophoto

Figure 3.20 represents instead the DBTR buildings sample, the grid layer, the AGEA 2012 orthophoto and the digitized buildings on the AGEA 2012 orthophoto.

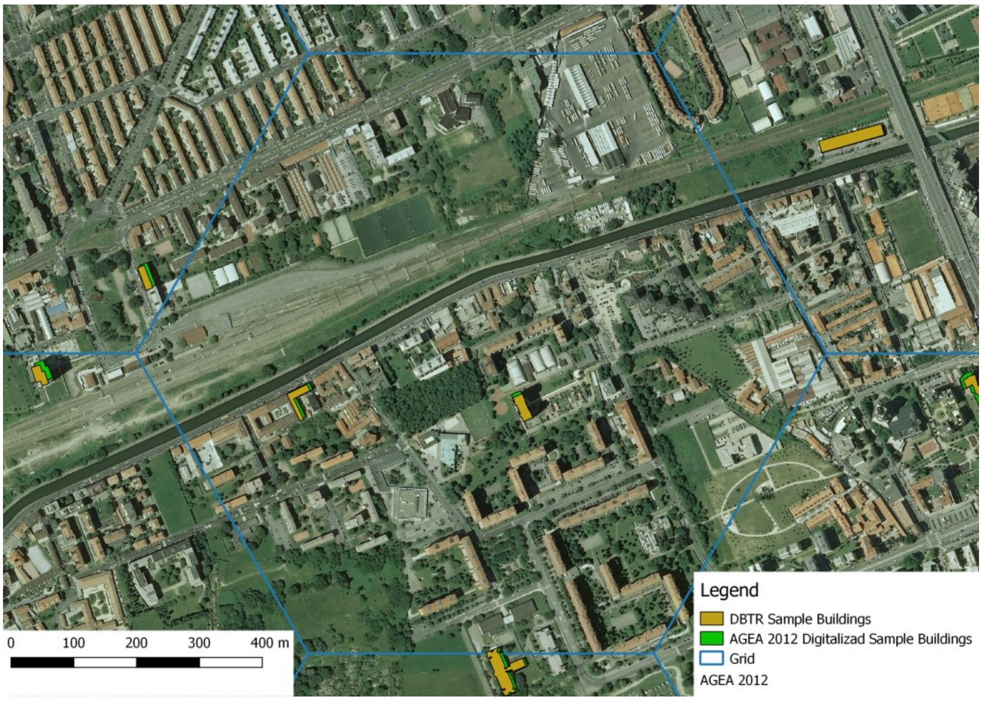


Figure 3.20 – DBTR and AGEA 2012 orthophoto buildings samples, grid and AGEA 2012 orthophoto

3.3.3.1 Acceptance of the sample

A total of 468 buildings are digitalized on AGEA 2012 orthophoto. The buildings corresponding to 34 DBTR buildings are not found on the orthophoto for different reasons, for example due to the presence of shadows or vegetation (see e.g. figure 3.21).

As specified in Subsection 3.2.3.2, AQL is set to 5% and thus the rejection limit is 36. The digitized buildings are 468, which is higher than the limit of 500-36=466. Hence the sample is accepted.

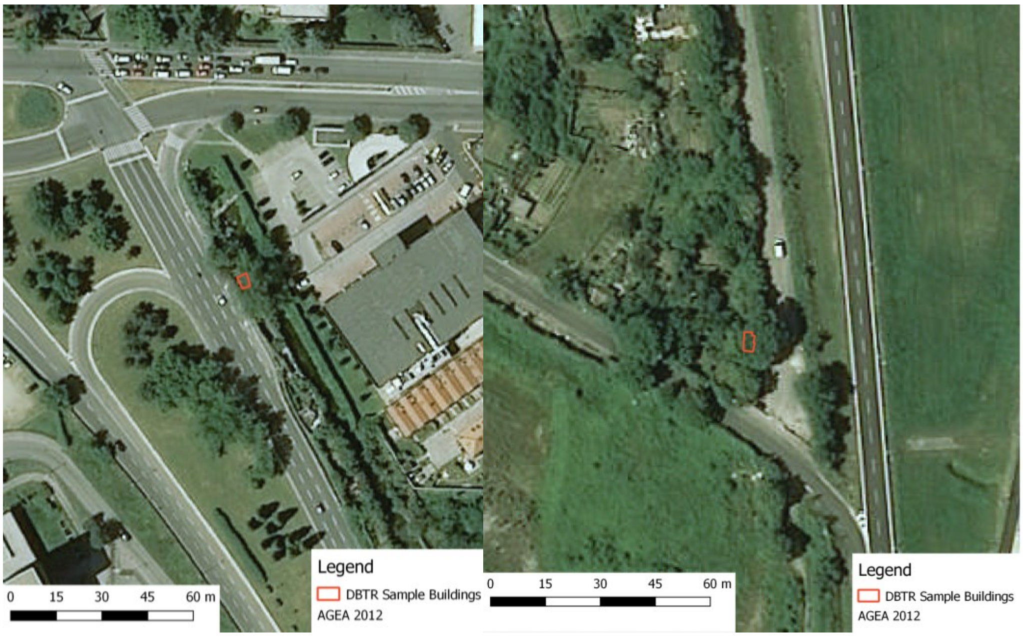


Figure 3.21 – Examples of DBTR buildings which are not detected on the AGEA 2012 orthophoto due to vegetation

In other cases, the homologous buildings do not really exist in the orthophoto itself. Figure 3.22 shows two examples of buildings available in the DBTR but not present in the orthophoto. This is due to the differences between the flight epoch and the date of the production of DBTR or they can be errors in the DBTR.

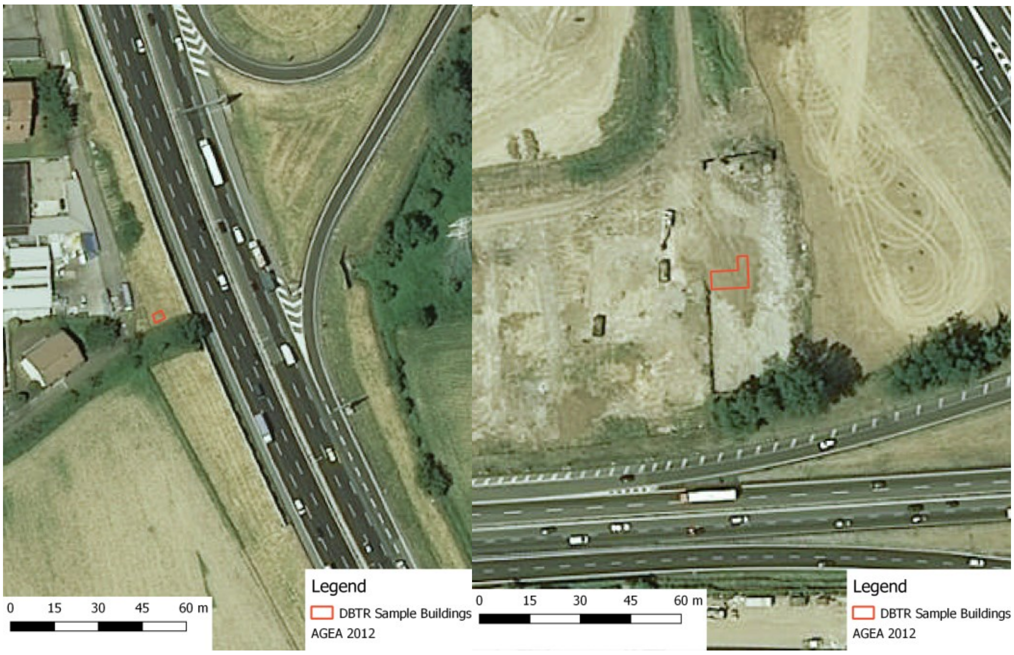


Figure 3.22 – Examples of DBTR buildings which do not exist in the orthophoto

This procedure is performed three times before the digitized buildings sample is compliant with the theshold set by ISO. In fact, the first two samples are rejected by the AQL and the process is restarted from the beginning (i.e. from the extraction of the random points).

3.3.4 Determination of the homologous points

Once the homologous buildings samples from both the DBTR and the AGEA 2012 orthophoto are extracted, they need to be compared. The strategy adopted consists of selecting three homologous couple of corners for each corresponding building of the samples: the closest homologous points, the furthest homologous points and a third couple approximately halfway between the two (see figure 3.23).

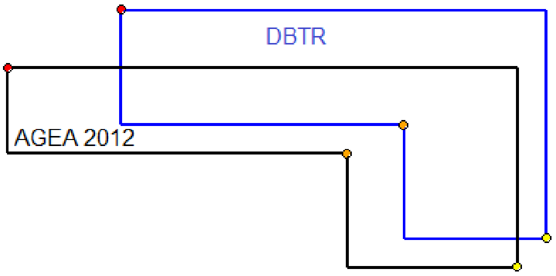


Figure 3.23 – Examples of homologous points selection for a building sample: in yellow the closest homologous points, in red the furthest homologous points and in orange a couple of homologous points with approximately halfway distance respect to the previous ones

In QGIS, the extraction of homologous points is facilitated by the command *Vector* > *Geometry tools* > *Extract nodes*. It creates a new point vector layer containing all the corners of a chosen polygon layer. Then the points which are of no interest (i.e. those different from the 3 points chosen) are deleted.

Figure 3.24 shows an example of the homologous points extraction: the red polygon represents the DBTR sample building with selected corners coloured in brown, while the blue polygon represents the sample building digitized on the AGEA 2012 orthophoto with selected corners coloured in light blue.

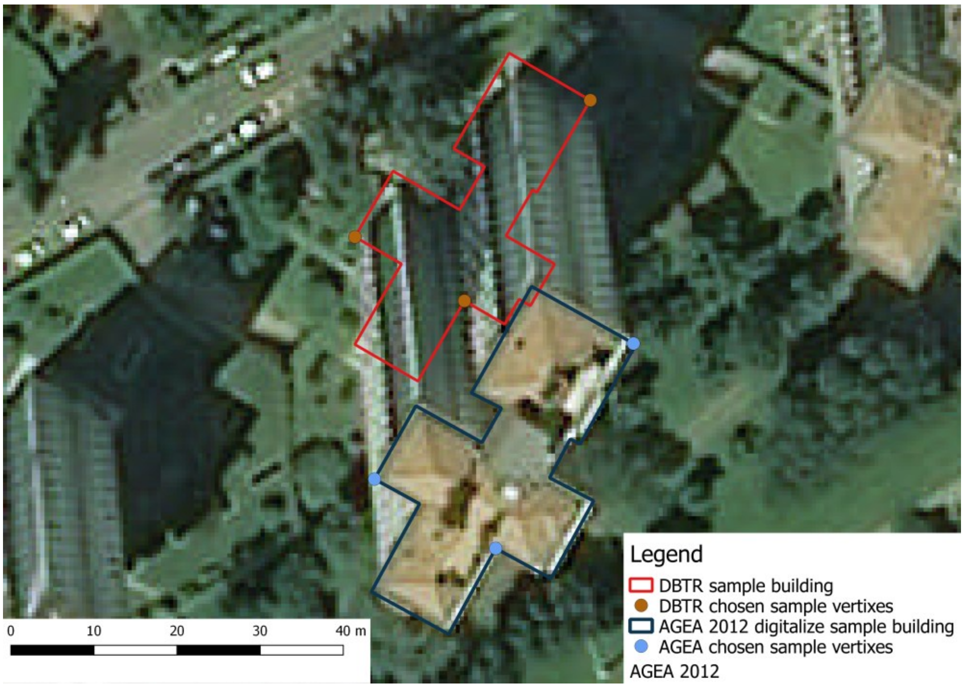

Legend
DBTR sample building
DBTR chosen sample vertixes
AGEA 2012 digitalize sample building
AGEA chosen sample vertixes
AGEA 2012
0
10
20
30
40 m

Figure 3.24 – Example of homologous points on the DBTR and AGEA 2012 orthophoto sample buildings.

To extract the coordinates of each selected corner (required to compute the statistical indexes, see Subsection 3.2.5), the QGIS command *Vector* > *Geometry tools* > *Add/Extract geometry column* is used. This instruction adds two columns to the attribute table of a vector point layer that represent the *X* and *Y* coordinates of the point in a particular coordinate reference system (CRS). In this study, the CRS is set to WGS 84\UTM zone 32 (EPSG: 32632). Hence, the results obtained are expressed in meters.

In some cases, less than 3 corners per building can be detected. Figure 3.25 shows two examples of buildings where only 2 corners could be identified on the orthophoto. In these cases, the problems are due to errors in the DBTR building dataset.

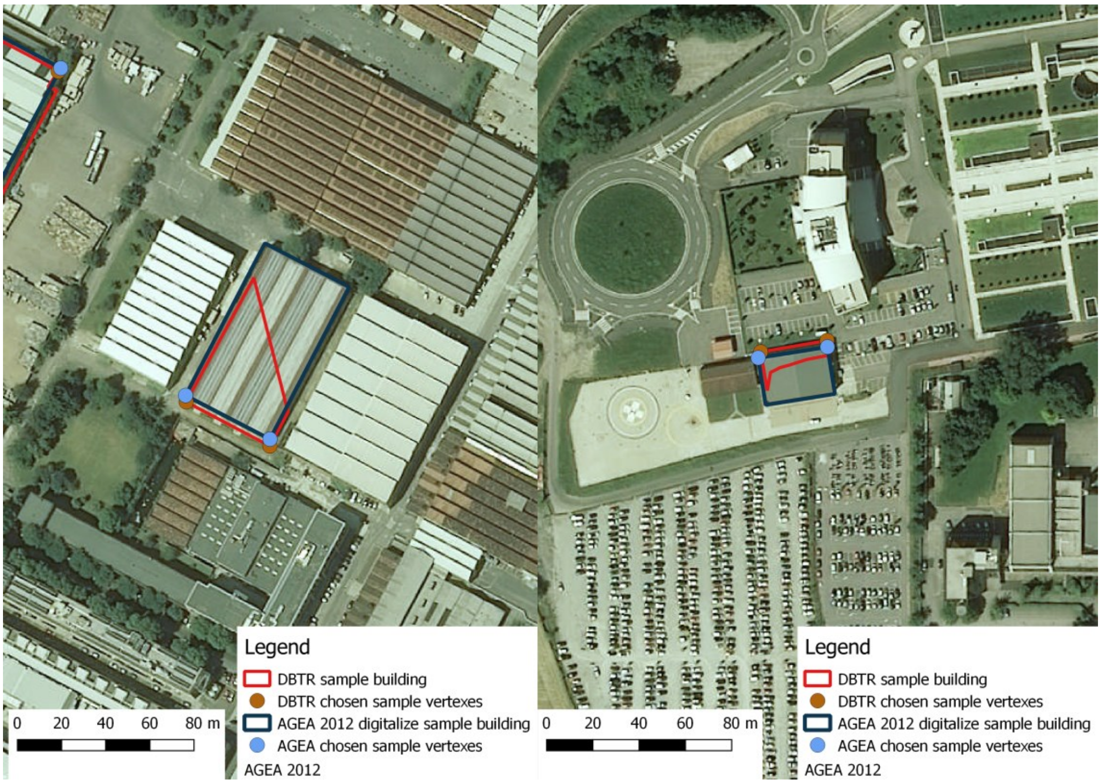

Legend
DBTR sample building
DBTR chosen sample vertexes
AGEA 2012 digitalize sample building
AGEA chosen sample vertexes
AGEA 2012
0 20 40 60 80 m
Legend
DBTR sample building
DBTR chosen sample vertexes
AGEA 2012 digitalize sample building
AGEA chosen sample vertexes
AGEA 2012
0 20 40 60 80 m

Figure 3.25 – Examples of DBTR buildings for which less than 3 homologous corners can be found on the AGEA 2012 orthophoto

At the end of the process, a total of 1’450 couples of homologous points are detected and their coordinates are stored.

3.3.5 Digitization error

As stated in Subsection 3.2.6, the process used to evaluate the digitization error is based on the theoretical framework proposed by Spalla and Galletto (2000) and presented in Subsection 3.2.6.

Similarly to the procedure used for the extraction of the sample on the AGEA 2012 orthophoto, a stratified random sampling is again used. Even in this case, it is the best way to sample because it allows a homogenous detection of points in the whole area of interest.

First, the area of Milan Municipality is divided in square cells (using again the MMQGIS plugin with the procedure described in Subsection 3.3.2.1) having a side of 3 km.
37 cells are found (see figure 3.26).

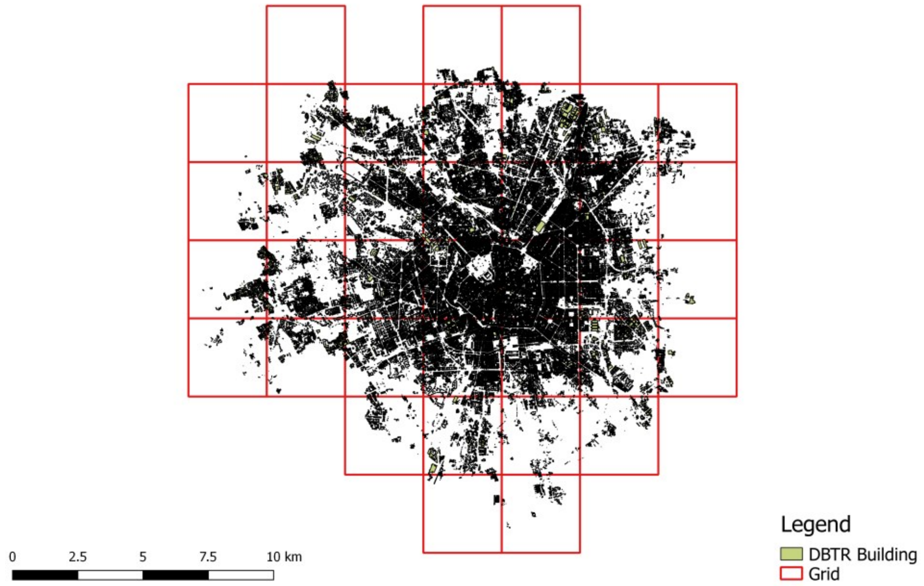


Figure 3.26 – Grid used to evaluate the digitization error superimposed on the DBTR building layer for Milan Municipality

Then, 3 points are randomly extracted in each cell and the closest DBTR building to each of them is selected (see figure 3.27). The procedure is the same as the one explained in Subsection 3.3.2.2.

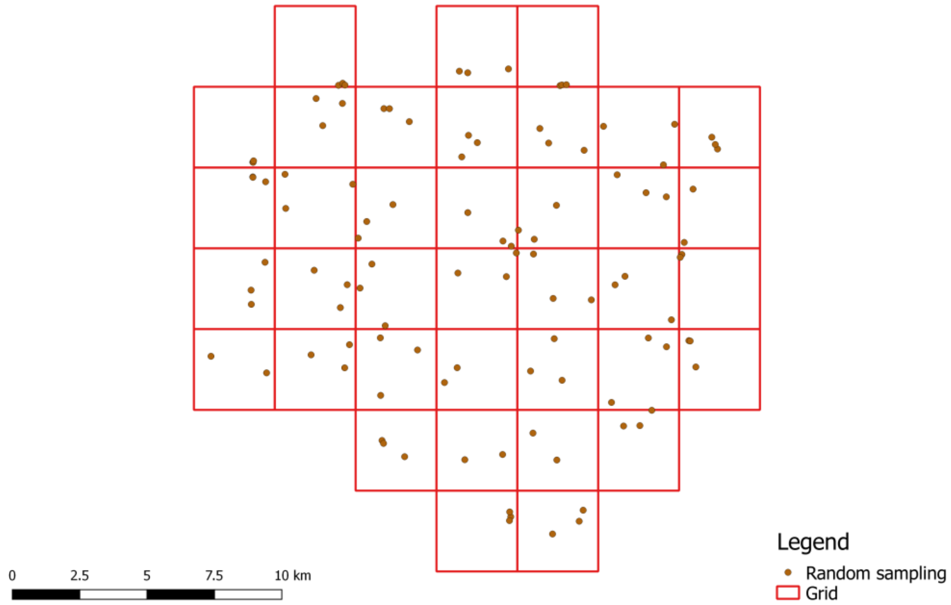


Figure 3.27 – Randomly sampled points in the square grid on Milan Municipality

A total of 105 buildings are selected. Again, in some cells only 2 buildings are found (see figure 3.28).

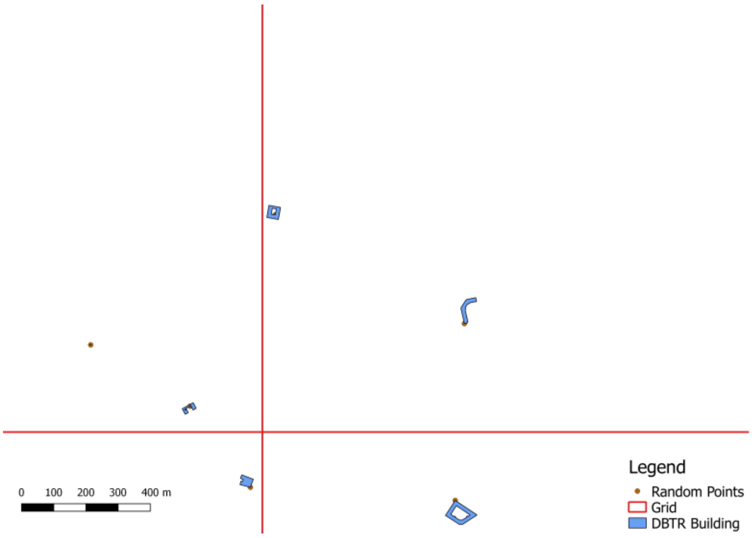


Figure 3.28 – Selected buildings for the evaluation of the digitalization error

After that, one corner is chosen for each of the buildings found. Figure 3.29 shows the final evaluation. The red shape represents the chosen DBTR building and the light blue dot is its corner selected as reference. The corresponding corner on the AGEA 2012 orthophoto are digitized two times (violet and yellow dots).

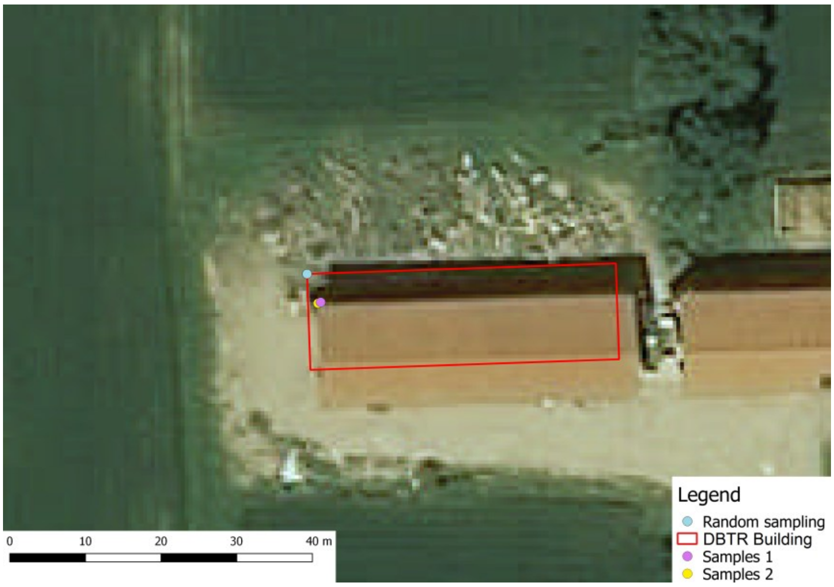


Figure 3.29 – Digitization of the corner (light blue dot) selected as reference in DBTR building (red shape) on the corresponding AGEA 2012 orthophoto corner (yellow and pink dots)

Then the coordinates of these two measures are extracted using WGS84/UTM zone 32 as the CRS. With this data, the true position (X_i, Y_i), the residual from the mean value (s_{ij}) and the mean square error (m_i) are computed. Finally, the digitization mean square error (σ_D) is computed. . This is found to be 0.5 m. As it is one order of magnitude smaller than the declared positional accuracy of the AGEA 2012 orthophoto (3 m), **digitization error is considered to be neglectable and thus it is not included in the final quality assessment.**

3.3.6 Statistical measurements of positional uncertainties

Through calculations made on the coordinates of the homologous points, the planimetric error is computed. The resulting values of the positional accuracy indexes (described in Subsection 3.2.5) are summarized in tables 3.6 and 3.7.

Table 3.6 – Statistics on the planimetric error *e* and its components e_X and e_Y in the *X* and *Y* directions

Index	e_x	e_y	e
μ	1.66 m	3.65 m	**4.32 m**
Me	1.04 m	2.19 m	3.03 m
σ	2.49 m	5.32 m	4.24 m
min	0.00 m	0.00 m	0.04 m
max	13.03 m	29.70 m	29.94 m
n	245	615	758

The average positional accuracy measured in the orthophoto is equal to 4.32 m, which is higher than the one declared by AGEA (3.0 m, see Subsection 3.1.1.2) and also the one suggested by the Italian thematic orthophoto specification (4.0 m, see Subsection 3.1.1.1). Considering the results found on *X* and *Y* axes, it appears that in the latter the positional accuracy is much worse (average of 3.65 m) compared to the *X* direction (average of 1.66 m).

The median of the positional accuracy is barely higher than the value of accuracy declared by AGEA. This means that more than half of the computed errors are larger than 3 m, as it is also demonstrated by the number *n* of positional uncertainties above a given threshold. In fact *n* is equal to 758, which is higher than half of the total number of homologous points (1’450).

Of particular interest is the standard deviation *σ*. On the bi-dimensional plane it measures 4.24 m, but again in the *Y* direction it has a higher value (5.32 m) than in the *X* direction (2.49 m).

The minimum *min* shows that in some cases the corners of the DBTR buildings and the buildings digitized on the 2012 AGEA orthophoto have an exact correspondence. In fact, in both the *X* and *Y* axes the minimum value is 0 m. On the bi-dimensional plane, the minimum value is still very low (0.04 m). Regarding the maximum *max* on the *Y* axis the value (29.70 m) is more than the double of the value computed on the *X* direction (13.03 m). The maximum value of the bi-dimensional error (29.94 m) is very close to the one related to the *Y* axis (29.70 m).

From table 3.6 it is also clear that the displacement of the orthophoto is much worse on the *Y* direction than the *X* direction. The values of *μ*, *Me*, *σ*, *max* and *n* on the *Y* axis are at least the double of those on the *X* axis.

Table 3.7 shows the values of the radius of the different confidence circles suggested by ISO (2013) (see Subsection 3.2.5.7). Compared to the declared accuracies of the orthophoto (see Subsection 3.1.1.2), they are all very high. The radius of the smallest confidence area, i.e. CE39.4, has a value of 4.15 m. It means that if a circle of radius 4.15 m is drawn around the position of the corner of a DBTR building, there is the only a 39.4% probability that its homologous point (i.e. the corresponding corner of the building digitized on the AGEA orthophoto) falls into that area. Looking at CE99.8, it can be concluded that, to be almost sure to find the homologous point of a point digitized on the orthophoto, a distance of 14.53 m must be considered.

Table 3.7 – Measures of the confidence areas for the probability values suggested by ISO

Index	**Value**
CE39.4	4.15 m
CE50	4.89 m
CE90	8.91 m
CE95	10.16 m
CE99.8	14.53 m

3.3.7 Evaluation of the assessment

Once the process flow suggested by ISO (see Subsection 3.1.3.2 and Section 3.2) is completed and gives a planimetric accuracy equal (in average) to 4.32 m, it can be concluded that the AGEA 2012 orthophoto is **not compliant with the accuracy declared in its metadata** (see figure 3.30).

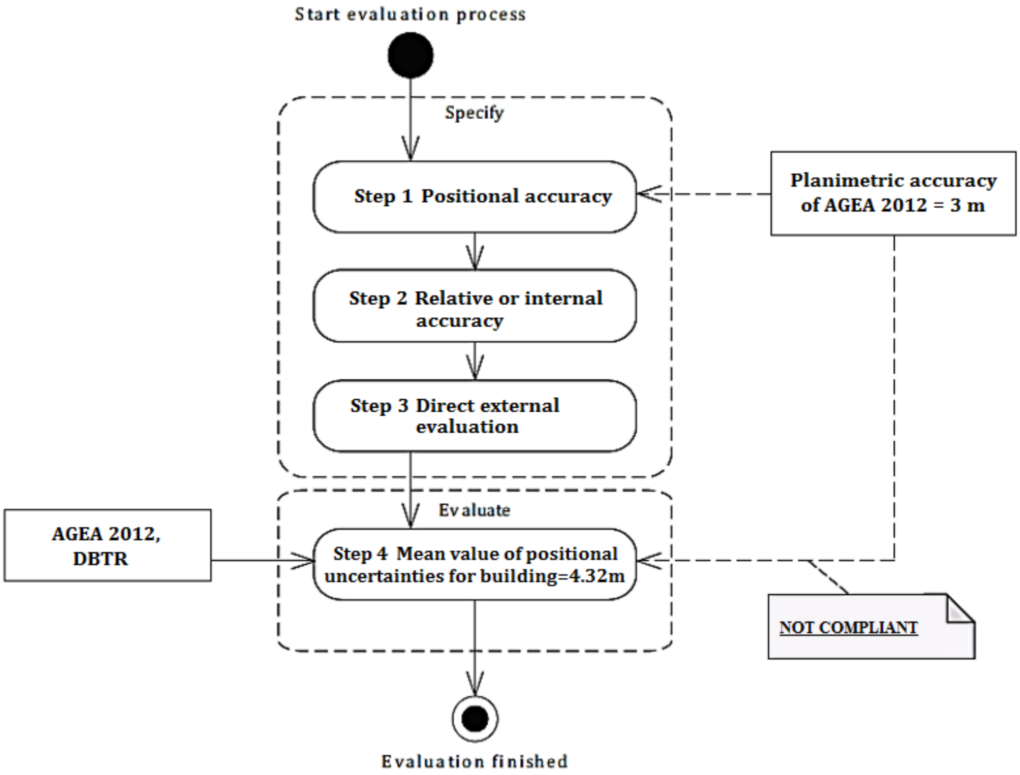

Start evaluation process
Specify
Step 1 Positional accuracy
Step 2 Relative or internal accuracy
Step 3 Direct external evaluation
Planimetric accuracy of AGEA 2012 = 3 m
Evaluate
AGEA 2012, DBTR
Step 4 Mean value of positional uncertainties for building=4.32m
NOT COMPLIANT
Evaluation finished

Figure 3.30 – Summary of the ISO process flow applied to the AGEA 2012 orthophoto (source: ISO 2013)

This result leads to the conclusion that in Milan Municipality the orthophoto has been poorly rectified in the production process. The scarce quality of the product is more and more evident as the height of the buildings increase. Figure 3.31 provides clear examples of the orthophoto distortions, in which the difference in displacement between the *X* and *Y* directions is visible. Readers should consider that, in an ideal (i.e. correctly rectified) orthophoto, the building facades should not be visible.

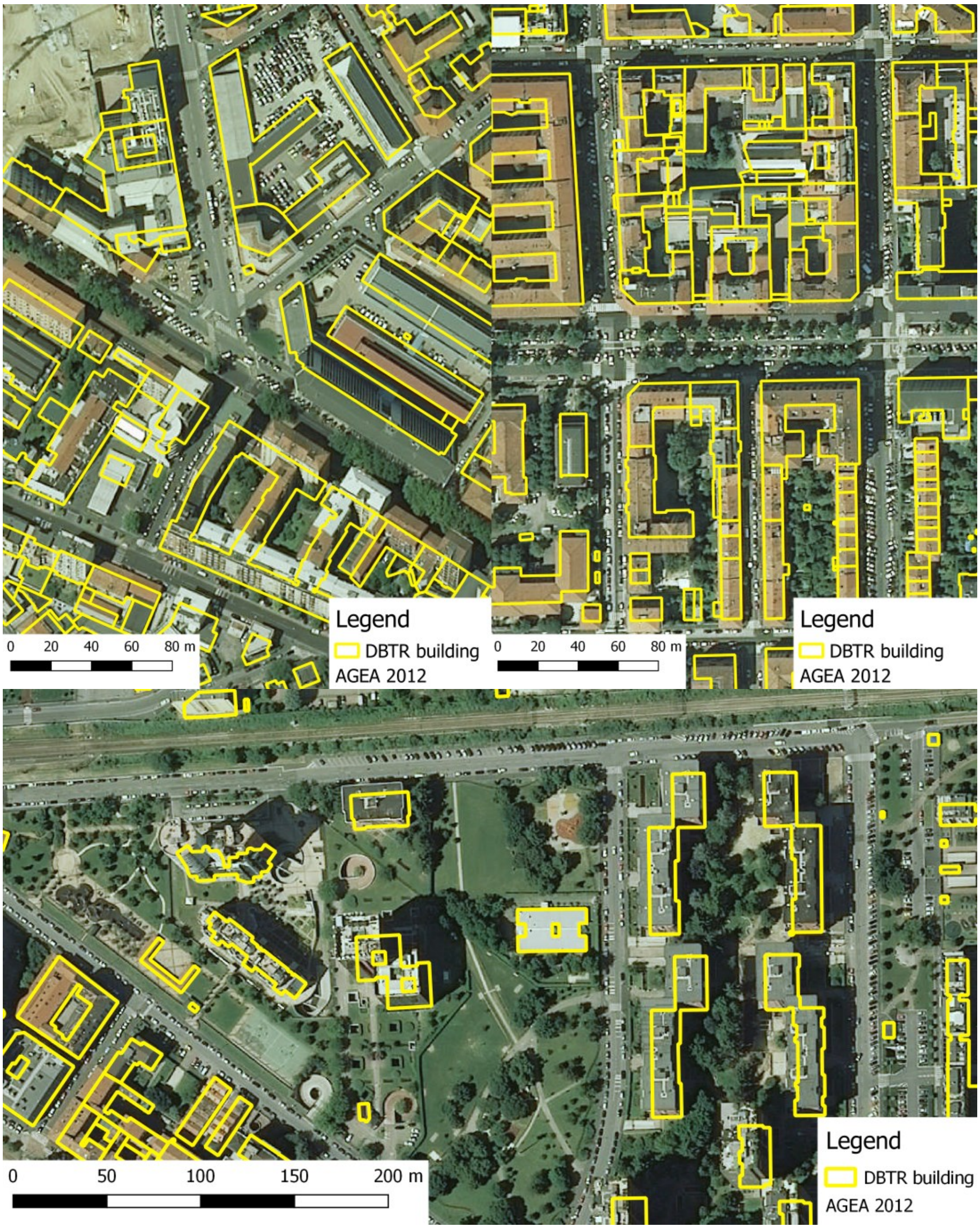


Figure 3.31 – Three example of visual comparisons between the AGEA 2012 orthophoto and DBTR building layer

A possible reason why – despite these results – the AGEA 2012 orthophoto has passed the accuracy compliance tests and thus has been considered suitable for release, is that the accuracy checks are typically performed using a random sampling approach (the same described in this work, see Subsection 3.1.1.1) but at a national level. In fact, the orthophoto is available for the whole Italian territory and the sampled points checked may have not been extracted in the area of Milan Municipality. Therefore, the results of this study are only valid for Milan Municipality, and further tests on other Italian areas are useful to evaluate whether they can be generalized (Brovelli et al. 2016).

3.4 Conclusions

This chapter has proposed a quality assessment to estimate the planimetric accuracy of the AGEA 2012 orthophoto on Milan Municipality, Italy through a comparison against the DBTR building layer. Both the datasets were created or last updated in 2012.

After a brief description of the datasets exploited in the analysis, the chapter has described the theoretical methodology suggested by based on ISO 19157 Geographic information data quality (ISO 2013) to assess the planimetric accuracy of a dataset using a direct external evaluation. This involves the choice of an area-based, stratified random sampling, and the computation of some the indexes useful to evaluate the assessment. Particular focus was put on the digitalization error.

Then, the chapter has illustrated how this methodology is practically applied to the case study of AGEA 2012 orthophoto, from the extraction of the sample used in the comparison up to the computation of the values of the chosen indexes. The chosen DBTR sample contains 502 buildings and 468 of them are detected and digitized on the AGEA 2012 orthophoto. A total of 1'450 homologous points are detected as roof corners of the buildings.

The assessment has shown that the average positional error (4.32 m) is larger than the one declared in the orthophoto metadata (3 m). Thus, based on the analysis on the buildings of Milan Municipality, it can be concluded that the AGEA 2012 orthophoto does not meet the product specifications. Visually, a poor rectification of the orthophoto confirms the results. In fact, it has been noticed that the taller is the building, the larger is the planimetric error.

From the methodological point of view, a small limitation consists in the fact that two variables have been neglected: the digitization error (estimated as 0.5 m) and the positional uncertainty of DBTR (20 cm) which was simply assumed as the ground truth. Both these values were neglected because they are one order of magnitude smaller than the accuracy target (3 m).

This quality assessment has also shown the importance of not taking the data quality for granted, as it might be different from the nominal one declared. In addition, and for the very same reason, the analysis has underlined the need for open (geo)data providers to introduce or refine the mechanisms for data quality control.

CHAPTER 4
LAND USE MAP QUALITY ASSESSMENT

This thesis aims at proposing two quality assessment procedures for different open geodata of Milan Municipality. In Chapter 3, the quality of a web service (orthophoto) was assessed, following a methodology suggested by ISO. This chapter describes another strategy that can be used to assess the quality of a vector dataset. In particular, it assesses the portion of Milan Municipality territory classified as "roads" in a regional land use map (DUSAF).

Similarly to the previous chapter, DBTR layers are used as ground true. Thus, two methods based on its layers will be proposed. Once the best method is detected, the quality of the street of DUSAF dataset for Milan Municipality is assessed.

This Chapter is divided in three parts. The first one regards the explanation of the case study. It describes the characteristics of the involved datasets and it introduces the software needed to process the data.

The second part concerns the methodology. This part supplies all the information needed to reach the result, from the definition of the methods used as references to the assessment procedure. It defines how the results are visualised and, finally, it introduces a spatial index for accuracy evaluation.

After that, results are presented in terms of tables and maps, allowing the determination of the best method and, finally, the detection of DUSAF quality.

4.1 Case study

This Section introduces the datasets involved in the assessment. First, there will be an overview on the analysed land use map and then there will be introduced the DBTR layers. Finally, there will be a summary of the used software.

4.1.1 DUSAF

The land use map named Use of agricultural and forestry soils (*Destinazione d'Uso dei Suoli Agricoli e Forestali*, DUSAF) is the Lombardy Region application of the European program CORINE Land Cover (CLC project), a tool used to analyse and monitor the soil usage (ERSAF 2007). DUSAF is a project started in 1998 and, at the time of writing (February 2017), it contains five regional land use map developed in different years (ERSAF 2016):

- 1998-1999 DUSAF 1.1 - obtained by photo interpretation of 2000 IT flight produced by Blom CGR (*Compagnia Generale Ripreseaeree*);
- 2005-2007 DUSAF 2.0 - obtained by an integration of photos taken from different regional databases. Since the photos have been taken at different times, the epoch of the land use map varies depending on the considered area (e.g. the epoch of Metropolitan city of Milan is 2006);
- 2007 DUSAF 2.1 - obtained in 2007 by photo interpretation of the whole Lombardy region and integrated with information taken by different regional databases.
- 2009 DUSAF 3.0 - available only for the cities of Brescia, Sondrio, Cremona, Milano and Monza e Brianza;
- 2012 DUSAF 4.0 - obtained by photo interpretation of AGEA 2012 orthophoto, available for the entire regional territory.

DUSAF is updated with almost a biennial frequency, through a photo interpretation of aerial photographs provided by AGEA, using of the geometrical structure invariance where no variations are registered (ERSAF 2016).

The DUSAF land use map is structured in 3 principal levels, consistently with the CLC project. Level I includes 5 major categories of coverage (Artificial surface, Agricultural areas, Forest and seminatural areas, Wetlands and Water bodies), progressively detailed into level II and III, named General Ambit. Level III is consecutively divided in other 2 local levels (level IV and V), which represent some specifications of the Lombardy's territory defined by the region itself, named Local Ambit (ERSAF 2016).

4.1.1.1 DUSAF 4.0

DUSAF 4.0 is a land use map of the Lombardy region updated in 2012, which has an informative scale of 1:10'000 (ERSAF 2016). As it is derived from the AGEA 2012 orthophoto, it maintains the same scale (AGEA 2012). DUSAF 4.0 is available as vector map in the Lombardy Region Geoportal (Geoportale della Regione Lombardia 2016) under an IODL 2.0 license (Italian Open Data License v2.0 2016).

The minimum detailed level of every theme corresponds to 1600 m², i.e 16 mm² on the map. On the other hand, the minimum linear dimension of a polygon is 20 m, corresponding to 2 mm on the map (ERSAF 2016).

Regarding the distribution of the first level of DUSAF 4.0 in the Municipality of Milan, most of the territory falls in the level Artificial surface, as it is possible to see from figure 4.1 and table 4.1, which shows the extension of each categories of coverage classified as level I in DUSAF.

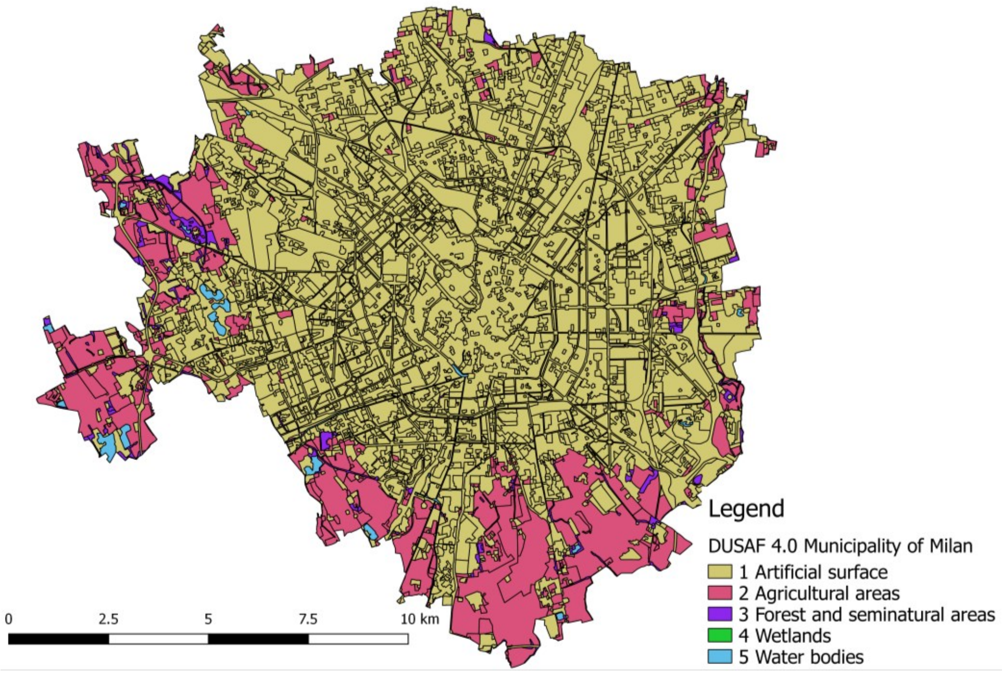


Figure 4.1 – DUSAF 4.0 level I distribution over the Municipality of Milan territory

Table 4.1 – Area of the level I of DUSAF 4.0 for Milan Municipality

LIVEL 1	Area [km²]	%
1 Artificial surface	143	78.5%
2 Agricultural areas	35	19.0%
3 Forest and seminatural areas	3	1.7%
4 Wetlands	0	0.0%
5 Water bodies	2	0.8%
TOTAL	183	

It is interesting to notice that almost 80% of the territory falls into the Artificial surface level, while almost all the remaining area are Agricultural areas. These zones are mostly distributed in the city outskirts, where some fields are still present.

Considering the distribution of the level I categories of Milan Municipality, in this study we decided to focus on the Artificial surface level and then to chose a lower level category on which perform the assessment. Table 4.2 shows all the levels in which Artificial surface is divided, indicating their extensions in percentage (respect to the whole Artificial surface level) and in square kilometres.

Table 4.2 – Artificial surface sub-levels extension with their relative percentages (respect to the extension of the Artificial surface level) and areas (in km^2)

GENERAL AMBIT				LOCAL AMBIT			
LEVEL II		LEVEL III		LEVEL IV		LEVEL V	
11 Artificial surface	29% 54 km^2	111 Continuous urban fabric	19% 34 km^2	1111 Dense urban fabric	18% 33 km^2		
				1112 Continuous urban fabric dense on average	1% 1 km^2		
		112 Discontinuous urban fabric	11% 19 km^2	1121 Scattered or nucleiform urban fabric	9% 17 km^2		
				1123 Sparse urban fabric	1% 2 km^2	11231 Farmhouses	0% 0 km^2
12 Industrial, commercial and transport units	32% 58 km^2	121 Industrial or commercial units	21% 38 km^2	1211 Industrial, artisanal, commercial and agricultural settlements and associated land	11% 21 km^2	12111 Industrial, artisanal, commercial settlements	11% 21 km^2
						12112 Agricultural settlements	0% 0 km^2
				1212 Public or private large implants of services	10% 17 km^2	12121 Hospital settlements	1% 2 km^2
						12122 Public or private implants of services	7% 12 km^2
						12123 Technological implants	0% 1 km^2
						12124 Cemeteries	1% 2 km^2
						12125 Military areas	1% 1 km^2
		122 Road and rail network and associated land	11% 19 km^2	1221 Road network and associated land	8% 14 km^2		
				1222 Rail network and associated land	3% 6 km^2		
		123 Port areas	0% 0 km^2				
		124 Airports	0% 0 km^2				
13 Mine dump and construction sites	3% 6 km^2	131 Mineral extraction sites	0% 0 km^2				
		132 Dump sites	0% 0 km^2				
		133 Construction sites	2% 3 km^2				

		134 Degraded areas not used or not vegetated	1% 3 km^2		
14 Artificial, non-agricultural vegetated areas	14% 25 km^2	141 Green urban areas	10% 19 km^2	1411 Parks and gardens	9% 16 km^2
				1412 Uncultivated green areas	1% 3 km^2
		142 Sport and leisure facilities	4% 7 km^2	1421 Sport centres	4% 7 km^2
				1422 Campground and touristic and receptive structures	0% 0 km^2
				1423 Entertainment centers	0% 0 km^2
				1424 Archaeological areas	0% 0 km^2

In this case study, the class named Road and rail network and associated land is the chosen level used in the evaluation. It is the third biggest level III category in Artificial surface, after the level Continuous urban fabric and Industrial or commercial units.

CLC describes the class Road and rail network and associated land as following (CORINE Land Cover 2000):

> “Motorways, railways, including associated installations (stations, platforms, embankments). Minimum width for inclusion: 100 m.”

While in the DUSAF documentation it has been slightly modified and it is defined as following (ERSAF 2007):

> “In this class are included areas related to road and rail network represented in the CTR including their associated lands (i.e. service areas, stations, parking lots, scarps, green shells). The minimum width for inclusion is 20 m.”

where CTR is the acronym of Technical Regional Map (*Carta Tecnica Regionale*, CTR), which is the basic cartography at medium scale of the Lombardy Region. It is interesting to notice that the minimum dimension is much smaller in the DUSAF map (20 m) (ERSAF 2007) than the parameter defined by CLC (100 m) (CORINE Land Cover 2000).

In the DUSAF map, the level Road and rail network and associated land is divided in two level IV categories: Road network and associated land (8% of Artificial surface level, corresponding to 14 km^2) and Rail network and associated land (3% of Artificial surface level, corresponding to 6 km^2). Figure 4.2 shows these levels: Rail network and associated land is displayed in red and Road network and associated land is represented in yellow. The light blue polygons shows other DUSAF's levels.

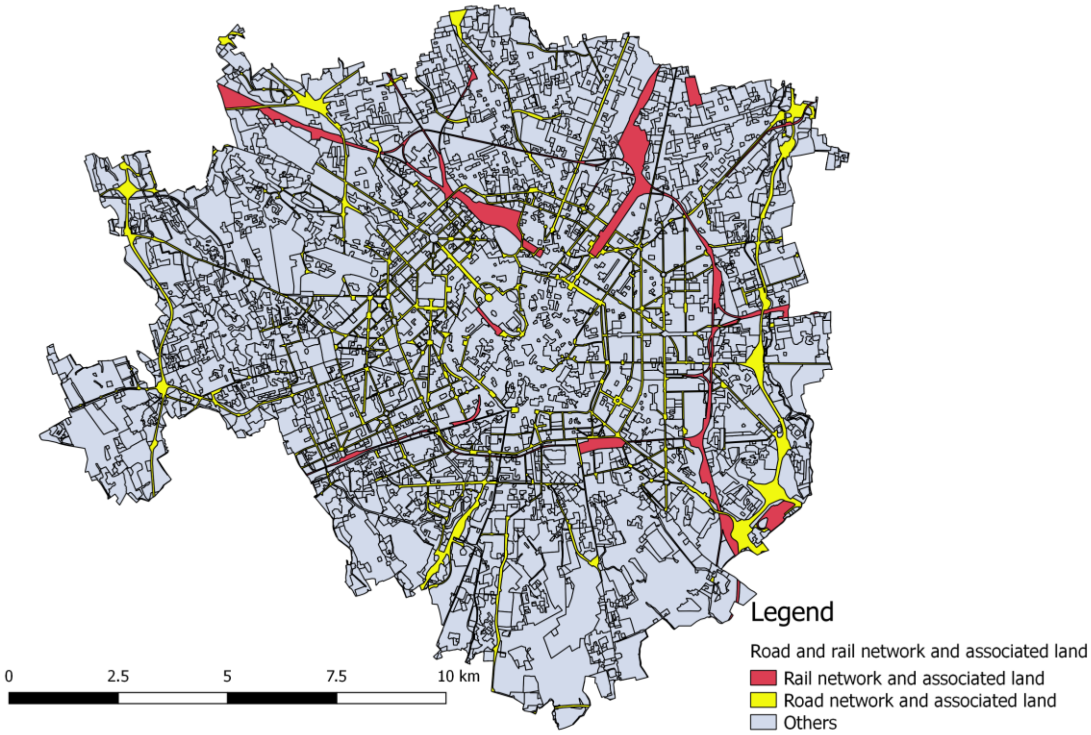

0
2.5
5
7.5
10 km
Legend
Road and rail network and associated land
🟥 Rail network and associated land
🟨 Road network and associated land
🟦 Others

Figure 4.2 – DUSAF levels Road network and associated land and Rail network and associated land in the Municipality of Milan territory

In this case study, we decided to perform an assessment on the level Road network and associated land, as it has an higher amount of surface on Milan Municipality territory respect to Rail network and associated land layer.

4.1.2 DBTR

In this case study, different layers of DBTR will be used to perform the assessment. Among all of them (already listed in Subsection 3.1.2), this evaluation will include the following (Italian Decree of November 10, 2011):

- Layer 01: Viability, mobility and transport;
  - Theme 0101: Streets.
- Layer 06: Vegetation;
  - Theme 0604: Urban green areas.
- Layer 10: Relevance areas;
  - Theme 1001: Transport services.

4.1.2.1 Theme 0101: Streets

Theme 0101 Street is part of Layer 01 Viability, mobility and transport, which describes all the information regarding transport, mobility and their elements, including even data on railways and other transportations (Italian Decree of November 10, 2011). As this assessment involves the evaluation of streets, particular attention is taken on this Theme.

The DBTR defines street as (Italian Decree of November 10, 2011):

> "the public area intended for pedestrian, vehicular and animal circulations"

It includes both the principal and the secondary mobility. The Geodetic Commission for cartography defined the distinction at 1:10'000 scale between principal and the secondary mobility on the basis of the road width (Italian Decree of November 10, 2011). If the road is larger than 7 m (two or more carriageways) it is defined as principal mobility, otherwise if it is smaller (one carriageway) it is denominate as secondary. Following this classification, all the highways and the ordinary streets are considered principal mobility (Italian Decree of November 10, 2011).

Figure 4.3 shows an example of how the street elements are simplified in the different Classes.

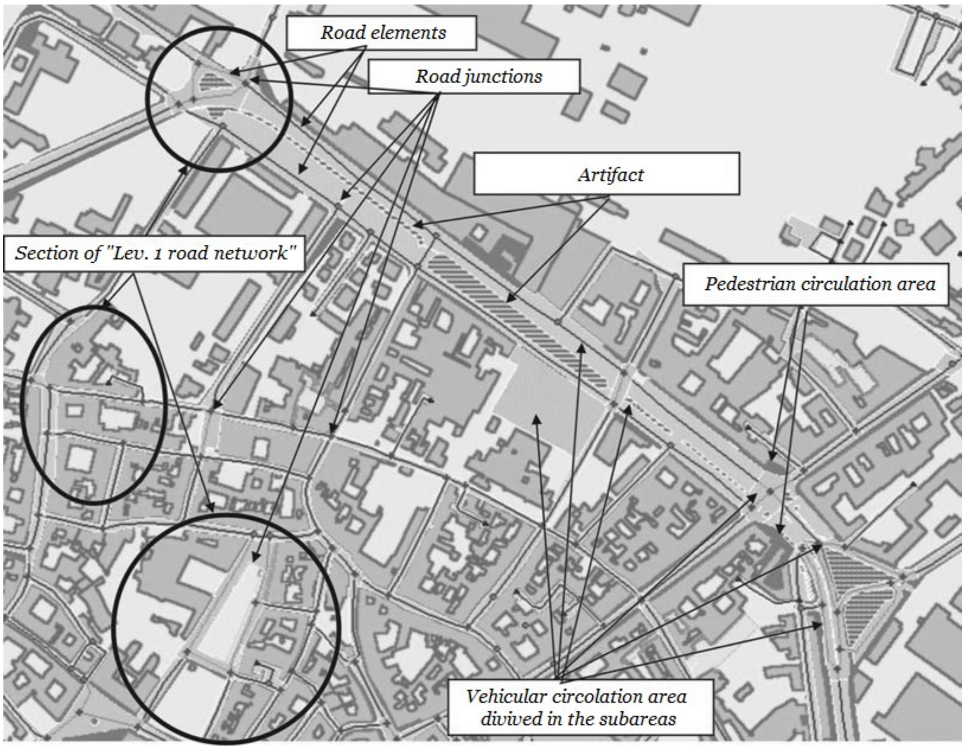

Road elements
Road junctions
Artifact
Section of "Lev. 1 road network"
Pedestrian circulation area
Vehicular circolation area divived in the subareas

Figure 4.3 – Classification of some road elements in the DBTR Theme 0101: Streets (source: Topographic Database Technical Specifications)

Theme 0101 Street is divided in 17 Classes and table 4.3 lists them all. The table includes: the name of all the Classes of the Theme 0101, if they contain information on Milan Municipality, the DBTR identification codes and the type of vector data that they include (i.e. point, line or polygon). As it is possible to see from the column "Available for Milan Municipality", 6 Classes are not available for the case study.

Table 4.3 – Classes of DBTR Theme 0101 Streets

Classes	Available for Milan Municipality	Code	Type
010101 - Vehicular circulation area	Yes	AC_VEI - 010101	Polygon
010102 - Pedestrian circulation area	Yes	AC_PED - 010102	Polygon
010103 - Cycling circulation area	Yes	AC_CIC - 010103	Polygon
010104 - Road area	Yes	AR_STR - 010104	Polygon
010105 - Secondary mixed circulation	Yes	AR_VMS - 010105	Polygon
010107 - Road element	No	EL_STR - 010107	Polygon
010108 - Road junction	Yes	GZ_STR – 010108	Point
010109 - Road segment	Yes	TR_STR - 010109	Line
010110 - Road intersection	No	IZ_STR – 010110	Polygon
010112 - Cycling element	Yes	EL_CIC - 010112	Line
010113 - Cycling junction	Yes	GZ_CIC – 010113	Point
010114 - Lev.1 road network	No	RT_ST1 - 010114	Polygon
010115 - Lev.2 road network	No	RT_ST2 - 010115	Polygon
010116 - Element of secondary mixed circulation	Yes	EL_VMS – 010116	Line
010117 - Junction of secondary mixed circulation	Yes	GZ_VMS - 010117	Point
010118 - Network of the secondary mixed circulation	No	RT_VMS - 010118	Polygon
010119 - Cycling network	No	RT_CIC - 010119	Polygon

In this assessment the procedure includes the Classes Road area, Road segment, Pedestrian circulation area and Cycling circulation area. They are described in the following section.

The Class **Road area** defines the area comprehended into the road borders. It is the portion of the plane formed by the carriageways, the sidewalks, the verges and the tracks. It includes different viability types, either pedestrian or vehicular or others.

The definition of Road area is derived from the Italian Rules of the Road (*Codice della Strada*), article 59 of the L.D. 82/2005 (Infrastructure and Transport Ministry, 1992), which affirms that (Italian Decree of November 10, 2011):

> "The road area includes the carriageway, the sidewalks, the verges and the tracks. The territory not included in these zones is not considered street (e.g. grassed edges, drains, open space contiguous to the street)".

Figure 4.4 shows an example of definition of the Class Road area.

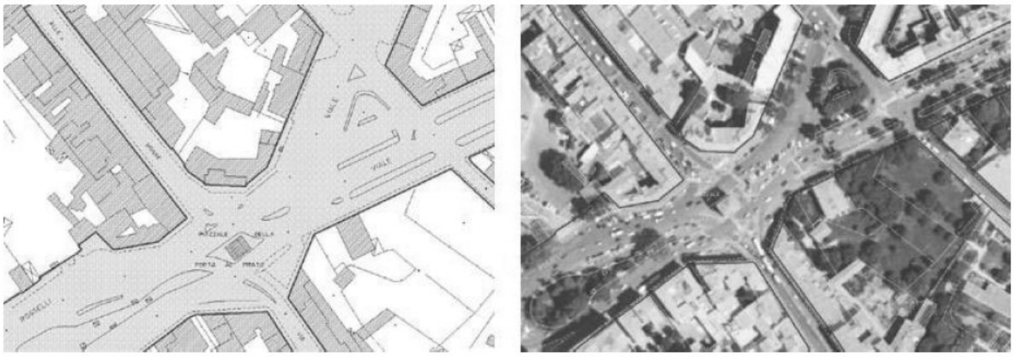


Figure 4.4 – Class Road area identification (left) respect to the equivalent area shown in an aerial image (right) (source: Topographic Database Technical Specifications)

Another important Class that will be used in this assessment is **Road segment**. This Class is obtained by merging one or more road elements that connect two intersections, according to the Geographic Data File (GDF) level 2 rules. The DBTR Class Road segment corresponds to the element named Road in the GDF standard, which is part of the 2D summary graph of the streets. The Class Road segment contains lines that represent the centre line of each street, even if more carriageways compose it (Italian Decree of November 10, 2011).

The "**Pedestrian circulation area**" DBTR Class represents the area intended to pedestrian circulation, including all the portions of street designed for that purpose (e.g. sidewalks, passages or parking pedestrian areas as colonnade or underpass, pedestrian crossovers, traffic islands, etc.). Figure 4.5 shows which zones are classified as Pedestrian circulation area.

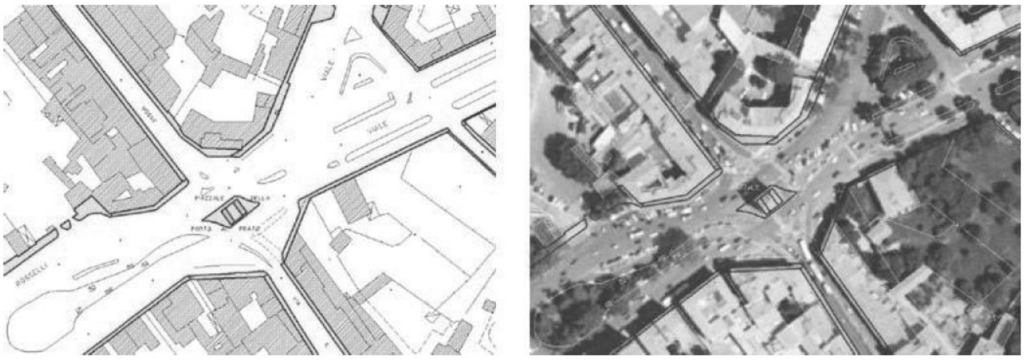


Figure 4.5 – Pedestrian circulation area identification (left) respect to the equivalent area shown in an aerial image (right) (source: Topographic Database Technical Specifications)

The pedestrian area can be located in a proper area or on road area. It has to be delimited by a horizontal signage. In this Class are not included the zone which were successively became pedestrian (e.g. limited traffic of the historical centres) (Italian Decree of November 10, 2011).

Finally, the **Cycling circulation area** Class represents the properly delimited portions of the road designated for the cycling circulation. They can be located as following (Italian Decree of November 10, 2011):

- Own location, with one or more direction of travel, physically detached from the vehicular and pedestrian area with adequate unsurmountable traffic dividers;
- On a reserved lane acquired from the carriageway, with one direction of travel conforms to the vehicular course. The cycle lane has to be located on the right respect to the vehicular one and it can be delimitated with longitudinal stripe or with lane delimiters;
- On a reserved lane acquired from the sidewalk, with one or more direction of travel, in case the width allows their construction. It has to be located near to the vehicular lane.

Figure 4.6 shows an example of the area classified as Class Cycling circulation area.

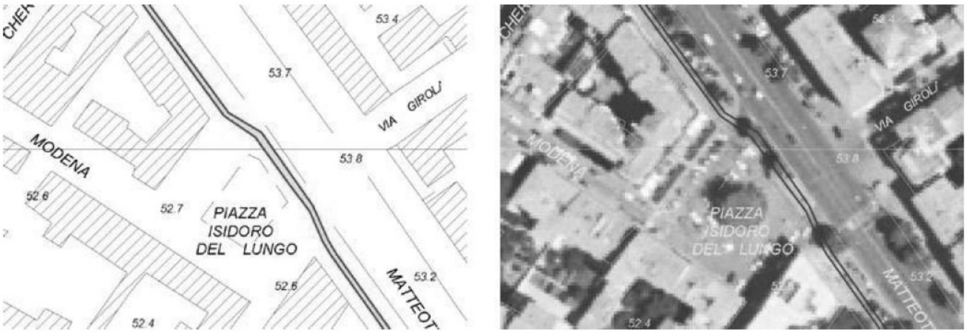


Figure 4.6 – Cycle circulation area identification (left) respect to the equivalent area shown in an aerial image (right) (source: Topographic Database Technical Specifications)

4.1.2.3 Theme 0604: Urban green areas

The theme Urban green is part of the Layer 06: Vegetation, which describes the vegetation areas. Layer 06 is divided in two main Themes: Agroforestry area and Urban green.
Theme Urban green includes information about parks, gardens, botanical gardens or, more generally, every case in which the vegetation has an esthetical purpose different from the agroforestry (Italian Decree of November 10, 2011). This Theme contains 3 classes and they are listed and described in table 4.4. They are all available for the Municipality of Milan.

Table 4.4 – Classes of DBTR Theme 0604 Urban green area

Classes	**Available for Milan Municipality**	**Code**	**Type**
060401 – Green area	Yes	AR_VRD - 060401	Polygon
060402 – Line of trees	Yes	FIL_AL - 060402	Line
060403 – Isolated tree	Yes	ALBERO - 060403	Point

Among all, only the Class **Green area** will be used in the assessment. This Class, in fact, includes information regarding green areas designed for ornamental or recreational purposes (e.g. flowerbeds, gardens, fields, urban tree-lined areas both public than private). Particular attention will be paid to the flowerbeds that, by definition, are described as following (Italian Decree of November 10, 2011):

> “small area of grass and flowers put to ornamental purpose, which is located both on private or street area”.

4.1.2.4 Theme 1001: Transport services

Theme 1001 Transport services is part of Layer 10 Relevance areas, which concerns extended portions of territory that contains objects related to different Layers and Themes. In particular, Theme 1001 Transport services includes the service areas of each transport classes located near their specific mobility tracks. It consists of objects, zones, artefacts and infrastructures of different types (Italian Decree of November 10, 2011). Table 4.5 lists the various Classes that compose Theme 1001 and describes briefly their characteristics.

Table 4.5 – Classes of DBTR Theme 1001 Transport services

Classes	**Available for Milan Municipality**	**Code**	**Type**

100101 – Street service area	Yes	SV_STR - 100101	Polygon
100102 – Railway service area	Yes	SV_FER - 100102	Polygon
100103 – Harbour service area	No	SV_POR - 100103	Polygon
100104 – Airport service area	Yes	SV_AER - 100104	Polygon
100105 – Other transport service area	No	SV_ATR - 100105	Polygon
100181 – Transports service area	No	SV_TRA - 100181	Polygon

Because this assessment regards the roads of Milan Municipality, the most interesting Class among those available is the **Street service area**. It contains the relevant areas of the street services (e.g. gas stations, rest areas, service station) and the paths that connect them to the street circulation (Italian Decree of November 10, 2011). Figure 4.7 shows an example of the contents of this Class.

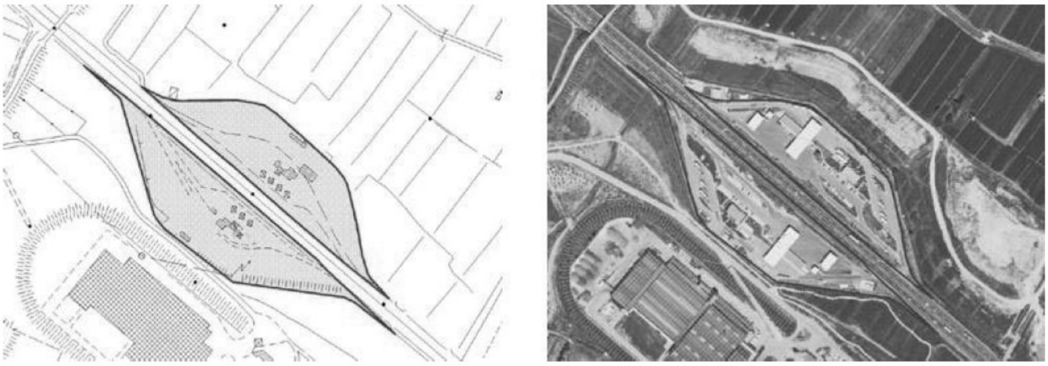


Figure 4.7 – Street service area identification (left) respect to the equivalent area shown in an aerial images (right) (source: Topographic Database Technical Specifications)

4.1.3 Software used in the evaluation

To perform the assessment, different software packages have been used. Beside some simple processing performed with QGIS, most of the procedure is accomplished using GRASS GIS and a Python script. Finally, in order to calculate the value of a specific spatial autocorrelation index (that will be introduce in Subsection 4.2.4) GeoDa and ARCGIS are used.

4.1.3.1 GRASS GIS

GRASS (Geographic Resources Analysis Support System) GIS is a free and open source GIS software suite used for geospatial data management and analysis, image processing, graphics and maps production, spatial modelling, and visualization (GRASS GIS 2017). It is part of the Open

Source Geospatial Foundation (OSGeo, www.osgeo.org). The version of GRASS GIS used in this assessment is the 7.0.3, which is the stable version for Linux at the time of development (September 2016) (available at https://grass.osgeo.org).

4.1.3.2 GeoDa

GeoDa is a free and open source software tool for spatial data analysis, developed by Dr. Luc Anselin and his team of the Centre for Spatial Data Science of the University of Chicago (US) (GeoDa 2017). It is available for the download at https://spatial.uchicago.edu/software. It is designed to facilitate new insights from data analysis by exploring and modelling spatial patterns. The program provides a graphical interface to explore spatial data analysis, such as spatial autocorrelation statistics and basic spatial regression analysis for lattice data (GeoDa 2017). In this assessment, the version of GeoDa was 1.8.12 as it was the last at the time of development (September 2016).

4.1.3.3 ArcGIS

ArcGIS is a GIS program developed by ESRI for working with maps and geographical information. The desktop version of ArcGIS is composed by ArcMap, ArcCatalog, ArcGlobe and ArcScene (ArcGIS 2017), which are components ideated for different purposes. In this case study, ArcMap 10.3.1 was used on Windows.

4.1.3.4 Quantum GIS

QGIS, already introduced in Subsection 3.1.4, is used to manipulate and merge the vector layers, in order to prepare them for the analysis in GRASS GIS. It is be also used to create the grid.

4.2 Methodology

The aim of this study is to compare the DUSAF Road network and associated land level (that from now on will be indicated as DUSAF) with the DBTR. This section will list all the steps needed to achieve the comparison. Another important topic that will be underlined is how the DBTR Classes are chosen in order to best fit the confrontation. In order to fulfil this purpose, two methods will be proposed. Once the best comparison is detected, the quality of the DUSAF layer can be assess.

To explain these concepts, this section is divided in 4 parts. The first one regards the definition of the two methods, the second part concerns the assessment procedure. Then the third part explains how the results will be visualise and, lastly, the forth regards the definition of the spatial autocorrelation and in particular to the Moran’s index.

4.2.1 Definition of the two methods

To perform the assessment, this thesis proposes two ways of comparison. In fact, one of the biggest difficulties in this evaluation is to detect the best DBTR Class (or combination of DBTR Classes) that matches DUSAF.

Nevertheless, before the introduction of the two methods, a clarification must be made. Between all the DBTR Classes introduced in Subsection 4.1.2, the Class Street service area will not be included in the assessment even if, by definition, it contains information related to the streets. As it possible to see from figure 4.8, the polygons included in the Class Street service area are very few and they are located in different positions respect to the area of DUSAF Road network and associated land. For these reasons, the Class Street service area is not considered in the two methods.

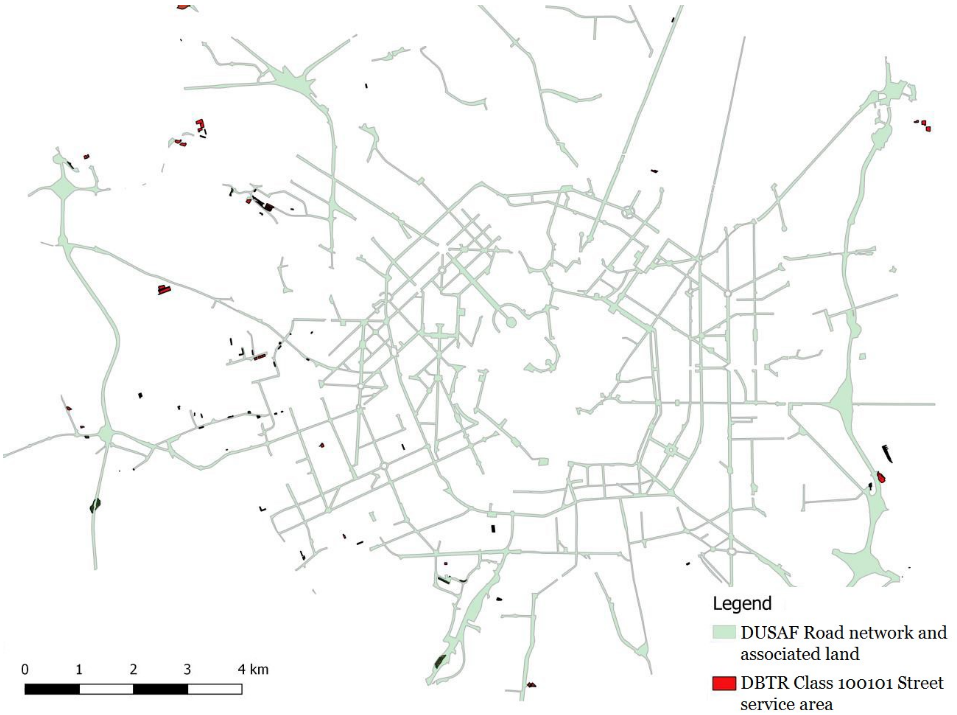

Legend
DUSAF Road network and associated land
DBTR Class 100101 Street service area
0
1
2
3
4 km

Figure 4.8 – DUSAF level Road network and associated land and DBTR Class Street service area

4.2.1.1 Method 1

Method 1 proposes the DBTR Class Road area as comparison for the assessment. In fact, by definition, it includes all the principal and secondary circulation (see Subsection 4.1.2.1). Another important reason for which the DBTR Class Road area is chosen as input layer is that the Lombardy Region indicated it as equivalent to the DUSAF (Tecnical specification 2016).

4.2.1.2 Method 2

Method 2 designs a combination of DBTR layers as comparison to DUSAF. In fact, even if the DBTR Class Road area theoretically includes the pedestrian and the cycling circulations, it has been noticed that in practice there are some lacks and differences. For this reason, we decided to add to the DBTR Road area portions of those layers. Specifically there will be added the following:

- Pedestrian circulation area (Class 010102): only the portions that have the attribute "position" (code AC_PED_POS) equals to "on road area";
- "Cycling circulation area" (Class 010103): only the portions that have the attribute "position" (code AC_CIC_POS) corresponding to "on road area".

To make the assessment more precise, we chose to subtract the portions of DBTR Green area (Class 060401) that have the attribute "type" (code AR_VERD_TY) equals to "flowerbed", because those areas can be located on the road area (see Subsection 4.1.2.3).

Nevertheless, before performing the addition and subtraction operations, a specific table regarding the possible overlaps between the DBTR layers have to be checked. It is included in the document named "Content specifications and physical scheme of the TopographicDatabase" (*Specifiche di contenuto e schema fisico di consegna del Data base topografico*) (Tecnical specification 2016). Considering the Classes involved in the assessment, table 4.6 summarizes the relations between them all.

Table 4.6 – Possibility of overlapping between object belong to different Classes

	Road area	**Pedestrian circulation area**	**Cycling circulation area**	**Green area**
Road area		\	\	\
Pedestrian circulation area	\		* note 4	DJ/TC
Cycling circulation area	\	* note 4		DJ/TC
Green area	\	DJ/TC	DJ/TC	

The DBTR Road area is not included in the table; hence, it does not have any overlapping specification. The cycling and pedestrian circulations can overlap but, as note 4 says, only if the cycling circulation has the position attribute different from "isolated" or if the pedestrian circulation has the position equal to "not on road area". Finally, the layer "Green area" has Disjoint or Touch (DJ/TC) constraint with both cycling and pedestrian circulations. This property imposes that the objects of the two Classes must be totally disjointed.

Hence, the constraints regard the relations between all the layers with the exception of the Road area. For this reason, the operations of addition and subtraction previously listed on the Road area Class can be performed without any restriction. The combination of those Classes will lead to the dataset used in the method 2 as comparison with DUSAF.

4.2.2 Assessment procedure

To perform the comparison, 3 datasets are needed: the DUSAF Road network and associated land level, the chosen DBTR input layer (different in the two suggested methods as explained in Subsections 4.2.1.1 and 4.2.1.2) and the DBTR Class Road segment.
The core of this procedure is to keep only the streets in the DBTR datasets with a width higher than 20 m (minimum dimension needed to be included in the DUSAF polygonal layers, see Subsection 4.1.1.1) and then compare this dataset with the DUSAF one. In fact, the DBTR datasets also contain the secondary circulation roads (see Subsection 4.1.2.1), that are not included in DUSAF. Hence, starting with the input datasets of the two methods, different subtraction operations will be performed in order to obtain some comparable products.

Figure 4.9 shows in green the DUSAF level and in purple the DBTR Class Road segment. It is possible to see that many streets in the DBTR Class are not included in DUSAF.

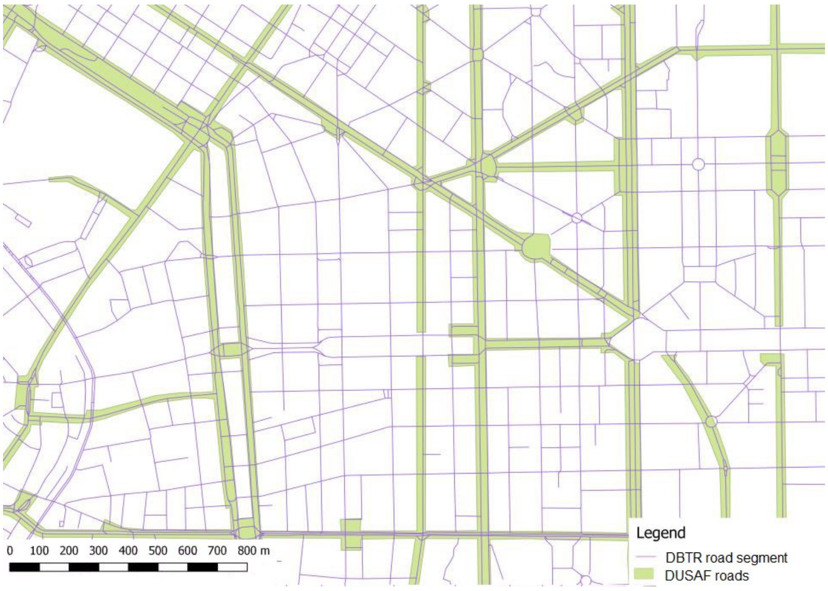


Figure 4.9 – DUSAF and DBTR Road segment

The first operation regards a subtraction between DUSAF and DBTR Road segment, in order to check if the not included streets are really smaller than the minimum dimension (see figure 4.10).

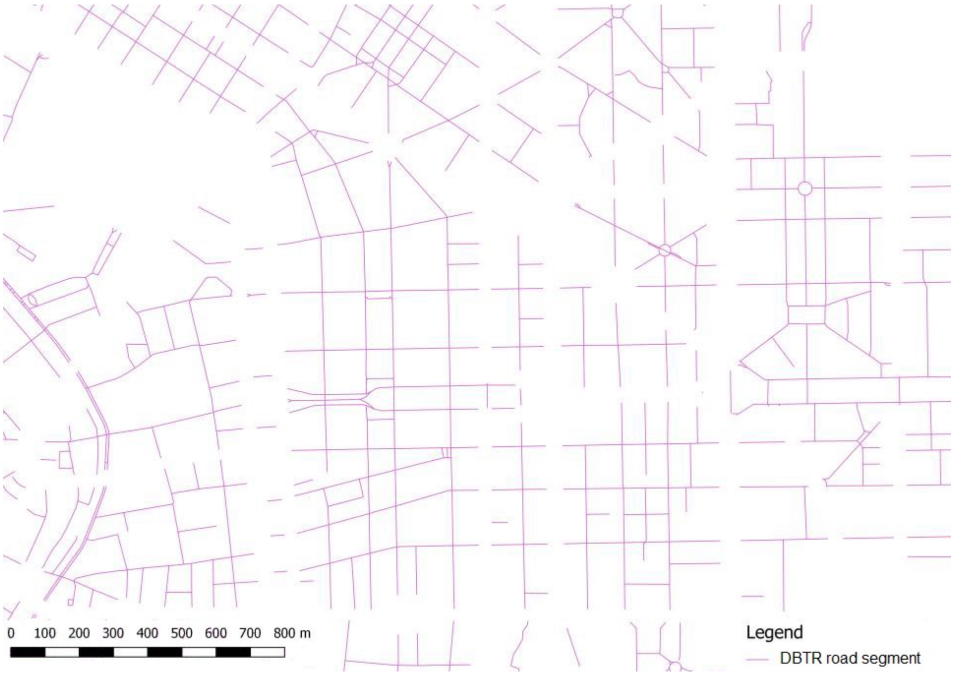


Figure 4.10 – DUSAF and DBTR road segment subtraction

Then, a buffer is performed on road segment. By definition, there are three types of buffers (see figure 4.11):

- No cap (image A in figure 4.11);
- Rounded cap (image B in figure 4.11);
- Square cap (image C in figure 4.11).

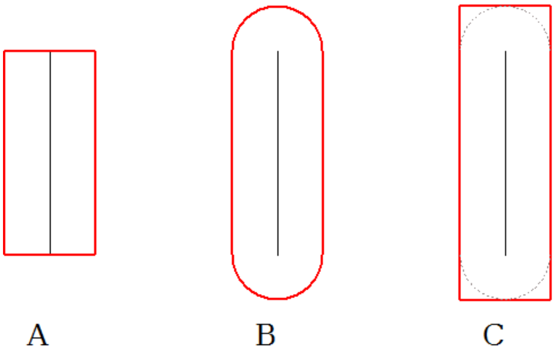


Figure 4.11 – Different buffer types: no cap (A), rounded cap (B) and square cap (C)

In this case study the “no cap” buffer is applied with a dimension of 10 m (i.e. it generates polygons of 20 m width). The result is shown in figure 4.12.

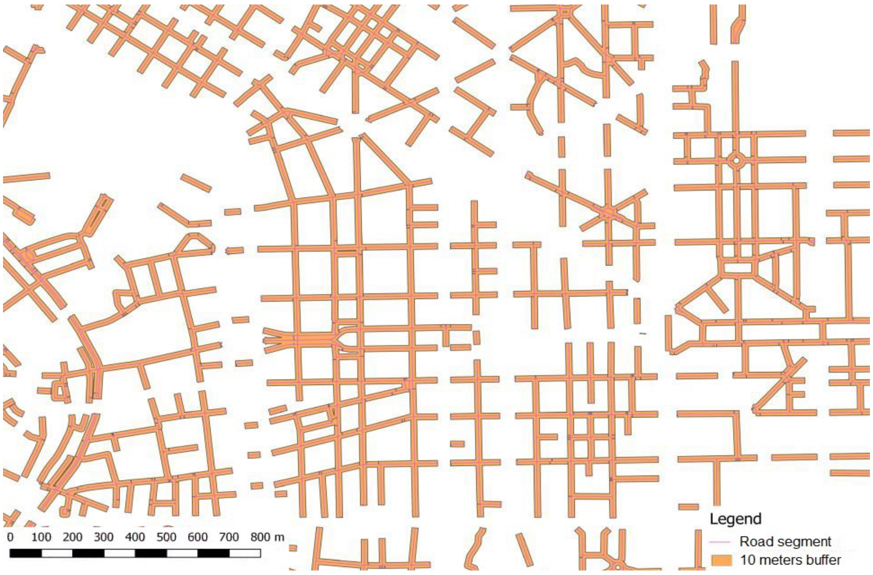


Figure 4.12 – No cap 10 m buffer on DBTR Class Road segment

Once the buffer is performed on the layer, it can be compared with the DBTR input layer (different in the two methods). From figure 4.13, which use method 1 input layer, it is possible to see that in many cases the buffers are visible under the DBTR (light blue polygon).

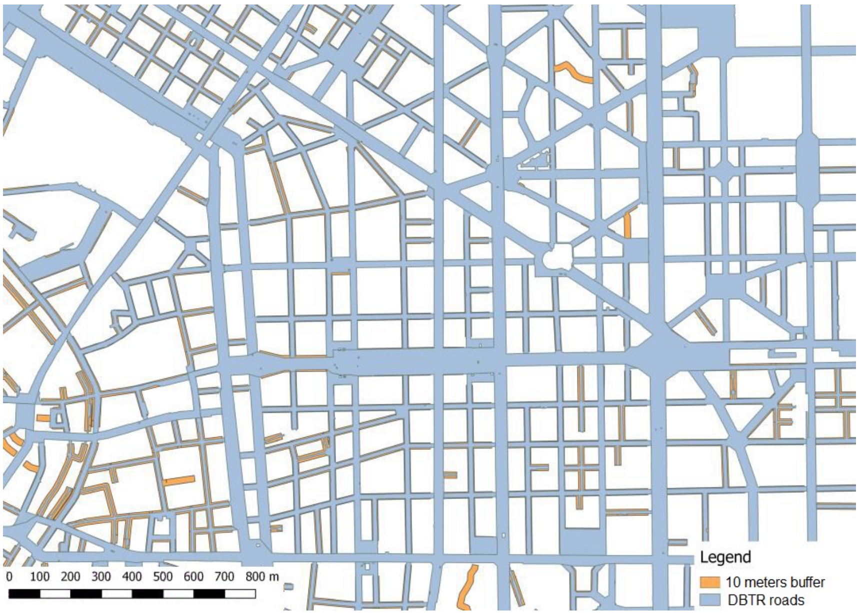


Figure 4.13 – 10 m buffer and DBTR input layer

Then a subtraction between the 10 m buffer and DBTR input layer is performed. In this way, only the roads with width larger than 20 m are kept. Figure 4.14 shows in brown the results of this operation (named in the legend "DBTR – buffer"), which is the layer that can be finally compared with DUSAF.

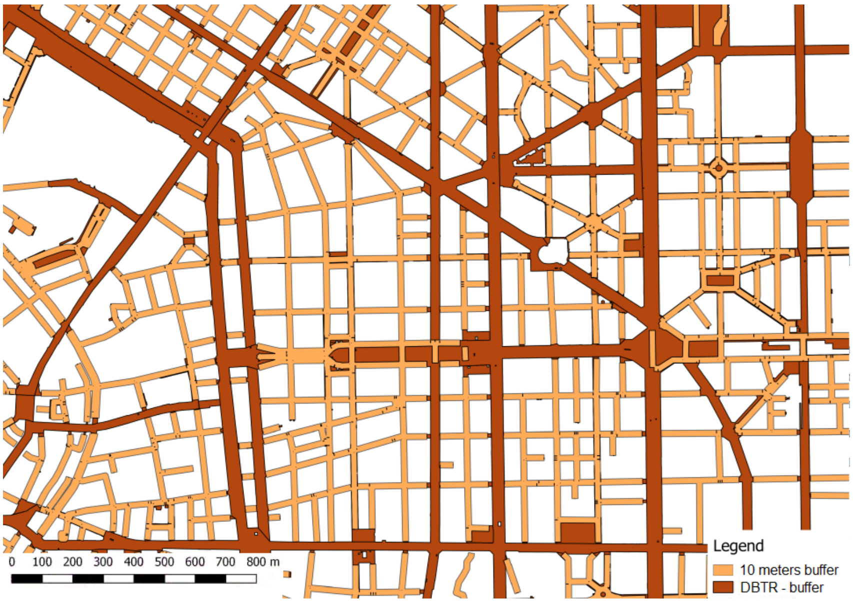


Figure 4.14 – DBTR roads comparable with DUSAF

Figure 4.15 shows the DBTR roads comparable with DUSAF layer (in brown) and the DUSAF one (in green).

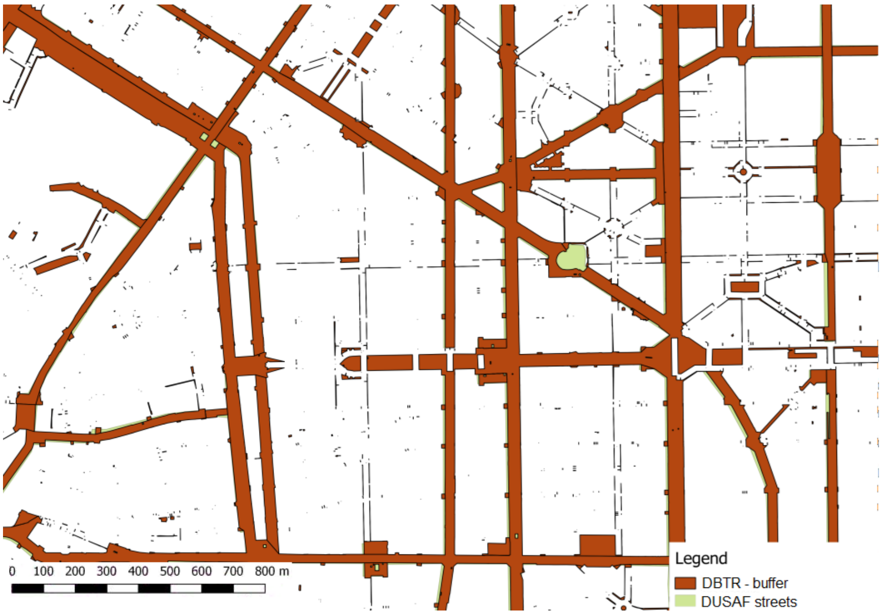


Figure 4.15 – DBTR roads comparable with DUSAF and DUSAF layer

Finally, DUSAF is subtracted from the "DBTR – buffer" layer and the outcome is displayed in purple in figure 4.16. These polygons represent area not included in the DUSAF level that, instead, should be comprehended in it, i.e. the result of this assessment.

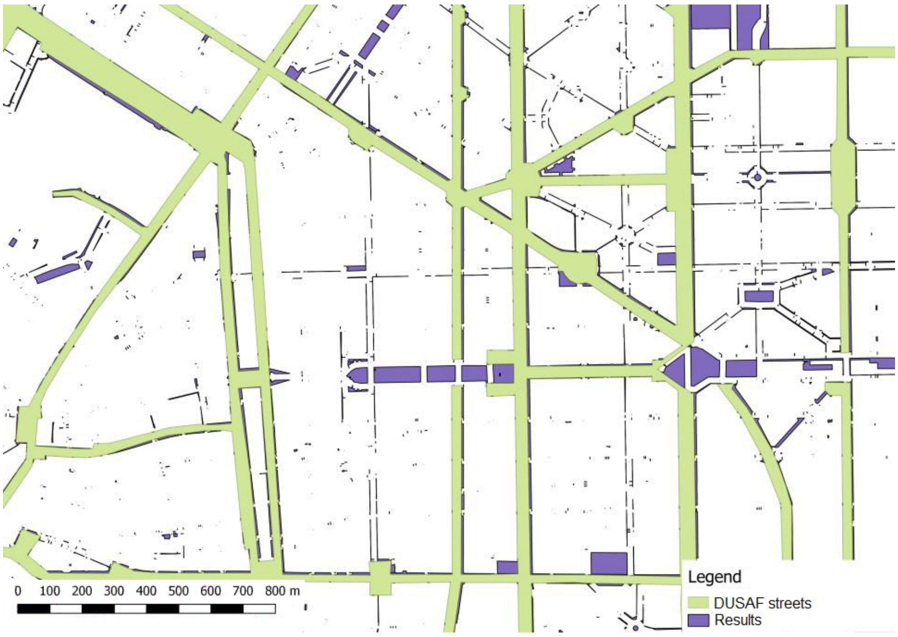


Figure 4.16 – Subtraction between the DBTR input layer and DUSAF

4.2.2.1 Chebyshev's inequality

To increase the probability of the amount of the overlay area between the DUSAF and DBTR, specific buffers will be applied on DUSAF. To determine the dimension of them, a statistical theorem is applied: the Chebyshev's inequality. The Chebyshev's inequality is a fundamental theorem of statistic which affirms that at least 1-1/ε² of data from any dataset must fall within ε standard deviation from the mean (with ε>0) (Čebyšëv 1867).

Let X be a random variable, the Chebyshev's inequality is expressed in equation 4.1 (Čebyšëv 1867).

$$P\,\{|X - \mu_X| \le \varepsilon\sigma_x\} > 1 - \frac{1}{\varepsilon^2} \qquad \text{Equation 4.1}$$

which can also be wrote as equation 4.2.

$$P\{\mu_X - \varepsilon\sigma_x \le X \le \mu_X + \varepsilon\sigma_x\} > 1 - \frac{1}{\varepsilon^2} \qquad \text{Equation 4.2}$$

For example, considering ε=2 the correspondent probability is > 75%. Hence the Chebyshev's inequality defines that at least 75% of the values of the distribution range between two standard deviations from the mean. For ε=3, the probability is > 89% and for ε=4, the probability is > 98.75% (Čebyšëv 1867).

In this case study, the Chebyshev's inequality definition is applied on the distribution of the DUSAF data. In fact, different ε will be applied in order to increase the probability that DUSAF would fall into the selected area. In practice it will be performed through a series of buffers on the DUSAF layer, with the dimensions suggested by the Chebyshev's inequality. Hence, knowing that the standard deviation of the DUSAF is 2 m (ERSAF 2007), if ε=2 the size of the buffer must be ε multiplied by the standard deviation, i.e. 2*2 m = 4 m.

Table 4.7 recaps the ε indexes that are used in this case study with the correspondent buffer dimensions that is performed on the DUSAF layer.

Table 4.7 – Chebyshev's inequality ε indexes, the correspondent probability (%) and buffer (in m)

ε	**Probability [%]**	**Buffer [m]**
1	0	2
2	75	4
3	89	6

As the buffer on the DUSAF affects the result of the procedure listed in Subsection 4.2.2, the process will be operated from the beginning 4 times (no buffer, 2 m buffer, 4 m buffer and 6 m buffer) for each method.

4.2.3 Visualization of the results

To better display the results, they are presented in a hexagonal grid. Then the differences between the compared layers are displayed (see figure 4.17) with the following indexes:

- False negative (FN), area of DBTR not overlapped on DUSAF;
- True positive (TP), area of overlapping between DBTR and DUSAF;
- False positive (FP), area of DUSAF not overlapped on DBTR.

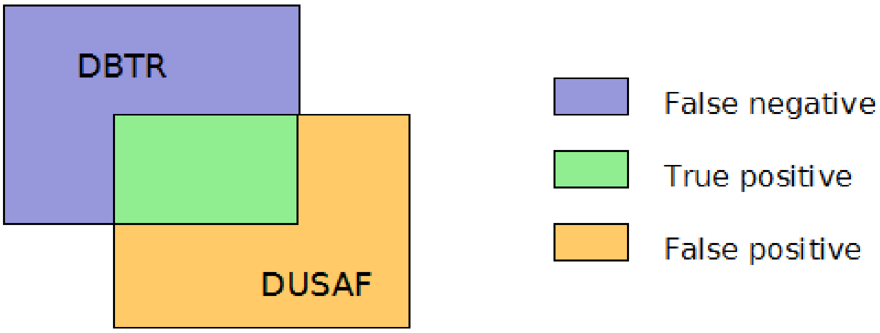


Figure 4.17 – False Negative, True Positive and False Positive considering DBTR and DUSAF area

4.2.4 Spatial autocorrelation

The last part of the methodology is the determination of the spatial autocorrelation between the FN, TP and FP indexes. The autocorrelation is a characteristic of a set of data and it measures the similarity between its elements. Generally, the autocorrelation statistics regards sequenced data, because it gives basic information about their order, which are not available in other descriptive statistics indexes (i.e. mean, variance). Nevertheless, it can be easily applied by a geographical point of view. In fact, the spatial autocorrelation exists when a variable exhibits a regular pattern over the space. It is present, for example, when similar values are cluster together in a map. More generally, the spatial autocorrelation statistic allows us to use statistical methods to measure the dependence among nearby values in a spatial distribution (Odland 1985).

To measure the spatial autocorrelation, different methods could be applied basing on the final aim. For examples, geographers prefer to use Moran's I or Geary's c index. Geologists and remote sensing analysts prefer the semi-variance. On the other hand, spatial econometricians estimate spatial autocorrelation coefficients through regression equations (Getis and Ord 1992).

In this case study, we applied the Moran's I to check if the grids representing the FN, FP and TP contain spatial autocorrelation between the data.

The Moran's I has been developed in "Notes on Continuous Stochastic Phenomena" by Patrick Alfred Pierce Moran (1950) and equation 4.3 shows its formula.

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \qquad \text{Equation 4.3}$$

where,

N is the number of element

w_{ij} is the weight of the j element respect to i element

Once a specific weight method is decided (e.g. influence sphere, inverse distance), the value of I ranges from -1 to 1. Figure 4.18 shows the different meanings of the results. If I = -1 there is negative spatial autocorrelation (random distribution), if I = 0 there is neutral spatial autocorrelation and if I = 1 there is positive spatial autocorrelation (clusters) (Moran 1950).

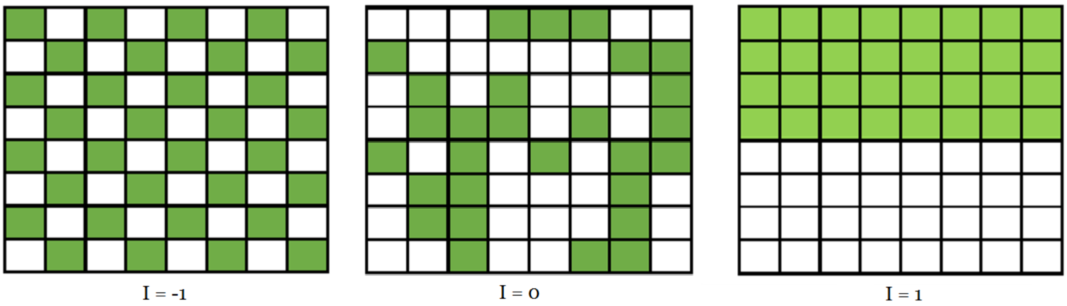


Figure 4.18 – Moran's I value meaning: if I = -1 there is negative spatial autocorrelation (random distribution), if I = 0 there is neutral spatial autocorrelation and if I = 1 there is positive spatial autocorrelation (clusters)

4.3 Results

This section explains operatively how the assessment has been performed, following the guidelines defined in the methodology.

The first part regards the determination of the input files. It explains how to obtain the datasets used as input in the two methods and how to operate the buffers on DUSAF. Then it shows the grid in which the territory is divided and it explains how the procedure is applied.

After, the results of the assessment are presented in different ways. First, the parameters measured on them are shown in a visual way using the grid, then by listing their numerical results in tables. After that, this Subsection proposes an analysis of the cells containing the higher amount of FN area, in order to underline why there are so many differences between the datasets.

Finally, this Subsection determines the best method to be used in the assessment with DUSAF and describes the quality of the DUSAF “Road network and associated land” against it. To conclude, an evaluation of the spatial correlation on the grid of the chosen method is performed.

4.3.1 Determination of the input files

This section introduces the files used as input in the evaluation: the DBTR datasets of the two methods and DUSAF layer with all its buffers.

4.3.1.1 Method 1

Method 1 proposes the DBTR Class Road area as comparison. It is displayed in figure 4.19. It is possible to see that the density of the roads is higher in the Milan city centre and it gradually decreases towards its outskirts.

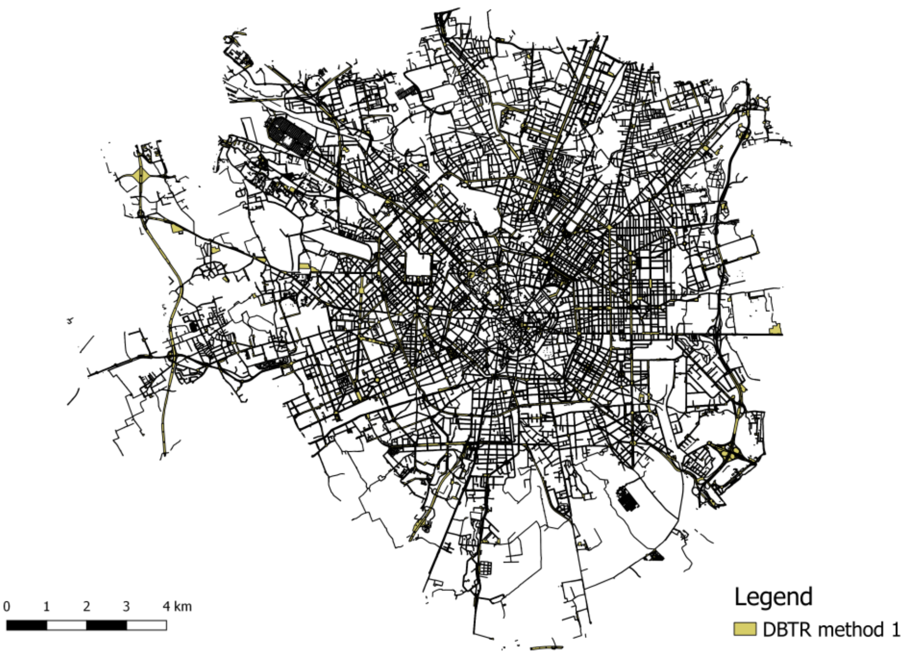


Figure 4.19 – Method 1 DBTR input dataset

4.3.1.2 Method 2

Method 2 proposes a combination of DBTR layers. Starting with the DBTR Class Road area, used as input layer in method 1, the pedestrian and cycling circulation located on road area are added and the flowerbeds (e.g. green areas that can be positioned on a road) are subtracted. The software used in this procedure is QGIS.

To reach this aim, the DBTR Class Pedestrian circulation area has to contain only the element located on road area. Operatively it is done using QGIS function *Select feature using an expression*. Using the expression NOT ("AC_PED_POS" = 'su sede stradale'), all the elements that have positions different to "on road area" ("*su sede stradale*") are selected and then erased. Figure 4.20 shows on the left the DBTR Class Pedestrian circulation area and on the right the portion of it located on road area. It is possible to see that most of the elements fall into this category.

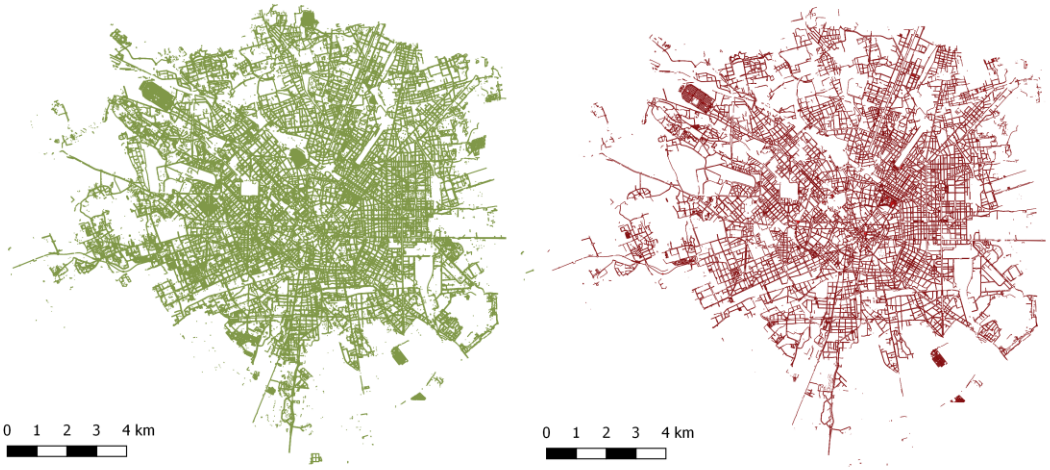


Figure 4.20 – DBTR Pedestrian circulation area (left) and pedestrian circulation area located on road area (left)

The same operation was done on the DBTR Class Cycling circulation area, using as expression NOT ("AC_CIC_POS" = 'su sede stradale'). Figure 4.21 shows on the left the entire DBTR dataset and on the right the portion located on road area. Even in this case most of the elements fall into this category.

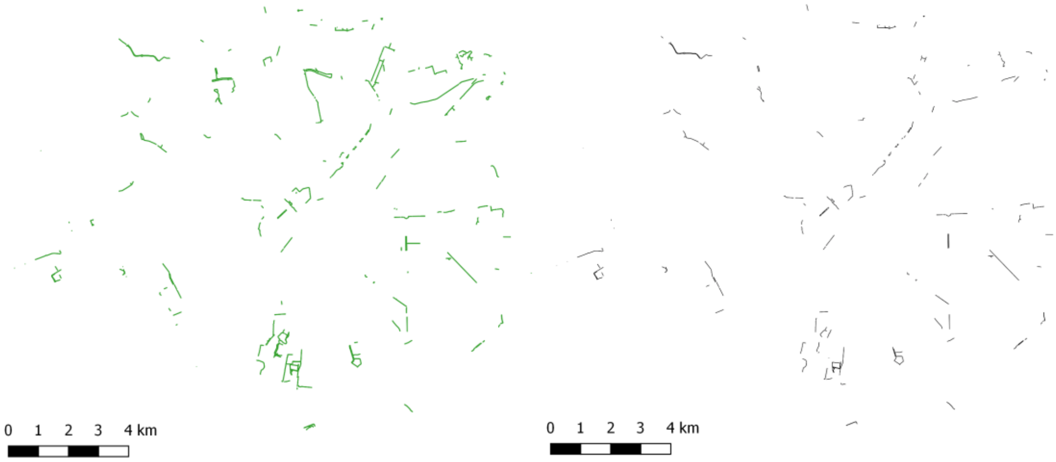


Figure 4.21 – DBTR Cycling circulation area (left) and cycling circulation area located on road area (right)

Finally, the same operation is performed on the DBTR Class Urban green areas. As only the flowerbeds (*aiuole*) need to be selected, the expression used is NOT ("AR_VRD_TY" = 'aiuola'). These elements are a small part of the entire dataset, as it is shown in figure 4.22, whereon the left the Class Urban green areas and on the right the elements classified as flowerbeds are displayed.

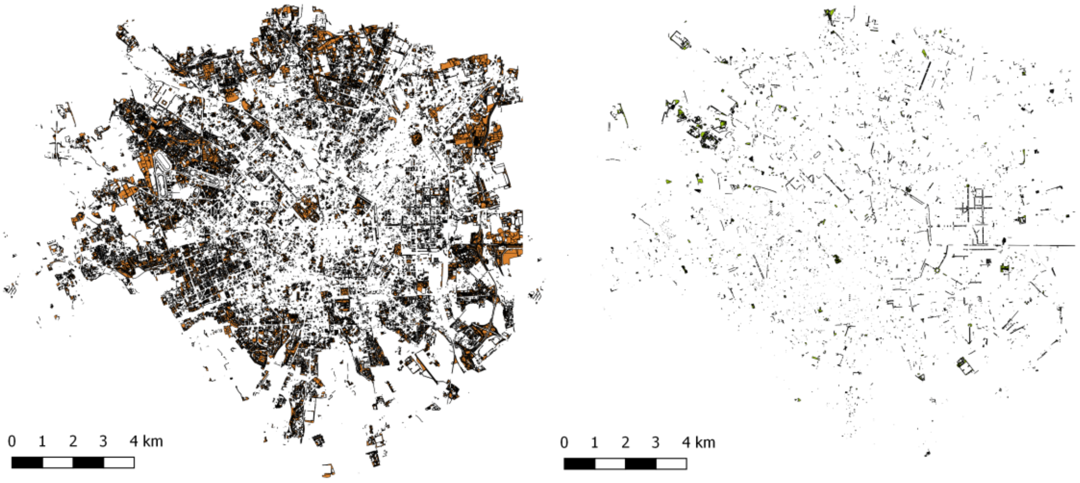


Figure 4.22 – DBTR Class Green area (left) and the area classified as flowerbeds (right)

After that, the used QGIS function is *Vector* > *Geoprocessing Tools* > *Union*, which allows to merge the DBTR Class Road area with the pedestrian circulation area and the cycling circulation area located on the road area.

Finally, the function *Vector* > *Geoprocessing Tools* > *Difference* allows to subtract the flowerbeds. Figure 4.23 shows the final product. It is possible to see that it is denser respect to the dataset used as input in the method 1 (see Subsection 4.3.1.1, figure 4.19).

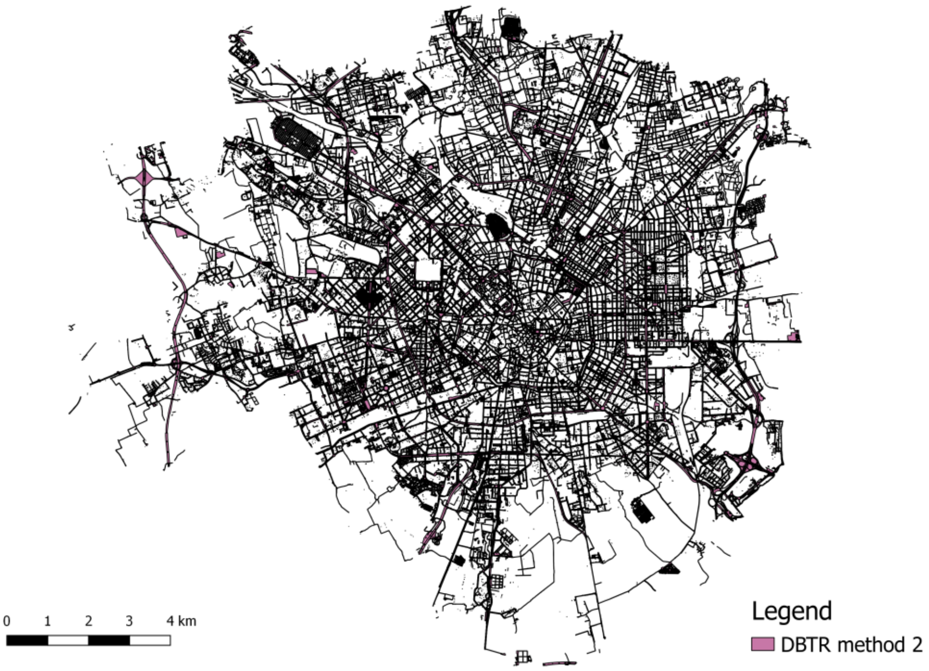


Figure 4.23 – Method 2 DBTR input dataset

4.3.1.3 DUSAF layer

The input products of the two methods are compared with DUSAF Road network and associated land level, shown by figure 4.24.

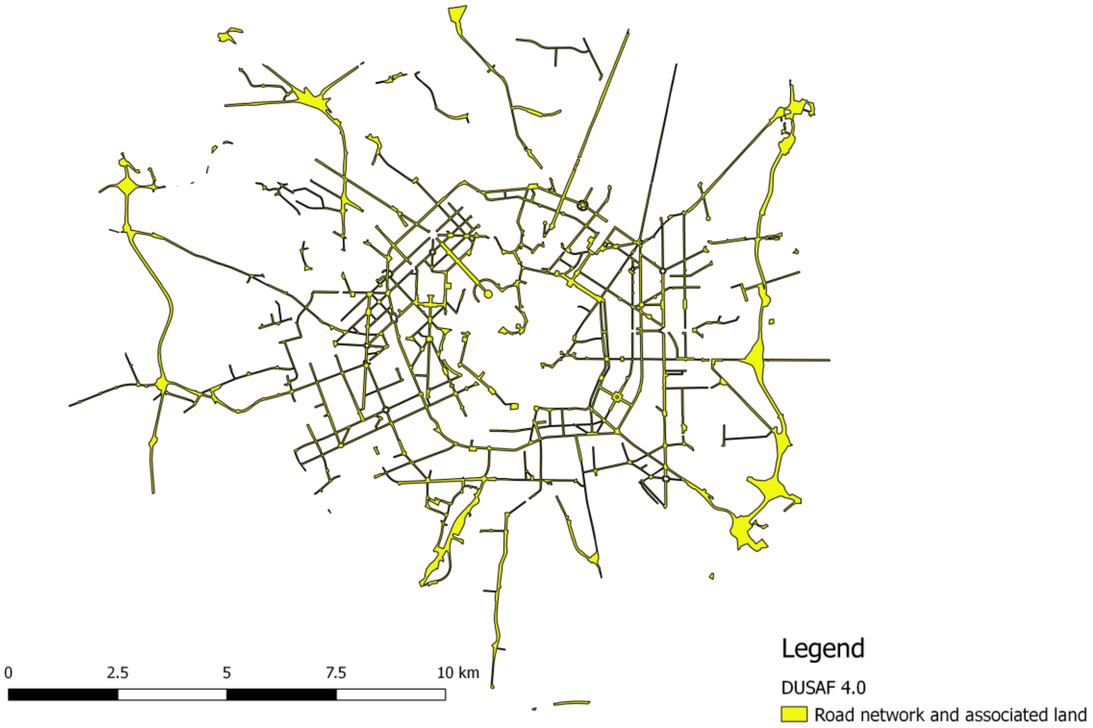


Figure 4.24 – Municipality of Milan DUSAF Road network and associated land level

Three different buffers are performed on it, with a dimension defined by Chebyshev’s inequality (see Subsection 4.2.2.1). They were developed with the QGIS function *Vector* > *Geoprocessing Tools* > *Buffer(s)*. Figure 4.25 shows the result: the DUSAF layer (green-ochre), the 2 m buffer (orange), the 4 m buffer (light blue) and finally the 6 m buffer (in blue).

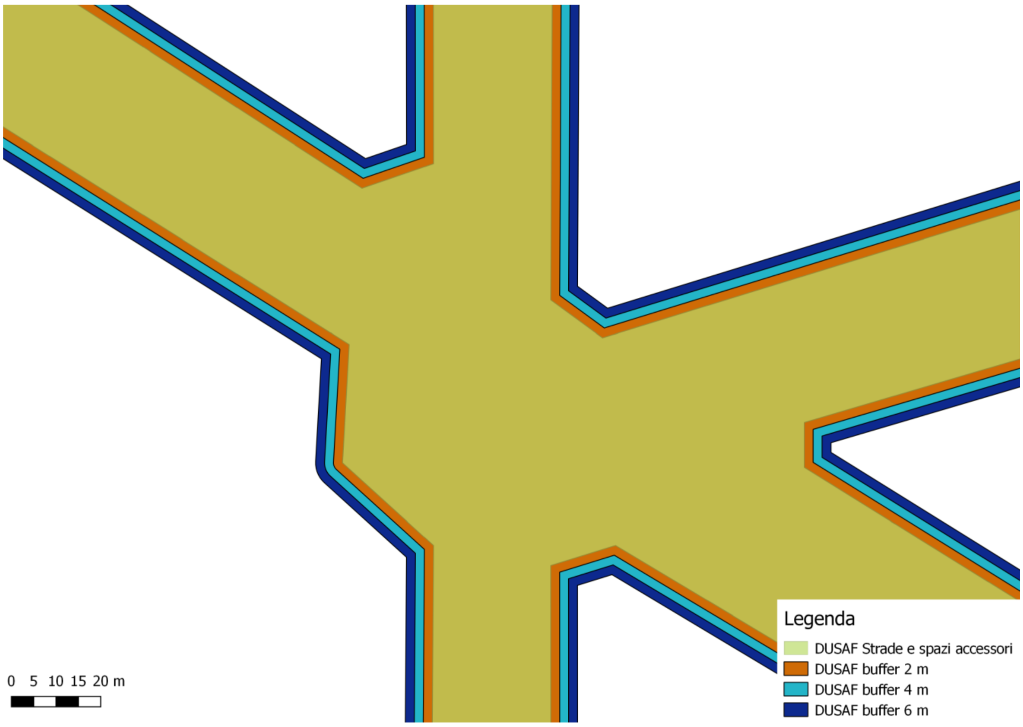


Figure 4.25 – Buffer of 2 m, 4 m, 6 m on DUSAF level

They will be used to check how much the increasing in the probability affects the assessment.

4.3.2 Creation of the grid

With the same procedure described in Subsection 3.3.2.1, the grid is created using the MMQGIS plugin in QGIS. The dimension of the diagonal of each hexagonal cell is 2'000 m, as shown by figure 4.26.

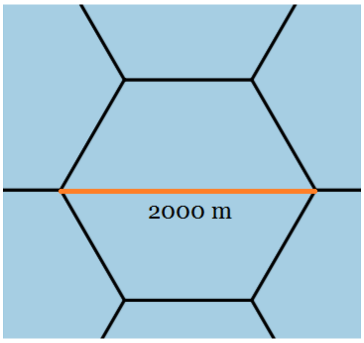


Figure 4.26 – Dimension of a cell of the chosen grid

98 hexagonal cells compose the grid, which covers all the Milan Municipality territory (see figure 4.27). This image also shows a univocal ID code for each cell.

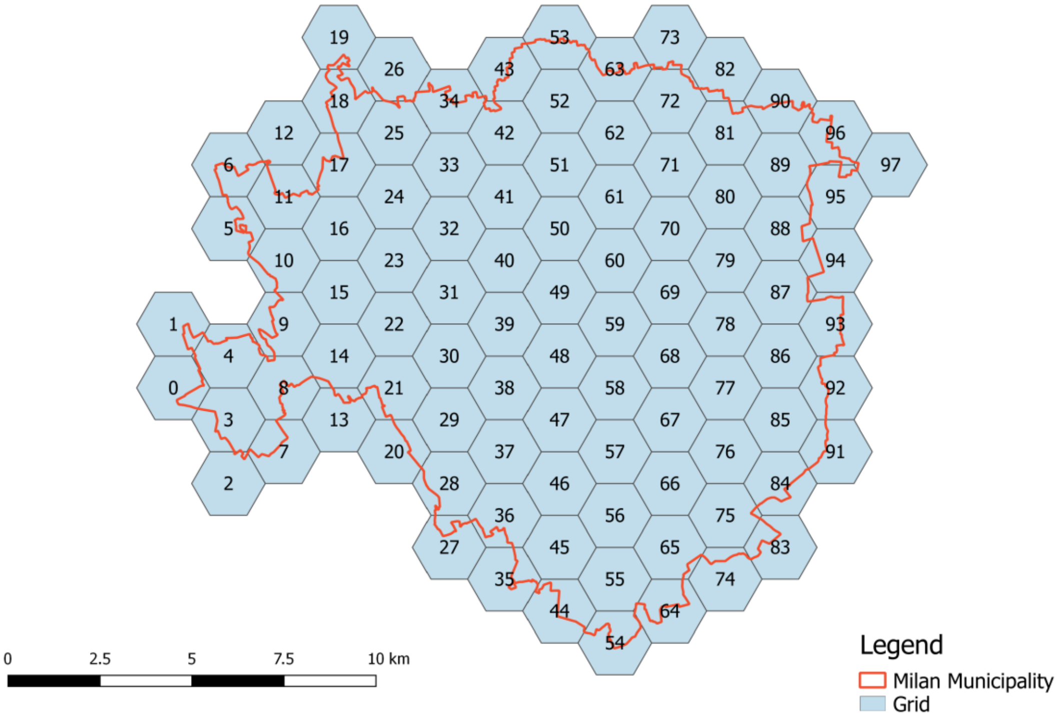


Figure 4.27 – Grid with ID code of each cells

4.3.3 Realization of the assessment procedure

Due to the fact that the assessment is a sequence of subtractions and buffer operations, which need to be performed many times, it was developed by means of a Python 2.7 script implemented in GRASS GIS. In fact, for each method the procedure had to be performed 4 times:

- DUSAF compared to the DBTR input dataset;
- 2 m buffer on DUSAF compared to the DBTR input dataset;
- 4 m buffer on DUSAF compared to the DBTR input dataset;
- 6 m buffer on DUSAF compared to the DBTR input dataset.

for a total of 8 times to reach all the results. The usage of a script allowed us to get faster and easier to the end of the assessment. The script is available in the Annex 1.

4.3.4 Visualization of the results

In this section, the results are displayed as maps. Using the grid described in Subsection 4.3.2, this Subsection shows a classification based on the amount of DBTR and DUSAF area in each cell and then it presents the results of the computed FN, TP, FP indexes for the two methods.

4.3.4.1 Input layers

The two methods approach the assessment with two different input dataset (see Subsections 4.2.1 and 4.2.2). Figure 4.28 displays on the left the input dataset of method 1 and on the right the one used in the method 2. In both the amount of area contained in each cell (in m^2) is classified in 6 classes: darker is the cell, higher is the value of the area. It is easy to see that in both cases the concentration of the area is higher in the Milan city centre. Method 2 cells generally contain higher value of area, as it is possible to see from the amount of cells that fall in the last category (11 in method 2 against 8 in method 1).

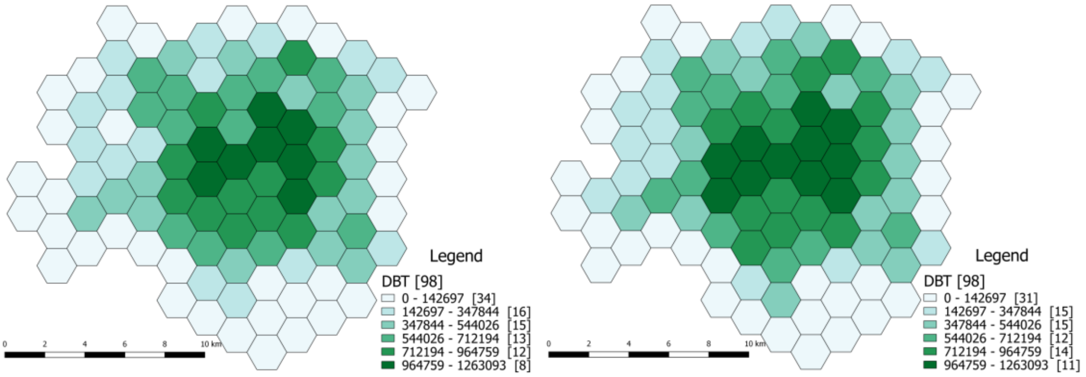


Figure 4.28 – DBTR input layer area density in m^2: method 1 on the left and method 2 on the right

The same categorization of the area is also used in figure 4.29, which shows the distribution of the area of DUSAF level and its buffers. The first thing that is possible to notice is that the values of the area are lower respect to the DBTR layers, in fact the last two categories are not reached in any case. Nevertheless, it is important to remember that in figure 4.28 are still present the secondary circulation roads, which will be erased in the python script. Hence, this image does not represent the map that will be directly compared with DUSAF, but it shows the starting point from which the datasets that fit the confrontation will be obtained.

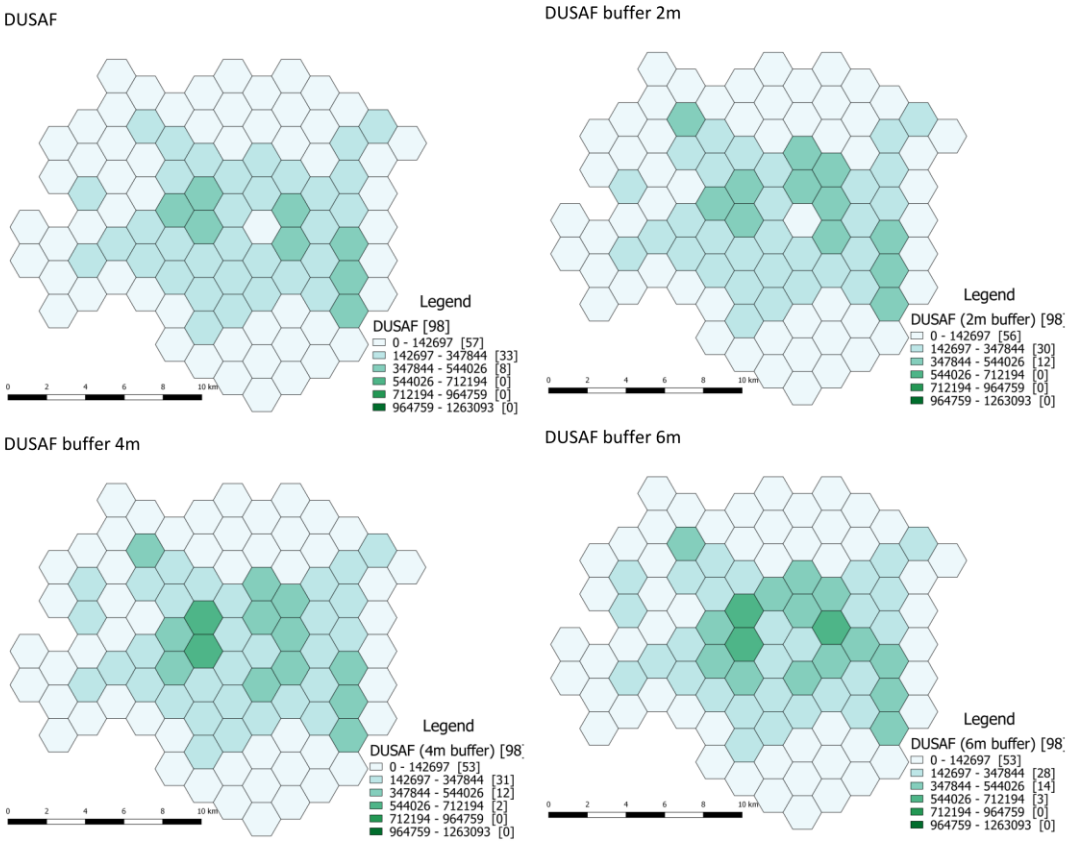


Figure 4.29 – DUSAF layer density

The buffers operated on DUSAF obviously increase the amount of area on each cell. Nevertheless, the distribution of the cells in the different categories does not change much. An interesting thing that can be noticed in all the maps is that, differently from the DBTR, the DUSAF area does not have a higher concentration in the Milan city centre. Only in the map showing the 6 m buffer on DUSAF, there is almost a circle of higher concentration area around the Milan city centre.

4.3.4.2 Method 1

To make a more readable map, the False Negative indexes are shown according to equation 4.4. The values that are shown in the maps correspond to the area of the DBTR that are not included in the DUSAF, divided by the total area of DBTR in the cell. The result is multiply by 100 in order to obtain a percentage.

$$FN[\%] = \frac{area \in (DBTR \setminus DUSAF)}{DBTR_{tot}} \cdot 100 \qquad \text{Equation 4.4}$$

Figure 4.30 shows the FN indexes obtained as results of the comparison. It is divided into 5 classes, from the lower percentage (white) to the higher ones (blue).

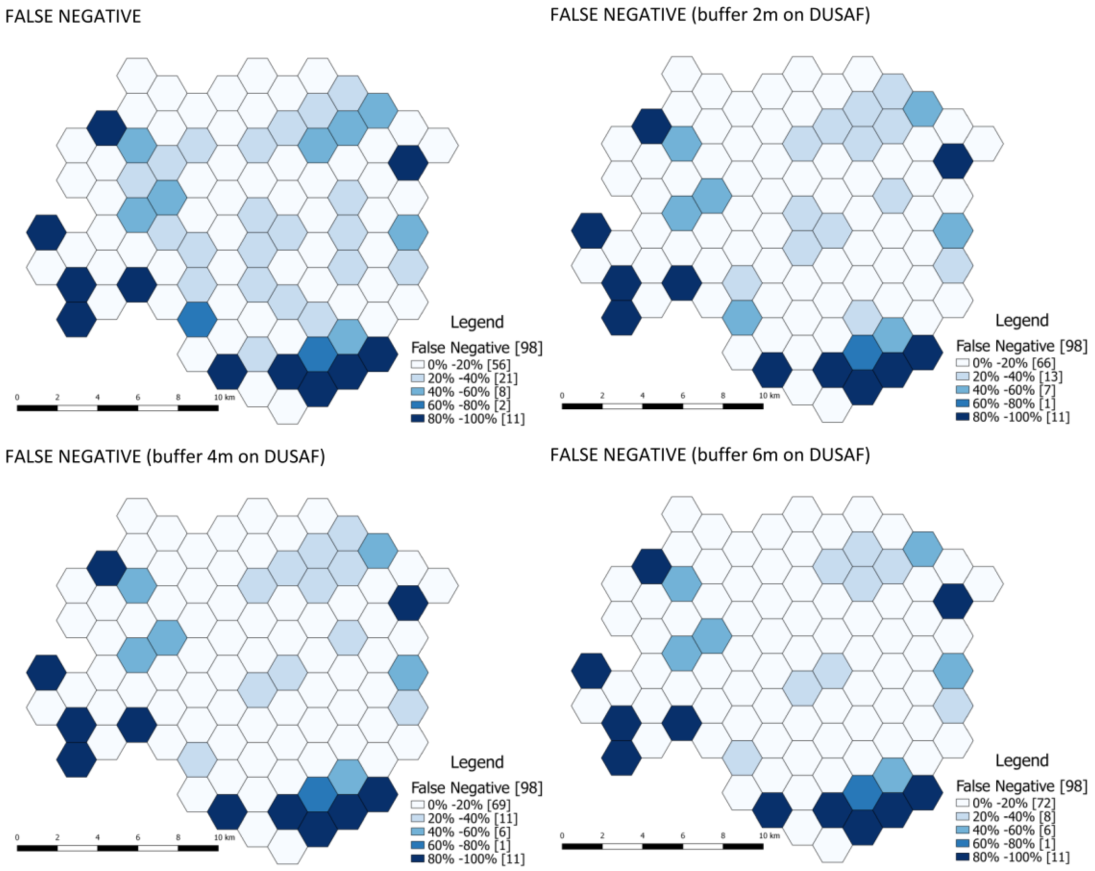


Figure 4.30 – Method 1 False Negative

Generally, it is possible to notice that the percentage of FN decreases when a buffer on DUSAF is applied. In fact, when a 6 m of buffer is performed, most of the cells fall in the first category (72 out of 98). This mean that using this method most of the DUSAF area covers the DBTR.

On the other hand, it can be noticed that the cells with higher value concentrate themselves toward the outskirts of Milan Municipality (especially in the south and west directions). In those locations the FN values range from 80-100%. This is probably due to DBTR areas that are not included into the DUSAF, even if in those cells both DBTR and DUSAF datasets have low quantity of area (see figures 4.28 and 4.29). This condition is equal in the all the showed maps. Theoretically, if DUSAF has some territories classified as street in those areas and a buffer is applied, the percentage of FN would decrease. As in this case the different buffers do not affect the results, it can be interpreted as lack of data in the DUSAF level.

The second index used for the assessment is the True Positive. The values that are shown in the map correspond to the intersection between the DBTR and the DUSAF area, divided by the area of DBTR. The result is multiplied by 100 in order to obtain a percentage (equation 4.5).

$$TP[\%] = \frac{area \in (DBTR \cap DUSAF)}{DBTR_{tot}} \cdot 100$$

Equation 4.5

Figure 4.31 shows the value of the TP index in the assessment considering the method 1. As the darker colour represents the higher percentages, it is possible to see that using this method the DUSAF level overlays well the DBTR one. Without any buffer, 53 cells out of 98 fall into the last category (TP ranges from 80 to 100%). Appling the different buffers, the situation gets better and better, passing from 61 cells in the 2 m buffer map to 65 of the 4 m buffer map and last to 69 with 6 m of buffer map.

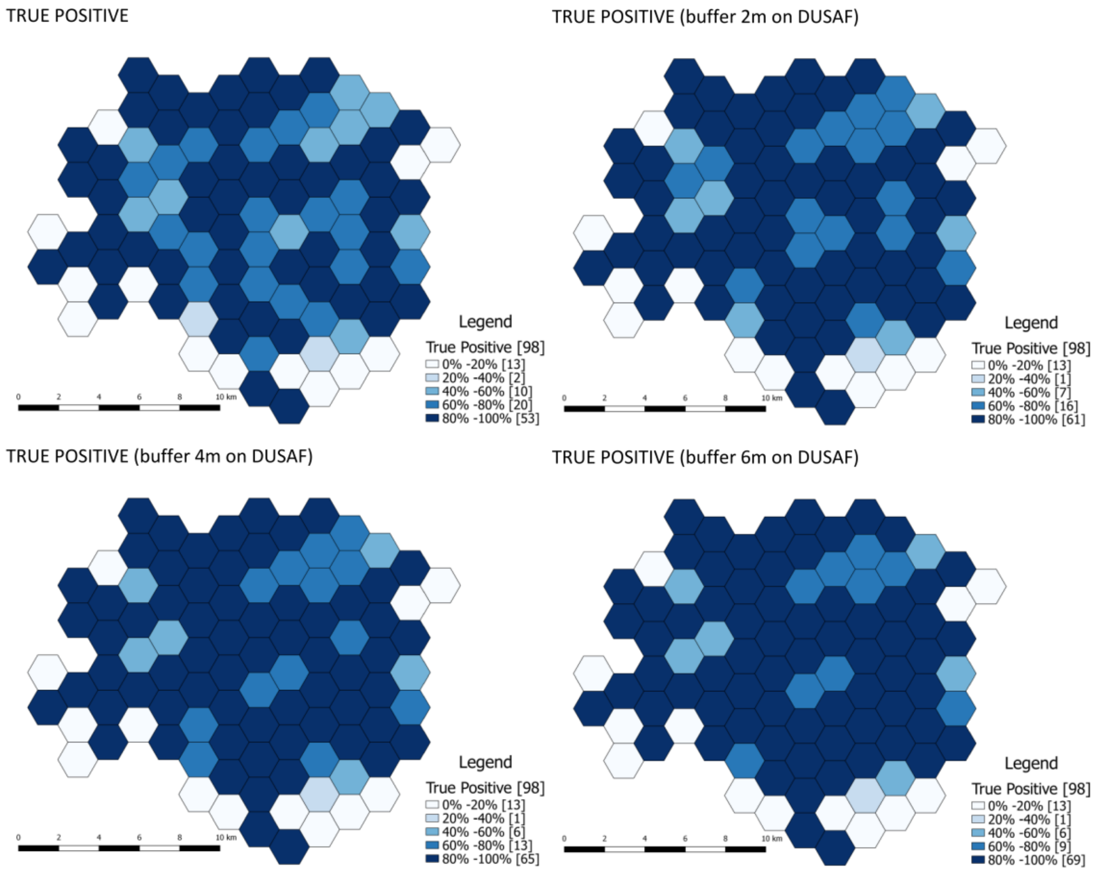


Figure 4.31 – Method 1 True Positive

Even in this case, it is possible to notice a problem of overlapping in the cells towards the outskirts of the grid. Similarly to the problem observed in the False Negative index, it can be noticed that the cells that have a high FN value also have a low TP value. This is due to the lack of DUSAF area in that landscape and, for this reason, there are not overlaid area measured by the True Positive index.

The last analysed index is the False Positive. Its values correspond to the DUSAF area not overlapped by DBTR. As before, the result is multiplied by 100 in order to obtain a percentage (equation 4.6).

$$FP\ [\%] = \frac{area \in (DUSAF \backslash DBTR)}{DBTR_{tot}} \cdot 100 \qquad \text{Equation 4.6}$$

Figure 4.32 shows the grid with a cataloguing on the FP, considering the method 1. From the first images is easy to see that most of the cells (73) have a very low False Positive index, which ranges from 0 to 20%. This underlines the goodness of the assessment, as there are few amount of DUSAF that are not overlaid by the DBTR input data.

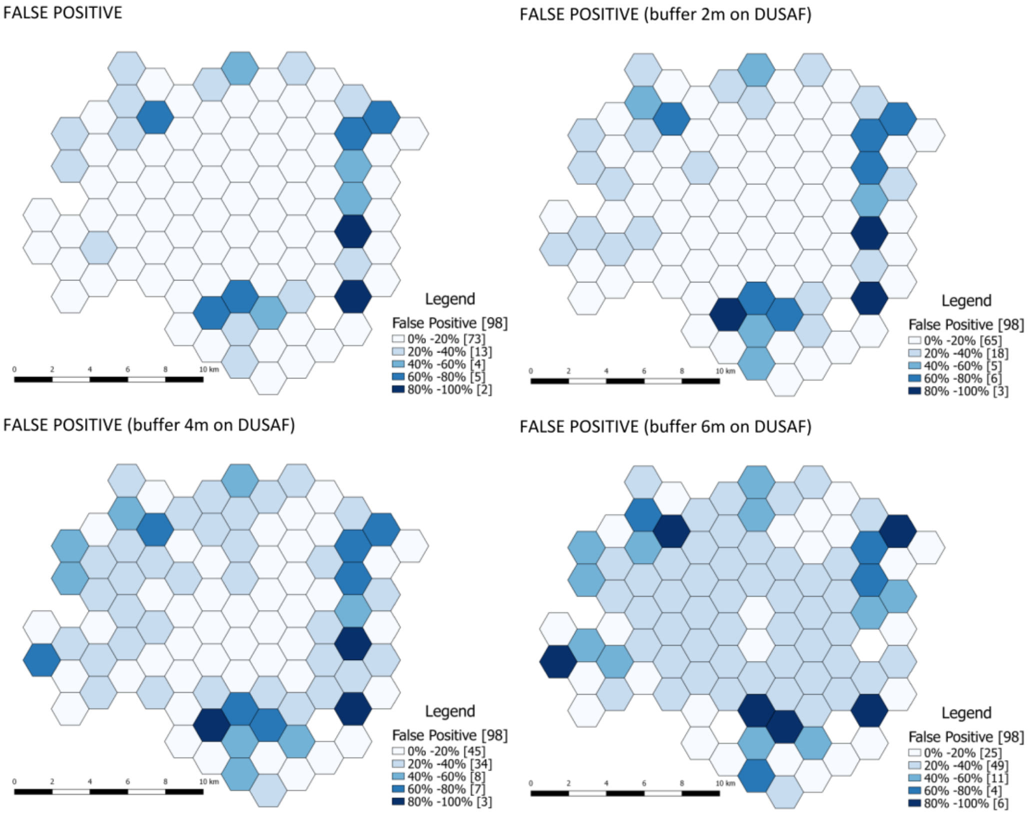


Figure 4.32 – Method 1 False Positive

Obviously when a buffers is performed, the False Positive index increases. It is easy to see that the classification of the cells became darker from the first (without buffer) to the last (6 m buffers).

Considering the cells that fall into the last category, there are very few values. In the best comparison (no buffer), there are only 2, in the worst case (6 m of buffer) there are 6 out of 98 cells.

4.3.4.3 *Method 2*

This section will describe the FN, TP and FP indexes measured on the result of method 2, computed through the same formulas explained in the previous subsection (see Subsection 4.3.4.2).

The first analysed index is the False Negative. It is shown in figure 4.33 and, as already stated, it represents the DBTR area not overlapped by DUSAF. Generally, it can be noticed that the situation is much worse with respect to the previous method. In all the maps, most of the cells fall into the second category, with FN values that range from 20% to 40%. Nevertheless, performing the buffers on the DUSAF layer produce a decreasing of the value of the FN indexes, as it can be seen from the increase of lighter cells.

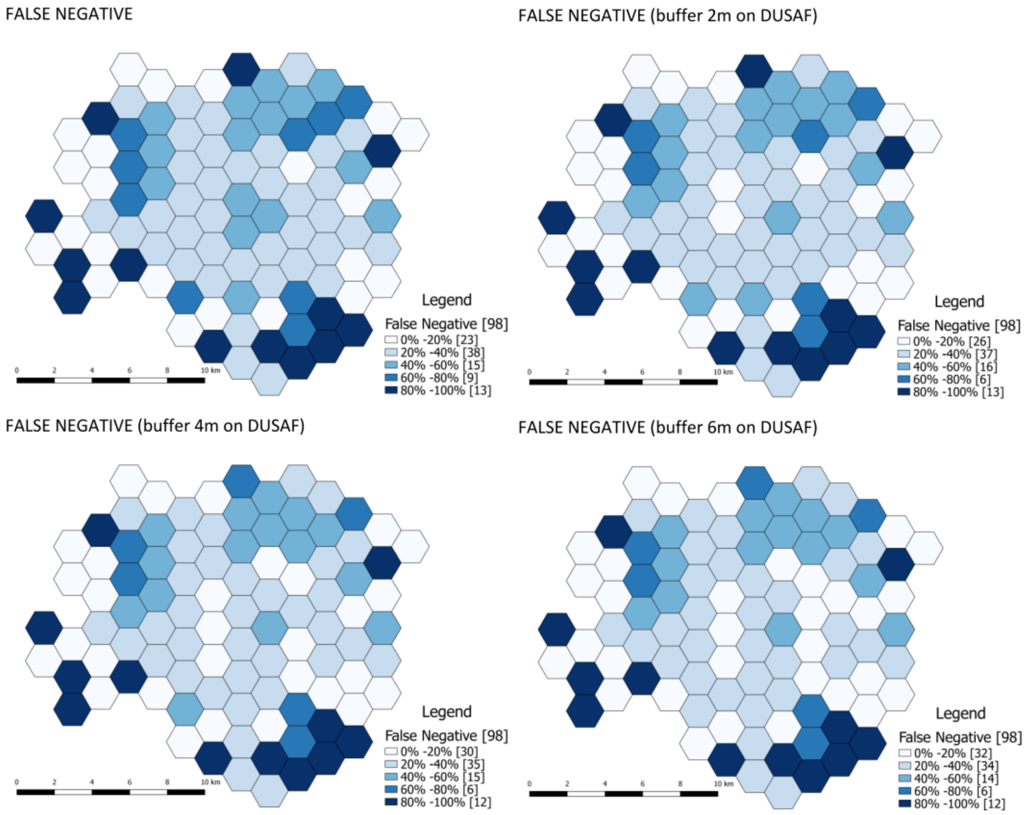


Figure 4.33 – Method 2 False Negative

Another thing that can be notice is that there are cells with higher value toward the southern and western borders. This condition also happens in the method 1 and it was explained as an error in the DUSAF level. Even in this case, the different buffers do not change the situation in those cells, which keep value of FN between 80%-100%.

From the comparison between DBTR and DUSAF (without any buffers), it is interesting to see that toward the border there are also cells that have the lowest FN index. As they mainly maintain the same class in all the maps, it can be affirmed that those cells contain low values of DBTR area and, for this reason, they have an FN value range between 0% and 20%.

The second analysed index is the True Positive, which represents the amount of overlapped area between DBTR and DUSAF. From figure 4.34, it is easy to notice that the situation is not so good. In fact, in all the cases the most common class is the one that contains TP values between 60%-80%. Because the buffers increase the TP value, it is registered an increase of the higher classes cells.

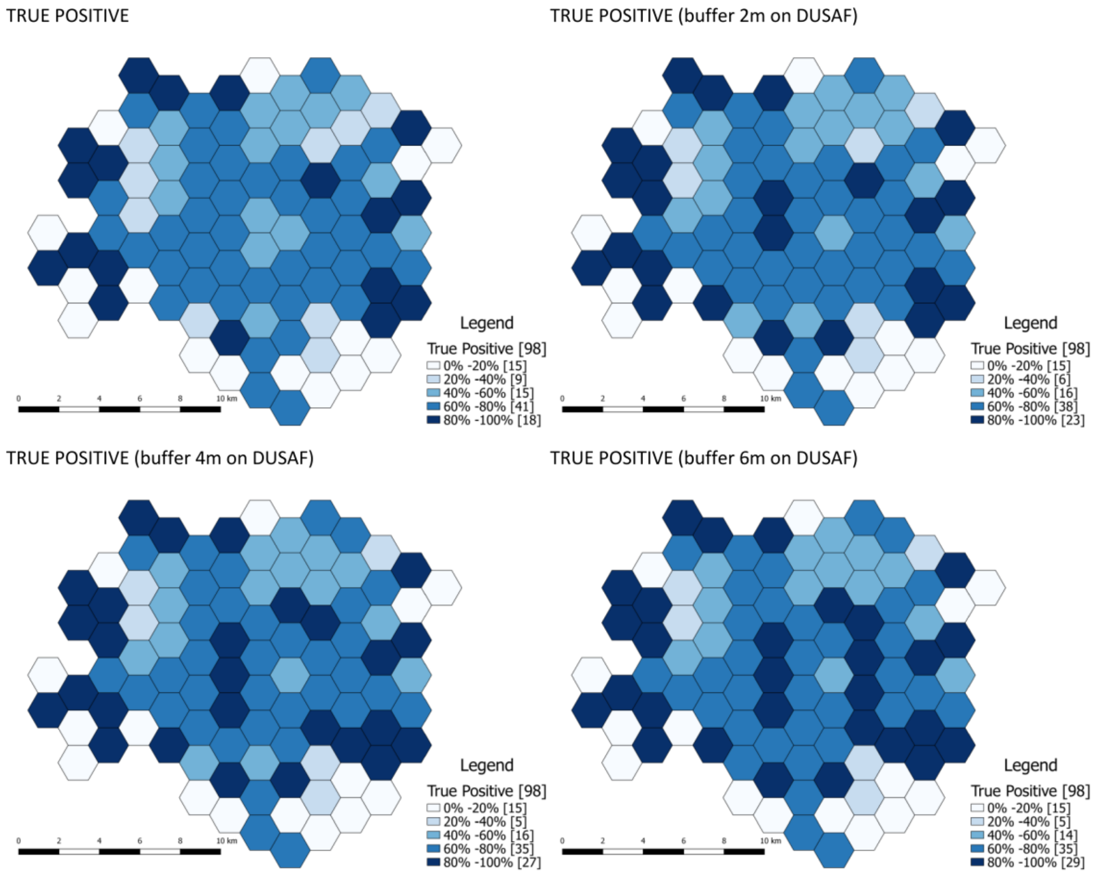


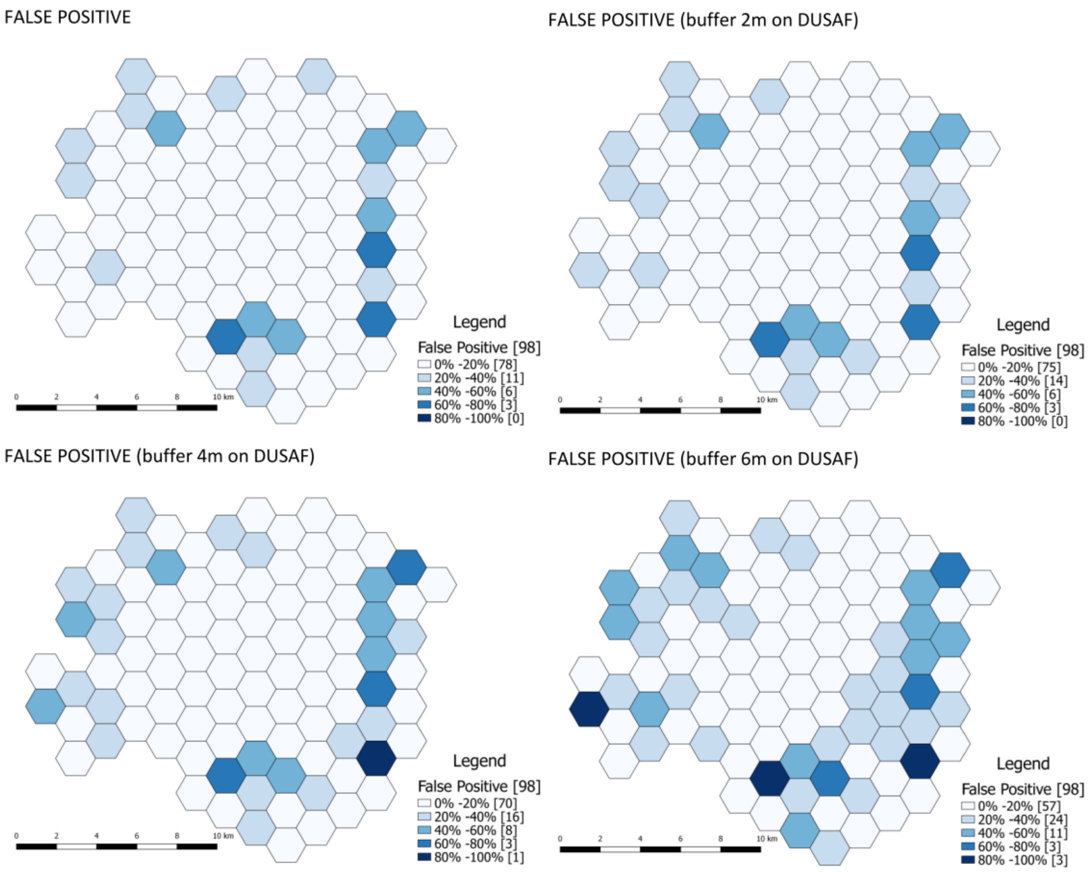
Figure 4.34 – Method 2 True Positive

Considering the 6 m buffer, 29 cells out of 89 fall into the last class, while in the method 1 there were only 69. It can be affirmed that, in this case, the DBTR and the DUSAF do not overlap well.

Finally, it can be notice that most of the highest TP indexes fall into the cells that registered the lowest FN index. This can be explained with low quantity of both DBTR and DUSAF areas that are located in the same positions.

The last index analysed for the method 2 is the False Positive. It shows the quantity of DUSAF area not included in the DBTR (see figure 4.35). The maps show that most of the cells fall into the first category, the one that contains FP values between 0% and 20%. The amount of cells in the other categories are few. Even if the buffer increases the value of FP index, the general situation do not change much, differently from the method 1. In fact, considering the 6 m buffer, there are only 3 cells that fall into the higher category, respect to 6 cells of the method 1. Hence, method 2 seems to fit better the assessment.

It is interesting to notice that the FP index in the Milan city centre is low. That means that the DUSAF area is well covered by the DBTR. In fact, in all the maps, the centre has FP values that fall into the first category.



FALSE POSITIVE

FALSE POSITIVE (buffer 2m on DUSAF)

FALSE POSITIVE (buffer 4m on DUSAF)

FALSE POSITIVE (buffer 6m on DUSAF)

Figure 4.35 – Method 2 False Positive

4.3.5 Comment on the results

Beside the description of the results using a spatial distribution (i.e. maps showing the grid), it is interesting to analyse even the numerical values of each index. As a matter of fact, these data could lead to better understand the result. This section illustrates all the measured values and, after that, it will analyse the 3 cells with the higher FN index for each methods. That subsection will help to better understand the dissimilarities between DBTR and DUSAF.

4.3.5.1 Numerical comparison between the two methods

Beside the spatial distribution of the results, it is also important to analyse the numerical values of each parameter involved.

The first considered value is the initial DBTR area, shown in table 4.8. As already stated before, method 1 simply uses the DBTR Class Road area, while the method 2 uses a combination of DBTR Classes (i.e. Road area + Pedestrian circulation area located on road area + Cycling circulation area located on Road area – Green area which type is flowerbeds). For this reason, the area of method 2 (43'130’038 m^2) is significantly higher respect to method 1 (32’455’006 m^2).

Table 4.8 – Numerical comparison between the area of the DBTR input datasets used in the two methods (in m^2)

	Method 1	**Method 2**
DBTR	32’455’006	43'130’038

The second analysed value is the DBTR areas comparable with DUSAF (see table 4.9). From now on, each parameter will be divided in 4 parts, considering the influence of the different buffers on DUSAF. The DBTR area comparable with DUSAF is the dataset generated by the script, as described in Subsection 4.2.2. From table 4.9, it is easy to notice that in both cases the area without buffer is less than half respect to the original one (e.g. method 1, the DBTR initial area is 32’455’006 m^2, while its version comparable to DUSAF without buffer is 14’110’645 m^2). Imposing the buffers on the DUSAF level, generate a slight increase on the DBTR area values in both methods.

Table 4.9 – Numerical comparison between the DBTR areas comparable with DUSAF involved in the two methods (in m^2)

	Method 1	Method 2
No buffer	14’110’645	19’595’703
2 m buffer	14'423’244	20’059’821
4 m buffer	14’610’687	20’212’445
6 m buffer	14'762’750	20’406’743

Table 4.10 shows the third listed parameter, the DUSAF area. Its values are equal in the two cases, as they represent the original layer and some buffers performed upon it. Respect to the previous indexes, in which the buffer marginally boosts the numbers, in this case they significantly increment the values.

Table 4.10 – Numerical comparison between DUSAF areas involved in the two methods (in m^2)

	Method 1	Method 2
No buffer	13’776’999	13’776’999
2 m buffer	15’014’009	15’014’009
4 m buffer	16'246’299	16'246’299
6 m buffer	17'473’877	17'473’877

The last three tables presented in this subsection regard the FN, TP and FP values. The first index, shown by table 4.11, is the False Negative. In all the cases, method 2 has values that are more than double respect to method 1 (e.g. the FN in method 1 with no buffer has 2’506’619 m^2 while method 2 has 6'574’715 m^2). Nevertheless, it can be noticed that in both methods the amount of area decreases in correspondence of the buffers, consistently with the Chebyshev’s inequality definition.

Table 4.11 – Numerical comparison between the False Negative areas of the two methods (in m^2)

	Method 1	Method 2
No buffer	2’506’619	6'574’715
2 m buffer	2’153’504	6'065’244
4 m buffer	1'906’785	5’563’597
6 m buffer	1’767’558	5'322’873

The second analysed index is the True Positive, listed in table 4.12. The differences between the values of the two methods are not so high. Although, they are still higher in all the cases of method 2. When the buffer are operated, the TP values lightly increase in the method 1, while in the method 2 they increment more.

Considering only the TP values, seems that the method 2 best fit the evaluation as the DBTR and DUSAF layers have more area in common. Nevertheless, to determine the better method, an overview between all the parameters is needed.

Table 4.12 – Numerical comparison between the True Positive areas of the two methods (in m^2)

	Method 1	Method 2
No buffer	11’604’027	13’020’989
2 m buffer	12’269’740	13’994’577
4 m buffer	12’703’902	14’648’848
6 m buffer	12'995’192	15’083’870

The last index is the False Positive, listed in table 4.13. It shows the smaller variations between the two methods. In all the cases the differences between each buffer is almost equal, around 200’000 m^2 higher in the method 2 values.

Table 4.13 – Numerical comparison between the False Positive areas of the two methods (in m^2)

	Method 1	Method 2
No buffer	5'087’058	5’256’824
2 m buffer	4'146’207	4’325’702
4 m buffer	3'340’547	3’526’206
6 m buffer	2’756’621	2'945’388

4.3.5.2 Most relevant cells

This section analyses the cells with the highest quantity of FN areas for both methods, in order to show differences between the DBTR and DUSAF datasets. It illustrates how DUSAF classified the area that should be included in the DUSAF but they are not.

The first part of this section analyses the 3 cells with the highest FN indexes in method 1; respectively cells with ID equal to 69, 40 and 30. Figure 4.36 shows the situation in the cell 69. It is easy to see that most of the FN areas are classified by DUSAF as Parks and gardens or Dense urban fabric.

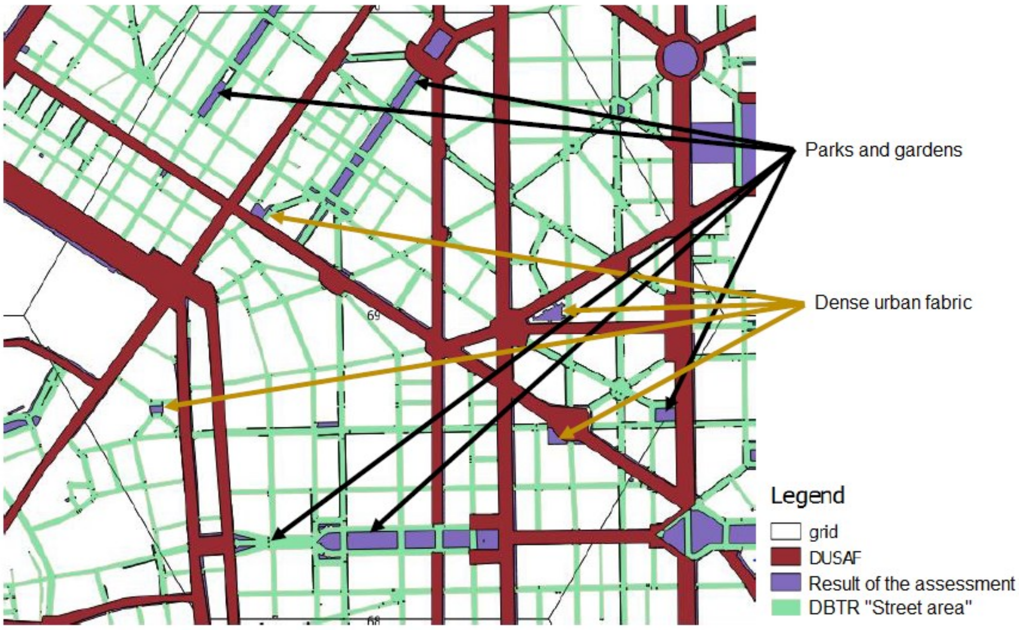


Figure 4.36 – DUSAF classification of the area from result of the assessment in cell 69 of the method 1

The second cell with the highest FN value is the number 40 (see figure 4.37). Even in this case there are many FN area classified as Parks and gardens in DUSAF. Some other areas are classified as Industrial settlements and Dump sites.

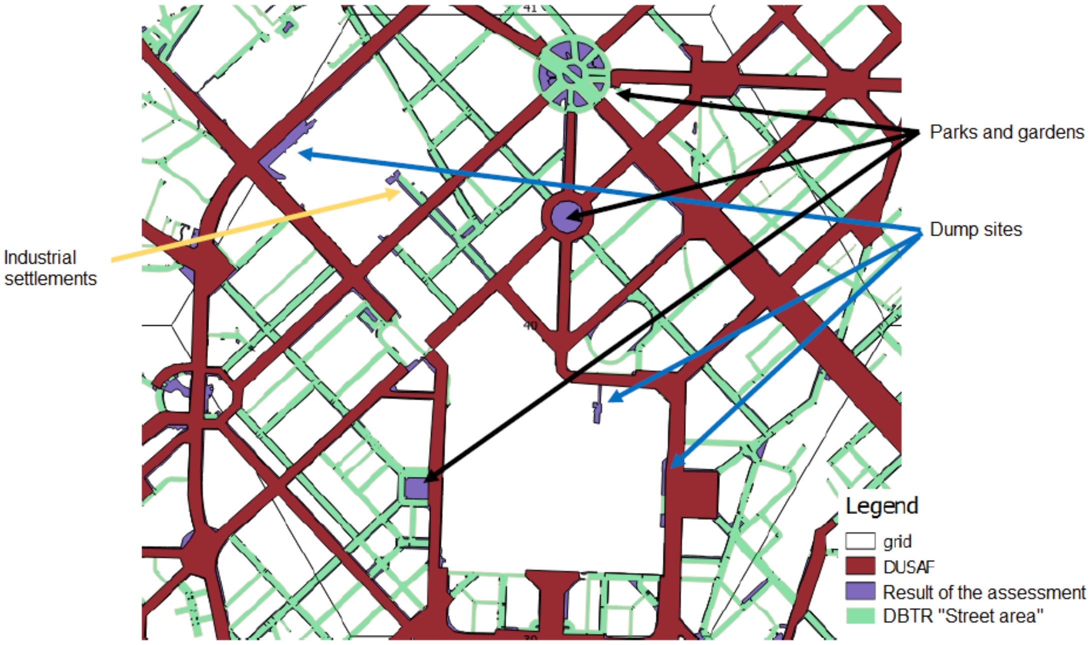


Figure 4.37 – DUSAF classification of the area from result of the assessment in cell 40 of the method 1

The third cell analysed for method 1 is the number 30 (figure 4.38). The FN area is, again, mainly classified by DUSAF as Parks and gardens. Other DUSAF levels found in the cell are Public or private implants of service, Dense urban fabric and in one little area of Sport centre.

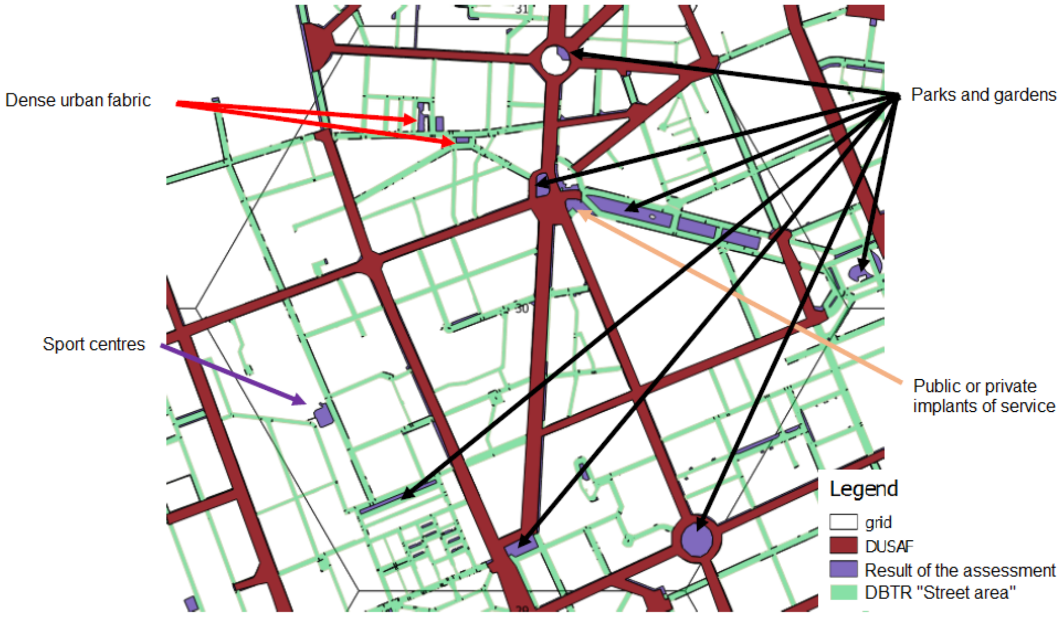


Figure 4.38 – DUSAF classification of the area from result of the assessment in cell 30 of the method 1

Generally, most of the differences regards area classified as Urban green. In fact, it happens that the trees located in the middle of two carriageways are considered road area by DBTR and Parks and gardens by DUSAF. These differences are due to the disagreement between the definitions of the classes.

On the other hand, the 3 cells with the highest FN indexes in method 2 are the 69, 25 and 40. Figure 4.39 shows the situation in cell 69. Looking at the FN layer (light green polygons) in this image, is easy to notice that there is a huge amount of area in the eastern part of the cell.

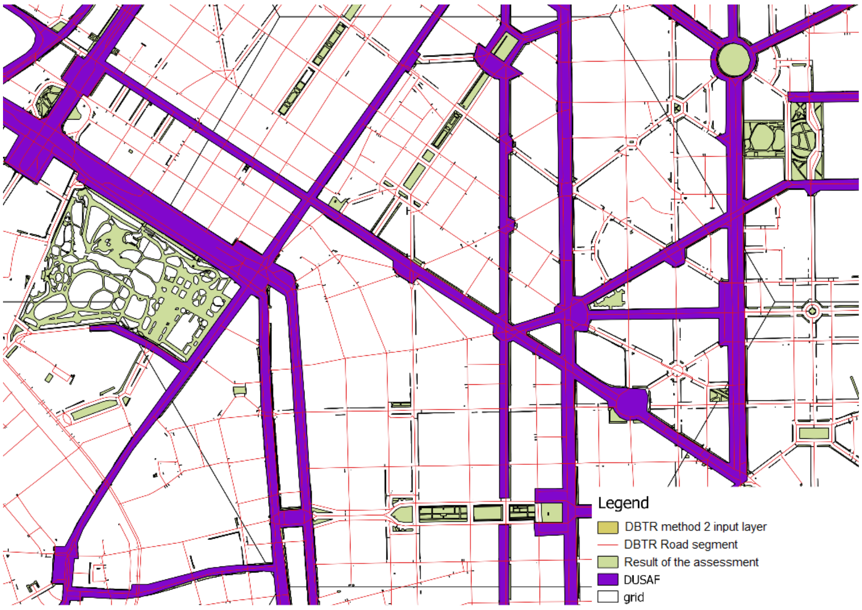


Figure 4.39 – DUSAF classification of the area from result of the assessment in cell 69 of the method 2

Focusing on this area, it is possible to see from figure 4.40 that it is not a street, but it represents the paths of a park (named Montanelli public gardens). DUSAF classifies correctly this area as Parks and gardens with some Industrial or commercial units in correspondence of the buildings. DBTR classifies wrongly this area, as it labels the location as pedestrian circulation positioned on road area. For this reason the results of the assessment contains a mistake, introducing an error of around 300 m^2 in the FN index.

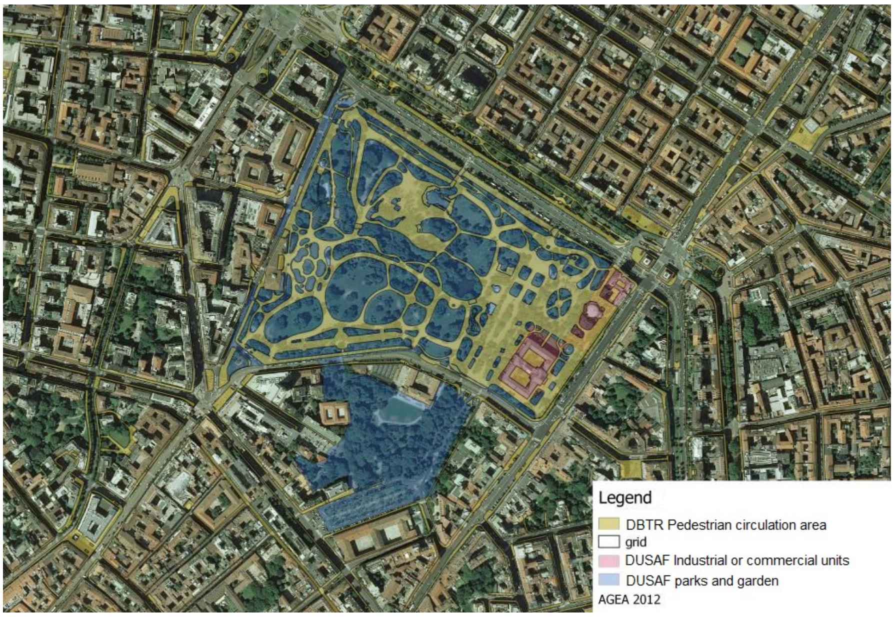

Legend
DBTR Pedestrian circulation area
grid
DUSAF Industrial or commercial units
DUSAF parks and garden
AGEA 2012

Figure 4.40 – DUSAF classification of Montanelli public gardens with AGEA 2012 orthophoto

The second cell with the highest FN value in method 2 is the number 25. Even in this case, it is easy to notice a huge mistake in results of the assessment in the lower part of the cell (figure 4.41).

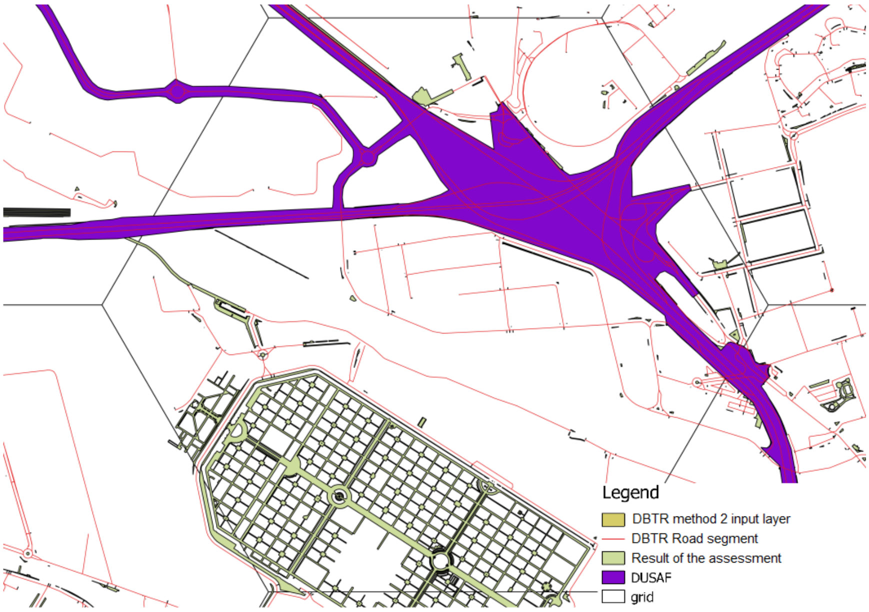


Figure 4.41 – DUSAF classification of the area from result of the assessment in cell 25 of the method 2

In fact, even in this case, this area is wrongly classified by DBTR as pedestrian circulation area located on a road area, even if it is a cemetery (Monumental Cemetery of Milan Municipality).

DUSAF, on the other hand, labels this area correctly as Cemetery (see figure 4.42). This mistake introduces 1389 m^2 of error in the FN index and it is the cause of its high value.

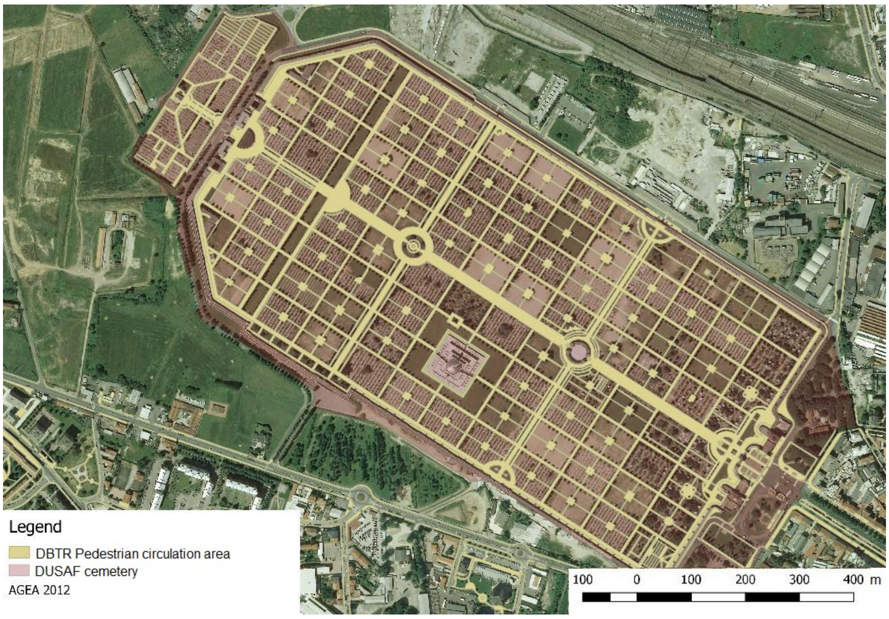


Figure 4.42 – DUSAF classification of Monumental Cemetery of Milan Municipality with AGEA 2012 orthophoto

The last analysed cell for method 2 is the number 40. Figure 4.43 shows that, there are some area in the results catalogued as Parks and gardens in the DUSAF layer, while in other cases there are area considered as Dump sites. Respect to the other cells presented before, there are not big mistakes.

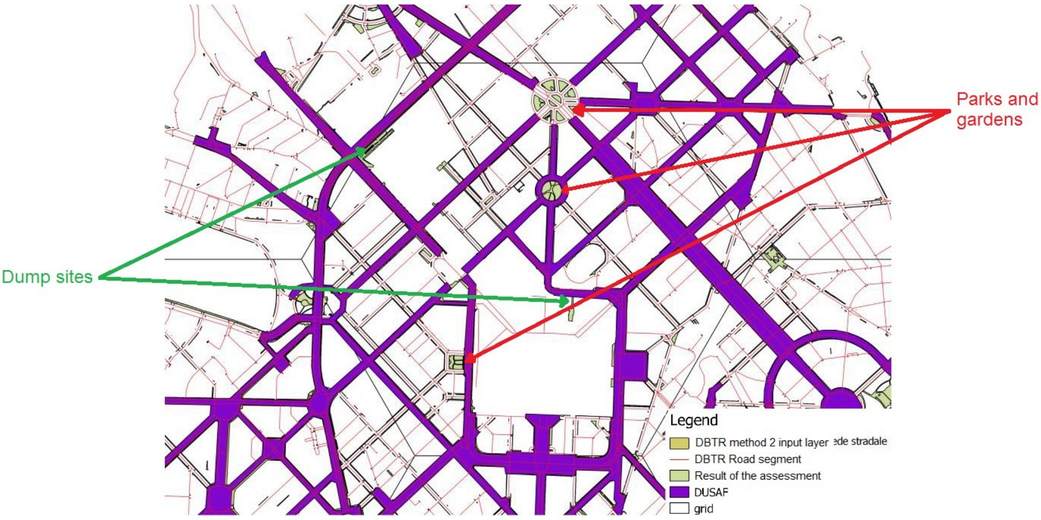


Figure 4.43 – DUSAF classification of the area from result of the assessment in cell 40 of the method 2

Finally, it is interesting to investigate the differences in the FN areas between the two methods. Figure 4.44 shows the chart of the absolute differences (in m^2) between them.

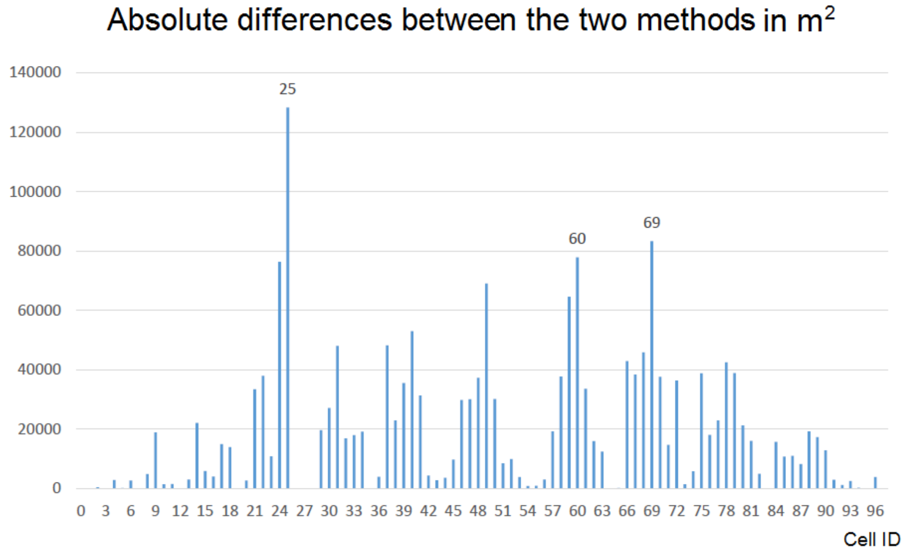


Figure 4.44 – Absolute differences between the two methods (in m^2)

It can be notice that the cell with the highest value is the 25, due to the presence of the Monumental Cemetery wrongly classified. On the other hand, the cells number 69 and 60 have such higher values due to the error in the cataloguing of the Montanelli public gardens.

4.3.6 Determination of the best method

The aim of this study is to provide a way to assess the quality of DUSAF. For this reason, between the two presented methods, the best one has to be chosen.

Considering the visualization of the results, it appears that the method 1 is better considering the FN and the TP (see Subsection 4.3.4.2), while method 2 has higher FP index (see Subsection 4.3.4.3). On the contrary, looking at the numerical results it is easy to notice that method 2 has higher results in all the cases (see Subsection 4.3.5.1). This condition can be considered as positive only for the TP index, in fact high values of FN and FP indicates a low precision in the assessment.

Another important statement is related to the wrong classification of some areas in the method 2, which introduce huge errors in the determination of the different indexes (see Subsection 4.3.5.2).

For these reasons, **method 1 uses the dataset that best fit the assessment for Milan Municipality**. Considering the visualization of the TP, method 1 shows that its DBTR and DUSAF datasets have the highest overlapping area respect to the amount of DBTR area involved (see Subsection 4.3.4.2). On the other hand, method 2 has the highest quantity of area in m^2 (see

Subsection 4.3.5.1), but still a lower percentage of it (see Subsection 4.3.4.3). Hence, it can be affirmed that considering the territory of the Municipality of Milan, **DUSAF Road network and associated land level has a good quality if compared to method 1**.

Choosing the method 1 reaffirms the statement of the Lombardy Region, which defines the input layer of this method as equivalent to the part catalogued as street in the DUSAF layer (see Subsection 4.2.1.1).

4.3.7 Spatial autocorrelation

In this case study, the spatial autocorrelation of some maps of method 1 are calculated. In particular, the FN, TP and FP indexes of the assessment result made with 6 m buffer on DUSAF are analysed. This choice is due to the fact that, for this method, they appear as the ones with higher clustered cells.

Operatively, the first thing to fix in order to estimate the spatial autocorrelation is the weight *w* (see Subsection 4.2.4). In this case study, the type of weight used is the sphere of influence. Each feature is analysed within the context of neighbouring features located within a specified distance. Neighbours within the specified distance are weighted equally, while the one positioned outside the specified distance do not influence calculations (weight equal to zero).

As this case study is two-dimensional, the sphere of influence is a circle and the distance used as threshold is its radius. Figure 4.45 shows how that value were calculated. For example, in order to include the six nearest cells, the threshold used is 2 times the apothem of the hexagon cell. To include the 18 nearest cells the threshold is equal to 4 time the apothem and, lastly to comprise the 36 nearest hexagons the limit is 6 time the apothem.

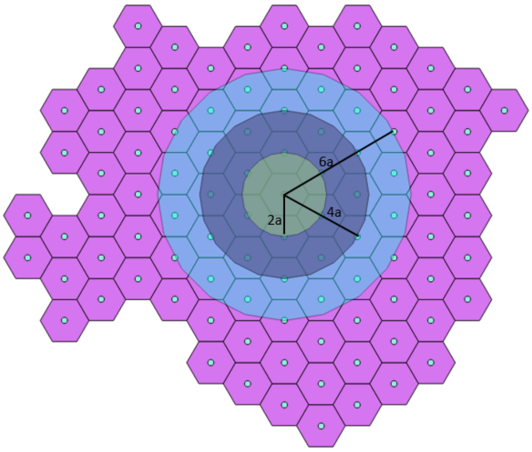


Figure 4.45 – Determination of the value of the threshold

As the apothem is equal to 866 m, table 4.14 shows the thresholds value (radius of the circle of influence) and their value rounded by excess. These are used in the estimation of the Moran Index in the two pieces of software.

Table 4.14 – Thresholds value (m)

	Value	**Excess rounded value**
2a	1732,05	1733
4a	3464,10	3465
6a	5196.15	5197

Appling these values of thresholds, figure 4.46 shows an example of which neighbouring features are selected (in yellow) respect to a specific cell (in green). The red cells have weight equal to zero.

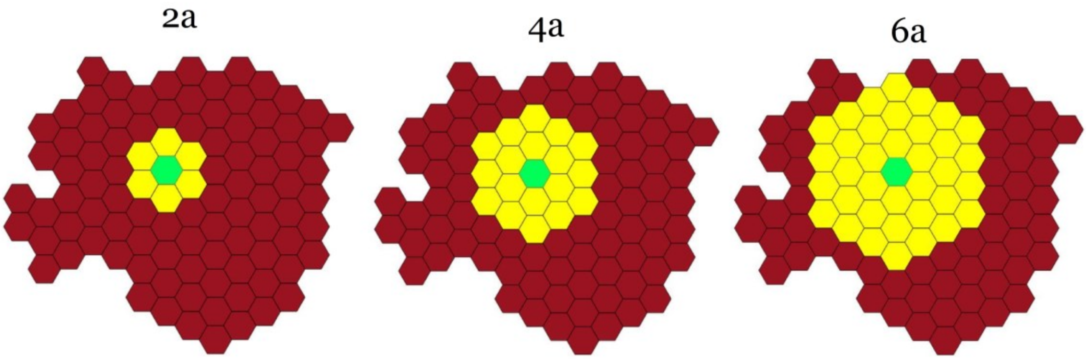
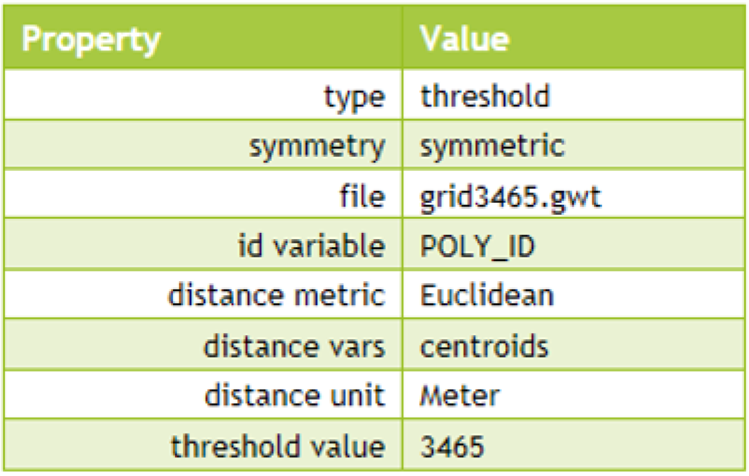


Figure 4.46 – Neighbouring features selected based on different threshold

4.3.5.1 GeoDa

To perform the Moran's I in GeoDa, the first thing to do is connect the data source (grid layer) with the program. The second step is defining the weight, through the function *Tools* > *Weight Manager* > *Create*. In the *Weights File Creation* interface, the *Weights File ID Variable* (i.e. the ID variable without duplicate) had to be specified. In this case study, a new ID Variable is generated through GeoDa (named POLY_ID). After that, a *Distance Weight* is selected and the "Euclidean" method is chosen as distance calculator. Then it has to be specified that the points used to estimate the distances are the centroid of each cells. Figure 4.47 shows the recap of the weight used in the determination of the Moran's I.



Property	Value
type	threshold
symmetry	symmetric
file	grid3465.gwt
id variable	POLY_ID
distance metric	Euclidean
distance vars	centroids
distance unit	Meter
threshold value	3465

Figure 4.47 – GeoDa Weights File

Finally, the Moran index can be determined on the chosen field (attribute of the grid, which contains the value of the indexes) through the function *Moran Scatter Plot* > *Univariate Moran I*. GeoDa returns both the numerical values and a plot representing the Moran index.

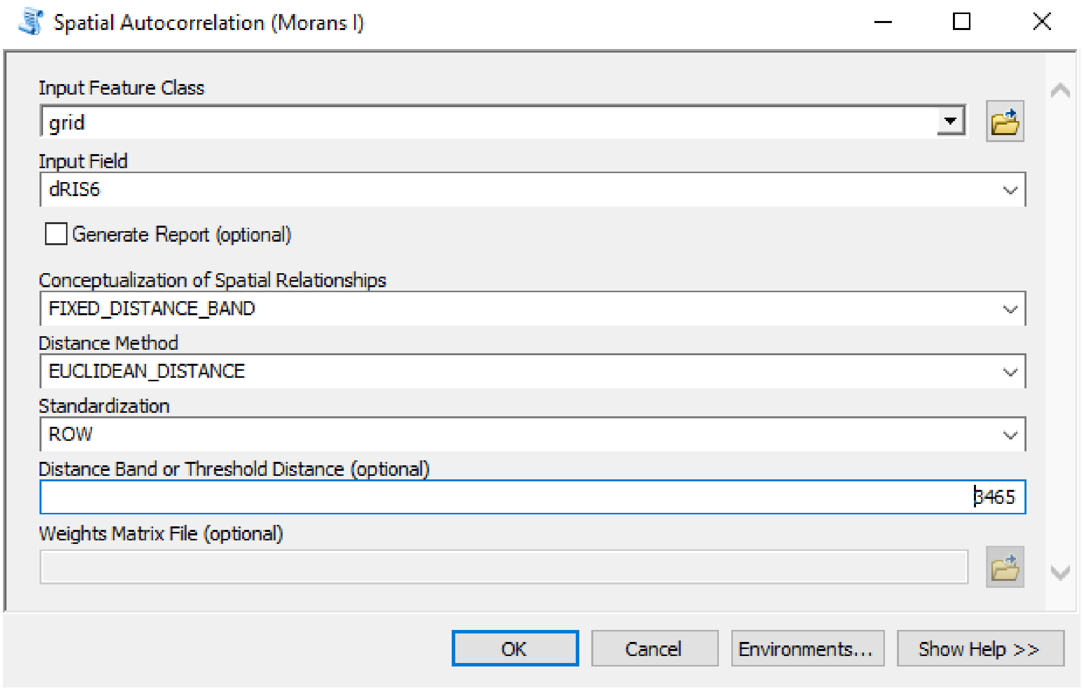
4.3.5.2 ArcGIS

Once the grid is added in ArcMap, the Moran Index I is calculable trough the function *ArcToolbox* > *Spatial Statistics Tools* > *Analyzing Patterns* > *Spatial Autocorrelation (Moran I)*. Figure 4.48 shows the interface that appears when this command is executed.

In the field *Input Feature Class* is specified the name of the layer and in the *Input Field* is indicated the attribute of the layer to analyse. The chosen spatial relationship is explicated in *Conceptualization of Spatial Relationship* as "Fixed distance band" with as *Distance Method* the "Euclidean distance". In the *Standardization* field is selected "Row" because the ArcGIS documentation suggested, as it affirms that (ESRI 2017):

> "Row standardized weighting is often used with fixed distance neighborhoods and usually used for neighborhoods based on polygon contiguity"

Finally, in the *Distance Band or Threshold Distance* field is specified the threshold value. This field use as unit of measurement the layer coordinate system one. Hence, as the grid is in UTM, the unit was meter.



Spatial Autocorrelation (Morans I)

Input Feature Class: grid

Input Field: dRIS6

□ Generate Report (optional)

Conceptualization of Spatial Relationships: FIXED_DISTANCE_BAND

Distance Method: EUCLIDEAN_DISTANCE

Standardization: ROW

Distance Band or Threshold Distance (optional): 3465

Weights Matrix File (optional):

Figure 4.48 – ArcMap Spatial Autocorrelation interface

Differently from GeoDa, ArcMap returns only the value of the Moran Index.

4.3.5.3 Comparison between the two software

This section compares the values of the Moran Indexes calculated through the two software. Both return the numerical value of the index, but with different significant digits (8 in GeoDa and 6 in ArcGIS).

Figure 4.49 shows the first analysed map, the False Negative obtained with 6 m buffer on DUSAF of the method 1. This image shows the grid and the three plots regarding the different thresholds (2a, 4a and 6a) obtained in GeoDa. The inclination of the purple line in the plot represents the Moran Index.

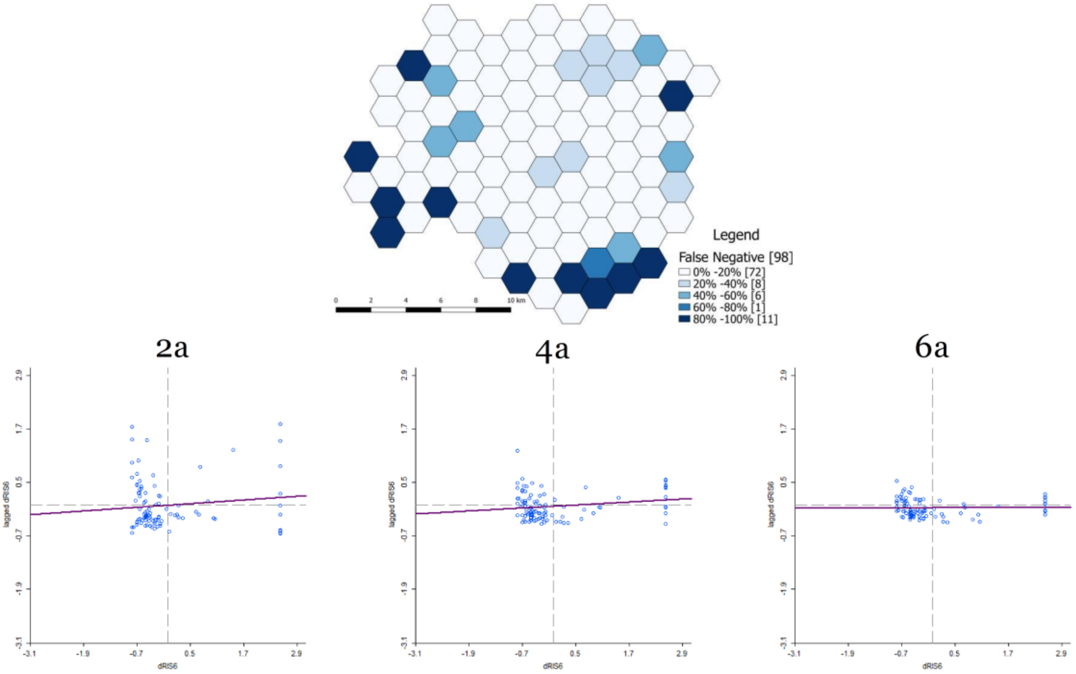


Figure 4.49 – GeoDa Moran Index plots of the False Negative index

The inclination of the purple line is almost horizontal. This statement is underlined by table 4.15, which lists the value of the Moran Index in the two software. It can be noticed that all the values are around 0. This means that for the FN grid there is neutral spatial autocorrelation (see Subsection 4.2.4). It also can be notice that the values are equal in the two software when the threshold is 2a, while in the other cases ArcGIS obtains higher results.

Table 4.15 – Moran Index values of the False Negative index

	2a	**4a**	**6a**
ArcGIS	0.067456	0.057597	0.020974
GeoDa	0.06745630	0.05588000	0.00512564

The second parameter on which the Moran Index is calculated is the True Positive. Figure 4.50 shows the map and the three plots of GeoDa. It can be noticed that the inclination of the line is higher with the 2a threshold, while in the other case is almost horizontal.

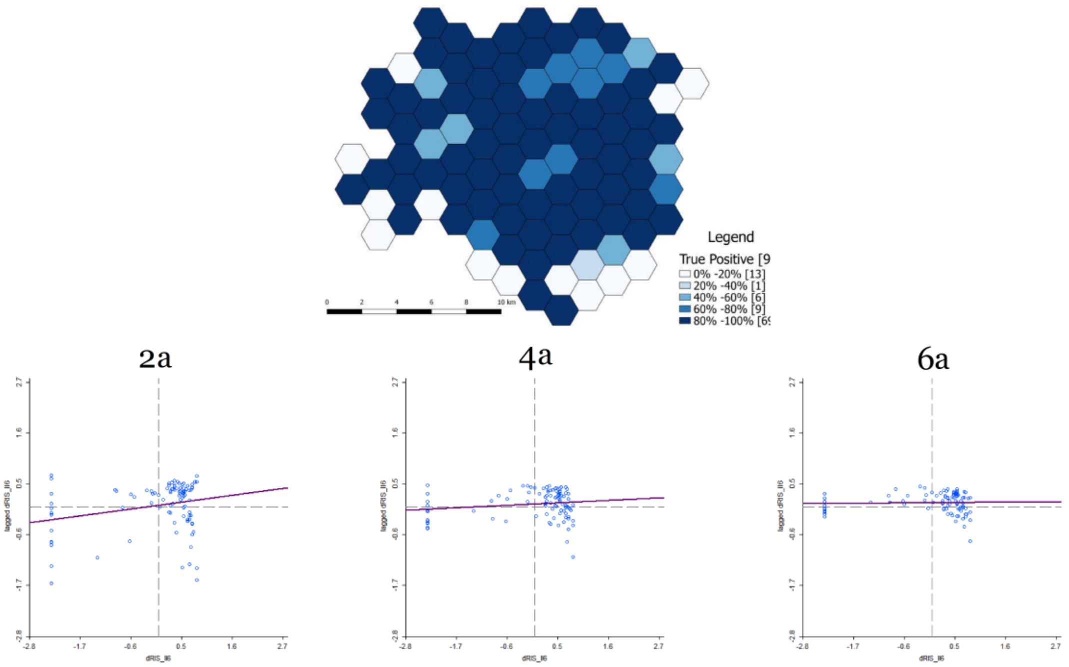


Figure 4.50 – GeoDa Moran Index plots of the True Positive index

Table 4.16 shows the Moran Index for the True Positive parameter. From this table can be see that when the threshold is equal to 2a, there is the greatest Moran’s Index, as can be seen also from the plot in figure 4.50. In this case, it can be affirmed that there is a slighter positive autocorrelation. For all the others cases, the Moran Index is around 0, hence there is neutral spatial autocorrelation.

Similarly to the FN case, the values of the Moran index obtained by the two software are equal when the threshold is 2a. In the other cases, ArcGIS obtain higher results.

Table 4.16 – Moran Index values of the True Positive index

	2a	**4a**	**6a**
ArcGIS	0.133243	0.060085	0.026853
GeoDa	0.13324300	0.04701200	0.00691789

The last analysed Moran Index is on the False Positive. Figure 4.51 shows the FP index and the three plots obtained in GeoDa.

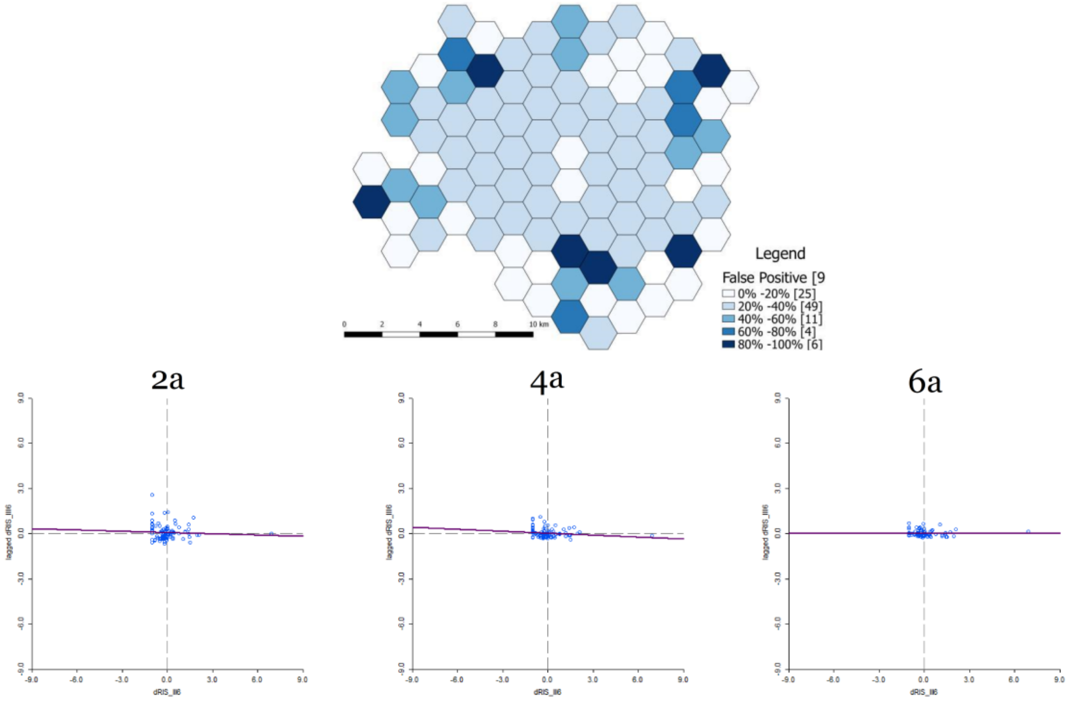


Figure 4.51 – False Positive GeoDa Moran Index plots

Even in this case, it is easy to see that the purple lines in the plots are almost horizontal, which means neutral spatial correlation. Table 4.17 underlines that statement, as all the values are near to 0. Generally, it can be noticed that all the values are negative, differently from the FN and TP.

Table 4.17 – False Positive Moran Index values

	2a	**4a**	**6a**
ArcGIS	-0.026874	-0.040862	-0.009876
GeoDa	-0.02687390	-0.04462120	-0.00184741

In this case study, in all the analysed cases there is a neutral spatial correlation. Only considering the TP index with 2a as threshold there is a slightly positive spatial correlation This is confirmed in both software even if ArcGIS gives always-higher values if the threshold considered is different from 2a.

4.4 Conclusions

This chapter has designed a procedure to compare two open geodata in the Municipality of Milan and has evaluated the quality of one respect to the other. The involved datasets are DUSAF, a regional land use map, and DBTR, the Topographic Database of Lombardy Region. This assessment has focused on the evaluation of the road area. Thus, the main objective of this study has been to determine the quality of DUSAF Road network and associated land level respect to the DBTR for Municipality of Milan.

It has started with a brief description of the case study and, then, the characteristics of the involved datasets have been illustrated, a fundamental operation needed to understand if the products contain comparable information.

The second part of the chapter regards the methodology. In this case study, it has been particularly difficult to detect the best dataset (or combination of datasets) to compare with DUSAF. For this reason, two methods has been proposed and described. Method 1 uses the DBTR Class Road area as input dataset, while the method 2 suggests a combination of DBTR Class as input (i.e. Road area + Pedestrian circulation area on road area + Cycling circulation area on road area – flowerbeds of Green area). After that, the procedure illustrates the theoretical concepts related to the Chebyshev's inequality and the Moran's I, as they explicate some choices made in the development of the methodology. Lastly, this section has also introduced how the results are visualised, defining the grid and the different indexes.

Then, this Chapter has described operatively how the results are reached following the methodology. It has illustrated how to obtain the dataset used in the two different methods, how to create the grid and it has briefly described the python script used to perform in the assessment. Finally it has shown the results in two different ways, by maps and by their numerical values. These information have led us to detect the best method of comparison between DBTR and DUSAF and, after that, have allowed us to assess the quality of DUSAF streets.

This assessment has demonstrated that DBTR Class Road area (method 1) is the best dataset to compare to DUSAF "Road network and associated land" level, as Lombardy region affirmed in one of its document (Tecnical specification 2016). Further, this Chapter has demonstrated that DUSAF Road network and associated land level has a good quality if DBTR Class Road area is used as reference.

Between the two methods, method 1 has the highest percentage of the True Positive index, i.e. the higher amount of overlapping area between the two layers, and the lowest False Negative index percentage, i.e. the amount of DBTR area not overlapped by DUSAF. Further, in all cases the increment of the probability of overlapping between the area of DUSAF and DBTR through some buffers, following the Chebyshev's inequality, has also increased considerably the TP area.

On the other hand, it has to be remember that in the method 2 there are some problems connected to some wrongly catalogued areas, which have introduced some errors in the comparison.

Finally, the analysis of the spatial correlation in the grid of method 1 with 6 m of buffer on DUSAF has detected a Moran's index almost equal to zero in all the cases (maps representing FN, TP and FP indexes). This value of the Moran's I indicates neutral spatial correlation, underlining the lack of both clusters and random distribution in the grid.

Unlike that presented in the previous Chapter, this study highlights that the analysed land use map has a good quality. These conclusions suggest that the quality of an open geodata can be heterogeneous. Therefore, users have better to check the quality of a dataset before using it.

CONCLUSION

This thesis underlined the importance of Open Data. In Chapter 1, it emphasized the influence that this type of information has in many different topics, paying specific attention on the subset of open geodata (VGI and authoritative). It illustrated the reasons of the spreading of Open Data over the last decade and their importance in the current worldwide situation. It also introduced the definition of spatial data quality, both internal and external, in order to better understand the following analyses.

Chapter 2 analysed the availability of Open Data for the Metropolitan city of Milan. From this set of information, the geodata were catalogued and sorted by different criteria. This evaluation allowed to answer the first of the two issues proposed as objectives of this study, i.e. the one regarding the availability of open geodata for the Metropolitan city of Milan. A huge presence of open geodata was found, with a positive tendency of publication which is in line with the general global trend.

Chapters 3 and 4 showed the quality assessment of two different openly-licensed geodata (of the first available as a web service and the second in a vector format) for the Municipality of Milan. The results aimed at answering the second issue set as objective of this thesis, i.e. the evaluation of the spatial data quality of the available geodata. Chapter 3 was focused on assessing the positional accuracy of the AGEA 2012 orthophoto (available as a WMS) through comparison against to the building layer of the Database Topografico Regionale (DBTR, Topographic Database of the Lombardy Region). Results showed that the positional accuracy of the positional accuracy of the roofs of the Municipality of Milan was not compliant with the nominal accuracy declared. Chapter 4 showed a comparison between the streets of the land use map named DUSAF (Use of agricultural and forestry soils) and the ones of the DBTR. The comparison between those layers returned good results, underlined by high values of overlapping areas. Hence, from the analyses undertaken in this work it can be concluded that the quality of the geodata for the Municipality of Milan is heterogeneous.

To conclude, this thesis has demonstrated the huge availability of open geodata for the Municipality of Milan; however, users must be aware that the quality of those data is heterogeneous and, sometimes, may be worse than what is formally declared.

LIST OF ACRONYMS

Accuracy	International Symposium on Spatial Accuracy Assessment in Natural Resources & Environmental Sciences
AGEA	*Agenzia per le Erogazioni in Agricoltura,* Agency for the Agricultural Supplies
AgID	*Agenzia per l'Italia Digitale*, Agency for Digital Italy
AGILE	Association of Geographic Information Laboratories in Europe
AIMA	*Azienda di Stato per gli interventi nel mercato agricolo,* National Company for the agricultural marker participation
AMAT	*Agenzia Mobilità Ambiente Territorio,* Mobility Agency for Environment and Territory
ANAC	*Autorità nazionale anticorruzione*, Anticorruption National Authority
AQL	Acceptance Quality Limit
ARPA	*Agenzia Regionale per la Protezione dell'Ambiente,*
ASCII grid	American Standard Code for Information Interchange grid
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
AT	Aerial Triangulation
BBox	Bounding Box
CLC	CORINE Land Cover
CNR	*Consiglio nazionale delle ricerche,* National Research Council
CP	Control Point
CRS	Coordinate Reference System
CRU	Climatic Research Unit
CSV	Comma-Separated Values
CTO	Chief Technology Officer
DBTR	*DataBase Topografico Regionale,* Topographic Database of the Lombardy Region
DEM	Digital Elevation Model

DORIS	*Détermination d'Orbite et Radiopositionnement Intégré par Satellite,* Doppler Orbitography and Radiopositioning Integrated by Satellite
DPA	Department for Public Administration
EAFRD	European Agricultural Fund for Rural Development
EAGF	European Agricultural Guarantee Fund
EO	Earth observation
ESA	European Space Agency
ETM	Enhanced Thematic Mapper
GDF	Geographic Data File
Geo	Group on Earth Observations
GeoTIFF	Geostationary Earth Orbit Tagged Image File Format
GIS	Geographic Information System
GLC	Global Land Cover
GMTED2010	Global Multi-resolution Terrain Elevation Data 2010
GOCE	Gravity field and steady-state Ocean Circulation Explorer
GPC	Ground Control Point
GPL	General Public License
GPS	Global Positioning System
GPX	GPS eXchange Format
GRASS	Geographic Resources Analysis Support System
GTFS	General Transit Feed Specification
HTML	HyperText Markup Language
ICA	International Cartographic Association
IGM	*Istituto Geografico Militare*, Geographic Military Institution
INSPIRE	Infrastructure for Spatial Information in Europe
IODL	Italian Open Data License
ISO	International Organization for Standardization
ISPRA	*Istituto Superiore per la Protezione e la Ricerca Ambientale,* Italian National Institute for Environmental Protection and Research

ISPRS	International Society for Photogrammetry and Remote Sensing
ISSDQ	International Symposium on Spatial Data Quality
ISTAT	*Istituto nazionale di statistica,*
JPEG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
LP DAAC	Land Processes Distributed Active Archive Centre
NAOMI	New AstroSat Optical Modular Instrument
NGO	Non-Governmental Organization
NMA	National Mapping Agency
OD	Open Data
OGP	Open Government Partnership
OSGeo	Open Source Geospatial Foundation
OSM	OpenStreetMap
PA	Public Administration
PAC	*Pubblica Amministrazione Centrale*, Central Public Administration
PAL	*Pubblica Amministrazione Locale*, Local Public Administration
PLoS	Public Library of Science
PSI	Public Sector Information
QGIS	Quantum GIS
RAPu	*Rete Archivi Piani urbanistici,*
RDF	Resource Description Framework
RNDT	*Repertorio Nazionale dei Dati Territoriali*, National Inventory of the Territorial Data
SHP	Shapefile
SIAN	*Sistema Informativo Agricolo Nazionale*, Agricultural Informative National System
SOSE	*Soluzioni per il Sistema Economico Pubblico e Privato,* Solutions for the Public and Private Economical System
SPOT	*Satellite Pour l'Observation de la Terre*, Earth observation satellite
SRTM	Shuttle Radar Topography Mission

TM	Thematic Mapper
txt	Text
USGS	United States Geological Survey
UTM	Universal Transverse Mercator
VGI	Volunteered Geographic Information
WCS	Web Coverage Service
WFS	Web Feature Service
WGS84	World Geodetic System 1984
WMS	Web Map Service
XLS	Microsoft Excel file
XLSX	Microsoft Excel file
XML	eXtensible Markup Language

BIBLIOGRAPHY

[1] Auer, Michael, and Alexander Zipf. “How do Free and Open Geodata and Open Standards fit together?” *From Sceptisim versus high Potential to real Applications*, 2009.

[2] Bédard, Yvan, and Denis Vallière. “Qualité des données à référence spatiale dans un contexte gouvernemental.” 1995.

[3] Benoît, David, and Pascal Fasquel. “Qualité d'une base de données géographique: concepts et terminologie.” 1997.

[4] Biallo, Giovanni. “Dati Geografici ed OpenData.” 2012.

[5] Boin, Anna T., and Gary J. Hunter. “What communicates quality to the spatial data consumer.” *Proceedings of the 2007 International Symposium on Spatial Data Quality (ISSDQ 2007)*, 2007.

[6] Bonney, Rick, et al. “Citizen Science: a developing tool for expanding science knoledge and scientific literacy.” *BioScience*, 2009: 977-984.

[7] Boulton, Geoffrey, Michael Rawlins, Patrick Vallance, and Mark Walport. “Science as a public enterprise: the case for open data.” *The Lancet*, 2011: 1633-1635.

[8] Braunschweig, Katrin, Julian Eberius, Maik Thiele, and Lehner Wolfgang. “The state of open data.” *Proceedings of the 24th International Conference on World Wide Web*, 2012.

[9] Brovelli, Maria Antonia, Marco Minghini, Monia Elisa Molinari, and Miriam Molteni. “Do open geodata actually have the quality they declare? the case study of Milan, Italy.” *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2016: 609-614.

[10] Carrara, Wendy, Wae San Chan, Sander Fischer, and Eva van Steenbergen. *Creating value through open data: Study on the impactof re-use of public data resources.* European commission, 2015.

[11] Čebyšëv, Pafnutij L'vovič. “Des valeurs moyennes.” *Journal de mathématiques pures et appliquées*, 1867: 177–184.

[12] Coleman, David J, Yola Georgiadou, Jeff Labonte, and others. "Volunteered Geographic information: the nature and motivation of produser." *International Journal of Spatial Data Infrastructures Research*, 2009: 332-358.

[13] Crăciunescu, Vasile, and Codrina Maria Ilie. "Free our maps." *W3C*, 2013.

[14] Davies, Tim. "Open data barometer: 2013 global report." *World Wide Web Foundation and Open Data Institute*, 2013.

[15] Fisher, Sander, Eva van Steenbergen, and Wendy Carrara. "Open data maturity in europe 2015, insights into the european state of play." (European commission) 2015.

[16] Flanagin, Andrew J., and Miriam J. Metzger. "The credibility of volunteered geographic information." *GeoJournal*, 2008: 137-148.

[17] Getis, Arthur, and Keith Ord. "The analysis of spatial association by use of distance statistics." *Geographical analysis*, 1992: 189-206.

[18] Gewin, Virginia. "Data sharing: An open mind on open data." *Nature*, 2016: 117-119.

[19] Girres, Jean-François, and Guillaume Touya. "Quality assessment of the French OpenStreetMap dataset." *Transactions in GIS*, 2010: 435-459.

[20] Goodchild, Michael F. *Fundamentals of spatial data quality*. Rodolphe Devillers and Robert Jeansoulin, 2010.

[21] Goodchild, Michael Frank. "Citizens as sensors: the world of volunteered geography." *GeoJournal*, 2007: 211-221.

[22] Grira, Joel, Yvan Bédard, and Stéphane Roche. "Spatial data uncertainty in the VGI world: Going from consumer to producer." *Geomatica*, 2010: 61-72.

[23] Haklay, Mordechai. "How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England." *Environ Planning*, 2010: 682-703.

[24] Halonen, Annti. "Being Open About Data Analysis of the UK open data policies and applicability of open data." *The Finnish Institute in London*, 2012.

[25] Harrison, Teresa M., and Lisa Falvey. "Democracy and new communication technologies." *Annals of the International Communication Association*, 2001: 25(1):1-43.

[26] Hecht, Robert, Carola Kunze, and Stefan Hahmann. "Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time." *ISPRS International Journal of Geo-Information*, 2013: 1066-1091.

[27] Hunter, Gary, and Kate Beard. "Understanding error in spatial databases." *Australian surveyor*, 1992: 108-119.

[28] Infrastructure and Transport Ministry. "Nuovo codice della strada." 1992.

[29] "ISO 19157 - Geographic information, Data quality." 2013.

[30] "ISO 2859 - Sampling procedures for inspection by attributes." 1989.

[31] "ISO 3534-2: Statistics, Vocabulary and symbols, Part 2: Applied statistics." 2006.

[32] "ISO 3951-1: Sampling procedures for inspection by variables." 2013.

[33] "ISO 9000 - Quality management system, Fundamentals and vocabulary." 2000.

[34] Korzybski, Alfred. "A Non-Aristotelian System and its Necessity for Rigour in Mathematics and Physics, in "Science and Sanity"." *Institute of General Semantics*, 1933: 747-761.

[35] Koski, Heli. "Does marginal cost pricing of public sector information spur firm growth." *Discussion Paper*, 2011: 1260.

[36] Lakomaa, Erik, and Jan Kallberg. "Open data as a foundation for innovation: The enabling effect of free public sector information for entrepreneurs." *IEEE Access*, 2013: 1:558-563.

[37] Magee, Andrew F., Michael R. May, and Brian. R Moore. "The dawn of open access to phylogenetic data." *PLoS One*, 2014.

[38] Miller, Paul, Rob Styles, and Tom Heath. "Open Data Commons, a License for Open Data." *LDOW*, 2008: 369.

[39] Moran, Patrick Alfred Pierce . "Notes on Continuous Stochastic Phenomena." *Biometrika*, 1950: 17-23.

[40] Morozov, Evgeny. *The net delusion: The dark side of Internet freedom*. PublicAffairs, 2012.

[41] Murray-Rust, Peter. "Open data in science." *Serials Review 34*, 2008: 52-64.

[42] Natale, Domenico. "Data quality e Open data." *Commissione UNINFO JTC1/SC7 Software Engineering*, 2011.

[43] *Nature*. "Closing the Climategate." 17 November 2010.

[44] Odland, John. *Spatial autocorrelation*. Regional Research Institute: West Virginia University, 1985.

[45] O'reilly, Tim. "What is Web 2.0: Design patterns and business models for the next generation of software." *Communications & strategies*, 2007: 1-17.

[46] Perritt, Henry H. Jr. "Open government." *Government Information Quarterly*, 1997: 14(4):397-406.

[47] Piotrowski, Suzanne J. *Governmental transparency in the path of administrative reform*. SUNY Press, 2008.

[48] Regione Lombardia. *Specifiche Tecniche aerofotogrammetriche per la realizzazione del Data base topografico alle scale 1:1.000 e 1:2.000*. 2007.

[49] Robinson , David G., Harlan Yu, William P. Zeller, and Edward W. Felten. "Government data and the invisible hand." *Yale Journal of Law & Technology*, 2009: 11:160.

[50] Salvemini, Mauro. "Gli open (geo) data sono per i cittadini o per chi provvede servizi ai cittadini?" *GEOmedia*, 2015: 19(1).

[51] Spalla, Anna, and Riccardo Galletto. *Fondamenti di cartografia numerica*. 2000.

[52] Tòth, Katalin, and Robert Tomas. "Quality of geographic information - simple concept made complex by the context." *Proceedings of the 25th internationa cartographic conference and the 15th general assembly of the Internationa Cartographic Association*, 2011.

[53] Van Noorden, Richard. "Confusion over publisher's pioneering open-data rules." *Nature*, 2014: 478.

[54] Zielstra, Dennis, and Alexander Zipf. "A comparative study of proprietary geodata and volunteered geographic information for Germany." *13th AGILE international conference on geographic information science*, 2010.

SITOGRAPHY

[1] AGEA. *Chi siamo: Agenzia per le Erogazioni in Agricoltura*. 04 October 2013. http://www.agea.gov.it/portal/page/portal/AGEAPageGroup/HomeAGEA/ChiSiamo (accessed December 09, 2016).

[2] AGEA. *Regole tecniche per la formazione, la documentazione e lo scambio di ortofoto digitali alla scala nominale 1:10'000*. 28 June 2012. http://www.centrointerregionale-gis.it/catalogue/seminario_28_giugno_2012/presentazione%20AGEA%20ortofoto_MA.pdf (accessed December 09, 2016).

[3] Alabaster, Jay. *Google crosses line with controversial old Tokyo maps*. 2009. http://www.japantimes.co.jp/news/2009/05/05/national/google-crosses-line-with-controversial-old-tokyo-maps/#.WNI5VDs1-00 (accessed September 25, 2016).

[4] *ArcGIS*. 2017. http://desktop.arcgis.com/en/ (accessed February 10, 2017).

[5] *CC0 Creative Commons*. 2016. https://creativecommons.org/share-your-work/public-domain/cc0/(accessed February 10, 2017).

[6] *CC-BY Creative Commons*. 2016. https://creativecommons.org/licenses/by/2.0/it/deed.en (accessed February 10, 2017).

[7] *CC-BY-NC-SA 3.0 IT Creative Commons*. 2016. https://creativecommons.org/licenses/by-nc-sa/3.0/it/deed.en (accessed February 10, 2017).

[8] *CC-BY-SA 3.0 IT Creative Commons*. 2016. https://creativecommons.org/licenses/by-sa/3.0/it/ (accessed February 10, 2017).

[9] *Creative Common* . 2016. https://creativecommons.org/licenses/?lang=en (accessed February 15, 2017).

[10] "CORINE Land Cover technical guide." *European Environmental Agency*. 2000. http://www.eea.europa.eu/publications/tech40add (accessed January 31, 2017).

[11] ERSAF. “Uso del suolo in regione Lombardia. Atlante descrittivo.” *Regione Lombardia.* 2007. http://www.territorio.regione.lombardia.it/shared/ccurl/889/766/Atlante%20Regione%202010.pdf (accessed January 31, 2017).

[12] ESRI. *Modeling spatial relationships.* 2017. http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/modeling-spatial-relationships.htm (accessed February 20, 2017).

[13] *European Data Portal, What we do.* 2016. https://www.europeandataportal.eu/en/what-we-do (accessed February 2, 2017).

[14] G8 leaders. *gov.uk.* 2013. https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex (accessed October 2016, 13).

[15] *GeoDa.* 2017. http://geodacenter.github.io/ (accessed February 10, 2017).

[16] *Geoportale della Regione Lombardia.* 2016. http://www.geoportale.regione.lombardia.it/ (accessed 11 2016, 11).

[17] *GRASS GIS.* 2017. https://grass.osgeo.org/ (accessed February 10, 2017).

[18] *Il Geoportale Nazionale.* 2016. http://www.pcn.minambiente.it/GN/ (accessed November 11, 2016).

[19] “Il Terzo Piano Nazionale d’Azione.” *Italia Open Gov.* 20 September 2016. http://open.gov.it/terzo-piano-dazione-nazionale/ (accessed December 2016, 15).

[20] Italian National Cartographic Geoportal. *Date ortofoto a colori AGEA anni 2009-2012: Geoportale Nazionale.* 2014. http://www.pcn.minambiente.it/geoportal/catalog/search/resource/details.page?uuid=%7B6CAE31BA-0BB5-4C2C-9837-147F334B2DCB%7D (accessed December 09, 2016).

[21] Italian National Cartographic Geoportal. *Pubblicati i servizi relativi alle ortofoto AGEA 2009-2012.* 2014. http://www.pcn.minambiente.it/GN/archiviati/archivio/2022-pubblicati-i-servizi-relativi-alle-ortofoto-agea-2009-2012 (accessed December 09, 2016).

[22] *Italian Open Data License v2.0.* 2016. http://www.dati.gov.it/iodl/2.0/ (accessed January 10, 2017).

[23] Open Data Insitute. *Supporting sustainable development with Open Data.* 2015. http://theodi.org/supporting-sustainable-development-with-open-data (accessed January 16, 2017).

[24] *Open Government Partnership.* 2 October 2016. http://www.opengovpartnership.org/about (accessed October 18, 2016).

[25] Open Knowledge Foundation. “Open Data Handbook Documentation, Release 1.0.0.” *Open Data Handbook.* 14 November 2012. http://opendatahandbook.org/guide/en/ (accessed January 15, 2017).

[26] Pollock, Rufus, and Katelyn Rogers. “How the UK Government Saved £4 million in 15 minutes with Open Data.” *http://opendatahandbook.org/.* http://opendatahandbook.org/value-stories/en/saving-4-million-pounds-in-15-minutes/ (accessed January 16, 2017).

[27] *Pubblic Library of Science*. 20 October 2016. https://www.plos.org/(accessed December 12, 2016).

[28] QGIS. *QGIS user guide*. 2014. http://docs.qgis.org/2.14/en/docs/user_manual/ (accessed January 26, 2017).

[29] Repertorio Nazionale dei Dati Territoriali. *Regole tecniche*. 03 July 2015. http://www.rndt.gov.it/RNDT/home/index.php?option=com_content&view=article&id=143&Itemid=241 (accessed December 09, 2016).

[30] Rogers, Katelyn. "Danish address registry." *http://opendatahandbook.org/*. 2010. http://opendatahandbook.org/value-stories/en/danish-address-registry/ (accessed January 16, 2017).

[31] *The international Open Data Charter*. http://opendatacharter.net/ (accessed March 16, 2016).

[32] *The open definition*. http://opendefinition.org/ (accessed April 16, 2016).

[33] "Tecnical specification." *Lombardy Region Geoportal*. 2016. http://www.geoportale.regione.lombardia.it/specifiche-tecniche (accessed February 14, 2017).

[34] *World Justice Project*. http://worldjusticeproject.org/open-government-index/open-government-partnership (accessed October 3, 2016).

APPENDIX 1
PHYTON SCRIPT

```
#__-*-_coding:utf-8_-*-

import sys
import math
import time
import grass.script as grass


 #Calculate the area of a input data

def area(data):
   feat_data = int(((grass.read_command("v.info",
   map=data,flags="t")).split("\n")[5]).split("=")[1])
   if feat_data>0:
      area_data = grass.read_command("v.to.db",map=data,option="area",flags="p")
      s_data=0
      l_data = area_data.split("\n")
      for item in l_data[1:-1]:
         if float(item.split("|")[1])<0:
            print item.split("|")[1]
         s_data+=float(item.split("|")[1])
   else:
      s_data=0
   return s_data

 #Procedure

def differenza(layer):
   #Difference between DBTR and DUSAF
   grass.run_command("v.overlay", ainput="linee", atype="line", binput=layer,
   output="Buffer1", operator="not", overwrite=True)
   #10m buffer
   grass.run_command("v.buffer", input="Buffer1", output="Buffer10",
   flags="c",distance=10, overwrite=True)
   #Differences between street area and 10m buffer
   grass.run_command("v.overlay", binput="Buffer10", ainput="area", output="diff",
   operator="not", overwrite=True)
```

```
    # Result (DBTR not DUSAF)
    grass.run_command("v.overlay", ainput="diff", binput=layer,
     output="Risultato_%s"%layer, operator="not", overwrite=True)
    #ResultII (overlap between DBTR and DUSAF)
    grass.run_command("v.overlay", ainput="diff", binput=layer,
    output="RisultatoII_%s"%layer, operator="and", overwrite=True)
    #ResultIII (DUSAF not DBTR)
    grass.run_command("v.overlay", ainput=layer, binput="diff",
    output="RisultatoIII_%s"%layer, operator="not", overwrite=True)


def main():

    grass.run_command("v.in.ogr", input="home/thesis/PythonScript/area.shp",
      overwrite=True, flags="o")
    grass.run_command("v.in.ogr", input="home/thesis/PythonScript/area.shp",
      overwrite=True, flags="o")
    grass.run_command("v.in.ogr", input="home/thesis/PythonScript/DUSAF.shp",
      overwrite=True, flags="o")
    grass.run_command("v.in.ogr", input="home/thesis/PythonScript/grid.shp",
      overwrite=True, flags="o")

    differenza("DUSAF")
    grass.run_command("v.buffer", input="DUSAF", output="DUSAF2", distance=2,
      overwrite=True)
    differenza("DUSAF2")
    grass.run_command("v.buffer", input="DUSAF", output="DUSAF4", distance=4,
      overwrite=True)
    differenza("DUSAF4")
    grass.run_command("v.buffer", input="DUSAF", output="DUSAF6", distance=6,
      overwrite=True)
    differenza("DUSAF6")

    for item in range (0,98):
        grass.run_command("v.extract", input="grid", output="gridd_%s"%item,
         where="CODE=%s"%item, overwrite=True)

        #DUSAF
        grass.run_command("v.overlay", ainput="gridd_%s"%item, binput="DUSAF",
      operator="and", output="Dusaf_nella_cella_%s"%item, overwrite=True)
        areaDusaf=area("Dusaf_nella_cella_%s"%item)
        grass.run_command("v.db.update", map="grid", column="DUSAF_STR_",
      where="CODE=%s"%item, value=areaDusaf)

        #DBTR
        grass.run_command("v.overlay", ainput="gridd_%s"%item, binput="area",
      operator="and", output="DBT_nella_cella_%s"%item, overwrite=True)
        areaDBT=area("DBT_nella_cella_%s"%item)
        grass.run_command("v.db.update", map="grid", column="DBT",
      where="CODE=%s"%item, value=areaDBT)



      #Results for DUSAF, 2m buffer on DUSAF, 4m buffer on DUSAf and 6m buffer on DUSAF
```

```
 grass.run_command("v.overlay", ainput="gridd_%s"%item, binput="Risultato_DUSAF",
operator="and", output="Ris_nella_cella_%s"%item, overwrite=True)
 areaDiff=area("Ris_nella_cella_%s"%item)
 grass.run_command("v.db.update", map="grid", column="RIS",
where="CODE=%s"%item, value=areaDiff)

 grass.run_command("v.overlay", ainput="gridd_%s"%item,
binput="Risultato_DUSAF2", operator="and", output="Ris2_nella_cella_%s"%item,
overwrite=True)
 areaDiff=area("Ris2_nella_cella_%s"%item)
 grass.run_command("v.db.update", map="grid", column="RIS2",
where="CODE=%s"%item, value=areaDiff)

 grass.run_command("v.overlay", ainput="gridd_%s"%item,
binput="Risultato_DUSAF4", operator="and", output="Ris4_nella_cella_%s"%item,
overwrite=True)
 areaDiff=area("Ris4_nella_cella_%s"%item)
 grass.run_command("v.db.update", map="grid", column="RIS4",
where="CODE=%s"%item, value=areaDiff)

 grass.run_command("v.overlay", ainput="gridd_%s"%item,
binput="Risultato_DUSAF6", operator="and", output="Ris6_nella_cella_%s"%item,
overwrite=True)
 areaDiff=area("Ris6_nella_cella_%s"%item)
 grass.run_command("v.db.update", map="grid", column="RIS6",
where="CODE=%s"%item, value=areaDiff)

#Results II for DUSAF, 2m buffer on DUSAF, 4m buffer on DUSAf and 6m buffer on
DUSAF
 grass.run_command("v.overlay", ainput="gridd_%s"%item,
binput="RisultatoII_DUSAF", operator="and", output="Ris_nella_cella_%s"%item,
overwrite=True)
 areaDiff=area("Ris_nella_cella_%s"%item)
 grass.run_command("v.db.update", map="grid", column="RIS_II",
where="CODE=%s"%item, value=areaDiff)

 grass.run_command("v.overlay", ainput="gridd_%s"%item,
binput="RisultatoII_DUSAF2", operator="and", output="Ris2_nella_cella_%s"%item,
overwrite=True)
 areaDiff=area("Ris2_nella_cella_%s"%item)
 grass.run_command("v.db.update", map="grid", column="RIS_II2",
where="CODE=%s"%item, value=areaDiff)

 grass.run_command("v.overlay", ainput="gridd_%s"%item,
binput="RisultatoII_DUSAF4", operator="and", output="Ris4_nella_cella_%s"%item,
overwrite=True)
 areaDiff=area("Ris4_nella_cella_%s"%item)
 grass.run_command("v.db.update", map="grid", column="RIS_II4",
where="CODE=%s"%item, value=areaDiff)

 grass.run_command("v.overlay", ainput="gridd_%s"%item,
binput="RisultatoII_DUSAF6", operator="and", output="Ris6_nella_cella_%s"%item,
overwrite=True)
 areaDiff=area("Ris6_nella_cella_%s"%item)
 grass.run_command("v.db.update", map="grid", column="RIS_II6",
where="CODE=%s"%item, value=areaDiff)
```

```
            #Result III for DUSAF, 2m buffer on DUSAF, 4m buffer on DUSAf and 6m buffer on
            DUSAF
             grass.run_command("v.overlay", ainput="gridd_%s"%item,
            binput="RisultatoIII_DUSAF", operator="and", output="Ris_nella_cella_%s"%item,
            overwrite=True)
             areaDiff=area("Ris_nella_cella_%s"%item)
             grass.run_command("v.db.update", map="grid", column="RIS_III",
            where="CODE=%s"%item, value=areaDiff)

             grass.run_command("v.overlay", ainput="gridd_%s"%item,
            binput="RisultatoIII_DUSAF2", operator="and", output="Ris2_nella_cella_%s"%item,
            overwrite=True)
             areaDiff=area("Ris2_nella_cella_%s"%item)
             grass.run_command("v.db.update", map="grid", column="RIS_III2",
            where="CODE=%s"%item, value=areaDiff)

             grass.run_command("v.overlay", ainput="gridd_%s"%item,
            binput="RisultatoIII_DUSAF4", operator="and", output="Ris4_nella_cella_%s"%item,
            overwrite=True)
             areaDiff=area("Ris4_nella_cella_%s"%item)
             grass.run_command("v.db.update", map="grid", column="RIS_III4",
            where="CODE=%s"%item, value=areaDiff)

             grass.run_command("v.overlay", ainput="gridd_%s"%item,
            binput="RisultatoIII_DUSAF6", operator="and", output="Ris6_nella_cella_%s"%item,
            overwrite=True)
             areaDiff=area("Ris6_nella_cella_%s"%item)
             grass.run_command("v.db.update", map="grid", column="RIS_III6",
            where="CODE=%s"%item, value=areaDiff)

# Remove the not useful layers
grass.run_command("g.remove",type="vect",name="gridd_%s,Dusaf_nella_cella_%s,DBT_nella
_cella_%s,Ris2_nella_cella_%s,Ris_nella_cella_%s,Ris4_nella_cella_%s,Ris6_nella_cella_%s"
%(item,item,item,item,item,item,item),flags="f")

          # Export the grid as shapefile
          grass.run_command("v.out.ogr", input="grid",
          output="/home/thesis/PythonScript/results/", format="ESRI_Shapefile",
          overwrite=True)




      if __name__ == "__main__":
          options,flags = grass.parser()
          sys.exit(main())
```