WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

EFMD
EQUIS
ACCREDITED

Master Thesis

Unveiling the analytical power of Open Street Map

Michael Astleitner

Date of Birth: 12.03.1987
Student ID: 1251189

Subject Area: Geographical Information Systems

Studienkennzahl: J066925

1.Supervisor: Prof. Dr. Axel Polleres
2.Supervisor: Monica Posada-Sanchez, MSc.

Date of Submission: April 24, 2017

Department of Information Systems and Operations, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

DEPARTMENT FÜR INFORMATIONS-
VERARBEITUNG UND PROZESS-
MANAGEMENT DEPARTMENT
OF INFORMATION SYSTEMS AND
OPERATIONS

Contents

List of Figures

List of Tables

Abstract

Open Street Map is becoming increasingly more popular among companies and citizens. As a product of volunteered geographic information, it offers a vast amount of potential functionality that can be applied for different usages. The data of Open Street Map is created and changed by non-domain experts. Therefore, it is prone to false and ambiguous information provided by those. This thesis will provide different comparisons between Open Street Map and geographic data sets, which are made available by European and local authorities, in order to evaluate the trustworthiness of Open Street Map. Comparisons will be done within the scope of predefined use cases and offer insights on aspects like shapes, descriptive and historical information of geographical objects within the European Union.

1 Introduction

In times of big data, where the amount of data is saved in various forms and repositories, the importance of having high quality analysis of this information is an important issue.[25] This highly depends on the quality that is provided by repositories that contain big data, since the data quality is the driving factor when it comes to the outcomes of the important analysis.[32] The last decades geospatial technologies have greatly evolved and became a part of every day use for citizens.[26][27] Furthermore, they become more and more important over time as citizens rely on them on a daily basis for their routine.

Accuracy is important for such maps, both its correctness and completeness. In order to keep those maps as accurate as possible, companies offer their services for money. Nevertheless, there is an alternative for developers and companies that want to use open source and therefore are not obligated to pay for the service of having a map in their application.

There are maps that are using information from volunteers across the world. Data, that is gathered in such a way for geographical information by the community, is called volunteered geographic information (VGI).

One of those VGI projects is Open Street Map (OSM), which is the core focus of this research. In order to analyze the current state of OSM in regards of their correctness and completeness, this thesis shows outcomes of comparing OSM data with data released from European authorities.

The work will show common problems in OSM, but also provide an outlook of the analytic potential that lies within OSM and their community behind it.

1.1 Motivation

The importance of geographical information has risen in the last decade due to the rapid growth and need of digital spatial data.[5] Therefore, their reliability and availability needs to be observed to satisfy user expectations and requirements.

In this regard, OSM offers a large potential in various fields that are desirable for many users, companies and governments. First and foremost, OSM can be freely accessed and used. This enables already a majority of users to enjoy services of various apps that use OSM for free or at least at a lower price in comparison to applications that use a commercial provider for accessing their geographical information.

In addition, as increasingly more users are participating in creating and editing data in OSM, the reaction time to environmental changes, compared to commercial providers, is potentially faster due to the large number of contributors on OSM (see figure 3). Interesting aspects of the motivational factors of why people are willing to spend their free time on collecting and providing geographical data is explained in [17], where further points on the benefit of real time updates due to citizens is pointed out in cases of e.g. natural catastrophes or roads being blocked. This is an advantage compared to data that is collected by businesses or authorities, as their data might take years until it is published and is consequently outdated the day it is published.

1.2 Problem statement

Businesses and authorities that provide geographical information, potentially offer reliable data, as their staff is trained to collect the given data.

This is an essential difference compared to OSM, where the VGI data is is generated and modified by contributors regardless their expertise. Those users are usually non-domain experts when it comes to collecting geographic information, therefore, it is prone to flaws in their completeness and accuracy. Moreover, geographic information consists of geometric and alphanumeric information. This adds another level of sophistication that OSM users should take care of and leaves another possibility for false, ambiguous or missing information that the OSM users might not capture.

1.2.1 Research questions

Scope of this master thesis is to evaluate the current correctness and completeness of information provided by OSM. The idea, on how to evaluate such quality criteria for OSM, is to formulate use cases that can be com-

pared against a source of high accuracy. Several geographical objects are identified, selected and further mapped to match the identical location of the other sources data environment.
The selection of those objects is arbitrary, as it is only important to have a sample size that is representative for an overall statement about the quality state OSM is currently in.
Furthermore, historical data, if available, will be analyzed to potentially demonstrate a certain development of OSM over time. As OSM is prone to ambiguous and missing information, which is shortly pointed out in subsection 1.2, having a historical course of the development of OSM could expose valuable information of user aspects.

Specifically, in this master thesis, the following research questions were formulated:

- Is Open Street Map offering as precise data of regional areas, within the European Union, as geospatial data of Eurostat does?
- Can Open Street Map be considered a reliable source for geospatial data when compared to official data from Eurostat and Copernicus?
- Do the users, which provide VGI, use the full potential of OSM to provide sophisticated information about the environment?

Together, all three research questions aim to identify the trustworthiness of OSM data by comparing it with data from an official source of European authorities.

1.3 Structure of the work

The remainder of this thesis is structured as follows:

Chapter 2 will explain GIS (Geographical Information System) and OSM in general and how the data is structured. After getting to know the structure of the data sources, a glimpse on how to query for the data will be illustrated. Furthermore, insights on how completeness and correctness of geographical information defined, will be provided. Throughout this chapter, a general overview, on what was already done by other researchers in regard of the quality of OSM and reasoning why certain tools, environments and data formats were chosen, will be discussed.

Chapter 3 introduces the authorities and sources used throughout the research. Moreover, the understanding of how they collect data will be fostered by providing insights on their methods.

Chapter 4 fosters the understanding on the difficulties of retrieving data and how to tackle those. Moreover, a closer look on a number of tools will be presented that were used to identify specific objects within the data.

Chapter 5 follows with analytically discussing the results of the geometric and alphanumeric information, which were initially identified. Consequently, to verify the completeness and correctness, geographical information of OSM will be compared with data sets from European authorities. Therefore, this chapter will include calculations and methods that are summarized in various outputs for the comparison and exploration of the geographical objects.
Furthermore, it reports comprehensive findings by further analyzing the data by using the outcomes of this chapter. Consequently, after the analysis is done, a link to the research questions, which are stated in subsection 1.2.1, will be made in order to answer them.

Chapter 6 concludes the work and offers aspects that can be of subject in future works.

2 Preliminaries

Herein, several environments that are used and mentioned throughout the thesis, will be introduced. In addition, insights on how those environments are structured and retrieved will be provided. This is of utmost importance in order to fundamentally understand how correctness and completeness are measured. Consequently, when allocating the collection of data for those environments, quality aspects will be more clear. This is needed for further understanding the decision making process of why certain environments and tools are used for fostering the research.

Combining all the preliminary details forms a good fundamental overview about OSM and GIS in general, which is necessary for understanding later steps that are taken for deeper analysis of geospatial data and user aspects.

2.1 Introducing GIS

There are several definitions for GIS. Therefore, it is of help to analyze some of the most common ones and derive the essence that is necessary for building a proper fundamental understanding of what is the purpose of GIS.

In [26] geographical information systems are defined as a special class of information systems that not only "keep track of events, activities, and things, but also of where these events, activities , and things happen or exist."

In a more concrete sense, [18] points out that geographic information systems are designed to capture, store, display, communicate, transform, analyze, and archive georeferenced information, that is, information tied to specific locations on the Earth's surface.

Furthermore, "they associate locations in space, and often in space-time, with properties such as temperature, population density, land use, or elevation, and are widely used today in support of research in geography, and in any other disciplines concerned with phenomena on or near the earth's surface." [19]

Those definitions are alike with the difference that they vary in their granularity and how they interpret the location. Meaning, that just stating that something can be found somewhere might not be enough. The second definition already makes clear that a geographical information system identifies a specific location on the surface of the earth. So now the location gets narrowed down to be found on the earth, which defines a limit of locations that can be identified, namely, the surface of the earth.

The third definition adds not only the surface, but also the elevation, including locations near the earth's surface. This extends the possibility of potential locations a lot, but still has a limit by staying near the earth. One

can now ask what exactly is meant by "near the earth", which is till the end of the earth's atmosphere.
Important bottom line of the definitions is a narrowed down practical relevance as stated by [11] that "geographic information is commonly broken into the components of space, time, and attribute."

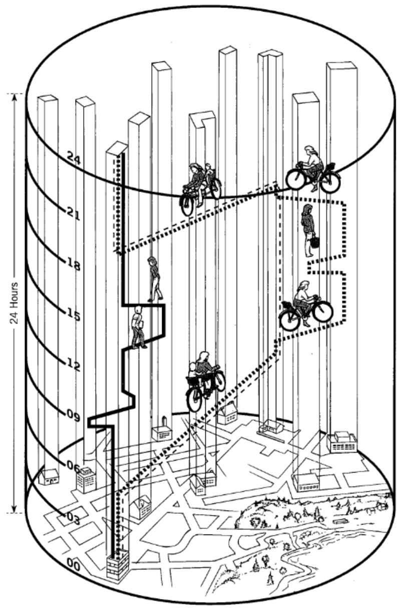


Figure 1: Space-time example redrawn by [30] from the original of [11]

Figure 1 shows the space-time represented as a cylinder, which illustrates nicely how GIS work. The bottom shows the map with its areas, streets and buildings. On the side of the cylinder the time is indicated with a 24 hour period, hence a day.
Inside the cylinder, activities of humans are displayed and rise up in space the further the time advances. Furthermore, when taking a closer look on the buildings of the map, one can see that pillars rise upwards from them. They demonstrate those buildings in space-time and whether any changes are happening like a new building is constructed or another one gets demolished within this time window of 24 hours.
This figure confirms what was analyzed by the definitions above. GIS are more complex, regarding the time and location aspect, as one might think at the first glimpse and further shows that the first definition is not enough to fully understand the broad range of information that is covered by GIS.

Projection

In order to flatten the world for using it as a map, several ways of how to project it have become famous. There are different characteristics on how to choose a map projection stated by [35] and the essence is summarized as follows:

1. Area: Most are designed as equal-area, meaning that any size of an object at one part of the map has the same size on any other part of the map.

2. Shape: An area must still be shown distorted in shape, but its small features are shaped correctly.

3. Scale: No map projection shows the scale correctly. Nevertheless, some lines on the map remain true and can be used, when chosen their location properly, to reduce errors somewhere else. Still, other large errors might still exist.

4. Direction: Maps do provide either the correct relative local directions to points or correctly for all points with respect to the center.

5. Special characteristic: Each projection provides characteristics that none of the others contain.

6. Method of construction: This characteristic is not that relevant anymore, as it was historically of importance when there were no computers that calculate complex formulas fast.

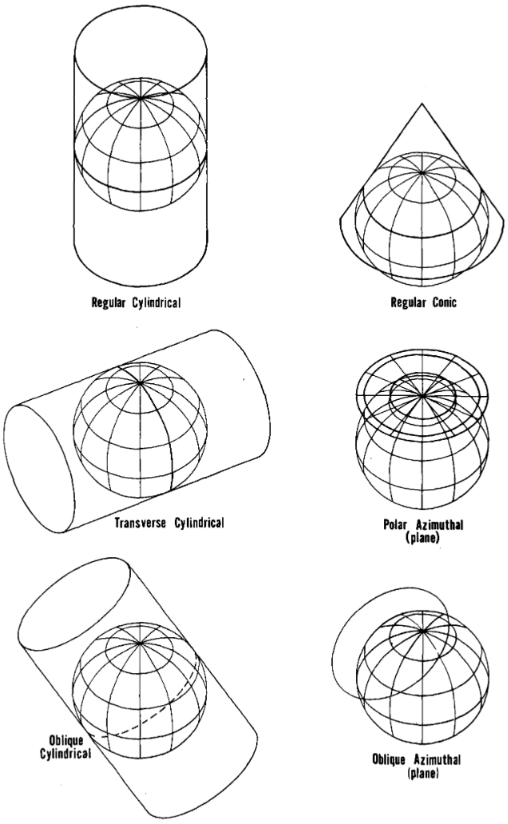


Figure 2: Various ways of projecting the world from [35]

The most used one being the **Mercator projection**, which is the top left projection shown in figure 2, namely, Regular Cylindrical. The reason for the popularity of Mercator is historically related and was originally mainly used for nautical purposes. [28] Mercator uses straight lines, therefore, has straight segments aligning parallel with the meridians.

2.2 Introducing Open Street Map

The Open Street Map, founded in 2004 by Steve Coast at the University College London (UCL), is a free and editable collection of geographical information created by volunteers. Geographical information on OSM not only contains points, polygons and lines, which are available as map images, it also includes the underlying data of those objects on the map.
Decades ago, the role of creating a map was considered to be done by a certain group of experts. Back then it was unimaginable that an non-domain expert could contribute to map the world as it was considered that only highly educated persons, preferably with an university-level degree, could

and should do so.
Nevertheless, this changed during the course of time and GPS (Global Positioning System) receivers became available and affordable for the non-domain expert. Furthermore, the accuracy level of those devices improved a lot and so people were able to check longitude, latitude and altitude of various locations.
However, at the beginning, uploading the collected information wasn't an easy task for most computer users until the GPX (GPS Exchange Format) got established and adopted by the GPS receivers. Furthermore, this allowed to better contribute to projects with user-generated maps, whereas OSM was probably the most popular one that gained the most from those developments and established its position as of today.[21]

The work of [4] further points out the various differences in the activity across all countries about their population participating in the mapping process in OSM. Some countries like Italy, the Netherlands, Kuwait, Croatia, and Liberia are at the top of the list, whereas the numbers of Asian and African countries have only a minor proportion of their population involved in the mapping process.

As of now, OSM is used by a large audience. Citizens, companies and governments use it on a daily basis for various areas of applications. Mostly provided through applications that citizens can use in order to find places (Points of interest), navigate (walk, bike, car etc.) by using GPS or track a certain progress. Those applications use OSM as a basemap in order to provide and support their services.
Another reason for its growing popularity is that it is free. Everyone can basically implement OSM and use it on their e.g. homepage or applications. A vast amount of companies build their business upon OSM and earn money by using OSM functionality. They offer services like consulting for customers who need certain information by using OSM.
Moreover, not only the fraction of users that are using OSM is growing. There are also the registered OSM users, which add or update data to OSM, generating VGI. Those numbers of contributors are increasing every year as well. In addition, also historical data is saved and gets more the more contributors there are that add data, thus a massive data set has summed up over the past years. In [36] this gets analyzed by using heat maps to visually display changes in the historic data.
However, with a growing number of users there is also a high potential of undesirable things to happen with the data, which add misinformation or ambiguity.

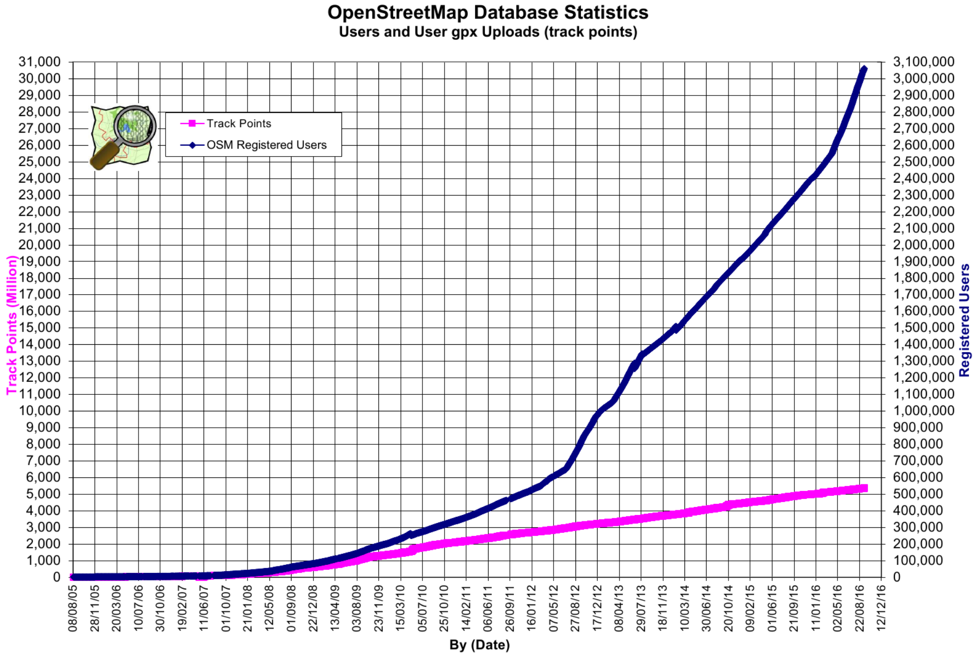


Figure 3: OSM statistics about accumulated user registration and GPX uploads[8]

This is underlined with the provided information of figure 3. The figure provides the accumulated growing number of contributors and GPX uploads. The number of contributors is growing very fast on a daily basis.

Milestones being:

- End of 2012 by reaching one million users
- February 2015 reaching around two million users
- End of August 2016 reaching around three million users

The gap between the million user jumps is clearly getting smaller, which confirms the growing trend in the interest of OSM.

Interestingly, the track points do not to grow in the same way the contributors do. Even though the user numbers are growing in a nearly exponential way, the track points are showing a more linear trend. This could be interpreted in various directions. One might say that at one point the majority

[8] http://wiki.openstreetmap.org/wiki/File:Osmdbstats1.png, last accessed 2017-02-13

of the world's geospatial information is collected and uploaded and therefore only maintenance is required. Another interpretation could be that the number of contributors is growing, but the majority might not be active and therefore only some core OSM users are constantly providing information via uploads.

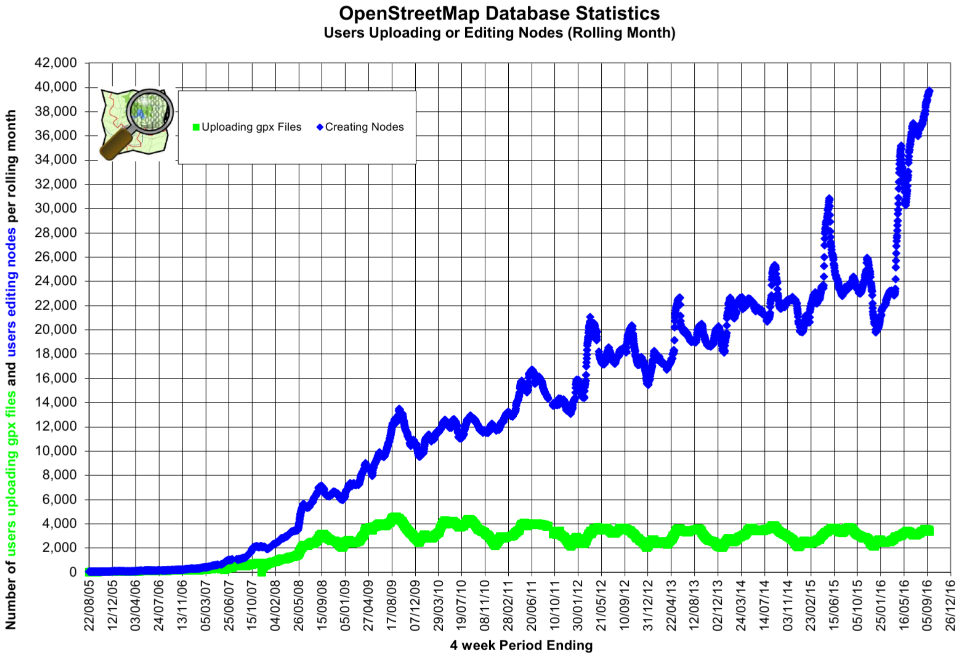


Figure 4: OSM statistics about contributors editing nodes or uploading GPX nodes[9]

Figure 4 provides interesting insights into the direction in which the user behavior is trending. In fact, the number of uploads of new nodes shows a constant pattern between 2000 and 4000 on a monthly basis. However, the number of nodes that get edited has a lot of ups and downs, but shows a general rising trend. This supports the statement above, that there is now more work that needs to be done in regards of the maintenance of the current geospatial information. The reason why the number of edits is growing could be due to actual environmental changes or corrections of the already existing information, because of some wrong input from the user before (creator). Another reason might be, because of the growing number of contributors provided in figure 3 that there are simply more users required for keeping

[9] http://wiki.openstreetmap.org/wiki/File:Osmdbstats1.png, last accessed 2017-02-13

OSM on a more up to date level.
Nevertheless, as the contributors can basically be anybody, as anybody can simply register at OSM and start adding and updating data right away, it is also prone to errors and ambiguity when more and more contributors are joining OSM.
OSM also does not have supervisors who double check the entered data and since not all users might be trained in observing their area, therefore, might enter wrong information, which needs to be corrected at some point.

2.2.1 Structure of Open Street Map

For users, of e.g. applications that use OSM, it is sufficient to see the map and being able use its features as mentioned in subsection 2.2. This is the intended way as users of applications, in general, don't like to bother with complex technical details, therefore, prefer usability within their applications they use.
However, beneath OSM, as for all software services and applications, there is an underlying structure in order to operate. This standardized structure is required to allow OSM to work properly. With the structure of OSM, it is possible to draw the maps. It describes the location of certain points and how they relate to others. Furthermore, when certain objects are drawn in order to form e.g. a building, it connects, by referencing, locations with each other in order to interpret and display the desired object.

Nevertheless, being able to draw and display objects on a map is not sufficient enough for the average user. When thinking back several decades, it used to be enough to have a map, for the user of it, to find a location or information they sought after.
Even then, if thinking of a street map or atlas, there was an index that enabled faster searching as it provided the page number and, if available, X and Y coordinate information for the grid on that page.

Nowadays, it is important to find locations fast and easy with geographical services. In addition, it is not required anymore, compared to older GPS systems for cars, to know the exact address or location for a desired result. In fact, knowing or remembering certain addresses gets less and less, as with today's applications it isn't really required anymore. Users simply want to search for a term that provides them search results from which they can chose from.
Therefore, users don't want to write addresses anymore, hence it is of the essence for services, which provide geographical information, to be able to

process arbitrary search terms and relate them to locations. A simple example of this could be that a user searches for e.g. a grocery store and retrieves a list of nearby stores. Furthermore, services should provide advanced searches or provide more sophisticated results such as i.e. opening hours of said stores. In order to make this possible, a standardized structure is a must. A closer look on how OSM realizes this, will now be explained.

Tag system

In principal a tag consists of two parts, namely a key (k) and a value (v). Tags are needed in order to describe specific features of map elements. In order to describe the features we interpret k and v as $k = v$, hence the key k contains a value v in order to describe the feature.

Elements being the core components of OSM and build the foundation OSM is built upon. There are only three types of elements namely *Node*, *Way* and *Relation*, which will be explained later in this subsection.

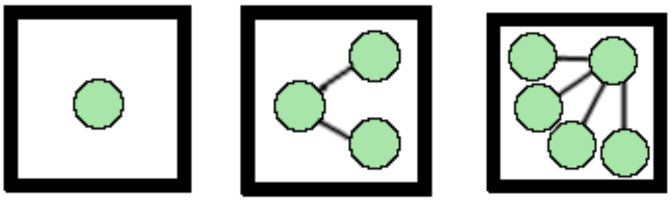
Features are needed for mapping a physical object on the map. OSM also allows a unlimited number of attributes to describe each feature. This is possible as the Tag system of OSM is free, therefore, arbitrary variations of descriptions can be created as long as they follow the principal of having a key and a value.

key=value	**Description**
opening_hours=Mo, Tu, Th-Su 10:00-18:00; We 10:00-21:00	A tag with a key opening_hours and a value of the weekdays and their respective opening hours
amenity=arts_centre	A tag that clarifies that this is an amenity of type arts_centre
addr:street=Albertinaplatz	A tag with addr (address) as namespace for specifying the object to be at the street called "Albertinaplatz"
addr:housenumber=1	A tag that further narrows down the address of the object with a housenumber of value 1
toilets:wheelchair=yes	A tag that provides a boolean value for the key of having toilets (namespace) for wheelchairs

Table 1: Tag system example of arts centre Albertina

The idea of table 1 is to provide an example of tags in form of a use case. The chosen use case being "Albertina", which is a arts centre in the middle of Vienna (Austria). The information provided in the table about the use case, is just a snippet of the real data, but enough for getting an understanding of the tagging system.

With the use of such sophisticated tagging, it is possible that OSM is fulfilling the modern requirements as stated in the beginning of this section. Furthermore, when using namespaces, as seen in the last three rows of table 1, the possibilities of providing very specific information are nearly unlimited.

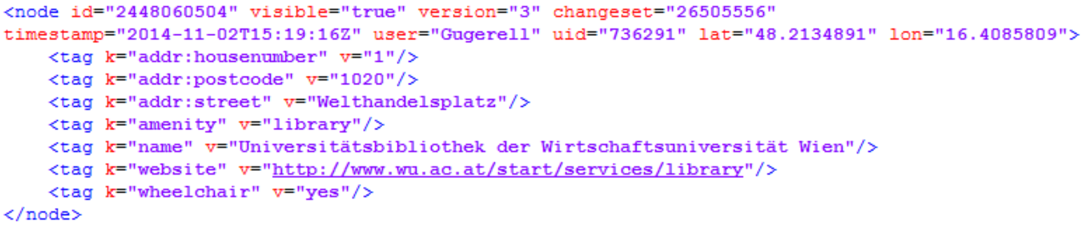


(a) Node (b) Way (c) Relation

Figure 5: Core Elements of OSM

Node

The node element consists at least of its node ID, longitude and latitude. When looking at figure 5a it visually represents a point (node) that can also optionally have an elevation level, meaning the altitude. With all three coordinates, or at least longitude and latitude, combined a node represents one point in space.
Nodes are the essential elements of the three core elements, because without them the other two wouldn't be possible to exist.



```
<node id="2448060504" visible="true" version="3" changeset="26505556"
timestamp="2014-11-02T15:19:16Z" user="Gugerell" uid="736291" lat="48.2134891" lon="16.4085809">
    <tag k="addr:housenumber" v="1"/>
    <tag k="addr:postcode" v="1020"/>
    <tag k="addr:street" v="Welthandelsplatz"/>
    <tag k="amenity" v="library"/>
    <tag k="name" v="Universitätsbibliothek der Wirtschaftsuniversität Wien"/>
    <tag k="website" v="http://www.wu.ac.at/start/services/library"/>
    <tag k="wheelchair" v="yes"/>
</node>
```

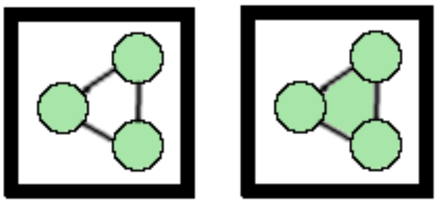
Figure 6: Example of OSM node XML structure (Use Case: "Library WU")

Figure 6 provides insights of the underlying XML structure of a node. In this example, it is possible to confirm what was stated before as it contains at least a node id, lat (latitude) and lon (longitude), which, by itself, would be sufficient for the node to fulfill its basic functionality.
However, by taking a closer look, various valuable information is provided within this one node that can be used to make the node better traceable and

provide users with potential desirable information. In this use case, specific information about the Library of the WU (Vienna University of Economics and Business) is provided in this one node. To simplify things in the description of the later example, we will name this node a "describing node".

Way

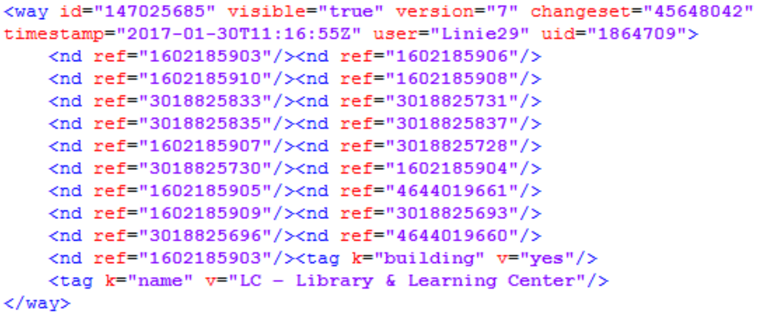
Connecting at least two nodes will result in a line from node A to node B and is called a way. A way can contain between 2 to 2000 nodes and is an ordered list that either can be open or closed. Figure 5b displays an open way that is connected via three nodes.



(a) Closed way (b) Polygon

Figure 7: Variation of ways

A closed way, as seen in figure 7a, is comparable to a round trip where the first node is at the same time the last node, hence it looks a polyline. When filling the polyline with a type like e.g. "grass", the result would be a polygon (also called area), which can be viewed at figure 7b.



```
<way id="147025685" visible="true" version="7" changeset="45648042"
timestamp="2017-01-30T11:16:55Z" user="Linie29" uid="1864709">
    <nd ref="1602185903"/><nd ref="1602185906"/>
    <nd ref="1602185910"/><nd ref="1602185908"/>
    <nd ref="3018825833"/><nd ref="3018825731"/>
    <nd ref="3018825835"/><nd ref="3018825837"/>
    <nd ref="1602185907"/><nd ref="3018825728"/>
    <nd ref="3018825730"/><nd ref="1602185904"/>
    <nd ref="1602185905"/><nd ref="4644019661"/>
    <nd ref="1602185909"/><nd ref="3018825693"/>
    <nd ref="3018825696"/><nd ref="4644019660"/>
    <nd ref="1602185903"/><tag k="building" v="yes"/>
    <tag k="name" v="LC - Library & Learning Center"/>
</way>
```

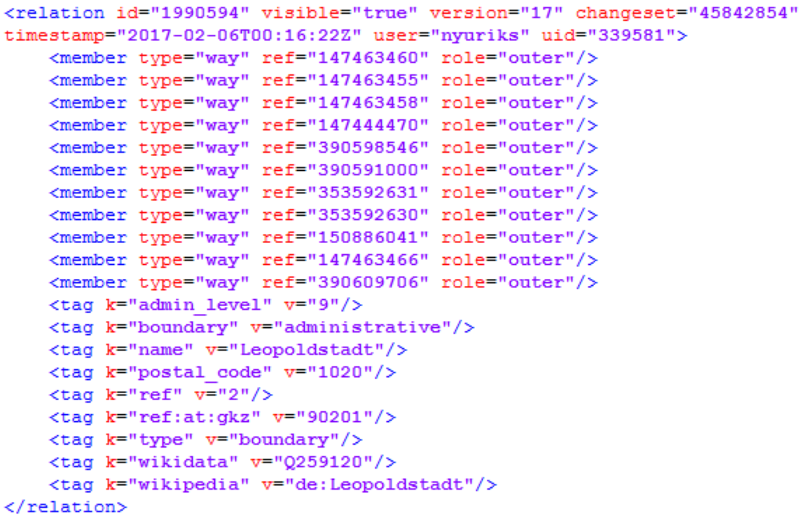
Figure 8: Example of OSM way XML structure (Use Case: "Library WU")

One might be surprised now when analyzing the represented closed way as XML structure in Figure 8. In fact, the use case is the same as in figure 6, but this time its the shape of the library building that is described by this

way.
The surprising factor may be that there is no longitude and latitude represented in the structure. When thinking about it, it gets clear that this wouldn't make much sense as a way consists of at least two nodes and therefore providing only one longitude and latitude pair would be of no use.
A way consists of several nodes and therefore it does not need its own longitude and latitude, because the nodes themselves carry this information with them.
Consequently, a way is an ordered list that references (ref) each node (nd) in order to create the way. As usual also some additional information are specific to that way can be optionally provided.
When analyzing the ID of the describing node of figure 6 with the IDs of the referenced nodes in the way, one notices that the describing node is not there. This is very common, as a describing node is usually placed in the center of an object (i.e. "building") and therefore of no use when displaying the shape of the object.

Relation

A relation consists of an ordered list of one or more nodes and/or ways. It can also contain more than one tag.
They are utilized for defining the geographical or logical relations between the other two core elements. This relation is visually presented in figure 5c and points out its connecting factor of the so called members of a relation. Furthermore, those members can have assigned roles that further describe the specific part a feature has in the entire relation.



```
<relation id="1990594" visible="true" version="17" changeset="45842854"
timestamp="2017-02-06T00:16:22Z" user="nyuriks" uid="339581">
    <member type="way" ref="147463460" role="outer"/>
    <member type="way" ref="147463455" role="outer"/>
    <member type="way" ref="147463458" role="outer"/>
    <member type="way" ref="147444470" role="outer"/>
    <member type="way" ref="390598546" role="outer"/>
    <member type="way" ref="390591000" role="outer"/>
    <member type="way" ref="353592631" role="outer"/>
    <member type="way" ref="353592630" role="outer"/>
    <member type="way" ref="150886041" role="outer"/>
    <member type="way" ref="147463466" role="outer"/>
    <member type="way" ref="390609706" role="outer"/>
    <tag k="admin_level" v="9"/>
    <tag k="boundary" v="administrative"/>
    <tag k="name" v="Leopoldstadt"/>
    <tag k="postal_code" v="1020"/>
    <tag k="ref" v="2"/>
    <tag k="ref:at:gkz" v="90201"/>
    <tag k="type" v="boundary"/>
    <tag k="wikidata" v="Q259120"/>
    <tag k="wikipedia" v="de:Leopoldstadt"/>
</relation>
```

Figure 9: Example of OSM relation XML structure (Use Case: "Leopoldstadt")

The same logic of ways, regarding the longitude and latitude, applies for relations. Figure 9 shows that no longitude or latitude would make sense as a relation is an ordered list of nodes and/or ways, which themselves carry the information of specific locations.

In this example there is no node in the relation. This is not a problem as the relation consists only of ways that include nodes, which carry the information of longitude and latitude for defining the location of connecting points for forming the ways.

The use case is about the district where the library of the WU is in, namely, "Leopoldstadt". The relation displays the border of the district, therefore, contains ways that form this border by relating them to each other in a ordered form.

Pointing out the mentioned role before that can be seen in this example as for all ways applies the i.e. role = "outer", which basically means that the role of a way is to be a part of a ring that forms the outer part of the polygon, namely, the border. Moreover, if there is an outer there is also an inner, which would define an inner part that is enclosed by a polygon. Relations that consist of an inner and outer part are called *Multipolygon relation*.

Another type of relation is called *Bus route*, which, as the name already indicates, is a relation that contains ways in a ordered sequence. The first node of every way within the relation represents a stop node where the bus stops.

2.2.2 The user aspect

According to [38], comparing the user aspects of VGI to other projects that collect GI, VGI has reached a critical mass that deserves consideration as field of inquiry within GIS science.

The work of [17] provides insight of why users are of such importance for mapping the world. He points out that many things, which are seen from above, can be automated for identifying geometries. However, he continues in stating that not everything can be seen from above and a further understanding is needed in order to properly describe the objects. Therefore, humans are trained at a young age in order to interpret things they see and give them reason.
Nevertheless, people might have difficulties when it comes to providing accurate information on the same level as a scientist in that field would. In fact, they understandably have less knowledge when acting within the geographical information domain.[13][2]
This, however, does not give the amateurs the credit they deserve as the author continues and makes following interesting statement that "drivers routinely trust driving directions given by local residents, for example, treating them in effect as professionals rather than amateurs" [17]. This points out that although in general the average user is an amateur when it comes to providing geospatial data, people still trust them when it comes to their domain they feel familiar with. Concluding from this perspective, OSM is a contribution of individual experts in their specific area making it a highly valuable tool for everyone to use. The author of [22] coins the term "citizen science", which he continues to define as "scientific activities in which non-professional scientists voluntarily participate in data collection, analysis and dissemination of a scientific project".[12] The question, although, still remains if the information they provide into the OSM database is sophisticated enough to reach the high level standards of quality known from other sources that are out there.

But how come that people are willing to sacrifice their spare-time in participating of collecting geospatial information and providing it for the rest of the world to use? The authors of [37] stated that "despite all available technology, people in modern societies feel more excluded from society, more isolated with respect to their communities, and more disenfranchised from the system of government and democracy. People have become more individualistic and pursue goals independently of each other." While this means on the one hand that individuals are seeking to do things on their own without the need of seeking the contact to another. On the other hand, it could be interpreted the other way around, because collecting geographical data and providing it for the society to see, could be interpreted as striving for attention of others. Furthermore, as OSM entries get edited by others, this could be considered some sort of communication. Concluding, that at this stage it still is not clear what drives people to participate. Is it because someone seeks attention or someone just likes to go out in order to collect and provide geographical information can't be said for sure. Suggesting that several reasons, being it ideological or personal, are behind each individuals thoughts of participation.

As mentioned in subsection 2.2, OSM is maintained and build upon its rising number of contributors as seen in figure 3. Furthermore, it is VGI that is provided by the contributors and therefore basically has no entrance barriers as anyone can simply register and add geographical information in OSM.

When bluntly looking at the numbers provided in figure 4, one might think that the rising number of edits is caused by the velocity of new users joining the OSM community.

However, the work of [29] from 2012 analyzes the contributors of OSM concludes that "the results show that only 38% (192,000) of the contributors carried out at least one edit in the OSM database and that only 5% (24,000) of all contributors actively contributed to the project in a more productive way. The majority of the contributos are located in Europe (72%) and each contributor has an activity area whose size may range from one soccer field up to more than 50 km^2."

This indicates that the majority of users (62%) to not participate in editing, and by editing the authors mean creating, editing and deleting, the OSM database. Basically meaning that the majority of contributors are inactive and did not go further than creating an account.

Additionally, it is not clear what happened to reverted edits, thus might letting them seem inactive.

These are interesting analysis that help to provide better interpretation in the numbers provided by figure 3.

It will be interesting to find out if the quality of OSM is in a good shape after knowing how to interpret the user numbers. Even though it has a lot of users, which have expert knowledge in their domain, the amount of inactive users might have an influence on some point of the quality of OSM.

2.3 Completeness

An object is complete if no information, which identifies the object, is missing. The research analyzes if objects are missing or if certain descriptive aspects of an object are missing. If not, then it can be considered as complete.
Therefore, the work differentiates between two ranks of objects that are not complete:

- *Partially complete:* Object is there but matches only to a certain extend, meaning that some data is mapped and identified, but others are missing.
- *Missing:*Object cannot be mapped and identified

This is done by using the information of the sources and try to map those with the information provided by OSM. Once the object is found, it will be examined in regard of its correctness. The evaluation progress of an object being correct or not will be discussed in subsection 2.4.

In order to illustrate the differences of what can be considered complete or not, an example of a building will be used. Note that this illustration does not necessarily represent the reality as it is just for understanding the stated above ranks of correctness.

Object	Name	Type	Toilets for wheelchair	Latitude	Longitude
Building A	Albertina	Museum	yes	48.2046365	16.3682605
Building B	Forbidden City	Attraction	yes	39.9174311	116.3907817
Building C	Eurospar Christl	supermarket	no	48.1878395	15.0867742

Table 2: Correctness illustration of complete data source

Object	Name	Type	Toilets for wheelchair	Latitude	Longitude
Building A	Albertina	type:museum	yes	48.2046365	16.3682605
Building B	Forbidden City	type:attraction		39.9174311	116.3907817
Building C					

Table 3: Correctness illustration of OSM

Complete
Investigating Building A in both tables (table 2 and 3) makes clear that both rows contain the same information. Exception might only be the column "Type", but this is just mapping issue and not related to completeness

as the information it contains is still the same, just has a different syntax. Therefore, source object of Building A is successfully mapped and identified with the target object in OSM and can be considered complete.

Partially Complete
When examining Building B in both tables, a missing value in column "Toilets for wheelchair" of table 3 can be identified.
Note that the object itself can still be identified and mapped as such with the coordinates, hence only some descriptive information is missing that makes it partially complete.

Missing
The last example, namely Building C, shows that in the source table 2 all information is available.
Nevertheless, information about Building C is not available at all in the OSM table 3. Therefore, the object cannot be mapped in any way with OSM and can be considered to be missing.

2.4 Correctness

Correctness and quality in general, can be a very subjective thing to measure. Interpretations between users can vary alot depending on their education and culture for identifying something to be correct or not. Correctness itself is defined as something that is true and accurate.
Furthermore, there should be a common understanding throughout participating parties of what can be considered correct and not.

Past analysis about OSM show problems with VGI as it is often ambiguous or misinterpreted. [3][33] Thus making it hard for users of OSM to trust the data. Approaches trying to identify users, who create wrong information, are provided by several authors. Their theories are based on the editing behavior of OSM users and aspects like:

1. Created/corrected edits
2. Lifetime of created/corrected tags
3. User reputation value

For this, time-stamps and their respective historical entries are analyzed, which contain data like the date, time and user name. These approaches have good aspects, but also run into potential situations that are causing interpretation flaws.

Therefore, some works will be introduced that show approaches of evaluation trustworthiness, including their advantages and disadvantages.

The authors of [23] introduce a model that accounts for the trustworthiness of features. They base this model by stating that rollbacks or deletions of features would result in changes of trustworthiness of a feature and the reputation of the contributor. By doing so, they identified four patterns, which should help to identify if an entry is trustworthy or not. Those patterns cover user behavior like how long a edit stays unchanged or if rollbacks are made by others or the original contributor.

However, it is not accounted for changes that are required due to environmental changes. In those cases a contributor would lose reputation even though his content was right till that point in time. Furthermore, even after a feature is corrected, it is not assured that the correction is really representing the truth about the object in the real world.

They continue in [24] by introducing a provenance vocabulary that is based on recurring editing- and co-editing patterns. This is done by interpreting those editing patterns. Consequently, they interpret a change intuitively as negative feedback. Moreover, the absence of such edits, are intuitively interpreted as positive feedback meaning that the data is correct.

Nevertheless, there are problems that are not covered with these assumptions. As discussed in subsection 2.2 contributors around the world have a different density in certain areas. Therefore, potential mistakes within the data could stay unnoticed there for a very long time till someone might actually review it and still then it is not sure if an error is identified by that other contributor. Furthermore, when intuitively providing negative feedback to changes of someone's edit, it is not accounted for changes that can actually be considered necessary. Those changes can include things like environmental changes or descriptive information when e.g. opening hours of an restaurant get changed.

Based on this work a follow up paper was created by [14] that extends their introduced methods by relating trustworthiness to the contributors reputation in respect to the relevance of editing and atomic editing operations.

On the one hand, these approaches are good and the authors have well thought about how to process the data in order to calculate user reputation, which can be related to a general trustworthiness status. On the other hand, they base the editing behavior by assuming contributors only apply correct changes and thus the reputation drops immediately for users whose content got edited.

Additionally, they would also need to consider areas with less user density and implement a system that e.g. lets contributors tag changes that are nec-

essary due to actual changes in the real world, which would then not lead into a reputation drop for the contributor who originally created/edited the content.
Therefore, this thesis uses the method of comparing data with sources that can be considered reliable in order to analyze correctness. Those sources are reviewed by experts, whereas OSMs data, or VGI content in general, is not.

Moreover, the work will differentiate between two ways of classifying correctness

- *Semantic correctness* covers the relation of an object in respect to its meaning and how it can be interpreted.

- *Geometric correctness* covers the positional aspects of an object and especially its shape. Consequently, all coordinates that belong to an object will be considered for evaluating the correctness of the shape.

2.5 Methodology

For making a statement about the completeness and correctness of OSM, it first needs to be decided on how to evaluate if information is true or false. Subsections 2.3 and 2.4 already provide insights of approaches on how this can be tackled and point out the problems with VGI in its current state.
Rather than analyzing the edits or check, by looking at the map, if a edit might be justified, one can also choose to make a field study and investigate by themselves. Taking a GPS capable device to see if the shape is according to the coordinates of the device and that also the description of the reviewed object is accurate. Practically, this approach is hardly feasible as it takes a lot of time and money to realize. Moreover, the problems of being outdated or that someone might change it again, are still persistent.
Another approach would be to find some method that could be automated and still be reliable regarding the investigation of the geographical object in respect to the shape and description. In addition, it should also minimize the time and money spent in comparison to other methods as pointed out before.
This work proposes a method that compares OSM data with a trustworthy sources.

For comparing OSM with data from another source, reliable information of these sources is needed. Furthermore, it is required to decide what will be compared as there are various data sets that provide different geographical

information of different domains. Examples of specific domains being data for transportation networks of a city, marine information, urban areas and land use allocation.
Consequently, the source has to be chosen in regard to the OSM data that should be analyzed. In order to have a look in various directions, when it comes to OSM geographical data, the following three aspects should be covered:

- Shapes: Source data should contain data about the shape of objects like ways (e.g. closed ways) and polygons, which then can be mapped to the exact location where their OSM counterpart is in order to analyze differences.
- Descriptive information: As OSM also includes a lot of information that describes the respective geographical objects, the source should also provide information of the same relevance. As an example, when an OSM object contains information about opening hours of a restaurant, the mapped object of the reliable source should also have this information in order to be comparable.
- Historical changes: OSM also stores historical information, which could also be compared with the reliable source in order to potentially find interesting aspects about the change in quality over time.

Based on these these defined observation points, it is now possible to look for reliable public data to use as source that can be compared against. It is important that those sources are reliable, or at least can be assumed or considered reliable, as they will be considered to be true in regards of their correctness and completeness.

2.5.1 Data sources

By searching for a trustworthy source that contains the needed information, a preparatory study was required for elaborating which sources to use.
in addition, it was necessary to define certain use cases that cover the defined aspects.
This subsection will cover the decision making process and reasoning why certain sources were chosen, hence also those that were considered, but not elected in the end.

Defining use cases

For investigating the defined aspects for comparing OSM and the data of a trustworthy source, it is of help to define use cases that support this goal. The need of defining more than one use case is due to the variety of results it could lead to. An example could be to analyze if there are quality differences between smaller and bigger geographical objects.
In addition, not all defined aspects might be available in a single use case, or at least not to a satisfying extent.
Consequently, it was decided to define the following use cases:

- Borders of European regions: By analyzing the borders of European regions the research focuses on the shape and historical data.

- Landuse of forests: For analyzing the landuse it is enough to narrow down the use case for only forests and focus on some samples. The landuse type gives information about the shape and descriptive information. Additionally, this is split into two use cases as during the research it got clear that there are differences between urban and land side areas.

- Population of European countries: To add a demographic use case it was decided to go with the population of European countries, which provides outcomes about descriptive and historical information.

- Address register of Austria: Data from the Austrian authorities is used in order to evaluate the completeness of addresses on some regions within Austria. This covers another aspect and source for descriptive information.

In fact, nearly all those use cases, except the Address register of Austria, offer the opportunity, due to their distribution, to explore qualitative differences of VGI within European countries. Nevertheless, this is not a core focus of this research, but can be interpreted as possible future works if any remarkable findings are identified.

OSM

As the core of this research, OSM was chosen to be the focus of its analytic potential. OSM is the target that will be compared against trustworthy sources. However, it is also of interest to know what comparable other maps, created through VGI, are out there.

Wikimapia is similar to OSM in a way that it's also generated via VGI. Its approach is that every user, including non-contributors, can participate and edit information. They also implemented gamification aspects in which users can earn badges and level up. However, as their user base is rather low, compared to OSM, it is hard to keep quality high.[20]

Google maps uses the tool Google Map Maker, which allows volunteers to participate in the editing of Google Maps. However, all these edits need to be reviewed by Google staff and therefore take long time till they are in the map. Nevertheless, the quality might be higher as every entry gets reviewed by an expert. There are also protected areas (read-only) where users can't contribute anything. Retrieving data that is needed for analyzing the defined aspects for this research is hardly possible with Google maps, as their API (Application Programming Interface) does not offer all the functionality. [10]

Eurostat

Eurostat, the statistical office of the European commission, offers a big collection of geographical data sets about Europe. Their main focus is to provide statistics about Europe, which already indicates that location is a key attribute, and by doing so, they use geographical maps to visualize their results better. A detailed introduction on Eurostat will be provided in subsection 3.1.
Several data sets were analyzed in order to identify the ones that could be used within the research scope.

In fact, various data sets could be used for analyzing the shape and historic data in order to compare it with OSM. It was decided to use the NUTS (Nomenclature of Territorial Units for Statistics) classification data set and a data set about the population of from Europe. Those two data sets cover all three defined aspects the research aims to analyze. Details about those data sets will be explained in subsection 3.4 and 3.5.

Another reason why Eurostat was chosen is, because they are the only ones that provide statistics, particularly, geographic data of Europe in the quality that was required for this research.[11] Consequently, it offers a good scale that can be used in order to potentially come up with additional find-

[10] https://developers.google.com/maps/documentation/javascript/usage, last accessed 2017-03-02

[11] http://ec.europa.eu/eurostat/en/about/overview/what-we-do, last accessed 2017-03-03

ings when comparing results between member states of the European Union.

Copernicus

In order to not only have one reliable source to compare the OSM data with, another one was searched for. By doing so, Copernicus identified as best fit as it also offers data of the European Union. More specifically, it was decided to go with the Corine (Coordination of Information on the Environment) Landuse data set provided by Copernicus, as it offers detailed information about the shape and type of it.

An alternative, that would also offer landuse data on the same scale, would be the LUCAS (Land Use/Cover Area frame statistical Survey) data set from Eurostat. However, with LUCAS it is not possible to get the shape for various landuse types. It offers information provided by a node that contains regional information within a limited area. In order for LUCAS to work, a manually created 2x2km grid is needed for interpreting the nodes that contain the areal information.[12] Concluding, that the LUCAS data does not offer as detailed information as Copernicus Corine landuse does, therefore, it was decided to go with Copernicus Corine landuse data set.
A deeper insight will be provided in subsection 3.6.

Bundesamt für Eich- und Vermessungswesen (BEV)

Another source is presented by the local authority called BEV, which is responsible for Metrology and Verification in Austria.
It was chosen to contact them due to past projects that already provided insights of the data they collect.
However, due to their limited responsible area, namely Austria, their data cannot be used for comparison across Europe, but will be used for comparing OSM data for Austria to also have a regional example use case.

Consequently, there are a lot of alternatives out there as every member state has its own local authority that collects geographical information.
It was decided to use their information about the Austrian address register and check against OSM completeness, if certain addresses can be found or not.

[12] http://ec.europa.eu/eurostat/documents/205002/7329820/LUCAS+Grid+Record+Descriptor/1df20fae-7fb2-4a89-aab6-8064a989527a, last accessed 2017-03-03

2.5.2 Research environment

A couple of tools were used for identifying and computing geographical objects. Insights will be provided why certain tool or environments were used and if comparable alternatives exist. Furthermore, strength and possible weaknesses will be pointed out that played a role in the decision making process when choosing those tools.

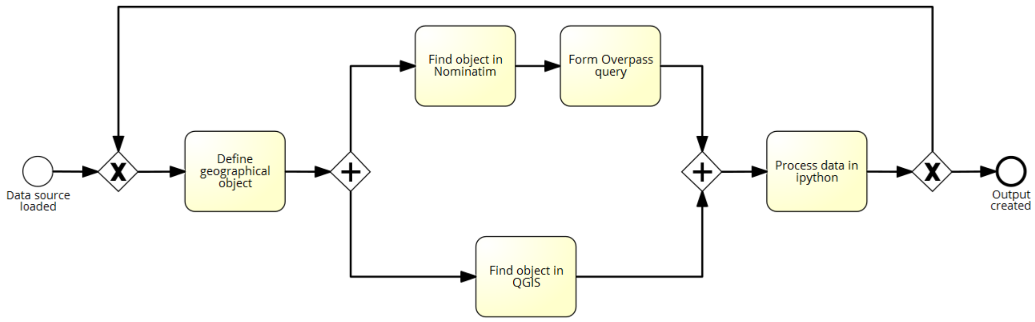


Figure 10: High level BPMN model of environment usage

Figure 10 provides a high level BPMN (business process model notation) process model of the environment. The figure shows the executed order of how the tools and environments were used. After the sources are loaded, the geographical object, which is subject of investigation, is defined. The OSM object will then be searched with Nominatim and once found the query will be formed in Overpass for fetching the data. Meanwhile, the same geographical object is searched with QGIS and define a mapping. Once the object is located, the data of how to map the objects will be added to the IPython environment in order to further process the data. After adding the information into the IPython environment, one can decide to define another object for adding to the analysis or print the output for further investigation.
A more low level explanation will be provided in section 4. Environments and tools that were used, will be explained in the paragraphs bellow. Moreover, understanding and reasoning of the decision process, of why certain environments or tools were chosen, will be explained.

Nominatim

Out of several alternatives like MapQuest Open, OpenCage Geocoder, MapZen Search (pelias geocoder), and LocationIQ, for investigating geospatial objects in OSM, Nominatim was chosen as it offers a free and convenient

solution. Furthermore, OSM itself uses the search functionality of Nominatim[13], for addresses and names, which strengthened the decision to go with it. Others did fall out due to license limitations or terms of use restrictions. It is especially helpful when trying to detect certain elements, as explained in subsection 2.2, within a given area. Moreover, the provided graphical user interface (GUI) offers tools that foster this.
When further checking a desired element it provides useful information of various relations of that element, which could be of interest like a "Child-of" or "Parent-of" relation.
This is especially helpful when it is not clear what to search for, but knowing a child or parent and going respectively up or down the relation in order to identify the desired object.

Overpass turbo and API

The Overpass API is read only and sends back information that corresponds to the customized query defined by a user. It is designed for only providing information within a fast reaction time, but still offers a powerful query language.
Overpass turbo is the GUI of the Overpass API. Furthermore, it is able to execute queries from the Overpass API and displays the results on an interactive map.
As the Overpass API does not support the GeoJSON format when fetching the data, it was necessary, at least for fetching shapes of an geographical object e.g. polygon (area), to use Overpass turbo. This was a strange experience while programming the data retrieval, because on the one hand Overpass API does not support the required GeoJSON and on the other hand it supports it via using the export functionality from the GUI of Overpass turbo.

Limitations:

- Downloading huge data: It works best for smaller sized downloads. There is no estimated time to how long the download takes when starting a query. This was also experienced during the research when trying out various queries. It even might happen that the whole browser gets slow or even crashes at some point when the query is downloading e.g. many features of a big region.

- Data history: It also has its limitations when it comes to historic data.

[13] http://wiki.openstreetmap.org/wiki/Nominatim, last accessed 2017-02-24

As stated in the description of the Overpass API[14], it may provide information of the state of an object at a certain time, but will not provide its full history. During the research this was also experienced when it comes to the describing elements of a feature, which will be gone into detail later in subsection 5.2.4. However, for the shapes of the object the historic data was provided as differences were experienced during the research.

QGIS

It is a free and open source GIS desktop application for various platforms. It provides specific functions for geographical data files, hence for conducting analysis, editing or viewing. It can be compared to ArcGIS, which is a more powerful tool. However, ArcGIS requires an expensive subscription in order for using parts of the service. Therefore, QGIS was chosen as it provides sufficient functions for the scope of the research.

It was mainly used for first explorations of the various sources and, as a follow up step, preprocessing those sources. Preprocessing consisted of removing information that was not needed in the research and also for converting the format of the source data into a more fitting format that could be used with Geopandas in IPython.
Those preprocessing steps need to be chosen carefully as certain actions take an excessive amount of time. An example for illustrating this might be to e.g. delete unnecessary columns first before deleting rows. That way the amount of time the rows would need to be deleted is reduced as an entire cell of that row is already gone due to deleting unnecessary columns first. This does apply as the number of columns is smaller than the number of rows in all the data sets that were used.
Finding the defined object can sometimes be a challenge as some data only provide very basic information like an ID, name and geographical information, whereas the name was not something that could practically be searched for as it was a kind of code name and not related to a specific findable geographic object.
Therefore, it was sometimes necessary to first find the defined object with the help of Nominatim. Afterwards, one could use the coordinates that are provided in the search result of Nominatim, as input for QGIS for navigating to the so identified geographical location. This allows to find the desired object also in QGIS and allows to further help mapping the objects for the

[14] http://wiki.openstreetmap.org/wiki/Overpass_API, last accessed 2017-02-26

later step of processing the data in IPython.

IPython and packages

IPython offers an enhanced interactive environment, which in a sense is an extended Python shell. The author of [31] underlines that, nowadays where computing power is not that big of a deal as it used to be, IPython is a valid tool for special purposes and provides examples of projects that successfully accomplished its goals by using it.

Another reason of why it was used in this research was the fact that the experience level was higher as for other environments, which some would have needed to be learned from scratch. Other environments being R, JScript and Matlab (license), whereas IPython offers better packages (e.g. Geopandas) for handling geographical information.
As mentioned, Geopandas is the core package that was used in the IPython environment. It offers a geo-dataframe in which the sources (e.g. Shapefile, GeoJSON, ... see subsection 2.5.3) could be loaded and handled with ease in order to further allow better analyzing and calculation of the data.

2.5.3 Formats

The importance of using adequate formats and the hurdles with others, will partially be pointed out in this subsection. Other sections, like section 4, will discuss related and additional hurdles regarding the formats. Furthermore, insights on how those formats are structured and what their respective advantages are will be provided. In general the preprocessing and converting of formats is a time consuming task, as the task itself takes very long and even more if data is being transformed for the reason of trying out data sets and if they could be used as subject of the research.
In fact, there were a couple of data, which came in all kinds of formats that needed to be dumped later on, because it didn't provide the information needed for the goal of the thesis.

Shapefile (.shp)

A shapefile is a file format with the extension .shp, which is developed and maintained by the company ESRI. Their technical description[1] points out that it is a vector data format for GIS software and was developed as partially open specification. The advantages of using a shapefile for geospatial activities are:

- It stores non-topological geometry and attribute information for the features within a data set. The feature is stored as a shape that contains the vector coordinates.
- A huge advantage when it comes to the size is that a shapefile does not store the processing overhead of the data structure, which enables it to be used faster for various purposes like drawing, editing, loading ...
- It also supports point, line, and polygon (area) features.

File	**Extension**	**Description**
Main	shp	Describes with each record a shape with a list of its vertices.
Index	shx	Each record contains the offset of the respective main file record.
dBase	dbf	Contains feature attributes with one record per feature. They are in the same order as in the main file and build a one-to-one relation between geometry and attributes, which is based on the record number.

Table 4: Core files of a shapefile [1]

Research use: For the research, shapefiles in general were used a lot and of the essence in a sense as they helped to make some analysis possible. When using published sources, provided by e.g. European authorities, it often is the case that it's in a kind of database format like .gdb or .mdb. Those, by itself wouldn't be the problem, but when having a limited processing power and memory, a shapefile is more efficient to use.
As pointed out before in the advantages, it is focused on the essence and therefore allows for faster processing.
QGIS was used in order to use its functionality to export information from the various database file formats into shapefiles, which then can be more conveniently loaded into the IPython notebook via using the Geopandas framework.
Another advantage being that one does not need specific database drivers or applications installed when using a shapefile. During the research there were some hurdles regarding the database related files that could be avoided when additionally providing shapefiles in the first place.

GeoJSON

GeoJSON is a format, with the extension being .geojson, for encoding a variety of geographic data structures and build upon the JavaScript Object

Notation (JSON). [7] When looking at the structure it very similar to JSON, but specifically designed to support geospatial information.
A GeoJSON object may represent a region of space (a Geometry), a spatially bounded entity (a Feature), or a list of Features (a FeatureCollection). GeoJSON supports the following geometry types: Point, LineString, Polygon, MultiPoint, MultiLineString, MultiPolygon, and GeometryCollection. [9]

Research use: The GeoJSON format was mainly used for the OSM data. Again, similar as for the shapefile, IPython with the Geopandas framework work very well and fast with the GeoJSON file format. Sadly the Overpass API didn't support GeoJSON, therefore, the files had to be manually exported and saved on the file system via using Overpass turbo. When searching for answers, a forum entry of the developers from overpass API was found in which they stated that it is not supported as the semantics seem not be clear.[15] Strangely enough, Overpass turbo allows the export into GeoJSON file format.
This is not that dramatic as they usually contain only a couple of kilobytes in the respect of the research focus of this thesis.
Alternatively, several approaches of transforming other types that are supported by the overpass API, but are barely usable with Geopandas, into GeoJSON failed.

XML

The Extensible Markup Language (XML) is based on open standards and aims to be simple and usable for various purposes across the internet.[8] It mostly focuses on document structure, which is pointed out by [15]. An example of the structure can be seen in figure 6, which also shows the capability to be human readable.

Research use: In this research it was only used in the beginning for getting a good overall picture about the data structure of OSM features. XML itself is more easy to read for humans and therefore provided an easier and faster understanding of the data structures of OSM.
It was also supported to be fetched from the Overpass API, but couldn't be used properly with the Geopandas framework, which is a core part of the development environment.
Several attempts to transform XML into GeoJSON failed as it is not that straight forward, so that in the end it was decided to not use it and directly

[15] https://github.com/drolbr/Overpass-API/issues/48, last accessed 2017-02-24

export GeoJSON instead as stated in the GeoJSON explanation above.

MDB and GDB

These two file extensions are associated with databases. MDB is related to be used via Microsoft Access, whereas GDB is a database file extension is a Garmin mapsource file and can contain data like route, tracks and ways.[16] Both database formats can carry geographical information with them. Nevertheless, the transformation into a shapefile allows it to be read by the Geopandas package of python and therefore more convenient to use as it can be load into a geo-dataframe. Handling the database formats as they are would require massive mapping efforts as those files can be rather big. For the source, which came as GDB file, for example it was 4 gigabyte of data which took already a long time for QGIS to convert into a shapefile. This included some crashes along the way, which is very time consuming.

CSV

The format of Comma Separated Salues (CSV) is used already out there for quite some time and is mainly used for exchanging and converting between spreadsheet programs. The format, as pointed out in [34], contain some main aspects like the separation through commas and that each record is located on a separate line.

Research use: CSV was used in the form of TSV (tab-separated values) that was provided for the population analysis by Eurostat as explained later in subsection 3.5. As no high sophisticated graphical information is needed for this kind of analysis, the form of having the data as TSV is appropriate. TSV and in general CSV, can easily be handled by IPython and Geopandas and therefore working with it was very convenient. Fetching the data and calculations are done very fast when using TSV with IPython.

3 GIS data of public authorities

After deciding on the sources to use, which is pointed out by the decision making process in section 2.5.1, a closer look on the organizations and their data sets will be provided in this section.
Furthermore, information how they collect and retrieve data will be given. This helps to foster the understanding of the decision process and how they ensure a high quality of their data.

3.1 Introducing Eurostat

Eurostat has its headquarters in Luxembourg and is the statistical office of the European Union. In fact, they are the only providers of statistics on a European level.
For generating their statistical analysis they do not themselves collect the data. Instead, the statistical authority from each member state is responsible for the collection.
After they analyze and verify the data they send it to Eurostat for further processing.
Afterwards, Eurostat is responsible consolidating the data and ensuring that it is comparable.

For this research Eurostat was used, because of their GISCO (Geographic Information System of the Commission) data. Within Eurostat, GISCO is responsible for meeting the European Commission's geographical information needs at 3 levels: the European Union, its member countries, and its regions.[16]

They offer certain geographical reference data like[17]:

- Administrative boundaries and statistical units
- Ports and airports
- Digital Elevation Model
- Population distribution
- Land Cover/Land Use information

[16] http://ec.europa.eu/eurostat/web/gisco/overview, last accessed 2017-03-03

[17] http://ec.europa.eu/eurostat/statistics-explained/index.php/Geographical_information_system_of_the_Commission_(GISCO), last accessed 2017-03-03

In this work, data from "administrative boundaries and statistical units" (see section 3.4) and "Population distribution" (see section 3.5) were analyzed.

3.2 Introducing Copernicus

Copernicus is a program, initiated by the European Union to be managed and coordinated by the European Commission, for improving information services withing Europe. By doing so, it uses data from satellites (Sentinels), which observe the earth, and in situ (non-space) data.
It was designed to meet user requirements for public authorities, service providers and other international organizations for its common goal to improve life quality for the citizens of Europe. [18]
The information they collect is free and open for everyone.

Copernicus uses the collected data to further processing and analysis for generating value-adding information. Another interesting aspect is that they keep their data comparable, meaning that it is possible to track historic changes over time with the data offered by Copernicus.

These value-adding activities are streamlined through six thematic streams of Copernicus services:

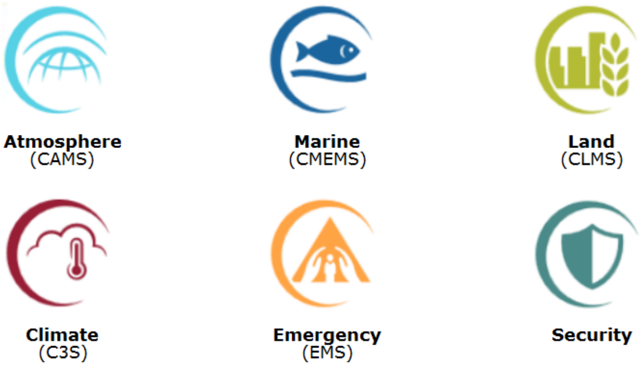


Figure 11: Copernicus reference data[19]

The main users of these six thematic streams shown in figure 11 are policymakers and public authorities. Using the information served by Copernicus they can base their decisions when developing new environmental legislation

[18] http://www.copernicus.eu/main/overview, last accessed 2017-03-04
[19] http://www.copernicus.eu/main/copernicus-brief, last accessed 2017-03-04

and policies.
For the research the Copernicus Land data (CLMS) was used, which offers information on land cover and will be further explained in subsection 3.6.

3.3 Introducing BEV

The Austrian office for Metrology and Surveying belongs to the "Bundesministerium für Wissenschaft, Forschung und Wirtschaft". The headquarters, besides 64 branches distributed across Austria, are in Vienna.
Their repsonibilites are base surveying, creation and maintenance of land register for the spatial allocation documentation and national topographic mapping.[20]
They offer a broad range of services, one of them being the Austrian address register, which is used for a use case in this research.

3.4 Eurostat NUTS classification

The Eurostat NUTS classification is used to divide the economic territories within the EU. As seen in figure 12 it consists of several layers that that offer different levels of granularity regarding the territories they display.

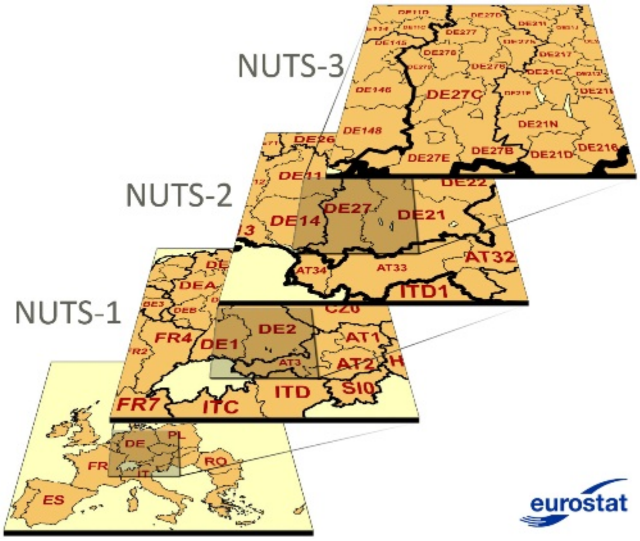


Figure 12: Eurostat NUTS[21]

[20] http://www.bev.gv.at/portal/page?_pageid=713,1605147&_dad=portal&_schema=PORTAL, last accessed 2017-03-15

[21] http://ec.europa.eu/eurostat/web/nuts/overview, last accessed 2017-01-24

Furthermore, those economic territories are marked in a way that they can be associated with the Member State they belong to.
On the lowest level of figure 12 there is the map that contains the Member States of the European Union. They are displayed with their respective acronyms like DE for Germany and AT for Austria.
Upon this structural foundation the NUTS layers are build. By adding alphanumeric characters, for each new layer one, those territories can be presented in their respective granularity level. This structure is at the same time the unique key for mapping certain regions to OSM data, but this will be covered later in subsection 4.1.1.
Three principals were used for creating the NUTS layers.[22]

First, certain population thresholds were defined as seen in the following table:

Level	**Minimum**	**Maximum**
NUTS 1	3 million	7 million
NUTS 2	800 000	3 million
NUTS 3	150 000	800 000

Table 5: NUTS principles and characteristics

Second, the administrative divisions of the respective member state are being preferred.
Third, it favours general geographic units as they are more suitable for indicators.
Note that for some member states have different granularity levels of their territories than others do. Therefore, not all have the lowest NUTS level i.e. NUTS 3 as their territories are kept bigger.

[22] http://ec.europa.eu/eurostat/web/nuts/principles-and-characteristics, last accessed 2017-03-14

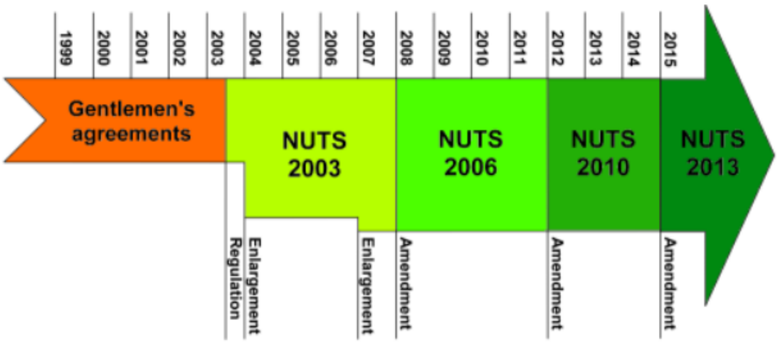


Figure 13: Eurostat NUTS history and release schedule[23]

For this research the NUTS data was used to compare the regional borders with their respective counterpart in OSM. However, for making them comparable, one has to know when the data is released or collected.
Figure 13 shows the history and release circle of the NUTS data. In fact, when analyzing the release times and their content it will be clear that comparing the data is very difficult. Meaning that it is impossible to know from the data when a certain feature within the data set was collected. Furthermore, the period of collecting the geographical information till the point it gets released is very long. Concluding that when the NUTS data is released it could already be outdated.
Nevertheless, Eurostat states that "the regulation also specifies stability of the classification for at least three years"[23]. This is important for their statistical analysis as it ensures that the data refers to the same regional unit for the specified amount of time.

3.5 Eurostat population

Eurostat also provides demographic data regarding population numbers across member states, but also collects data from non-member states.
Data is collected via NGIs (National Statistical Institutes) from each member state, which are collected at their national and regional areas.
The population data is updated several times over the course of a year. Each year, on the first of January, Eurostat publishes population data that is based on estimates on the total number of births, of deaths and of the net migration from the last year.
During the course of the year, those estimates get updated by data provided from the European States. This data gets updated on a monthly basis by

[23] http://ec.europa.eu/eurostat/web/nuts/history, last accessed 2017-03-04

calculating the numbers form births, deaths, immigration and emigration.[24]

3.6 Copernicus Corine Land Cover

The Copernicus land monitoring service offers detailed information, which is needed for the investigation of land cover shapes and their describing properties. In fact, it provides geographical information on land cover and their related variables.
All kinds of applications are supported by their collected data such as water management, forest management, agriculture and food security, spatial planning etc. [25]

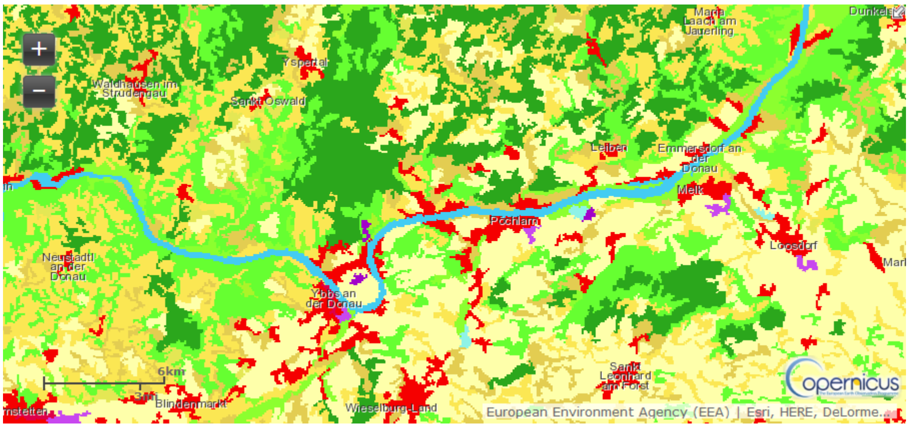


Figure 14: Copernicus Corine Land Use[26]

For an example on how the land cover data looks like, figure 14 gives a good first impression.
It shows a part of the Corine Land Cover of the CLC2012 data set that was used for this research. The data itself is collected via satelites and in-situ.
In-situ means that field surveyors are collecting information by observing and registering the objects on the ground.
However, the major use of in-situ data, within the CLC project, is to amend the satellite data during the process of production and further verify results that is offered from space-born data.[27]

[24] http://ec.europa.eu/eurostat/web/population-demography-migration-projections/methodology, last accessed 2017-01-27

[25] http://www.copernicus.eu/main/land-monitoring, last accessed 2017-03-05

[26] http://land.copernicus.eu/pan-european/corine-land-cover/clc-2012/view, last accessed 2017-01-27

[27] http://land.copernicus.eu/user-corner/technical-library/Addendum_finaldraft_v2_August_2014.pdf, last accessed 2017-03-05

The Corine Land Cover is mainly produced out of visual interpretation from the high resolution satellite images. This is done by the respective countries, whereas some also implemented semi-automated processes for, which is a mixture of various techniques like satellite images, in-situ, generalization and GIS integration.[28]
In fact, CLC uses the LUCAS data from Eurostat, which was mentioned as an alternative to CLC in section 2.5.1, in order to help improve their data accuracy by interpreting the photos via a photointerpreter made by the LUCAS project.[10]
Furthermore, the CLC kept improving their quality and topicality over the last couple of years. Production times are getting shorter, therefore, the time window till upcoming releases is reduced with the last one being a two years difference.

There is a classification for each of the types of the Corine Land Cover. In fact, there are five major classes, which each can be further narrowed down to realize a higher granularity level.[6]
The classes for CLC being:

- Artificial surfaces
- Agricultural areas
- Forest and semi-natural areas
- Wetlands
- Water bodies

Underneath those classes and sub-classes there is a code. Symbolized by the colored shapes as seen in figure 14, the code offers the possibility to search for the objects in the database regarding their various types the code represents.
A more detailed view of these classes and codes, on how they can be applied when working with them, will be shown in section 4.1.2, which covers a certain use case for analyzing the CLC with OSM.

3.7 BEV address register

BEV offers various data sets that contain geographical information of Austria.

[28] http://land.copernicus.eu/pan-european/corine-land-cover/view, last accessed 2017-03-17

For this research the data set of the Austrian address register was chosen. The data set itself is collected by the branches of BEV, which is spread across Austria.
It contains the required information that is needed for mapping and creating queries against OSM.

In general, for common use, an address consists of various attributes like:

- Street name
- House number (sometimes with staircase and door number)
- Postal code
- Region name

In the case of the address register of BEV much more information is provided. The data set is very sophisticated.
They further, besides the information mentioned in the list above, provide information like longitude and latitude, building categorization, EPSG (European Petroleum Survey Group) Code and many more that add information about a certain address, which is in principal a point (node) on the map.

EPSG Codes are key values that reference certain projections as seen in figure 2. Their importance for the mapping progress will be explained in section 4.

4 Mapping geospatial objects

The mapping process of how to relate identical features, from two different data sources, is of the essence. Only with sophisticated mapping a comparison is possible.
This section covers the mapping processes behind all the use cases that are covered in this thesis. Furthermore, this helps to foster the understanding of how the output in section 5 was generated and avoids misconceptions of how certain information was retrieved.
The use cases themselves use different sources that are compared against OSM. Therefore, a different mapping approach is needed for every one of those use cases.
This requires the exploration of the source data structure and also of the respective objects in OSM.
Understanding both structures is key for relating them with certain common features.

In general, there are different techniques and indicators for retrieving information about objects.
This research covers manual approaches as well as automated data retrieval. This is dependent on the use case and feasibility as in some cases automation would be too time consuming or complex in order for getting meaningful results. Moreover, in all of the use cases a manual investigation is always required for getting an overall picture of the various data structures.
On some occasions the manual inspection via certain tools was necessary for detecting faults in initial assumptions or the query itself.

The various data sets, of different sources, come with diverging projections. The system that is used for identifying the projection is called coordinate reference system (CRS) and is used to locate geographic entities.
It uses identifiers to uniquely and unambiguously identify a certain projection. The CRS is essential for every GIS in order to function and fulfill its purpose.
There are various identifiers formed, but the most popular being the EPSG, which is also used as an identifier in all of the use cases of this thesis.
Consequently, for making the data sets comparable with OSM, a conversion into the right projection, hence EPSG identifier is needed.

4.1 Use cases

The following use cases will point out how geospatial objects from various sources were mapped to find them also on OSM. Furthermore, insights of the structures of the sources will be provided, which is if the essence for understanding the reasoning behind the mapping procedures done throughout this research.
The use cases analyze different aspects of geographical objects, therefore, each use case has its own ways of mapping the information in order to achieve its specific goal.
However, all of them focus in general on the aspects defined in section 2.5 for further providing answers to the defined research questions in subsection 1.2.1.

4.1.1 Shapes of European regional areas

The areas (polygons) of European regions is provided by the NUTS data set of Eurostat. As detailed discussed in subsection 3.4 the NUTS data set has a standardized code table for every region of each member state within the European Union.

The NUTS data set can be download from the Eurostat homepage as a MDB file. When loading this MDB file into QGIS the following columns could be identified:

- *OBJECTID* is the primary key and an continuing number per entry. Values are provided as Integer64.
- *NUTS_ID* contains the information of how to call a certain region and includes the acronym of the respective country the region is in. (e.g. DE11, AT212, ...) Values are provided as String.
- *STAT_LEVL_* is a value between zero and three that indicates the NUTS level the respective area is in. Note that this information is indirectly also provided with the NUTS_ID column as the amount of numbers appended to the text indicate the level. Values are provided as Integer.
- *SHAPE_Leng* is a hidden field that is created when the data is uploaded into a Database. It describes the length when query asks for geometry. Values are provided as Real.

- *SHAPE_Area* is also a hidden field that is generated when the data is uploaded into a Database. It returns the shape when the query asks for the geometry. Values are provided as Real.

Furthermore, there is other data provided when opening the data set in QGIS, because the relations between the entities get broken down. Therefore, when identifying an area with the identification tool of QGIS, much more information can be viewed about the object like square kilometer, perimeter (in km) and some parameters that are of internal Eurostat use like information of outposts they have nearby.

Anyway, as the data set has to be imported into IPython for using it with Geopandas, it was converted into a Shapefile. The MDB format itself wouldn't be a problem for IPython to load, but for Geopandas. Additionally, it avoids getting specific data base drivers for the operating system to handle it.
Conveniently, QGIS can do this by converting the NUTS data set, which is currently loaded as MDB format, into a Shapefile by just exporting it. This can take some time depending on the size of the file, which in this case is only a couple of megabytes and therefore done very quick.
Preprocessing of the data set, before exporting it as an Shapefile, was not required in this case as the size wasn't that big and the data itself is useful as is.

Now that the data set of NUTS was loaded into IPython it was necessary to find a way of mapping the regions to their respective counterpart.
Eurostat does provide detailed NUTS maps, as PDF format, for each country that is in the data set.[29] Those maps contain every region for every NUTS level within the respective country. Moreover, they contain the NUTS code and the name of each region, which makes it possible to more conveniently find them also in OSM.

Having the region name, Nominatim was used to identify it via the search function. However, as sometimes the region names are not translated the search results sometimes did not provide the wanted inforamtion.
Consequently, another way for these cases had to be found in order to identify the region also in OSM. Nominatim provides Child-Of and Parent-Of information if an object is within our surounding another object. When knowing the region from the NUTS maps it was now possible to e.g. locate a city within that region and use the Child-Of relation to identify the administrative border where the city is in.

[29] http://ec.europa.eu/eurostat/web/nuts/nuts-maps-.pdf-, last accessed 2017-03-21

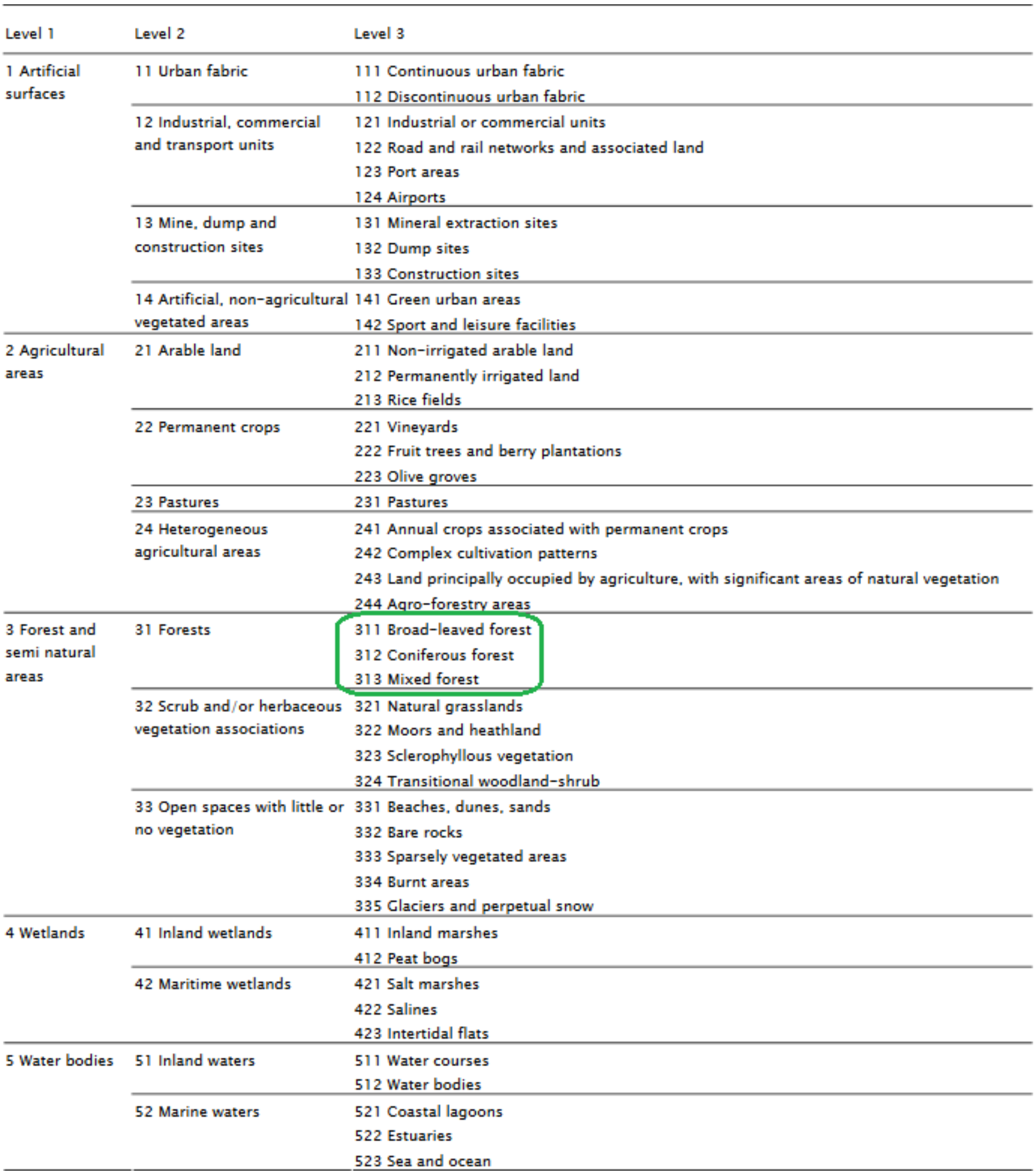
These two apporaches ensured that all the counterpart regions were found in OSM.

For mapping the regions the OSM ID and the NUTS ID could be used. However, as Overpass isn't always that stable and also didn't provide the format needed by Geopandas, the regions were preprocessed in the sense of exporting them as GEOJSON and loading them with Geopandas by providing the relative path to the exported file.

Now that the data can be loaded into IPython by using Geopandas, it is now possible to start comparing the NUTS and OSM regions as will be of further discussion in subsection 5.2.1.

4.1.2 Landuse forest on land side

For this use case the Corine Landuse data set from Copernicus was used. As briefly descirbed in subsection 3.6 every land use has a classification with a deeper granaularity behind. This level of granularity can partially be seen at figure 14 where the colors indicate a specific code.
Furthermore, when selecting a specific land use, these codes can be used for gaining information of the specific types of land use as shown in the following figure.



Level 1	Level 2	Level 3
1 Artificial surfaces	11 Urban fabric	111 Continuous urban fabric
		112 Discontinuous urban fabric
	12 Industrial, commercial and transport units	121 Industrial or commercial units
		122 Road and rail networks and associated land
		123 Port areas
		124 Airports
	13 Mine, dump and construction sites	131 Mineral extraction sites
		132 Dump sites
		133 Construction sites
	14 Artificial, non-agricultural vegetated areas	141 Green urban areas
		142 Sport and leisure facilities
2 Agricultural areas	21 Arable land	211 Non-irrigated arable land
		212 Permanently irrigated land
		213 Rice fields
	22 Permanent crops	221 Vineyards
		222 Fruit trees and berry plantations
		223 Olive groves
	23 Pastures	231 Pastures
	24 Heterogeneous agricultural areas	241 Annual crops associated with permanent crops
		242 Complex cultivation patterns
		243 Land principally occupied by agriculture, with significant areas of natural vegetation
		244 Agro-forestry areas
3 Forest and semi natural areas	31 Forests	311 Broad-leaved forest
		312 Coniferous forest
		313 Mixed forest
	32 Scrub and/or herbaceous vegetation associations	321 Natural grasslands
		322 Moors and heathland
		323 Sclerophyllous vegetation
		324 Transitional woodland-shrub
	33 Open spaces with little or no vegetation	331 Beaches, dunes, sands
		332 Bare rocks
		333 Sparsely vegetated areas
		334 Burnt areas
		335 Glaciers and perpetual snow
4 Wetlands	41 Inland wetlands	411 Inland marshes
		412 Peat bogs
	42 Maritime wetlands	421 Salt marshes
		422 Salines
		423 Intertidal flats
5 Water bodies	51 Inland waters	511 Water courses
		512 Water bodies
	52 Marine waters	521 Coastal lagoons
		522 Estuaries
		523 Sea and ocean

Figure 15: CORINE Land Cover (CLC) nomenclature[30]

The code table of figure 15 used by the Copernicus CLC is also used by OSM.[31] Consequently, this would supposedly make it very convenient for the mapping process.

[30] land.copernicus.eu/eagle/files/eagle-related-projects/pt_clc-conversion-to-fao-lccs3_dec2010, last accessed 2017-03-22

[31] http://wiki.openstreetmap.org/wiki/Corine_Land_Cover, last accessed 2017-03-11

However, in OSM the code (Integer) itself is not used, but the descriptions (String) are findable in the values of key (value=key principal)within the descriptive information of an geographic object. To find the descriptions in the values, one can query the following keys:

- *landuse*
- *leaf_type*
- *natural*

For the use case in the thesis the focus will be on the level 2 class "31 Forests", which more specifically is "311 Broad-leaved forest", "312 Coniferous forest" and "313 Mixed forest". The mapping for OSM requires the key *landuse* and *leaf-type*. The *natural* key is not needed as this is used for other types as "water" or "wood". However, for the use case in subsection 5.2.2 there will also be the *natural* listed as it is sometimes within the analyzed landuse area of OSM.
The mapping between OSM and Copernicus CLC was identified as the following:

Description	**Copernicus CLC**	**OSM**

Broad-leaved forest	311	landuse=forest
		leaf_type=broadleaved

Coniferous forest	312	landuse=forest
		leaf_type=needleleaved

Mixed forest	313	landuse=forest
		leaf_type=mixed

Table 6: Mapping of land use "forest" between Copernicus CLC and OSM

From the table 6 already several things can be derived. For defining a forest in OSM the value landuse with the key forest is always needed. Furthermore, it never stands alone as it should always be with with the leaf_type further described within the geographical object.
Another difference in the information needed to match the description is the amount of information that is needed for Copernicus CLC when comparing it with OSM. For Copernicus CLC there isn't any misinterpretation possible as there are only three values possible, which are uniquely referring to their respective description.
However, for the OSM part there is more information needed as not only the landuse, but also the leaf_type is needed. Moreover, as OSM is an open

standard not all fields are necessarily required and need to match what is described in their wiki. Thus, makes it possible that a forest can't uniquely be identified to a certain type or even leaves room for errors and misinterpretations as i.e. the second forest type already shows. It is named officially, according to the code table, a "Coniferous forest", whereas in OSM the *leave_ type* tag needs the value "needleleaved", which could possibly be misinterpreted by contributors who enter information into OSM.

The data set of the Copernicus CLC had about 4 gigabyte and was downloaded as GDB format. Consequently, for avoiding longer loading time and a better transition for geopandas, the data set was preprocessed.
The preprocessing was done by using QGIS and the functionalities that come with it. As the data set initially contains all the land use types, the ones that are not needed were removed. This is done by creating a query that would drop all rows that do not contain the codes "311", "312" and "313". Afterwards, the data was exported as Shapefile so that Geopandas can easily load the data as data frame in IPython.

For the OSM data there is the same problem as stated in subsection 4.1.1 that Overpass API does not support GeoJSON. Therefore, the data was exported and downloaded as GeoJSON to make it loadable into IPython via Geopandas.
Now the data can be accessed and loaded from both sources and is ready for further processing to be compared against each other.

Note that the data one downloads from OSM is everything that is retrieved as output from the defined query. Meaning that several i.e. forests could be combined and downloaded as one GeoJSON file. This information is necessary and will be further explained in subsection 5.2.2.

4.1.3 Landuse forest in urban areas

During research of subsection 4.1.2, differences between land side and urban areas were noticed. Therefore, it was decided to create a second use case, which also uses the Copernicus CLC data set, but focuses on manual comparison (subjective interpretation) of some urban areas where forests appear.
In addition, also satellite images of another source, namely Google Maps, were used to further support the manual analysis.

Consequently, not much data mapping was required to make land uses findable for queries. However, for finding the land uses for Copernicus CLC, coordinates where used for locating the exact position for analyzing the area. For OSM and Google Maps this was not needed, as the names of the land uses, which are normally some parks in urban areas, can all be found via their respective search functionality.

4.1.4 Population of European countries

The data set of the population is provided by Eurostat via a TSV file. This is convenient as it can be directly loaded into IPython, when providing the tabulator as separator.
As the use case only focuses on descriptive data i.e. population number, no further preprocessing steps were required as the data set itself is small and focuses on the core elements that are needed to compare population numbers for certain countries over time.
The matrix of the TSV file contains country acronyms as row header and year as column header. The cells are then filled with alphanumeric values that represent the population number.
The reason they are alphanumeric is, because some values have a letter appended (flags) that adds additional information to the number like forecast, estimate, provisional .... Those are needed if certain countries failed to provide current population numbers on time.
Nevertheless, those will not be further considered in this research, therefore, the letters will just be removed so that the value can be converted into an Integer value.

For the OSM part a manual analysis, of where to find the population data, was needed.
When using Nominatim to search for a country the search found all the desired countries and provided them as relation.
However, the relation of the country does not carry the population information. This gets inherited, and therefore needs to be searched for separately, by the node that represents the country. This was found when using the Parent-of relation that is offered by Nominatim.
The tag that was used is named "population" and the value is an Integer number representing the amount of inhabitants from the respective country.
The only exception, in which Nominatim didn't provide the Parent-of relation to list the node, was "France". Therefore, a query against Overpass API was created that searched within the relation of France for nodes with the name France. This worked pretty well as only one node got returned, which indeed was the relevant one containing the population information.
Afterwards, the queries were created and adjusted to also fetch for historical

population data.

Overpass query example:

```
 [date:"2016-01-01T23:59:00Z"];
node["population"](1683325355);
out body;
```

As with the other use cases, the outputs were exported and stored as GeoJSON format. Alternatively, as in this case the shape of object is not needed, also a direct query could have been created that directly fetches the data via the Overpass API and saves the value, of the population for a certain year, in a list.

4.1.5 Austrian address register

The Austrian address register is provided by the BEV via a CSV file. In fact, the whole data of the address register contains more CSV files, but for this research only the main one is necessary. From this CSV the following features and its values were needed.

PLZ	**HAUSNRZAHL1**	**HAUSNRBUCHSTABE1**	**RW**	**HW**	**EPSG**
1010	1.0	A	2303.63	341286.14	31256
1040	7.0		2972.65	339850.54	31256
3680	29.0		-97476.58	343352.28	31256

Table 7: Austrian Address Register data example

The above table provides an extract of how the data is formatted and provided by the BEV.

- *PLZ* is the postal code and needed to select the specific rows out of the CSV for comparing certain areas/districts with OSM. The information is obtained as an String value.
- *HAUSNRZAHL1* provides the information about the house number. The information is provided via a Float value.
- *HAUSNRBUCHSTABE1* contains the information if certain houses have a letter attached to their number. This is important as those two columns are separated and need to be adjusted for the query later. The information is provided via a String value.
- *RW* directly translated means "right value" of the coordinate, which is the longitude and provided as a Double value.

- *HW* directly translated means "high value" of the coordinate, which is the latitude and also provided as a Double value.

- *EPSG* is explained in the introduction of section 4 and provided as a Integer value.

For the use case, which is described at section 5.2.5, the focus was to have a small sample size of several districts. Therefore, the PLZ from the data source was used to fetch all the rows of the concerning PLZ into a data frame.
Afterwards, some data cleansing was done due to the fact that some HAUSNRZAHL1 had empty values. An empty value is displayed as NaN (Not a Number) and therefore needs to be removed as the query can't handle it otherwise.
Having the data frame cleansed it was now possible to use it and form a query against the Overpass API. While doing this it was also necessary to convert the provided EPSG 31256 into EPSG 4326, because Overpass API is working with this projection. Additionally, the HAUSNRZAHL1 had to be converted into Integer value due to the fact that it was converted as a float after fetching it from the CSV and there are no house numbers with decimal places.
After getting rid of the decimal places it was converted into a String as it is necessary to concat the HAUSNRZAHL1 with HAUSNRBUCHSTABE1, which is already a String. This is necessary as only then the complete house number is formed. This can also be seen in the table 7 as it shows that some houses need that text to specify it. However, this will be covered in more detail in section 5.2.5.
Now with this preprocessed data it is possible to form the query against the Overpass API and check for the house number with the allocated coordinates. Given that the coordinates reference a certain place (node) in OSM, a radius of about 15 meters was added that account for the GPS inaccuracy. Additionally, it is assumed that identical house numbers aren't in the range of 15 meters to each other.
Putting this in a loop for each row of the data frame makes it possible to do this in a convenient way and collect the outputs to base the analysis.

As some districts have a big number of house numbers the Overpass API kept crashing at several tries. This was caused by the server being too busy with other traffic and it also seemed to block after a certain amount of requests. The request problem could be solved with implementing a two minute delay after a certain amount (e.g. 500) rows were analyzed.

A data frame that contains about 4000 house numbers needs about 50 minutes to finish. So having a reliable algorithm that could finish is quite of some importance as those numbers sum up and one has to restart the code again.

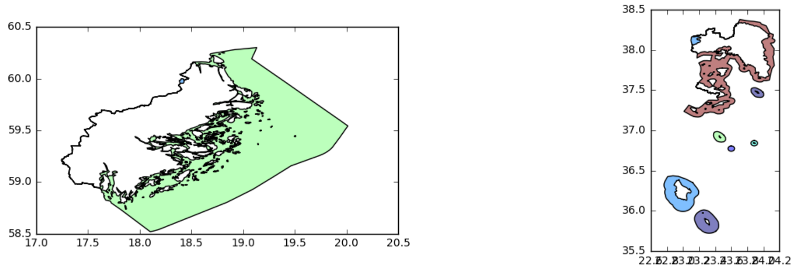
4.2 Unattended and problematic incidents

One of the problems that were faced when dealing with the data was not enough processing power. By this the processing power for operating QGIS is meant. It quite often happened that the program crashes when dealing with big data or lots of operations at the same time. This is caused by the fact that when doing e.g. some preprocessing by deleting some columns or rows, it does not process immediately.
Instead, it will be visualized immediately and processed when clicking the Save button. The problem was faced a number of times when trying out different data sets and with the ones that were actually used in the end.
Moreover, this is a very time consuming and tedious task to do over and over again. An alternative could therefore either be to have a higher processing power or try out different (commercial) tools like ArcMap.

Another problem that occurred quite often was that the Overpass API would crash when querying lots of information or even at random. Those errors and how they can partially be avoided can be seen in subsection 4.2.1.
Luckily, the OSM dump with Overpass API of the university of economics and business in Vienna could be used to fetch data. This was viable for data that is not so time critique as for example the address register data.
For other use cases, in which the historical information should also be compared, this was not useful, because the universities dump was only taken at a certain time with no historic data included.
Furthermore, even when querying against the universities Overpass interface, there were still crashes due to some load issues. Nevertheless, those were seldom happening and in that sense it helped to improve the reliability of fetching the data for certain use cases.
Concluding, that in general Overpass API seems to have a scalability issue that it takes longer and does not run consistently when dealing with more data.
When talking about longer times it is meant that when e.g. running a query 500 times with a time of 5 minutes the time for running the same query 1000 times is not 10 minutes, but rather 15 minutes, which indicates performance and scalability problems.

Use case: European regions

For some European regions the comparison was not feasible. The regions were both identified in the NUTS dat set and searched by their name via Nominatim in order to get the relation.
Then both regions get overlapped and the colored areas are what is not matching. More of the analysis and how it is done in detail gets explained later in section 5.



(a) Region in Sweden

(b) Region in Greece

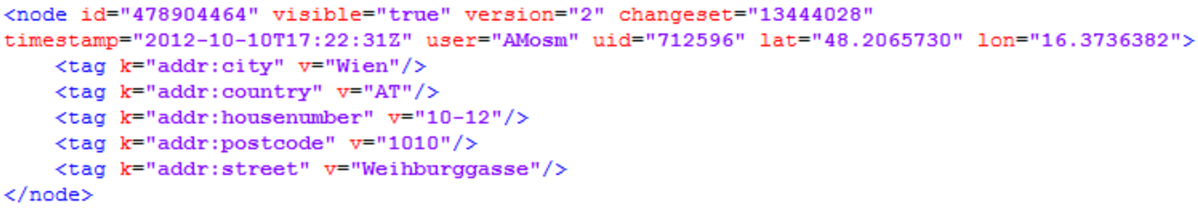
Figure 16: Examples of dropped regions

The two figures above show regions within Europe that are problematic to compare with NUTS. The reason being is that in OSM the regional territory ends somewhere below sea level, whereas in NUTS only the areas above sea level are covered in the data set.

Consequently, one would have to explicitly remove the water polygons from the OSM region in order to make it comparable in a more useful way. However, it shows that it isn't clear where a regions border should be placed or what it should contain. It makes sense that OSM also includes the sea area as territory to the respective region, but on the other hand this data might be hard to retrieve for the contributors. Thus, making the correctness questionable and would need to be further investigated how these data is retrieved, as it is under water area, and what is the correct way of drawing regional borders.

Use case: Address register

Although the completeness of the address register is good and offers potential to be even better when further tweaking with the query, there is a situation that is not handled in this research. Further sampling is needed in order to identify more patterns of how house numbers are provided in OSM, but this example makes clear what the problem is.



```
<node id="478904464" visible="true" version="2" changeset="13444028"
timestamp="2012-10-10T17:22:31Z" user="AMosm" uid="712596" lat="48.2065730" lon="16.3736382">
    <tag k="addr:city" v="Wien"/>
    <tag k="addr:country" v="AT"/>
    <tag k="addr:housenumber" v="10-12"/>
    <tag k="addr:postcode" v="1010"/>
    <tag k="addr:street" v="Weihburggasse"/>
</node>
```

Figure 17: Problematic house number

The snippet of the data structure of a certain building shows that it contains a range of house numbers. This by itself wouldn't be a problem if the numbers would be presented in a proper way (e.g. a node for each number within the way that represents the building).
However, in this case the house numbers are provided in the format "10-12", which not only makes it hard to identify the 10 and 12 uniquely, but also contains a hidden information about house number 11.
The provided source from the BEV only contains single house numbers and a letter if needed. Therefore, the mapping here is a challenge.
For now, even though the information is there, the query can't find it and recognizes those i.e. three house numbers as missing.

4.2.1 Data preprocessing

The initial idea was to form queries within the IPython framework, so that one can just execute the program and generates the output.
However, due to several reasons this was not possible for most cases.

One of the reasons was that the Overpass API kept crashing if query fetched too much data. Moreover, there were two main reasons for crashes:

- Server load too high: Appears when the load of the Overpass API service itself is currently too high, meaning that too many concurrent users are using it and it can't handle the request anymore. This can happen at random times and also with small queries.
- Too many requests: Was returned when the query is fetching lots of information via the Overpass API. This could partially be lowered when hard-coding delays after a certain amount of data was fetched.

Another reason, of why preprocessing was required, is the fact that the Overpass API does not support the GeoJSON format, which lead to manually exporting the OSM output as GeoJSON and storing them.

Other preprocessing steps were done to smaller the file size in order to minimize loading times. This was done by deleting data that was not required and generating Shapefiles with the help of QGIS. Further details have already been can be read in the section 4.

5 Comparison

For the data that is about to be compared, GPS systems were used to retrieve the coordinates. The GPS signals accuracy is about 10 meters. Therefore, the data can technically hardly be correct as the GPS is inaccurate and therefore adds uncertainty to the data.
A possible approach to solve this would be to account for the 10 meter accuracy by taking it as offset into account.
However, after some manual tryouts during the research it was decided to remove these offsets again, because they do not help in making the results more accurate. As by adding a certain offset to the data could even lead to adding bias. Even when defining that when the offsets touch each other for counting it correct, it could lead to false results as the actual data might be right and therefore those areas wouldn't match.

Another aspect, which is of relevance for the sake of comparison, is the sample size that should be taken. Considering in taking only a couple of examples could be too less of information in order to provide a reliable statement of the outcomes. However, using too many samples results in more time spent for the analysis part and also might not even be necessary for most cases as the outcomes might not vary that much and more information would not provide further insights.
Having the data ready and prepared, it will be shown in this section what differentiates the OSM data and the Eurostat NUTS, Copernicus CLC data sets. Furthermore, in order to provide understandable visuals, the data will be linked together.
Demographic information of the population will be compared between OSM and Eurostat by using the NUTS regions of Eurostat and finding the respective region in OSM. Furthermore, to have a comparable value between the samples, an accuracy value is introduced, which shows the relation between various area calculations.

5.1 Mapping of different structured data

In order to compare data, one has to map it correctly. This mapping process is different for every data set that was used during the cause of this research. These processes have similarities to a typical ETL (Extract Transform Load) process. Difference being that an ETL process normally fetches data from more than one data source in order to store it in a data warehouse. For this research the principal work flow was used as a method to integrate the data.

Extract

Data was offered in different formats and sizes. Therefore, data needed to be extracted for making it accessible within the development environment.

Transform

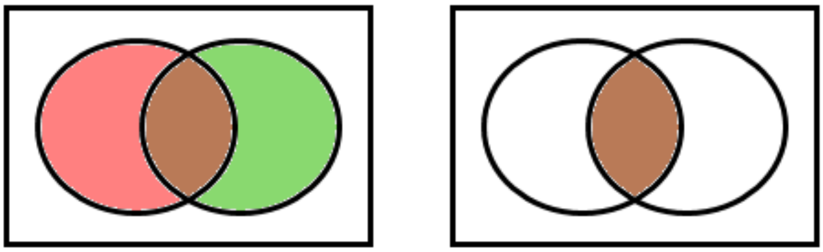
Preprocessing and cleansing was needed for making the data further processable with the development environment.
Further analysis showed that the data comes in different projections. Consequently, not only the preprocessing part was needed, but also the converting of projections in order to create a valid transformation that could be used to compare against OSM.

Load

Finally, once all previous steps are fulfilled, the data is ready to be compared against the data of OSM.

5.2 Analyze differences

For analyzing the differences, objects need to be compared against each other. In order to compare a source A against another source B, they need to be prepared and mapped properly.
There are different methods of how to compare two sources. For the figures and description to follow the red object will be A and the green object will be B.



(a) Union

(b) Intersection

Figure 18: Union and intersection of two shapes

Figure 18a shows the union $A \cup B$ of two objects. The areas where both objects overlap (shown in the figure as brown colored) is called intersection $A \cap B$ and shown in figure 18b by leaving only the intersecting area left.

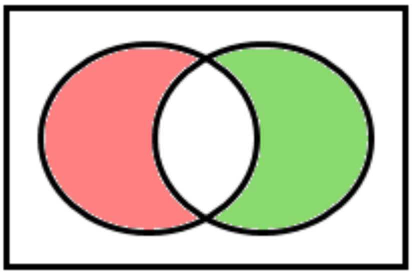


Figure 19: Symmetric Difference

In order to compare the difference of both objects, it is necessary to calculate the symmetric difference $A \bigtriangleup B$ as shown in figure 19. Therefore, we abstract the intersection (figure 18b) from the union (figure 18a)

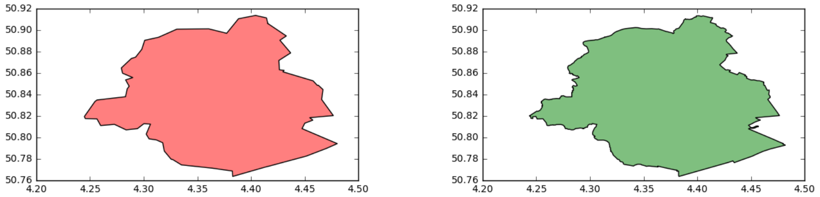
Symmetric Difference: $(A \cup B) \setminus (A \cap B) = A \bigtriangleup B$

For the use cases that are about to be compared, the objects, which were talked about before, are the shapes of the polygon. Those will be compared with the support of the calculation methods described above.
The two shapes will be positioned upon each other in order to make the overlapping areas of the polygon comparable. Once the overlapping (intersecting) areas are calculated they can be abstracted from the union and leave only those areas left that do not match.
Those areas will be colored in the same way, where one color will represent the polygon of the official source and the other will represent the OSM area.

From this point on, various outputs and analysis can be derived as a couple of areas are identified. Those can be compared against each other and brought into relation.
This does not necessarily require these polygons to carry over measurement information. All that is needed for this are the geographical information, which are provided by longitude and latitude. Those are points that are connected in order to form the polygons.
Using them is enough to calculate distances and areas in different units.
For this research the measurement unit will be kilometer (km) and used in the use cases that are about to follow in this section.

5.2.1 Regional areas

The first use case is about comparing region borders of some regions within European countries. For comparing the data, Eurostat NUTS is used, which was released in 2013, and the retrospective historic data of OSM of that specific region.
On the one hand, this analysis focuses on the comparison with Eurostat, which is for the comparison with official released data from an state authority. On the other hand, OSM regions were compared of how the borders might change over time. Changes of regional borders do normally not happen and therefore it is interesting to see if there are changes in OSM.
Changes in borders would imply that one region gets bigger, whereas the neighbour of that region would get smaller by this amount. This is due to the fact that the polygons are connected and do not have a, so to say, empty area in between them.



(a) NUTS 2013 Brussels

(b) OSM 2013 Brussels

Figure 20: Shape of Brussels (Belgium)

The figures above show the shape of Brussels (Belgium), which is one example out of several that were picked within the European Union. The figure 20a being extracted from the NUTS data set and the figure 20b is fetched from OSM.
Moreover, the NUTS data is displayed in red and the OSM data in green. This information is needed for later to see which areas are not overlapping when comparing them.

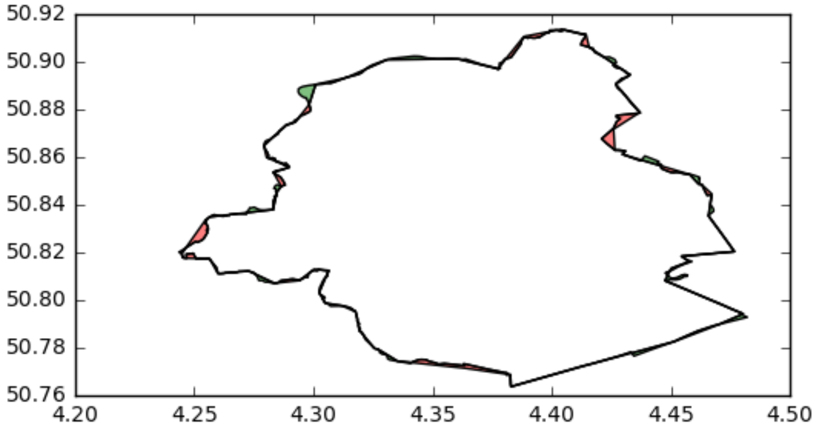


Figure 21: Symmetric Difference Brussels 2013

After mapping both sources of Brussels with each other, the symmetric difference results in the output that can be seen in figure 21.
The core overlapping area is matching pretty well as the majority of the area is white. Only some areas at the border can be identified to be different. This is indicated by colored areas in red or green, which represents the source as stated before.

Country	Region	sq. km ES	sq. km OSM	intersection	Symm. Diff.	Acc symm. diff.	int. acc.
belgium	Arr. de Bruxelles - Capitale	165.287	164.706	163.249	3.49506	2.11454	2.14094
bulgaria	Sofia	1354.75	1354.84	1344.84	19.9164	1.47012	1.48095
czech	Hlavni mesto Praha	496.204	495.888	492.041	8.00974	1.6142	1.62786
denmark	Hovedstaden	2575.83	2565.5	2550.83	39.6492	1.53928	1.55436
germany	Berlin	891.657	890.8	884.956	12.546	1.40705	1.4177
ireland	Dublin	971.576	976.601	964.012	20.2917	2.08854	2.10492
greece	Florina Regional Unit	1936.86	1937.83	1929.46	15.7775	0.81459	0.817716
spain	Comunidad de Madrid	8533.79	8531.42	8504.94	56.478	0.661817	0.664061
france	Paris	107.069	107.531	105.949	2.70211	2.5237	2.55038
croatia	Grad Zagreb	643.243	636.788	628.527	22.977	3.57206	3.65569
italy	Roma	5361.79	5373.47	5345.27	45.1408	0.841898	0.844501
latvia	Riga	307.916	306.261	303.978	6.22105	2.02037	2.04655
lithuania	Vilniaus apskritis	9830.3	9837.51	9787.63	92.5513	0.941491	0.945595
luxembourg	Luxembourg	2619.25	2620.79	2609.06	21.9195	0.836863	0.84013
hungary	Budapest	526.667	527.032	522.757	8.1845	1.55402	1.56564
malta	Malta	247.334	246.678	242.307	9.3994	3.80028	3.87913
netherlands	Gelderland	5183.05	5181.76	5168.2	28.4196	0.548318	0.549894
austria	Wien	411.756	414.636	409.015	8.36203	2.03082	2.04443
poland	Miasto Warszawa	518.844	518.953	515.822	6.15481	1.18626	1.1932
romania	Alba	6314.38	6347.26	6230.15	201.34	3.1886	3.23171
slovenia	Gorenjska	2135.89	2135.42	2128.06	15.1801	0.710716	0.713329
slovakia	Bratislavsky kraj	2050.55	2051.33	2043.29	15.296	0.745947	0.748596
united_kingdom	London	1617.92	1637.14	1611.41	32.2437	1.99291	2.00097

Table 8: Region comparison

Table 8 provides a representative collection of the analyzed shapes of several regions of different countries across the European Union. The first two columns contain the information from what region in which country was analyzed.
The next four columns, namely "sq. km ES", "sq. km OSM", "intersection" and "Symm. Diff.", are calculated in square kilometers and offer information of the size of the areas..

- *sq. km ES* provides information of the total square kilometer from the NUTS data set
- *sq. km OSM* provides information of the total square kilometer from the OSM
- *intersection* provides infromation of the intersecting parts of both sources in square kilometer
- *Symm. Diff.* provides the calculated square kilometers of all the areas that are not overlapping, namely the symmetric difference

For the last two columns the relations were calculated and shown in percentage. The first of the last two columns is "Acc. Symm. diff" and shows how much percentage the symmetric difference, compared to the total square kilometer if the NUTS region, is.

$\underline{\text{Accuracy value NUTS:}}\ \frac{SymmetricDifference}{SquarekilometerNUTSregion} \times 100$

The last column is "int. acc." and calculates the relation between the summentric difference and how much percentage it has compared to the intersecting areas.

$\underline{\text{Accuracy value intersection:}}\ \frac{SymmetricDifference}{intersection} \times 100$

Discovering some of the samples offered in table 8 points out differences between countries and regions.
The data is from the year 2013 as this was the year of the NUTS release that was used here. Overpass offers to query for historic information, which contains the historic information of changes in the shape of geographical objects.
The regions that were observed are of different sizes. Consequently, it is also necessary to calculate the relative aspects of the comparison.

Comparing the region of Germany (Berlin) with the one from Denmark (Hovedstaden) shows that, although the German region is about 3 times smaller, the relative difference of the symmetric difference to the NUTS data is almost the same. Same is true when the data is brought into relation with the intersecting areas.
Pointing out that the size isn't the reason for a bad accuracy of the OSM data, if such a bad result would exist that is.

However, there are results, therefore regions, that have better results than others. One example of this can be analyzed when looking at the region of Netherlands (Gelderland) and Malta.
The region within the Netherlands is about 20 times bigger then Malta, but the accuracy is higher, even though the square kilometers that are not matching are in total more. This points out the necessity of bringing the data into relation, because a square kilometer off for a small region i.e. Malta is more punishing for the accuracy than it is for the region in Netherlands.
With a difference of 3,252%, between the symmetric difference and NUTS, and a 3,33% difference, between the intersecting areas, they show also the widest spread outcomes of the samples that were analyzed.

Important to note again is that those non-intersecting areas are not one big area. They sum up and are summarized as results in the table 8.
For better understanding this, it helps to take a look again at figure 21 and observe the colored areas. They can be very small and sometimes hard or even not distinguishable from the border.

Country	Region	sq. km OSM 2013	sq. km OSM 2016	Difference
belgium	Arr. de Bruxelles - Capitale	164.706	164.645	-0.0605427
bulgaria	Sofia	1354.84	1353.57	-1.26985
czech	Hlavni mesto Praha	495.888	495.89	0.00221451
denmark	Hovedstaden	2565.5	2565.44	-0.0593794
germany	Berlin	890.8	890.739	-0.0611209
ireland	Dublin	976.601	973.953	-2.64814
greece	Florina Regional Unit	1937.83	1937.44	-0.395293
spain	Comunidad de Madrid	8531.42	8531.77	0.342255
france	Paris	107.531	107.534	0.00280448
croatia	Grad Zagreb	636.788	642.158	5.37026
italy	Roma	5373.47	5373.23	-0.242296
latvia	Riga	306.261	306.273	0.0115608
lithuania	Vilniaus apskritis	9837.51	9839.89	2.3798
luxembourg	Luxembourg	2620.79	2624.2	3.40747
hungary	Budapest	527.032	526.955	-0.0774413
malta	Malta	246.678	246.692	0.0138871
netherlands	Gelderland	5181.76	5181.75	-0.0161926
austria	Wien	414.636	414.634	-0.00173027
poland	Miasto Warszawa	518.953	519.033	0.0797907
romania	Alba	6347.26	6347.35	0.0859933
slovenia	Gorenjska	2135.42	2135.4	-0.0188426
slovakia	Bratislavsky kraj	2051.33	2051.36	0.0228111
united_kingdom	London	1637.14	1637.13	-0.00721411

Table 9: Region comparison OSM 2013 v. 2016

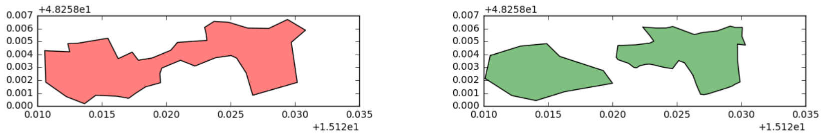
Table 9 lists the same regions as table 8, but now the OSM data from 2013 is compared with the OSM data of 2016. The goal of why the comparison is not with the data from Eurostat NUTS is to show if there are updates happening within those three years in OSM.
In the offline world, borders normally do not change, or at least not everywhere within a three year time span. Therefore, this analysis could contain some interesting results regarding any changes in borders of certain regions.

When looking at the table, one can see that there are changes done on every region. However, some of them are rather small and close to zero. Keeping the GPS inaccuracy in mind could explain those changes if different contributors evaluate locations differently over the three years.
It is fair to assume that all those changes might be due to slight adjustments to reflect the true border. Only exception could be interpreted in the change of Croatia as the 5,37 square kilometer are the most with a total area of about 642 square kilometer also in relation rather high.
Furthermore, this change is a little below one percent and when keeping Croatia's result from table 8 in mind (about 3,6%), could lead to a substantial improvement for the next comparison with a new release of the NUTS data set.
This also supports the assumption that those changes in the OSM data of

borders are for bringing them closer to a higher accuracy level.

5.2.2 Forest landuse at land side

The second use case is about comparing landuse of some forests within the European Union.



(a) Copernicus AT Landuse

(b) OSM AT Landuse

Figure 22: Shapes of landuse in Austria near Muenichreith

Figure 22a, which is colored red, shows the official data of a forest near an Austrian village named Muenichreit. Same does figure 22b, but this data is colored green and fetched from OSM. Due to the fact that the Copernicus data is from 2012, the historic data of OSM from 2012 was also taken in order to make it comparable.
From those two figures it is already clear that they are different in some aspects. The most obvious one being that from the Copernicus data, the forest is connected, whereas in the OSM data there are two separate forests.
This also means, in that case, that for the Copernicus data there is only one ID needed to identify this object in the database, whereas in OSM it is identified as two objects and requires two IDs. Therefore, a query needs to be formed that contains both OSM IDs in order to get this shape for comparison.
Analyzing the shape line by line, more differences can be found. In order to have a more comfortable way of comparing both shapes, the theory of symmetric difference is also applied here for calculating and showing the differences of where those two shapes are not intersecting.

Forests, especially very small ones, do most of the times not have a name to search for. Therefore, mapping forests from OSM with Copernicus is a taxing task and requires the use of longitude and latitude in order to find the counterparts.
However, after identifying a forest in OSM the IDs can be retrieved and used

for forming a query.

Overpass query example:

```
[date:"2013-12-31T23:59:00Z"];(
rel(2192257);>;
rel(2192363);>;
rel(2186660);>;
rel(2192292););
(._;>;);
out;
```

The first line of the query shows how to access the historical data by providing the exact date and time it should be accessed.
Line two to five contain four relations that are searched for in order to fetch them and create one single output that highlights those relations.
This patching together is required in order to make them better comparable, when the same forest in Copernicus i.e. contains only one area. Otherwise the comparison would hardly be feasible if only parts are compared with the total forest of the other source.
This will be more clear when looking at the concrete examples that are about to follow in this subsection later on.

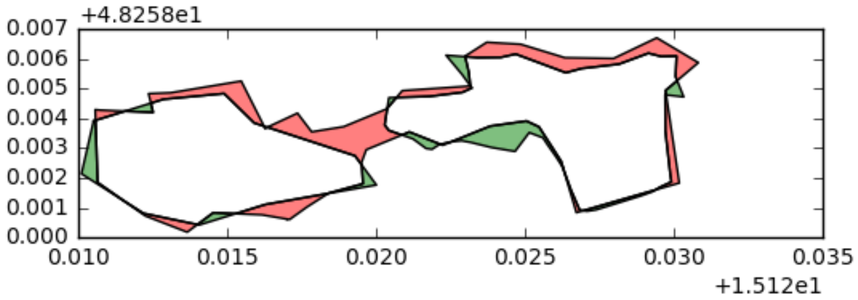


Figure 23: Symmetric Difference Landuse AT

Figure 23 shows visually the symmetric difference of this area where the forest/forests is/are located. The white area provides the information of where both sources intersect and the colored polygons are the respective polygons as explained above for figure 22a and 22b.

There are different finite types of landuse that are listed in figure 15. When investigating the table, one can see that there are certain codes that match a more specific type of a certain landuse. In this use case the landuse type "forest" was chosen for collecting and comparing. Forest has the code "31" and its respective sub types, which have the same prefix, but differentiate via a different suffix.

For the comparison, the data set from Copernicus CLC, which is released in irregular intervals, of the release year 2012 was chosen. Therefore, in order to make the comparison viable, the historic OSM data from 2012 was used. OSM uses the same code for specifying the landuse types, hence specific mapping wasn't required for comparison. However, the Corine landuse code isn't used as OSM tag or key. Instead, the name i.e. description of the land type is used to identify it.

This is different in the Copernicus CLC data set, because there the code is used and querying can be easily applied. For OSM one could query for the name as key (type = String), but needs to be specific in the tag that is used for querying the key.

This could potentially lead to ambiguous uses of the tag and key syntax, as there might be more possibilities of how OSM users understand those. This is the reason for not only comparing the shape of the area, but also the indicators of it when comparing it to the Copernicus CLC data set.

Location	sq km OSM	sq km Copernicus	sq. km Intersection	Symm. Diff.	symm. diff. Accuracy	intersection accuracy
AT_south_of_Muenichreith	0.468415	0.521556	0.436822	0.116327	22.3039	26.6304
ESP_north_of_madrid	8.6904	8.62711	8.20771	0.902081	10.4564	10.9907
DE_south_of_trier	0.33094	0.570776	0.281381	0.338952	59.3845	120.46
ITA_south_of_trieste	0.203763	0.505767	0.186145	0.337239	66.6788	181.17
LUX_south_of_luxembourg	1.60327	1.58615	1.44641	0.392329	24.7346	27.1243

Table 10: Landuse comparison with intersection

The information of table 10 allows to analyze landuse of certain areas that contain forests. The samples were picked at random from several countries for getting a better understanding of the quality and completeness regarding landuse types and shapes.

The columns and needed calculations are the same, just with different data, as in subsection 5.2.1.

The table shows that the differences regarding the shape of the landuse type forest, is in relation way higher than the ones for regions, which is shown in table 8 .

This already provides some insights in the quality differences of certain OSM areas. According to this information it looks like the borders are more accurate than the landuse areas. Also the variety is higher when looking at the symmetric difference accuracy and intersection accuracy. For the landuse

examples of DE and ITA the intersection accuracy is even above 120 and 180 percent, which means that the non-overlapping areas are, twice and even more, larger then the overlapping areas.
This in essence could be interpreted as that different forests exist in both sources.
Such inaccuracy is a tremendous find regarding the shape size and form when comparing these polygon areas.
However, when looking at the sizes of the forests it shows that they have almost the same size, exception being DE and ITA.
Concluding, that it could be just the shape that's wrong and needs adjustment and not that there are new forests created due to huge differences.
For DE and ITA, as their areas show such a big difference, it could be possible that one or the other contains data that is not accounted for in the other source.
Meaning that e.g. an area that contains both grass and trees is accounted as forest in the one and as grass area in the other source. However, up to this point this is an assumption and needs further investigation.

Location	ID	Landuse	leaf_type	natural	square km	square km tot	relation total
AT_south_of_Muenichreith	way/13864914	forest			0.220816	0.468415	47.1411
AT_south_of_Muenichreith	way/177365061	forest			0.247599	0.468415	52.8589
ESP_north_of_madrid	way/29021369	forest			8.6904	8.6904	100
DE_south_of_trier	way/420800966	forest	mixed		0.33094	0.33094	100
ITA_south_of_trieste	way/156314070			wood	0.203763	0.203763	100
LUX_south_of_luxembourg	way/31009206	forest	mixed		1.39646	1.60327	87.1006
LUX_south_of_luxembourg	way/131343117	forest			0.15631	1.60327	9.74942
LUX_south_of_luxembourg	way/28358164			water	0.000621102	1.60327	0.0387396
LUX_south_of_luxembourg	way/28358165			water	0.00194033	1.60327	0.121023
LUX_south_of_luxembourg	way/200931208	grass			0.00154534	1.60327	0.0963863
LUX_south_of_luxembourg	way/200931209	grass			0.0463962	1.60327	2.89384

Table 11: OSM Landuse detailed

More often than less an area has to be created by combining different forest shapes. This means that when analyzing a landuse area it happens that it is "Parent-of" several forest shapes. Those need to be combined in order to get the area as a whole and make it comparable.
The reason for such a separation can be indirectly seen in table 11 when looking at some examples that have more or different indicators. In order to tell what percentage a certain forest is made out of, this separation is necessary to provide information of different "leaf types" or "natural" and make them account for when analyzing the forest as a whole.

Relation total: $\frac{Squarekilometer}{Squarekilometertotal} \times 100$

The "relation total" expresses the relation of a single shape of OSM data

compared to the total square kilometer of that forest area it is contained in, so to speak a "Child-of" relation. For some areas this is 100 % as they only contain one shape and therefore only have one ID that needs to be used to find those areas.
According to what is stated in table 6 only the "Landuse" and "leaf_type" should be required to describe a forest area in OSM.
However, as all IDs were fetched from the area that was subject of the analysis it was found that not every ID contained information in those tags.
After further investigation it was found that those areas were not specified as forests, but "wood" and "water" areas were found in these cases in the tag "natural".
As also the same area in Copernicus was analyzed it was decided to not leave this information out, because it might also be visible in Copernicus and would otherwise cause empty areas.

Location	code	Landuse	km	tot sq km	percentage
AT_south_of_Muenichreith	312	Coniferous forest	0.521556	0.521556	100
ESP_north_of_madrid	311	Broad-leaved forest	1.93189	8.62711	22.3933
ESP_north_of_madrid	312	Coniferous forest	3.9893	8.62711	46.2415
ESP_north_of_madrid	313	Mixed forest	0.828037	8.62711	9.59808
ESP_north_of_madrid	313	Mixed forest	1.02671	8.62711	11.901
ESP_north_of_madrid	312	Coniferous forest	0.481965	8.62711	5.58663
ESP_north_of_madrid	311	Broad-leaved forest	0.369199	8.62711	4.27952
DE_south_of_trier	311	Broad-leaved forest	0.570776	0.570776	100
ITA_south_of_trieste	313	Mixed forest	0.505767	0.505767	100
LUX_south_of_luxembourg	311	Broad-leaved forest	1.58615	1.58615	100

Table 12: Copernicus Landuse detailed

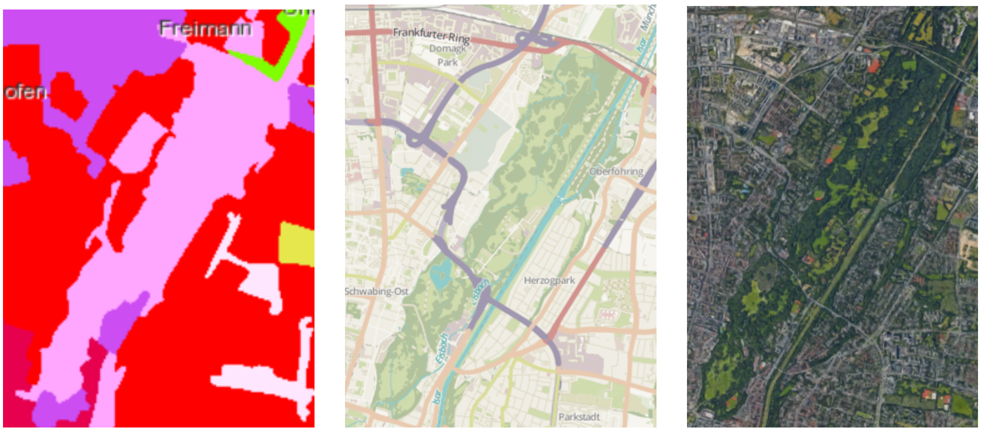
The table 12 shows the Copernicus forest areas, which are summarized in table 10. The percentage indicates that most areas are 100 %, meaning that those can directly be used for comparing it with the OSM data.
However, an exception is the the forest in Spain (ESP) as six areas were needed in order to make it comparable with the OSM data. The reason for this can partially be found when looking at the code or landuse columns. They indicate that the forest contains different types of leafs and therefore was split upon those. With this separation one can analyze what types to what percentage of leafs are in this forest. When compared with the OSM data in table 11 it is clear that the Copernicus CLC data set is more sophisticated. OSM data provides barely any information about the specific types of leafs or naturals. Most of the time just "forest" is entered as landuse type, which is sufficient when searching for forests, but useless when searching for certain kinds of forests.

The Copernicus CLC data set does not even contain the broader term "forest". Every feature uses the specific code of the parent "forest", which indicates that the term "forest" in general could be avoided, because it is to unspecific and obsolete if the code or specific Landuse is provided.
Given that Copernicus CLC data is released from an official source, it is assumed to be the most accurate data. Furthermore, this would mean that using the term "forest" alone, without any "leaf type" like in the OSM examples, is inaccurate per default.
Concluding, that there always has to be a "leaf type" when "landtype=forest" in the OSM data is used.

5.2.3 Landuse urban area

While investigating and identifying samples to use for subsection 5.2.2 another interesting aspect was identified. Therefore, the decision was made to dedicate a separate use case for these samples and have a more critical view on the Copernicus CLC data set.
As it seems there are differences between forest areas on land side and urban areas. Consequently, to make it visually more appealing, this use case was done via manual observing the areas by using several sources.
Copernicus CLC, OSM and Google Maps were used for analyzing parks within cities. Moreover, four samples were analyzed, which offer insights on the correctness of those areas.
The observations were made by zooming in on the map to the specific areas and interpret what was seen.
Google Maps was added as third source as it offers satellite images and is peer-reviewed by experts. Even though the review of the Google Maps data isn't needed at that point as the satellite images are enough to draw conclusions from.

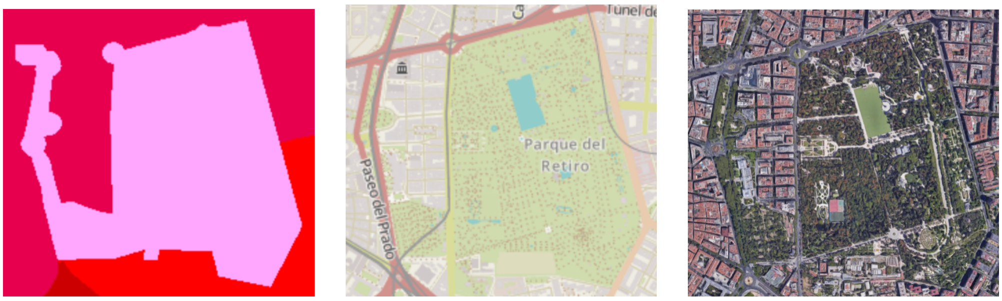
In the examples that are about to follow it is important to understand that the pink color in the Copernicus data set is accounting for "Green urban areas" and has the code "141", which can be also seen in the code table at figure 15.



(a) Copernicus (b) OSM (c) Google Maps

Figure 24: Comparison of "Englischer Garten" in Munich(DE)

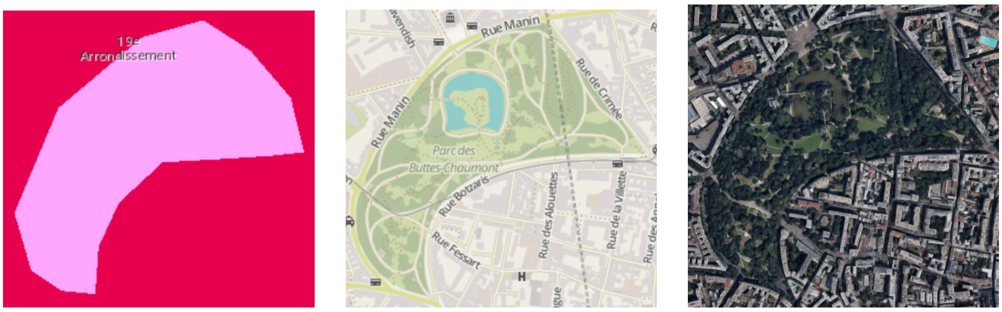
The figures of figure 24 show already a major aspect that was discovered during the research. The Copernicus CLC data at that region mostly contains "Green urban areas", which is innacurate when looking at the respective area of OSM and Google Maps.
In OSM, the dark green shapes show some forest types, which is confirmed when analyzing the same shapes at the Google Maps satellite images.
Thus, making the Copernicus CLC inncaurate for the use in urban areas.



(a) Copernicus (b) OSM (c) Google Maps

Figure 25: Comparison of "El Retiro" in Madrid (ESP)

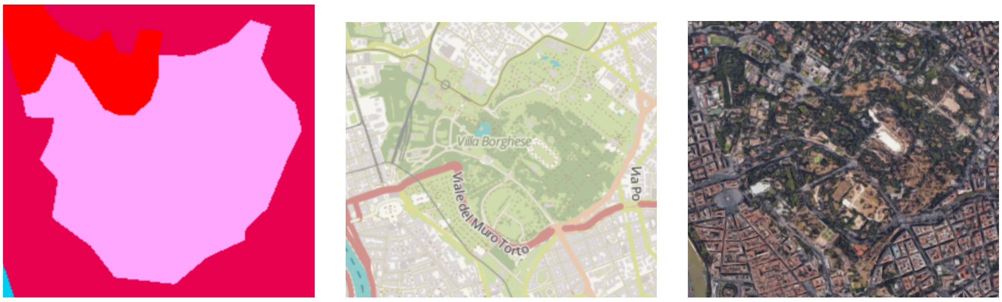
This can further be seen in the sample of figure 25. Here, the park "El Retiro", even has a water type area that indicates a small pond. This can be seen in both OSM and Google Maps, but not in the Copernicus CLC data. Furthermore, there is no differentiation again between forest or grass as it simply is all categorized as "Green urban areas" again.



(a) Copernicus (b) OSM (c) Google Maps

Figure 26: Comparison of "Parc des Buttes Chaumont" in Paris (FR)

The park in Paris (figure 26) and the one in Rome (figure 27) have similar outcomes as Copernicus always qualifies those whole areas under the same category, but they cannot be simply put there as they clearly contain other types of land cover.
Nevertheless, it sometimes is very hard to see at the Google Maps images if there are forests or water areas, but for the most parts it is at least clear that there are bigger forests within those parks.



(a) Copernicus (b) OSM (c) Google Maps

Figure 27: Comparison of "Villa Borghese" in Rome (ITA)

This puts the Copernicus CLC data into a new light regarding their trustworthiness of data quality. However, one could argue that for simplification sake's the parks in cities were categorized that way. This could still then be considered a flaw in the data set for urban areas as other sources, e.g. Google Maps and OSM, have those information in their data.
Concluding that those sources might be of more use for urban areas than the Copernicus CLC data is.

5.2.4 Population

For having a demographic example that could also be used to analyze the historic descriptive data of OSM, the population data from Eurostat was chosen to compare against OSM.

The Eurostat population data is released at the beginning of every year and its data is collected by their respective European Member States.

The use case aims to see differences in the time line of the population data between those two sources.

Country	ES 2014	OSM 2014	diff	%	ES 2015	OSM 2015	diff	%	ES 2016	OSM 2016	diff	%
Austria	8506889	8205533	301356	96	8576261	8205533	370728	95	8700471	8205533	494938	94
Belgium	11203992	11035948	168044	98	11208986	11035948	173038	98	11289853	11035948	253905	97
Czech	10512419	10516125	-3706	100	10538275	10516125	22150	99	10553843	10516125	37718	99
Germany	80767463	81879976	-1112513	101	81197537	81879976	-682439	100	82162000	81879976	282024	99
Spain	46512199	46157822	354377	99	46449565	46157822	291743	99	46438422	46157822	280600	99
France	65889148	65073482	815666	98	66415161	65073482	1341679	97	66661621	65073482	1588139	97
Croatia	4246809	4284889	-38080	100	4225316	4284889	-59573	101	4190669	4284889	-94220	102
Hungary	9877365	9930915	-53550	100	9855571	9930915	-75344	100	9830485	9930915	-100430	101
Italy	60782668	59619290	1163378	98	60795612	59619290	1176322	98	60665551	59619290	1046261	98
Slovakia	5415949	5404322	11627	99	5421349	5404322	17027	99	5426252	5404322	21930	99

Table 13: Population

During the creation of table 13 it got clear very soon that something is off with the OSM data. The population number never changes in OSM from 2014 till 2016.

Investigating behind the reason for this brought no concrete result. Therefore, it is assumed that it has to do with using the Overpass API and its limitations regarding historic data as pointed out in subsection 2.5.1.

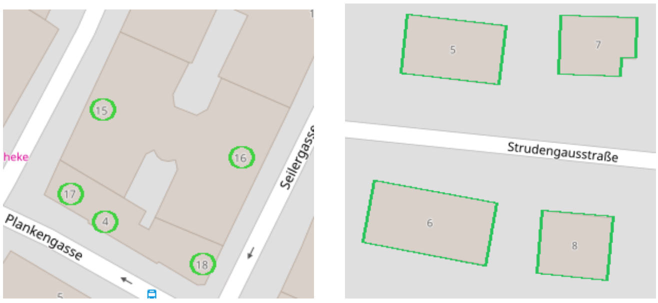
Nevertheless, after manually trying older dates, it was found that the population number does change, but one has to go as far as the year 2012 to get alternating population results in OSM.

5.2.5 Austrian address register

The goal of the comparison was to analyse the completeness of addresses in OSM, by using the Austrian address register.

Nevertheless, during the research several interesting aspects came up, which led to investigate beyond completeness and attempt to find reasoning in the initial results that were gathered.

Initially, a region on the land side of Austria was analyzed as the amount of houses are less and therefore faster to query for. The house numbers were found only within ways, therefore, a query was formed that searches all the ways of that region, depending on their respective coordinates, if their related house number exists.



(a) Houses with nodes as house number

(b) Houses with ways as house number

Figure 28: Locating house number in OSM

Figure 28 provides insights of how house numbers are saved in OSM. It is indicated by the green circles (nodes) and rectangles (ways). When comparing OSM with the Address Register of BEV, the results were on some examples very bad when only taking ways, as seen on figure 28b, into account. Therefore, after manually investigating different houses the reason behind these bad results was clarified. Some houses have their house number saved in the way that draws the border of the house, whereas others have a node within that way that describes the house further and contains the house number. This was found after testing the initial query against a district of Vienna.

Consequently, nodes were added to the query in order to make the house numbers findable. This is important as it influences the completeness significantly as seen in the examples below.

It makes clear that having no standardization in such cases is a problem. This is even more significant when taking into account that this is appearing within the same country and even within the same districts.

When trying to identify the best solution, it gets clear when analyzing figure 28a. having several house numbers within one building makes nodes the best option as only then they can be clearly identified and their location (entrance) seen clear.

PLZ	**housenumbers**	**ways**	**nodes**	**total**	**total completeness**
1010	2560	121	2162	2283	89%
1030	4130	1416	1766	3182	77%
1040	1721	550	1012	1562	90%
1060	1612	521	951	1472	91%
3680	767	698	1	699	91%
8700	4306	2101	588	2689	62%

Table 14: Address results without house number letters

The initial analysis of various districts was done without appending the letter of the house number to it. Those results are listed in table 14, whereas the first four contain results from districts of Vienna and the last two are of some land side areas.
The original district that was mentioned in the introduction of this subsection has the PLZ "3680" and contained 767 house numbers. Note that the quantities in the *housenumbers* column are already cleansed, meaning that the rows that contained NaNs were removed.
It gets clear why the research was wrongly assuming that house numbers are stored in the descriptive information of ways as nearly all of the 696 found matches are ways. Consequently, the accuracy, indicated by the *total completeness* column, with 90% undermined this.
Looking at the first district of Vienna, which is the first row of table 14, caused the manual investigation as only 121 matches were found out of 2560 house numbers, which is a rather bad accuracy with about 4%.
Therefore, the table provides both information, either nodes or ways, of where the house numbers were found.

Afterwards, in the attempt to get an even better results, the letters of the house numbers were appended to the query. Reasoning being that the letter adds more precision to a house number and after manually analyzing some houses, it was confirmed that the letters are also added to the house number in OSM.

PLZ	**housenumbers**	**ways**	**nodes**	**total**	**total completeness**
1010	2560	118	2121	2239	87%
1030	4130	1386	1757	3143	76%
1040	1721	540	1014	1554	90%
1060	1612	511	955	1466	90%
3680	767	695	1	696	90%
8700	4306	1987	582	2569	59%

Table 15: Address results using house number letters

Table 15 provides results of the query when adding letters to the respective house numbers. Surprisingly, it didn't improve the majority of the results and even made most of them worse.
This indicates, when searching more sophisticated, it does not find the house numbers in OSM. Concluding that the letter is not always attached to the house number in OSM.
However, the results kept quite balanced, which also means that some house numbers were found that weren't found before. Furthermore, some that were found before aren't now because they didn't required letter appended initially.

Overall, the results make clear that a standardization is needed in the information of house numbers should be added. Out of the research it is proposed to go with saving the house number, or information about specific houses in general, within the nodes. This is supported by the fact that one building can have several house numbers, which indicate different entrances.

Nevertheless, even though the results got better after adding nodes to the query, there are patterns that hardly can be identified, as summarized for this use case in subsection 4.2, and could be used for further works.

6 Conclusion

This section will summarize the key findings and discuss suggestions for further works. OSM, hence VGI in general, enjoys great popularity of citizens that want to contribute their knowledge for others. Those contributors are specialists in their region and OSM enables them to share their region specific knowledge.
However, as those contributors are non-domain experts in the field of GIS, the data they enter is prone to errors and ambiguous interpretations. Moreover, this is fostered by the fact that OSM uses open standards, which offers contributors even more possibilities to enter data incorrectly.

The evaluation of the trustworthiness of OSM was done via comparison with data sets from official sources provided by European authorities. Authorities, namely Eurostat, Copernicus and BEV, publish those data sets for everyone to use. Several use cases have been defined to analyze quality aspects like shape, descriptive and historical information of geographical objects.

OSM maintains a high degree of correctness when dealing with larger areas such as regional borders, which were the topic of one use case. The comparison of shapes was done by applying the symmetric difference of both sources and calculate non-intersecting square kilometers. The outcomes show that they differ from the Eurostat NUTS data set only slightly (between 0,66% and 3,65%).
Nevertheless, our analysis in the second use case shows that the results are not as promising when considering land use shapes and types of forests.
The CLC data set from 2012 was used for comparing OSM and Copernicus. First, both sources do not necessarily have one ID for one forest area, which means that those have to be patched together for making them comparable. This makes sense, if the forests contains several different leaf types and each smaller shape could represent a relation of the contribution of a certain type that way. Second, the results after applying the symmetric difference degraded significantly as non-intersecting parts were up 180% different from the intersecting part. This points out the problems that occur when entering land use shapes. It is difficult to estimate where a forest border should end, because the landuse types are most often mixed e.g. with trees, grass, rocks, water and so contributors might have different views on how to input the data for representing a forest.
Additionally, the second use case focuses also on the types and if those match with the Copernicus CLC data. The problem that was identified is the inac-

curate or sometimes missing tag information in OSM. Most of the time the contributors only enter that it is a forest, but not what specific leaf type information this forest carries. This is done with high accuracy in the Copernicus CLC. In fact, Copernicus CLC does not even use the general term forest as it is obsolete by default when providing a specific forest leaf type

While analyzing land types, the investigation process identified interesting aspects at urban areas, which lead to create an own use case for this. It was found that land types for urban areas e.g. parks only have a general type "Urban grass area" in Copernicus CLC, whereas in OSM this is more specific as also trees and water areas were shown. Therefore, the satellite images of Google Maps were used for confirming this detailed structure.
This puts the Copernicus CLC data into a new light, as it cannot be used for a detailed comparison in urban areas.

A use case for looking into demographic information and their historic changes was created by using the Eurostat population data set that is published at the beginning of every year.
While there are changes observed in the Eurostat data set the research could not observe any changes in the OSM historic data as the population number never changed.
This was most likely the cause of using the Overpass API, which was confirmed to not save all the historic information. Therefore, the output of the sample countries are not representing the actual situation regarding the correctness of OSM population data.

Last but not least, a use case that analyzes Austrian addresses was created by using the Austrian address register provided from BEV. The goal was to check the completeness of certain Austrian districts within OSM, regarding the addresses that are inside the districts. The results show that the completeness, with some exceptions, is around 90%, but offers some opportunities for improvement. This is due to the fact that there are certain cases that contain the right information, but the way the contributors added the information is not standardized and therefore requires query adjustments.

Overall, OSM proved to be useful for analyzing bigger objects, but not for smaller objects regarding the geometric aspects. For the alphanumeric information the results were mixed as for forests they appear rather unsatisfying, whereas for the address register they are good. However, the address register use case was only evaluated for Austria.

6.1 Future works

All of the use cases offer potential to be extended via introducing additional steps for a deeper analysis. Being the reason to improve the results by removing problematic patterns as discussed in subsection 4.2 or focusing on single use cases as more questions got raised after their analysis.
For most further approaches the same or similar process model could be used as seen in figure 10.

Scan whole regions for their land use. In order to achieve this goal more processing power and a better interface are needed. During the implementation phase a lot of crashes happened due to high traffic and restrictions on the Overpass server. This can partially be avoided, but still happened occasionally, when using a dump of OSM on a private Server without restrictions.
Afterwards, the Copernicus CLC data set could be completely queried for certain regions to retrieve all e.g. forest shapes and compare them with their counterparts in OSM.
This could lead to interesting findings if forests are missing and the state of correctness for certain regions.

Query improvement for address register. To achieve better results for comparing the house numbers of the Austrian Address Register with OSM a query optimization is required. As the solution that is used in this thesis searches per house number, it was found that in OSM sometimes ranges are provided.

Gain access to historic population data of OSM. As Overpass does not provide proper historic population data, another source could be used that provides this information. This would require a comparison of various sources with samples that fetch historic population information from OSM.

Compare OSM with future NUTS releases. As NUTS is published in irregular intervals, but at least every two or three years, this could be used to conduct the same analysis again and compare the results. It would be interesting to see if OSM keeps continuing to strive against a better correctness that comes close to the data of NUTS.

Investigate deeper the bad results of forest land use comparison. Interesting aspects might be found of why nearly all results of forest land use are not matching with their counterpart in OSM. Most obvious one being

that contributors are unsure of where forests start and end.
Decision making processes of where does a forest end and another land use type start could be analyzed by surveys that contain pictures of such border cases and let the respondents draw lines of where they think the border is. Additionally, the same survey should be handed over to experts in order for making it possible to compare those results.
This might offer also the potential to not only conduct an empirical research, but also qualitative research that requires interviewing OSM contributors and experts from Copernicus.
Some of the problems are pointed out in their technical guide[32] were they state that changes are only accounted for if a certain size (in ha) is reached.

[32] land.copernicus.eu/user-corner/technical-library/CLC2006_technical_guidelines.pdf, last accessed 2017-03-29

References

[1] 1997, 1998 Environmental Systems Research Institute, Inc. Esri shapefile technical descriptio. http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf, 1998. [Online; accessed February 24, 2017].

[2] Aggrey Agumya and Gary J Hunter. Determining fitness for use of geographic information. *ITC Journal*, (2):109–113, 1997.

[3] Ahmed Loai Ali, Falko Schmid, Rami Al-Salman, and Tomi Kauppinen. Ambiguity and plausibility: managing classification quality in volunteered geographic information. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 143–152. ACM, 2014.

[4] Jamal Jokar Arsanjani, Alexander Zipf, Peter Mooney, and Marco Helbich. An introduction to openstreetmap in geographic information science: Experiences, research, and applications. In *OpenStreetMap in GIScience*, pages 1–15. Springer, 2015.

[5] Graeme F Bonham-Carter. *Geographic information systems for geoscientists: modelling with GIS*, volume 13. Elsevier, 2014.

[6] M Bossard, J Feranec, J Otahel, et al. Corine land cover technical guide: Addendum 2000. 2000.

[7] Tim Bray. The javascript object notation (json) data interchange format. 2014.

[8] Tim Bray, Jean Paoli, C Michael Sperberg-McQueen, Eve Maler, and François Yergeau. Extensible markup language (xml). *World Wide Web Consortium Recommendation REC-xml-19980210. http://www. w3. org/TR/1998/REC-xml-19980210*, 16:16, 1998.

[9] Howard Butler, Martin Daly, Allan Doyle, Sean Gillies, S Hagen, and T Schaub. The geojson format. Technical report, 2016.

[10] George Büttner and Barbara Kosztra. Clc2006 technical guidelines. *European Environment Agency, Technical Report*, 2007.

[11] Nicholas R Chrisman, David J Cowen, Peter F Fisher, Michael F Goodchild, and David M Mark. Geographic information systems. *Geography in America*, pages 353–375, 1989.

[12] Jeffrey P Cohn. Citizen science: Can volunteers do real research? *BioScience*, 58(3):192–197, 2008.

[13] Michael Curry. *Digital places: Living with geographic information technologies.* Routledge, 2008.

[14] Fausto D'Antonio, Paolo Fogliaroni, and Tom Kauppinen. Vgi edit history reveals data trustworthiness and user reputation. 2014.

[15] Philip Fennell. Extremes of xml. *XML LONDON 2013*, 2013.

[16] Rosmadi Ghazali, Zulkiflee Latif, Abdul Rauf Rasam, and Abd Manan Samad. Integrating cadastral gis database into gps navigation system for locating land parcel location in cadastral surveying. In *Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium on*, pages 469–473. IEEE, 2011.

[17] MF Goodchild. Citizens as voluntary sensors: spatial data infrastructure in the world of web 2.0. ijsdir 2: 24–32, 2007.

[18] Michael F Goodchild. Geographic information systems. *Progress in Human geography*, 15(2):194–200, 1991.

[19] Michael F Goodchild. Geographic information system. In *Encyclopedia of Database Systems*, pages 1231–1236. Springer, 2009.

[20] Michael F Goodchild and Linna Li. Assuring the quality of volunteered geographic information. *Spatial statistics*, 1:110–120, 2012.

[21] Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.

[22] Muki Haklay. Citizen science and volunteered geographic information: Overview and typology of participation. In *Crowdsourcing geographic knowledge*, pages 105–122. Springer, 2013.

[23] Carsten Keßler, Johannes Trame, and Tomi Kauppinen. Provenance and trust in volunteered geographic information: the case of openstreetmap. 2011.

[24] Carsten Keßler, Johannes Trame, and Tomi Kauppinen. Tracking editing processes in volunteered geographic information: The case of openstreetmap. In *Identifying objects, processes and events in spatio-temporally distributed data (IOPE), workshop at conference on spatial information theory*, volume 12, 2011.

[25] Steve LaValle, Eric Lesser, Rebecca Shockley, Michael S Hopkins, and Nina Kruschwitz. Big data, analytics and the path from insights to value. *MIT sloan management review*, 52(2):21, 2011.

[26] Paul Longley. *Geographic information systems and science.* John Wiley & Sons, 2005.

[27] David J Maguire. An overview and definition of gis. *Geographical information systems: Principles and applications*, 1:9–20, 1991.

[28] Mark Monmonier. *Rhumb lines and map wars: A social history of the Mercator projection.* University of Chicago Press, 2010.

[29] Pascal Neis and Alexander Zipf. Analyzing the contributor activity of a volunteered geographic information project–the case of openstreetmap. *ISPRS International Journal of Geo-Information*, 1(2):146–165, 2012.

[30] Don Parkes and Nigel J Thrift. *Times, spaces, and places: A chronogeographic perspective.* Wiley, 1980.

[31] Fernando Pérez and Brian E. Granger. IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29, May 2007.

[32] Barna Saha and Divesh Srivastava. Data quality: The other face of big data. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 1294–1297. IEEE, 2014.

[33] Simon Scheider, Carsten Keßler, Jens Ortmann, Anusuriya Devaraju, Johannes Trame, Tomi Kauppinen, and Werner Kuhn. Semantic referencing of geosensor data and volunteered geographic information. In *Geospatial semantics and the semantic web*, pages 27–59. Springer, 2011.

[34] Yakov Shafranovich. Common format and mime type for comma-separated values (csv) files. 2005.

[35] John Parr Snyder. *Map projections–A working manual*, volume 1395. US Government Printing Office, 1987.

[36] Johannes Trame and Carsten Keßler. Exploring the lineage of volunteered geographic information with heat maps. *Proceedings of GeoViz: Linking Geovisualization with Spatial Analysis and Modeling, Hamburg, Germany*, pages 10–12, 2011.

[37] Georg Trogemann and Michiel Pelt. Citizen media–technological and social challenges of user driven media. In *Proceedings of the BroadBand Europe conference*, pages 11–14, 2006.

[38] David L Tulloch. Is vgi participation? from vernal pools to video games. *GeoJournal*, 72(3-4):161–171, 2008.