MSc thesis in Geomatics
Faculty of Architecture and the Built Environment

Comparison of Remotely Sensed and Volunteered Geographic Information for water reservoirs

Maria Moscholaki
November 2020

TU Delft

Deltares
Enabling Delta Life

COMPARISON OF REMOTELY SENSED AND VOLUNTEERED GEOGRAPHIC INFORMATION FOR WATER RESERVOIRS

A thesis submitted to the Delft University of Technology in partial fulfillment of the requirements for the degree of

Master of Science in Geomatics for the Built Environment

by

Maria Moscholaki

November 2020

The work in this thesis was carried out in the:

3D geoinformation group
Department of Urbanism
Faculty of the Built Environment & Architecture
Delft University of Technology

Deltares
Independent institute for applied research in the field of water and subsurface
Delft

Supervisors:	Assistant Prof. Clara Garcia-Sanchez
	Balázs Dukai
	Gennadii Donchyts (Deltares), Christine Rogers (Deltares)
Co-reader:	Dr. Ravi Peters

"Thousands have lived without love, not one without water."

— W. H. Auden

ABSTRACT

The surface water extent around the world is constantly changing due to natural factors (e.g. geology and over-abstraction of water), climate change (e.g. higher water evaporation due to warmer climate) or human activities (e.g. reservoir construction). Water reservoirs are important for the management of the ecosystem, as both humans and the natural environment depend highly on them for their existence and well being. Flood control, agricultural irrigation, electricity generation, drinking and municipal water supply are only some of their main uses. Considering this, it is of high importance to have accurate maps that depict the reservoir outlines to determine their surface extend and storage capacity. However, their extend is not always well defined or there are discrepancies between various surface water datasets.

This dissertation aims to provide an answer about which datasets match better as well as identifying the problematic areas by performing a quality control analysis. The main challenge of this thesis is that all available datasets have certain limitations regarding their coverage and quality. The waterbody delineation from satellite images is affected by the atmospheric conditions (e.g cloud obstructions) or topographic elements that create artifacts and influence the correct classification of water pixels. OpenStreetmap (OSM) on the other hand, has uncertain quality over locations, as the data is freely supplied by volunteers. Moreover, HydroLAKES which was created based, amongst others, on the Global Reservoir and Dam Dataset (GRanD), is still incomplete.

In this thesis, an intercomparison of accuracy algorithm that can perform large scale analysis is created, by using the country of Angola as a use case, five datasets as input (Global Surface Water, Sentinel 2, OSM, HydroLAKES, GRaND) in both raster and vector format and the cloud processing platform of Google Earth engine.

The identification of similarities or mismatches between the datasets is performed in terms of positional accuracy. Two quality measures have been considered for the pairwise comparison of features: percentage of overlap and Hausdorff distance. In addition the completeness of the datasets respectively to the total common water area of the water reservoir datasets is reviewed.

The results of this research shows that large scale analysis for the comparison of accuracy between water reservoir datasets of different formats is possible. The pre-processing of the input Satellite data is semi-automated. The created automated algorithm for the main analysis offers information for all corresponding features between datasets. More specifically, statistics about the shape similarity, the percentage of overlap and the water area completeness of the datasets are being presented.

ACKNOWLEDGEMENTS

This document marks the completion of my journey at TU Delft, and therefore I would like to take the opportunity to express my gratitude to the people that supported me along the way. First of all, I would like to thank my supervisor Clara Garcia Sanchez for her continuous guidance and useful but most importantly encouraging comments. The second person that that I would like to thank is my second supervisor, Balasz Dukai, for providing his expertise in the field of Geomatics and his constructive feedback in general. Thanks as well to my co-reader Ravi Peters for his valuable comments during the finalization of my thesis.

Special thanks, to my mentor Gennadii Donchyts at Deltares, who played a significant role in my graduation by assigning to me a project that is completely inline with my interests but most importantly by sharing with me all this knowledge about remote sensing and teaching me by example his high level thinking and understanding of scientific concepts. Also, many thanks to Christine Rogers, for her fruitful advises, the technical support and for always being present and ready to help me with this graduation topic.

To my fellow students, with which I have shared all this lovely experiences and to my mother and sisters that always support me, thank you very much. Without you I would not be able to fulfill my goals.

Lastly, this thesis is dedicated to my dear father.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

1 INTRODUCTION

Water is one of the most vital elements on earth. It is of high importance for the preservation of all forms of life, humans, animals, and plants [Khodaei and Nassery, 2008]. To manage these water resources, accurate maps that provide reliable information on the spatial distribution, seasonal and annual changes of surface water are essential [Santoro et al., 2015].

The accurate knowledge of the available water stocks in the world can provide answers to questions related to water availability, flood hazards, food and health safety, agricultural and industrial usage, but also support hydrological and ecological processes. Water reservoirs, either man made or created by nature, are used to generate electricity and prevent floods. There are millions of water reservoirs in the world and their storage capacity can have a big impact on the living of people and the surrounding ecosystems. The increase of water demand due to the growing population combined with the water reservoir fluctuations because of climate changes, have made the monitoring of their dynamics necessary, now more than ever. An accurate representation of the truly permanent geometry of the existing water reservoirs can provide a clear picture of the area of surface water, the storage volume or any changes that take place.

There are many available sources of information. Satellite imagery has been extensively used for water detection purposes. The large variety and amount of Earth Observations, together with the processing cloud platform of Google Earth Engine that offers enormous computational resources, provided us with the ability for planetary scale mapping. Moreover, initiatives where citizens are allowed to participate in data collection, also known as Volunteered Geographic Information (VGI), made worldwide mapping efforts that provide free geographic data, possible. OpenStreetMap (OSM), a type of crowd-sourcing editable world map, is currently the biggest freely available geodata platform, that has been used in a wide range of Geographic Information Systems (GIS) and applications as an alternative source information, or supplementary with other official authoritative datasets [Brovelli and Zamboni, 2018].

1.1 PROBLEM STATEMENT AND MOTIVATION

The amount of Earth Observations and other geospatial information is constantly increasing. The foremost advantage of remote sensing-based techniques is that they provide an effective way of monitoring the surface of Earth continuously on a global scale. This is due to the ease of data access that is offered freely and openly in different temporal and spatial resolutions [Jakovljević et al., 2019; Huang et al., 2015; Avisse et al., 2017]. However, there are some factors that affect the final accuracy and lead to the miss-classification of water pixels (error of omission and commission). Optical earth observation imagery is easily affected by cloud obstructions, terrain and cloud shadows, snow, ice and "dark" vegetation as they present similar spectral properties with surface water [Thissen, 2019].

On the other hand, OSM is based on the collection of Geographic Information gathered and updated by volunteers [Bhattacharya, 2012]. These data are provided from sources such as Global Positioning System Global Positioning System (GPS) devices, cadastral data, through manual digitizing (editing) on medium and high-

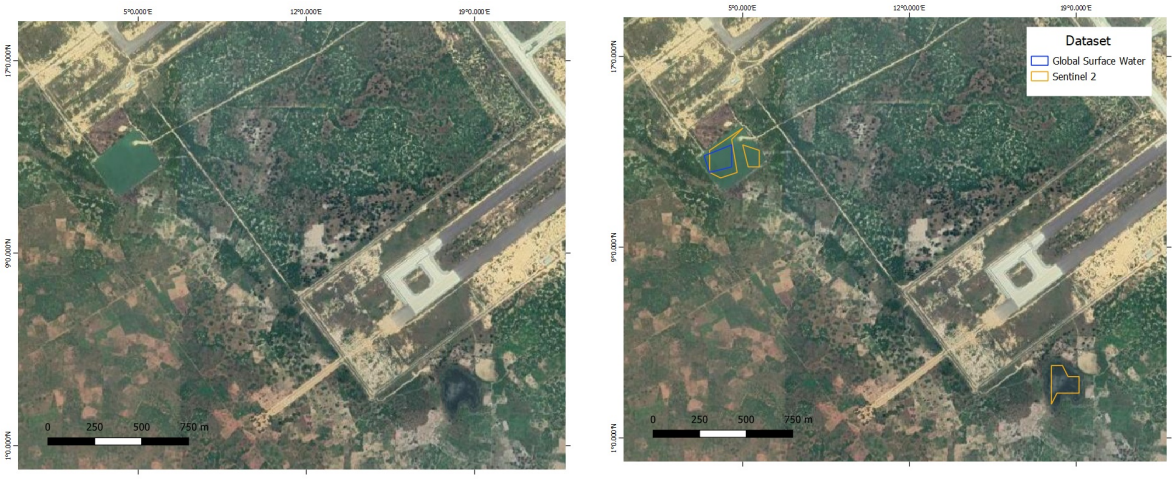
resolution satellite and aerial imagery or from knowledge about an area [Goetz and Zipf, 2013; Barron et al., 2014]. The most significant advantage of this provider is its global coverage and up to date nature. Many studies, however, are questioning the OSM data quality [Kato, 2018], as they are created without any formal qualifications. This is the main reason why the use of these georeferenced data have not been extensively adopted by GIS professionals [Mooney et al., 2010a]. Other vector surface water datasets, such as HydroLAKES and the Global Reservoir and Dam Database (GRaND) were created in a joint international effort to compile the existing water reservoir and dam data and gather all this information in one reliable database. Even though, data dissimilarities and record gaps were corrected during the development of these databases, they are still incomplete as information about many existing water reservoir is missing [Lehner et al., 2011]. Examples of the various dataset inconsistencies are given in the following figures. More specifically, Figure 1.1 shows an example of a reservoir that is only present in the OSM database, Figure 1.2 only detected with Sentinel 2 and Figure 1.3 another found only with Global Surface Water (GSW) and Sentinel 2. Figure 1.4 gives an example of a reservoir noticed from OSM and Sentinel 2 but not GSW possibly because of its lower resolution. In Figure 1.5 a water feature was located with Sentinel 2, GSW and HydroLAKES but not OSM. Figure 1.6 identifies a false positive OSM water feature registration and lastly, Figure 1.7 illustrates the geometric differences of a feature present in all five datasets.

(a) Satellite Image **(b)** OSM water feature

Figure 1.1: Example of OSM water feature

(a) Satellite Image **(b)** Sentinel 2 water feature

Figure 1.2: Example of Sentinel 2 water feature



(a) Satellite Image **(b)** GSW and Sentinel 2 water features

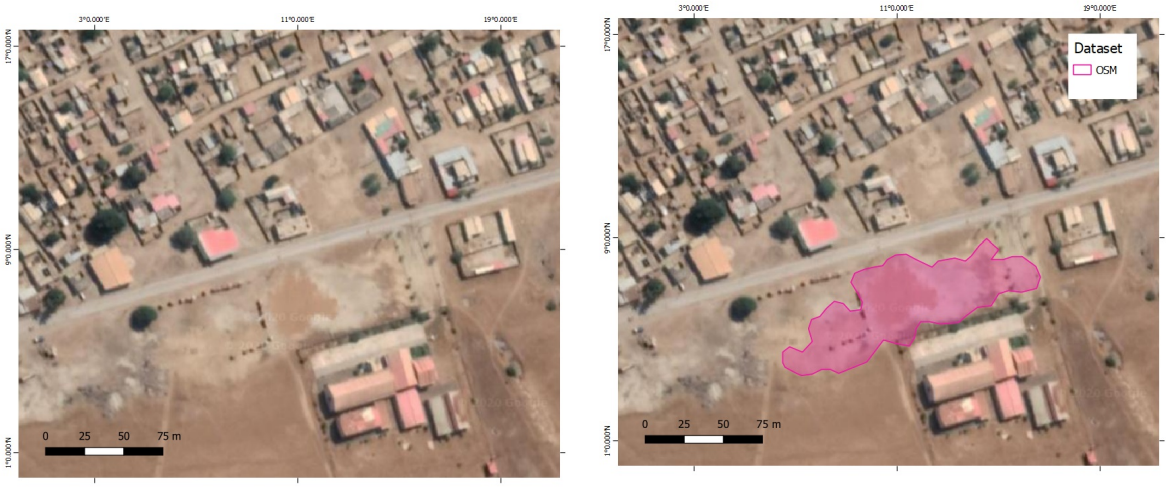
Figure 1.3: Visual comparison of extracted waterbodies from different data sources

(a) Satellite Image **(b)** OSM and Sentinel 2 water features

Figure 1.4: Visual comparison of extracted waterbodies from different data sources

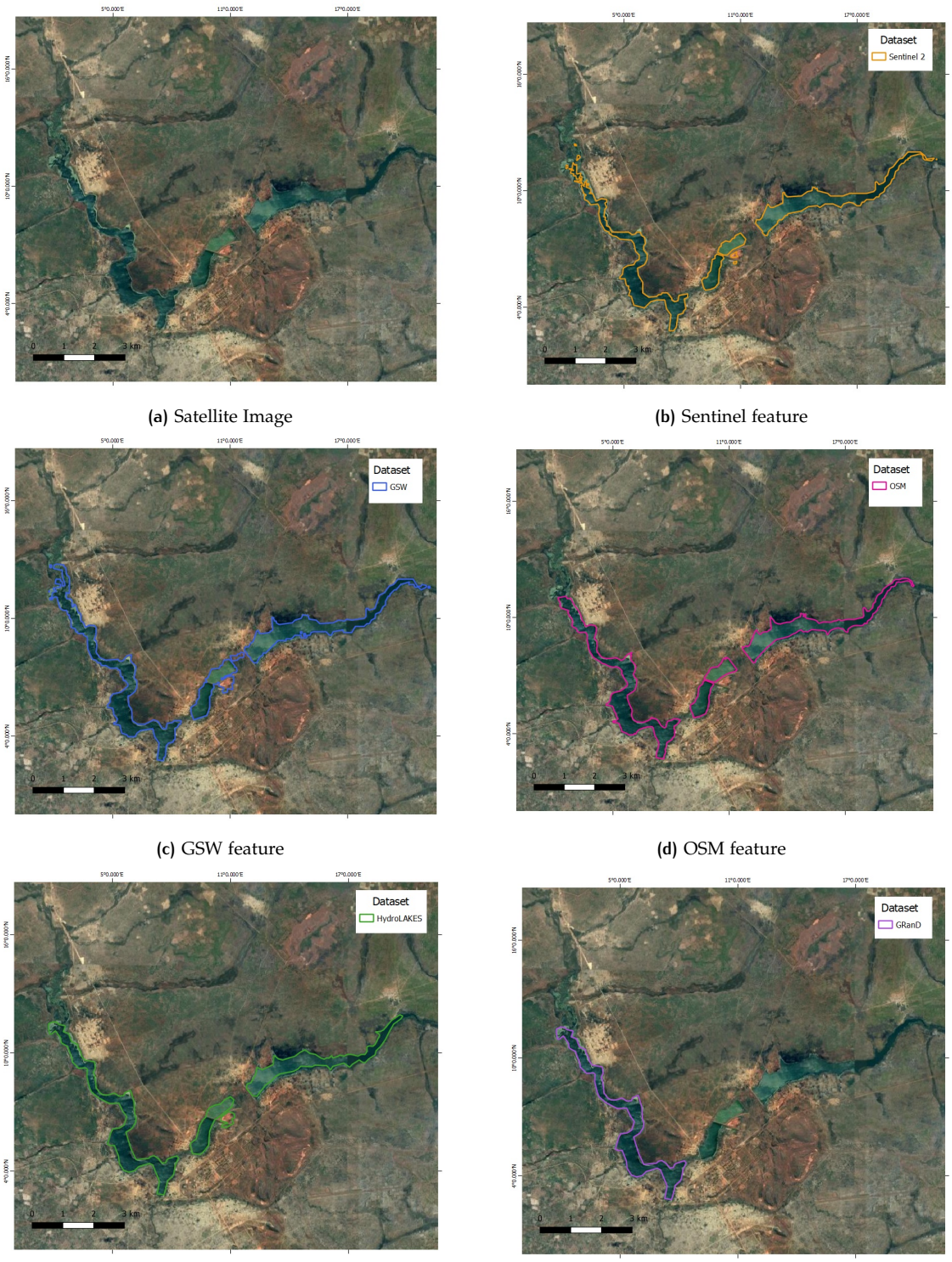
(a) Satellite Image **(b)** Sentinel 2, GSW and HydroLAKES water features

Figure 1.5: Visual comparison of extracted waterbodies from different data sources



(a) Satellite Image **(b)** OSM feature

Figure 1.6: Example of false positive water feature



(a) Satellite Image

(b) Sentinel feature

(c) GSW feature

(d) OSM feature

(e) HydroLAKES feature

(f) GRaND feature

Figure 1.7: Visual comparison of extracted waterbodies from different data sources

The dynamic nature of the water extend both in space and time along with the limitations mentioned above in the various water detection methods and datasets, make it very hard to create an accurate high-resolution waterbody map [Yamazaki et al., 2015]. In order for these datasets to be completely reliable, we would need an objective confidence map of every water mask. However, assessing the level of trust as to whether the data is accurate is not easy, as all datasets contain uncertainties and even in situ observations, are only point-based and cannot give a representative idea of the spatial distribution of water in time and in large scale. Therefore, in

order to see how well these datasets match or mismatch, a comparison process for the water reservoir geometries at larger scale is needed. This can increase the confidence of the outline water reservoir geometries by analysing the correlation of the remotely sensed and official vector datasets with OSM.

1.2 RESEARCH QUESTIONS

Having the introduced problem statement in mind, the formulation of the following main research question emerges:

What are the spatial differences between Earth Observation based and Volunteered Geographic Information for water reservoirs and how can they be addressed in an automated way at a large scale?

Answering this research question would arouse a positive contribution in various fields that are not only relevant to the scientific/research community but also to policy and decision making of public administrations and various environmental strategies. To achieve that, the following sub-questions need to be addressed:

- What are the differences in terms of spatial coverage?
- What are the differences in terms of positional accuracy?

1.3 RESEARCH SCOPE

The focus of this thesis is to create a comparison of accuracy algorithm that can operate in a large scale using different water reservoir datasets. A workflow has been developed that is tested in a smaller extend than the whole world, and more specifically in the country of Angola. However, the created algorithm can be used also for bigger datasets.

A methodology that explores the positional accuracy and completeness of OSM water masks compared to GSW, Sentinel 2, GRaND and HydroLAKES water datasets is being applied. Focus has been given only to polygon water features from the OSM database, as linear primitives have not been considered. The OSM dataset has been filtered out to contain only features that show a possible relevance to water reservoirs. Nevertheless, the algorithm can work for any two polygon datasets given as input.

It is important to highlight that the performed analysis does not ensure the quality of the data, as it doesn't utilize ground truth information, i.e. information provided by empirical observations. It focuses on the evaluation of the data by performing an intercomparison of datasets. Therefore, the main goal of this research is to assess the level of agreement amongst different sources and identify unreliable or missing water reservoir areas depending on their dis/similarity as indicated by the diverse chosen quality criteria.

This research was initiated by Deltares, which is an independent institute for applied research in the field of water, subsurface and infrastructure. It runs several projects throughout the world such as reservoir planning, design and operation. Deltares has created its own water reservoir dataset by combining information from different sources. The developed algorithm of this thesis shall facilitate the improvement of this database.

The main limitation of the developed quality analysis tool, is the semi-automated nature of the pre-processing of the Satellite data. It does not require heavy downloads and everything is processed on Google through it's Earth Engine. However, due to the use of large scale remote sensing data, the pre-processing of the Sentinel

2 and GSW water datasets together with the quality analysis cannot be performed on the fly, but instead the created water mask has to be first exported. Nonetheless, the algorithm could be applied on a global level in an automated way for both the pre-processing and main analysis, as part of the suggested future work of this thesis.

Another point that is out of the scope of this graduation project is the refinement of the created Sentinel 2 water dataset by implementing additional operations to avoid potential classification errors. Lastly, due to time constraints, fixing the mismatches and creating water reservoir polygons based on the combination of all input datasets, is not done in this study.

1.4 THESIS OUTLINE

This thesis is structured as follows:

- Chapter 2 presents the theoretical background that is essential to understand the scientific concepts analyzed throughout this thesis. An introduction to the basic remote sensing principles are explained and various data processing algorithms are introduced.
- Chapter 3 reviews the related research that has been done so far in the comparison of accuracy of different datasets, some concepts of which have also been adopted in the implementation of this study.
- Chapter 4 describes the approach of the methodology that was chosen to answer the research questions as a workflow. It presents the mathematical and other concepts on which the pre-processing and main analysis was based.
- Chapter 5 mentions briefly the data and tools of this research.
- Chapter 6 offers a detailed description of the implementation of the methodology. The final results together with the potential challenges and factors that affect them are being provided.
- Chapter 7 offers an insight of how a certain variable of the algorithm affects the analysis results.
- Chapter 8 discusses the conclusions and answers to the research questions of this thesis. Finally, it presents recommendations for future work.

Additionally, Appendix A contains an assessment of the reproducibility of this research.

2 THEORETICAL BACKGROUND

The implementation of this research is based on different scientific concepts and technologies. This chapter aims to provide an overview of the theoretical background that is related to this study, which will facilitate the comprehension and understanding of the concepts and results described in the following chapters. In Section 2.1, an insight of the basic remote sensing principles is given. More specifically, Section 2.1.1 describes how the acquisition of information with remote sensing works, as well as the different Spaceborn remote sensors that are widely used for water detection purposes. Next, Section 2.1.2 gives a general introduction to various surface water detection methods. Section 2.1.3 addresses the challenges when analyzing satellite imagery to extract information about existing waterbodies.

In addition, Section 2.2 presents a number of raster and vector data processing algorithms used throughout this research. Starting with Section 2.2.1 and Section 2.2.2, data transformation algorithms referring to the change of the shape of object in images and the conversion from vector to raster format are explained respectively. In Section 2.2.3, a resampling method with bicubic interpolation, reviews how the resolution of an image can be modified. Section 2.2.4 describes how all possible relation between two sets of data can be described by using the Venn Diagrams. Lastly, Section 2.2.5 explains the steps of line simplification operation.

2.1 REMOTE SENSING PRINCIPLES

2.1.1 Optical Earth Observation Sensors

Remote sensing techniques are based on the principle of measuring the reflected and emitted radiation from the Earth's surface and the atmosphere by using several parts of the Electromagnetic Spectrum (EM) that is not visible to the human eye (see Figure 2.2[1]). Satellites contain sensors that measure this electromagnetic radiation (see Figure 2.1[2]). There are two types of sensors, passive and active. Passive sensors (digital cameras and multispectral scanners) are instruments that receive and measure the reflected sunlight emitted from the sun (Sensor 1). This reflected energy takes place only during day time, when the sun illuminates the Earth. Passive sensors (thermal scanners) utilize also the energy that is naturally emitted, such as thermal infrared, which can be measured throughout the day and night but only in cases where the amount of energy is big enough to be detected (Sensor 2).

1 https://www.pgc.umn.edu/guides/commercial-imagery/intro-satellite-imagery
2 https://www.omnisci.com/technical-glossary/remote-sensing

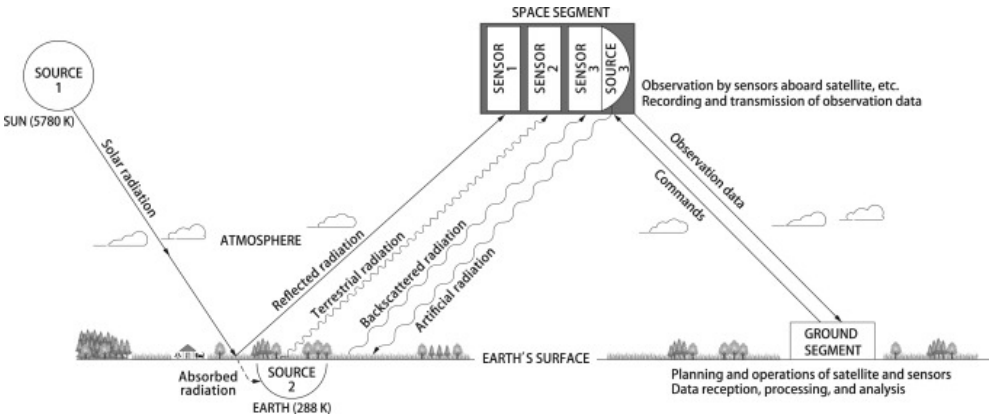


Figure 2.1: Remote Sensing Process

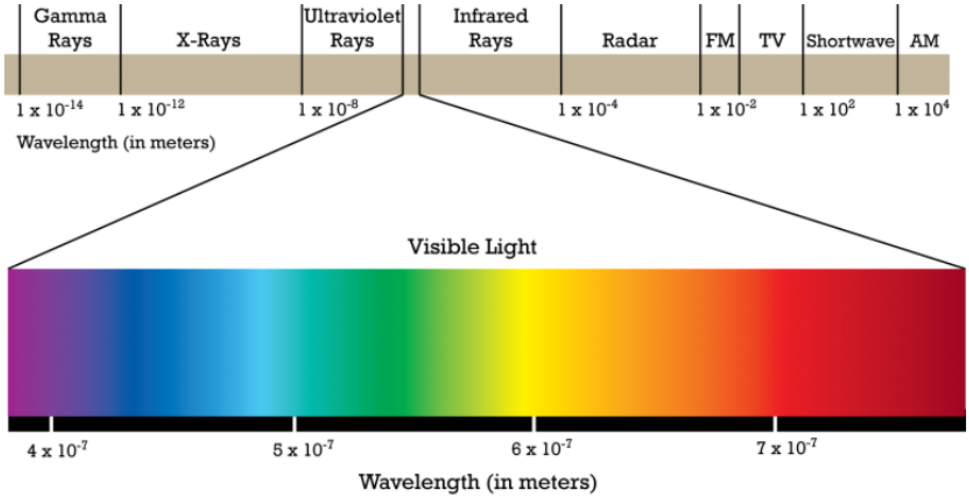


Figure 2.2: Electromagnetic Spectrum

Active sensors (Radar and Lidar) on the other hand, emit their own electromagnetic energy and record the energy that is scattered back in the direction of the instrument (Sensor 3). This energy is in the visible and near-infrared part of the EM spectrum which is produced by lasers or microwaves in order to collect data. Microwaves particularly, can work day and night and are highly weather independent because microwaves are much larger than the water particles present in the air. As a result the waves are not affected by them. Therefore, even in an area with tropical climate where clouds cover the atmosphere, the radar can see through them. Same thing also occurs in dense vegetation and in upper soil areas where radars are able to penetrate them.

Optical sensors (imaging and thermal systems) use the visible, near-infrared, and shortwave infrared spectrums. In order to work better however, they need good weather conditions. Optical remote sensing has shown a big progress in the last decades. Because of the high spatio-temporal availability of data, they have been extensively used for water detection purposes [Huang et al., 2015]. As the years go by, more satellite data with better spatial, temporal, and spectral resolutions are acquired [Donchyts, 2018]. The spectral resolution describes the range of the EM in which a satellite can sample the reflected radiance whereas the spatial resolution, refers to the pixel size of a satellite image or in other words the ground area depicted in this pixel. Lastly, the temporal resolution describes the time needed for a satellite to orbit and revisit the same area.

The spatial resolution plays an important role in the process of surface water detection, as it determines the level of detail of the detected water feature. High

resolution images offer higher level of detail. On the other hand, high temporal resolution offer the possibility of more frequent data. Amongst the various optical sensors, the ones that have been used mostly for water mapping purposes, are the ones with medium resolution (10-30 m) [Donchyts et al., 2016]. More specifically, multispectral satellite imagery originating from Landsat 8 (30m) has been available since 11\02\2013. Sentinel 2A, launched in 2015, is another, more recent, valuable source of information for waterbody mapping as it offers a 10m resolution.

Sensor group	Satellite/sensor	Number of bands	Spatial resolution (m)	Temporal resolution (day)	Maximum swath at nadir (km)	Scale of application[b]	Data distribution policy (costs)	Data availability
Coarse resolution sensor	NOAA/AVHRR	5	1,100	0.5	2,800	R-G	no	1978--
	MODIS	36	250–1,000	0.5	2,330	R-G	no	1999--
	Suomi NPP-VIIRS	22	375–750	0.5	3,040	R-G	no	2012--
	MERIS	15	300	3	1,150	R-G	no	2002-2012
	Sentinel-3 OLCI	21	300	2	1,270	R-G	no	2016--
Medium resolution sensor	Landsat	4–9	15–80	16	185	L-G	no	1972--
	SPOT	4–5	2.5–20	26	120	L-R	yes	1986--
	Aster	14	15–90	16	60	L-G	no	1999--
	Sentinel-2 MSI	13	10–60	5	290	L-R	no	2015--
High resolution sensor	IKONOS	5	1–4	1.5–3	11.3	L-R	yes	1999--
	QuickBird	5	0.61–2.24	2.7	16.5	L	yes	2001--
	WorldView	4–17	0.31–2.40	1–4	17.6	L	yes	2007--
	RapidEye	5	5	1–5.5	77	L-R	yes	2008--
	ZY-3	4	2.1–5.8	5	50	L-R	yes	2012--
	GF-1/GF-2	5	1–16	4–5	800	L-R	yes	2013--

Figure 2.3: Commonly used Spaceborne Remote Sensors for Surface Water Detection [Huang et al., 2018]

2.1.2 Surface water Detection methods

The reflectance of a surface depends on its material, and it varies with the wavelength of the electromagnetic energy, which is what makes it possible to identify Earth's surface features differently by analyzing their spectral reflectance signatures. Water detection is mainly based on its characteristic of significantly lower reflectance in the infrared part of the EM compared to other landcover types [Huang et al., 2018].

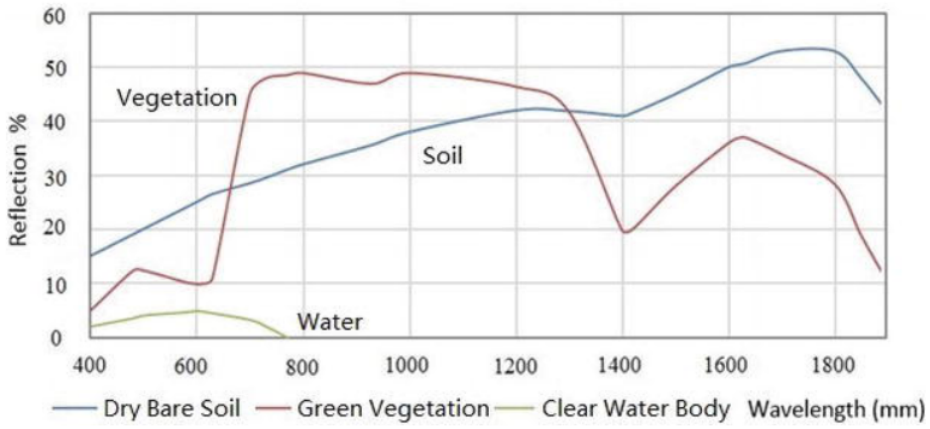


Figure 2.4: Reflectance of different Land types [Zhu et al., 2018]

Several surface water extraction methods have been developed in the past, to separate water from non water features. Water indices are considered an easy and effective way of extracting water. McFeeters [1996] created the Normalised Difference Water Index (NDWI) which is found from the normalized difference between the green and Near-Infrared (NIR) bands, calculated using Equation 2.1. This way

each pixel is assigned a value between -1 and 1, with water pixels having positive values [McFeeters, 1996].

$$\textbf{NDWI} = \frac{GREEN - NIR}{GREEN + NIR} \tag{2.1}$$

Another variation of the NDWI, is the Modified Normalised Difference Water Index (MNDWI) of Xu [2006] in which NIR was replaced by the Short-Wave Infrared (SWIR) band [Ogilvie et al., 2018]. This water index is considered to be more reliable in urban areas than the NDWI [Donchyts et al., 2016]. However the limitation of MNDWI is that it cannot discriminate so easily water from snow or cold clouds [Thissen, 2019]. It is expressed by Equation 2.2. The resulting positive values represent the water features because of their higher reflectance in GREEN and SWIR bands, while non-water features are smaller or equal to zero.

$$\textbf{MNDWI} = \frac{GREEN - SWIR1}{GREEN + SWIR1} \tag{2.2}$$

Recently a new Automated Water Extraction Index (AWEI) was introduced by Feyisa et al. [2014], which shows improved and more accurate classification results in areas with dark surfaces or shadows (Equation 2.3).

$$\textbf{AWEI} = 4 \times (GREEN - SWIR1) - (0.25 \times NIR + 2.75 \times SWIR1) \tag{2.3}$$

Some studies propose also the use of the Normalised Difference Vegetation Index (NDVI), which can be used to exclude dark vegetated areas (see Equation 2.4) by setting a high threshold value. Water pixels in this case have negative values and with the vegetation index we can detect water and floods in some areas [Domenikiotis et al., 2003].

$$\textbf{NDVI} = \frac{NIR - RED}{NIR + RED} \tag{2.4}$$

The various water indices have performance differences and their use is also challenged by the need for an automated optimal threshold method [Li et al., 2019]. Therefore, alternative methods have been used to extract information from remotely sensed image data. Frazier and Page [2000] used a simple process in which a single infrared band is density sliced to acquire a waterbody map, by dividing the range of brightness values of this band to intervals and then assigning a color to each interval [Campbell, 2002]. A histogram is generated from the values of the pixels of this map. The different colors assigned to each interval allow the separation water from non water features. Other very common classification techniques are the supervised (with training samples) or unsupervised (without training samples) classification methods [Manavalan et al., 1993; Ozesmi and Bauer, 2002]. These were used to generate land cover maps, from which water maps could be extracted. However, these techniques are based on rules that are not easily formed and possibly not robust enough to be applied on a global scale [Huang et al., 2018].

2.1.3 Surface water Detection challenges

Cloud obstructions are a significant problem when analyzing satellite imagery. Donchyts et al. [2016] and Hansen et al. [2013b] proposed the creation of cloudless composite images that are based on average cloud-free reflectance values [Thissen, 2019]. Another approach was introduced by Donchyts [2018], who uses multiple cloud-free images and a probability density function to accurately detect large-sized water reservoirs that present only small changes in their shape. The view angle of the satellite, and the position of the sun, have been used by Zhu and Woodcock [2014] for cloud shadow and snow detection. This technique was adapted by Tan et al. [2013] in combination with a Digital Elevation Model (DEM) for terrain shadow detection.

The nature of waterbodies is dynamic as they change over time and between seasons. Considering this, it is clear that the static vector maps might not always give a representative idea of the true extend of the surface water. A way to create better quality water maps is by using multi-temporal images, which are equally important with the analysis of higher resolution imagery for more accurate water body mapping. Mueller et al. [2016] implemented an algorithm to map the surface water extend across Australia, by analysing 25 years of Landsat imagery using a decision tree classifier and logistic regression that compares the water classification results with ancillary datasets. This way it was possible to identify the areas where the occurrence of water is more persistent (e.g reservoirs) and where more temporal (e.g floodplains). Yamazaki et al. [2015] created a global 90 m resolution water body map from multi-temporal Landsat satellite images. Feng and Bai [2019] created a global land cover map produced through integrating multi-source instead of mulitemporal, satellite imagery datasets.

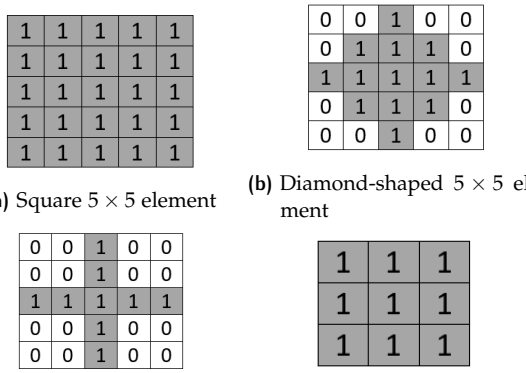
2.2 DATA PROCESSING ALGORITHMS

2.2.1 Morphological Operators

The morphological operators are commonly used in image processing for manipulating raster images and more specifically in applications such as hole filling, thinning and thickening, boundary extraction of features and identification of connected components [Srisha and Khan, 2013]. The shape or morphology of objects in a binary image can be changed with four basic morphological operators: erosion, dilation, closing or opening.

These morphological operations are done by applying to the input image a structuring element, which is a small binary matrix. The matrix is scanned as a sliding mask over each pixel in the input image to change its values and create an output image of the same size. The effect of the interaction between the input image and the structuring operator depends on the characteristics of the matrix (shape and size). The structuring elements can have varying sizes and different arrangement of one, zero or none values within the matrix which affects its shape (see Figure 2.5).

The 3x3 square is probably the most common structuring element used in dilation and erosion operations. Larger structuring elements will produce more extreme effects. With larger structuring elements, it is quite common to use an approximately round-shaped structuring element, as opposed to a square one. Kernels of different shapes (e.g. square, circle, cross, diamond) can be acquired by placing the values within the matrix in certain positions [Thissen, 2019].



(a) Square 5×5 element

(b) Diamond-shaped 5×5 element

(c) Cross-shaped 5×5 element

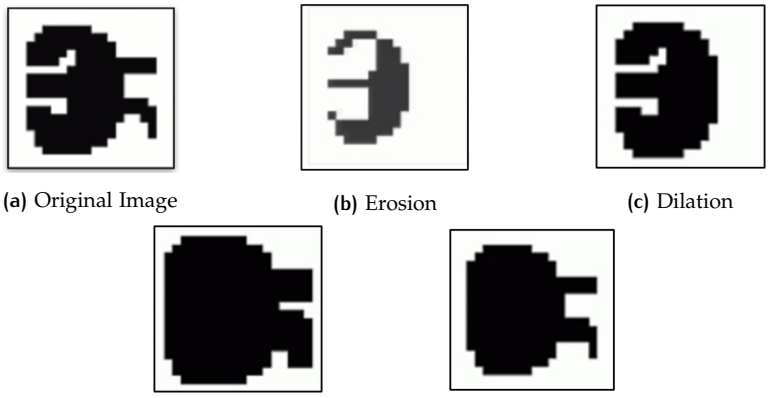
(d) Square 3×3 element

Figure 2.5: Representation of structuring elements

The four operators can be explained as follows:

- Erosion: The basic effect of erosion on an image, is to shrink image regions that represent features and to remove structures of certain shape such as branches or connections depending on the selected structuring element. As the areas of the object pixels shrink holes within those areas become bigger (see Figure 2.6b).
- Dilation: This operator enlarges the object areas of an image. As these areas grow in size holes and gaps within these regions become smaller. The effect of the operator on the input image depends on the shape and size of the structuring element (see Figure 2.6c).
- Opening: This operation combines both an erosion and dilation. It similar to erosion as it tends to smooth the contour of an image object and remove thin protrusions, but it less destructive (see Figure 2.6d).
- Closing: Closing is a dilation followed by an erosion operation. It keeps the general shape of the original object and is often less destructive while smoothing it, filling holes and eliminating thin gulfs (see Figure 2.6e)

The effect of these operations is shown in the following illustrations (Figure 2.6):



(a) Original Image

(b) Erosion

(c) Dilation

(d) Opening

(e) Closing

Figure 2.6: Morphological Operators [Bhatia and Goel, 2011]

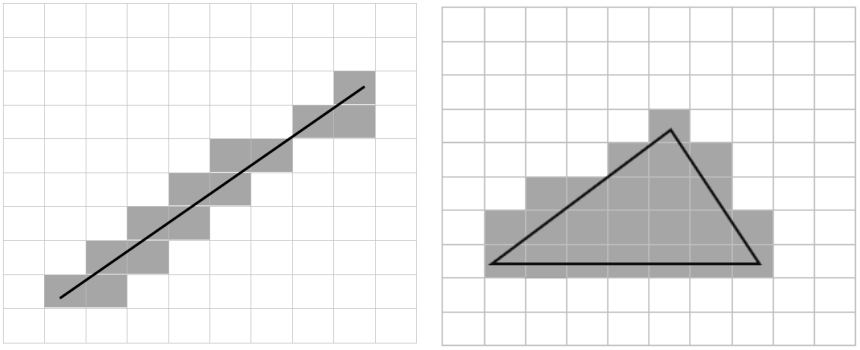
2.2.2 Rasterization

The raster model is a grid-like structure used widely for storing geographic data (e.g. satellite or aerial imagery). According to this model spatial objects are usually

represented in uniformly spaced cells or pixels organized in a structure of rows and columns, where each element represents one cell/pixel. It is usually used for continuous data such as temperature, or electromagnetic radiation and each cell describes a value that contains information about that specific phenomenon in this location. The size of each cell defines the resolution of the raster image, measured in linear units of distance in metric terms (e.g meters, feet) or in degrees of latitude and longitude [Goetz and Zipf, 2013].

The advantages of raster representation is that it is a very simple data structure and efficient for performing overlay processes (Section 2.2.4). Moreover, map algebra tasks such as arithmetic (e.g. addition) and statistical operations (e.g. mean, median etc.) are quick and easy to perform and also the time required for operations like interpolation or resampling is reduced. Lastly, it is considered to require less storage memory when dealing with continuous data. On the other hand, raster data come also with a number of disadvantages. More specifically, when dealing with dense grids, higher memory and computational resources are needed. Apart from that, it is difficult to adequately observe fine details or small features depending on the cell resolution of the raster (pixel size).

In some cases, vector-to-raster data conversion is required, as most datasets are in vector format (points, lines and polygons). This process is called rasterization, during which, point, linear or polygonal primitives are decomposed into pixels, i.e. the interior of these features is filled with pixels (see Figure 2.7). The Bresenham's line algorithm is a raster conversion algorithm, which is based on finding which pixels on a raster image need to be selected in order to approximate as closely as possible a line that connects two points or a polygon connecting multiple points. The result of this transformation is a binary image (0 and/or 1 values) . This algorithm is used widely for drawing geometric primitives onto an image as it utilizes low demanding, in terms of computational power, operations.



(a) Linear Primitive **(b)** Polygonal Primitive

Figure 2.7: Bresenham's Line Algorithm

2.2.3 Resampling with bicubic interpolation

Conversion from raster to vector might cause inconsistencies between the created polygons, due to the varying resolution of the datasets. A way to avoid this, is by resampling the raster image. Image resampling is a mathematical process of creating a new version of the raster cell grid with a with a different width and/or height in pixels. The value of each cell in the new raster will be computed by sampling or interpolating in a neighborhood of cells of this pixel in the original raster object

[Sachs, 2001]. Bicubic interpolation is considered to be slower in computation speed, but it is supposed to have better smoothing results when upsampling.

$$f(x,y) = \sum_{i=0}^{3}\sum_{j=0}^{3} a_{ij}x^i y^j \tag{2.5}$$

where x and y the coordinates of the new location, f(x,y) the value of the pixel and a_{ij} the 16 coefficients for the 16 neighbors.

2.2.4 Venn Diagrams

Overlay operations are used when we want to combine and find the logical relationships between two or more sets of spatial or other data. Venn diagrams are widely used in GIS environments, as an indication of all the possible logical relations between sets when performing overlay operations. These diagrams depict elements that are modelled as members of a collection. This way can identify the set-based geometry of the Euclidean 2D space and perform operations on the different sets. There are four basic operations in Venn Diagrams that represent the Intersection, Union and Complement between sets. To do this the sets are overlapped in every possible way (see Figure 2.8).

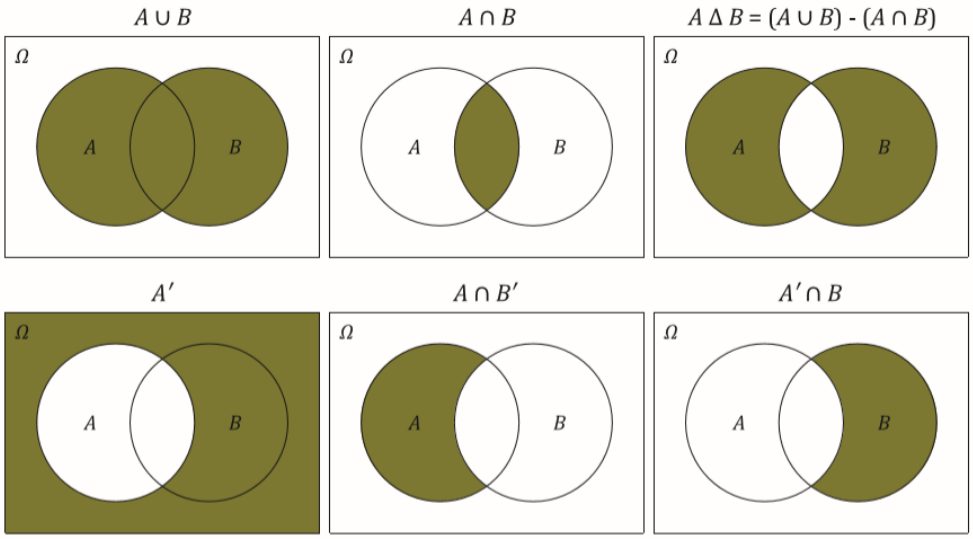


Figure 2.8: Venn Diagrams representing the union, intersection, symmetrical difference, absolute and relative complement of sets [Cardinal, 2019].

2.2.5 Line Simplification Algorithm

Line simplification is an operation used widely in cartographic generalization on lines and area boundary features to remove unnecessary detail, i.e. unnecessary points while retaining the most imortant ones and preserving the essential shape of the line [Ivanov et al., 2000]. The advantages of simplification of linear objects are the reduced storage space and plotting time, faster vector processing (e.g rotation, scaling etc.) and vector to raster conversion . The Douglas–Peucker algorithm is one of the traditional methods to simplify linear objects [Douglas and Peucker, 1973]. The algorithm uses a recursive function (Figure 2.9[3]). A line is of a set of points. Initially, only the first and last points are kept and are connected with a straight

3 http://resources.esri.com/help/9.3/arcgisengine/java/gp_toolref/data_management_tools/how_simplify_line_data_managemer

line. Afterwards, it computes the distance of all intermediate points perpendicular to the straight line. The point (P) that is furthest away from the line segment that has the first and last point as endpoints is then chosen. If this point is closer to the line segment than a tolerance value (threshold), then all points between the first and last point of this line segment can be discarded. Otherwise, this point is kept and the process is being repeated for the line segment between the first and furthest point P and the furthest point P and last point. The process is repeated, until no more vertices of the original curve are available, for which the distance to the curve is greater than the predefined threshold. After the completion of the recursion, a new curve is being acquired that consists of all the points that have been marked as kept.

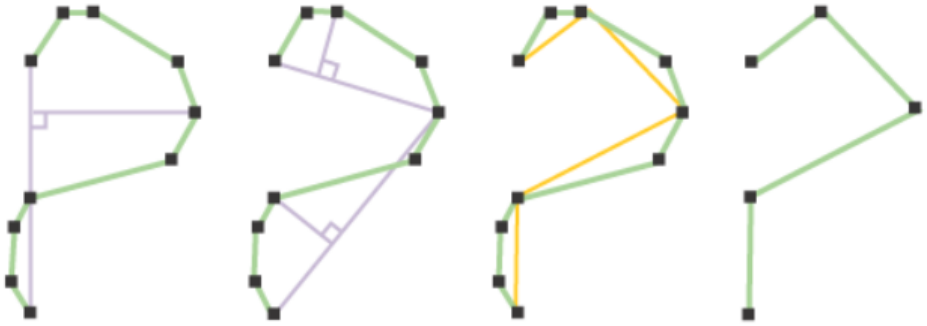


Figure 2.9: Stages of recursive function of line simplification

The tolerance value dictates the degree of simplification. The choice of the threshold value can be made non-parametric by using the error bound due to digitization [4]. When the tolerance is used for many features, a trial and error process may be required to find one threshold that is suitable for all features.

4 https://en.wikipedia.org/wiki/Ramer%E2%80%93Douglas%E2%80%93Peucker_algorithm

3 RELATED WORK

OSM geographic information has assisted several mapping procedures. However at the same time, OSM data have been criticized about their inherent variable quality amongst locations. Because of this, several studies over the past years have put these data to test, to quantify the differences with other datasets. A number of automated and semi-automated procedures have been developed, that make use of different OSM features in various locations but also varying quality measures and software.

Haklay [2010] compared the OSM road network in England with the Meridian dataset of the national mapping agency for Great Britain, by performing an SQL query analysis. This dataset is highly accurate as it is based on topographical measurements. The outcome if this survey showed that, when it comes to positional accuracy, OSM performs very well and in some cases even better that the Meridian dataset. However, it still is very incomplete regarding the registration of the existing streets. The evaluation of the positional accuracy of these linear features was conducted according to the Goodchild and Hunter [1997] comparison method, which creates a buffer around the features of the high quality dataset and the percentage of the tested object that falls within this buffer is calculated (Figure 3.1). To assess the completeness, a grid in the extend of the whole UK road network was created, and then for each cell the total length of the Meridian and OSM datasets was compared.

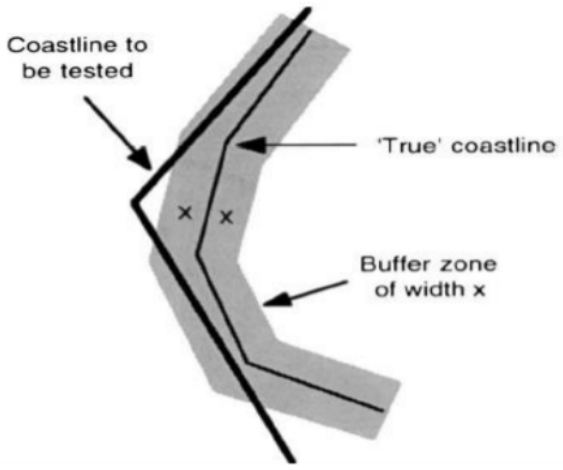


Figure 3.1: Goodchild and Hunter buffer comparison technique [Goodchild and Hunter, 1997]

A similar survey was performed by Girres and Touya [2010], in which the quality of the OSM spatial data in France were assessed, by using as a reference a national topographical dataset. Because of the complexity of the analysis they used a quality control procedure of various quality components, carried out in a geographic database. More specifically, the geometric accuracy was evaluated by choosing the euclidean distance for point primitives, the Hausdorff and average distance for linear features and the surface distance and granularity for area primitives. Moreover the semantic accuracy, the logical consistency, currentness and lineage were analyzed to show the level at which the OSM database is in agreement with the real

world. The quantitative quality assessment, which was performed on a number of manually selected features that were matched in both datasets, showed that there is a big heterogeneity in the geometric accuracy of the compared features. As for the completeness, when performed in terms of the number of objects is still lacking information (10%), whereas when considering the length or area of an object, a 40% match was achieved.

Brovelli et al. [2017] developed an automated comparison algorithm of OSM and authoritative road datasets that was based on a grid based evaluation of the completeness and positional accuracy of road objects. The algorithm, which is available as an open source module of GRASS GIS, showed good results in the tested area of Paris. Later in 2018 Brovelli and Zamboni [2018] performed a map matching and similarity check analysis for buildings to estimate the completeness of building registrations in the OSM database.

Bhattacharya [2012] attempted to find similarities and dissimilarities between the OSM and the Dutch topographic map Top10 NL. The quality assessment and object matching was performed in terms of the positional accuracy, the shape of OSM polygons and the lineage, i.e. the time related information about the collection and evolvement of the data. An interesting part of the positional accuracy evaluation, is the use of the difference of angle of the direction between two objects, as extra geometric quality measure. Following the same mindset, Fairbairn and Al-Bakri [2013] assessed the positional and shape quality (geometric similarity analysis) of OSM and other large scale data with ultimate goal to evaluate if the integration of these type of datasets is feasible in terms of accuracy and precision.

Mooney et al. [2010b] created an automated quality control algorithm, without considering any of the used datasets as reference. The comparison process was tested on OSM natural features (e.g lakes, ponds,forests) and various authoritative datasets in several European countries. As a quality metric for the shape similarity between objects, the authors used a turning function. This tool is used to describe the shape of a feature in a discrete format. More analytically, the shapes total length is normalized to 1, by dividing the length of each edge segment to the perimeter of the shape. Basically it represents 2D shapes in 1D. The spacing between points in the OSM polygons plays a significant role, as the larger the amount of points the more complete the turning function. The metric returns a value in the range [0,1] where 1 corresponds to identical polygons and 0 to completely dissimilar.

A matching process of linear features in Britain was developed by Koukoletsos et al. [2012]. OSM linear data were put to test against a reference dataset, to assess the heterogeneity of the two sources. The study area was first divided in $1km^2$ grid tiles. The analysis was performed at both segment and feature level, forming different steps of the process. The reason for this, is that a feature in one dataset may correspond to more than one in the other, creating errors and obstacles in the matching process. Therefore, both reference and tested features were decomposed into segments. As criteria for the segment based matching they used two geometric constraints. Since the OSM information is manually digitized on maps that are provided amongst others from sources such as acgps, the OSM feature accuracy is considered to depend on the GPS receiver accuracy. Based on that, the distance and angular tolerance, forming the two constraints, are calculated by taking into consideration the GPS accuracy. Ones the distance tolerance was formed, the segments within the searching area were analyzed to identify sections of similar orientation. Moreover, the analysis focused on identifying segments without any attributes, matching the names of features, classifying features as matched and non matched and finally estimating the data completeness according to the percentage of the matched data in each tile of the grid. The outcome of their survey showed that OSM data in Britain prove to be more complete in the urban than rural areas.

The quality of the OSM bulding data in Munich was analysed by Fan et al. [2014] relatively to the completeness, semantic accuracy, positional and shape accuracy of these features against a topographical authoritative dataset of buildings. As an out-

come of their research they found that the OSM data have high completeness and semantic accuracy, whereas their positional accuracy did not perform equally well (an average of 4m offset between datasets was noticed). To the contrary, the OSM building footprints presented high shape similarity to the German reference dataset. The assessment of the completeness was based on choosing as matched objects only the ones that presented an overlap of over 30% than the smaller part in the two datasets. In the cases of 1:1 correspondence they key points of the reference dataset where chosen and the for both objects was calculated. The minimum bounding rectangle of OSM was then shifted to the center of the authoritative minimum bounding rectangle, and if the edges of the former matched the edges of the latter, the match was considered successful. The positional accuracy was assessed according to the average distance of the corresponding points between the two datasets, and their shape similarity based on the turning function.

In the same concept, Müller et al. [2015], compared the OSM buildings in Switzerland with an authoritative dataset. The centroid distance and turning function were used as quality criteria. The research has shown that even though Switzerland has some gaps, its overall performance especially in the urban regions is quite good if not better that the reference data in some parts.

Donchyts et al. [2016] produced a 30 m resolution surface water mask for rivers by using Landsat satellite imagery, the Shuttle Radar Topography Mission DEM and OSM data in the Murray Darling River Basin in Australia. The goodness of fit between OSM and the LANDSAT water masks was assessed in terms of the completeness (based on total water area) and the positional differences of the two datasets using the Goodchild and Hunter [1997] method of increasing overlay polygons. As a result of his research it was concluded that 50% of the OSM linear water features agreed with the water extend extracted from Landsat 8 and the drainage network created from the Shuttle Radar Topography Mission DEM.

3.1 CONCLUSIONS

Many surveys have examined the quality of OSM features over various locations, however mostly in a small extend. Moreover, even though features such as buildings, roads or rivers have been assessed for their quality, there is no research that focuses specifically on the quality assessment of water reservoir features. The described methods and algorithms, which consist of various quality metrics and software, can also be applied in the case of water reservoirs. Their main characteristic nonetheless, is that firstly they usually compare the tested OSM features with authoritative datasets and secondly this comparison is mostly on a local scale.

Remote sensing is considered a valuable source of information over the last decades, as it has offered freely, large amounts of satellite data at a global scale. Although, these data have their limitations, their main advantage is that they offer the possibility of taking into consideration short or long term changes of surface water, by combining images from different moments in time. At the same time, although other water reservoir datasets in vector format exist (e.g GRaND), they are are still lacking information. Furthermore, the accuracy of their data is questionable because of their monitoring process or tea fact that they are a static representation, in which the changes of the reservoirs in time or within seasons have not been considered. Therefore the need emerges to explore different databases to examine how much they agree regarding their geometry and extend.

To conclude, while OSM spatial data have been put to test in the past, a general algorithm that compares different formats of data in one environment and at large scale does not yet exist. Therefore there is still room for investigating and developing new algorithms. The Google Earth Engine (GEE) platform offers the possibility of large scale analysis as it can store and process these constantly growing volumes

of satellite data, while also enabling the processing of vector data at the same time. This research will not make use of ground truth data, as no specific water reservoir dataset will be considered a reference, but it will attempt to create an algorithm that compares the accuracy of different formats of water reservoir datasets by utilizing both satellite imagery and vector data.

4 METHODOLOGY

The following methodology aims to compare and assess the quality of the different surface water datasets. To do this, we firstly focus on the pre-processing procedures that were followed for the preparation of the input data. More specifically, Section 4.1.1, describes the steps and the methods used to generate the Sentinel 2 water features. Next, Section 4.2 presents the chosen accuracy measures for the comparison analysis process. Starting with the completeness analysis (Section 4.2.1) which demonstrates the percentage of common water area between datasets within the specified study area. The Section 4.2.2 refers to the 'percentage of overlap' between two datasets, which is a positional accuracy measure for linear and polygonal features that shows the spatial offset between two features. Lastly, Section 4.2.3 and present the Euclidean which is necessary for the computation of the final quality metric, the Hausdorff distance (Section 4.2.4), as measures of dissimilarity of the shape and position of two features. Figure 4.1 provides an overview of the workflow of this research. Further details of each step shall be explained in Chapter 6.

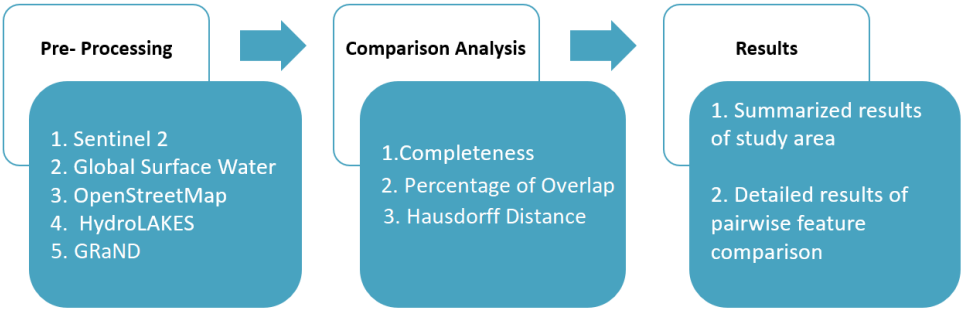


Figure 4.1: Workflow of study

4.1 PRE-PROCESSING

The implementation of the methodology described in Section 4.2, required initially a pre-processing step for the selected datasets. This preparation varied amongst the datasets, but in general it had two main objectives. Firstly, the Sentinel 2 satellite images in raw format, provide no information about the surface water. Therefore the Sentinel 2 water mask data had to be generated from the start. In order to detect and classify water pixels, certain steps need to be followed as described in Section 4.1.1. Secondly, as the datasets have different formats, an essential step is the conversion into one matching format so that the data can be compared. More specifically, all input datasets were converted into vector format. Furthermore, the comparison analysis in terms of completeness and percentage of overlap was performed in vector format. However the Hausdorff Distance contained also a conversion from vector to raster format as part of its computational workflow. Other steps such as filtering, cropping and other minor transformations of the input datasets are explained thoroughly in Chapter 5 and Chapter 6.

4.1.1 Sentinel 2 surface water dataset

Sentinel-2 being the latest temporal resolution and highest spatial resolution data available, is very useful for detailed water surface boundary extraction. The presence of water in a pixel can be described with historical observations of the same area in different moments in time. The approach is based on the sampling of different cloudless historical images, to compute a single image that presents where water occurred during this period. The values of this composite image represent the 20^{th} percentile of the values of the GREEN and SWIR bands for each pixel in all sampled images. The 20^{th} percentile is the value below which 20% of the observations fall within, in other words 80% of the observations are above this value. The percentile images appeared according to Donchyts et al. [2016] to describe better the water dynamics that the median of all values, which was used in other studies [Hansen et al., 2013a]. This was confirmed by visual inspection only, and it is explained based on the fact that surface water, changes sharply depending on the topographical conditions that appear locally. The choice of the 20^{th} percentile is explained in more detail in Section 6.2.1.

Cloudless composite images can be generated by employing percentile images to estimate the average cloud-free reflectance values. Cloud coverage is estimated by exploiting the statistical properties of the image and more specifically the reflectance property of clouds. The main idea is that the more bright the pixel appears in certain bands, the more likely it is that it is covered by clouds. High brightness in the SWIR band is ideal for cloud detection. Moreover, pixels with low reflectance in the B2 (Blue) and high reflectance in the B10 (Cirrus) bands are highly possible of being cirrus cloud. The percentile images are computed according to Donchyts et al. [2016] on a per-band basis using Top of Atmosphere Top of Atmosphere (TOA) reflectance values, to avoid the confusion created by different atmospheric correction algorithms of satellites. TOA measurements is the reflectance measured by the sensor without performing any type of corrections for clouds, atmospheric aerosols and gases.

To acquire cloud free images, a quantile analysis of the distribution of the reflectance of each pixel in the entire image is performed, just to choose that part that is considered cloud free. In more detail, the images are first sorted from cloudy to less or no cloudy, by calculating per image, the reflectance value for which a specific percentage of the pixels (percentile) falls below. Thereafter, once the sorted listed of the images is formed, they can be categorized relatively to their cloud coverage depending on a threshold value indicated from the mean annual cloud frequency acquired by MODIS satellite imagery. Finally, after classifying those images into cloudy and cloudless, only the cloud-free images are being kept for further processing.

The generation of the surface water mask based on cloud-free Sentinel 2 images requires a certain process, based mainly on two image processing algorithms as proposed by Donchyts et al. [2016]. The steps in short can be described as follows (see Figure 4.2):

1. Computation of spectral water index
2. Computation of edges using the Canny edge detector
3. Buffering of detected edges
4. Computation of a threshold value for buffered area by using the Otsu thresholding method

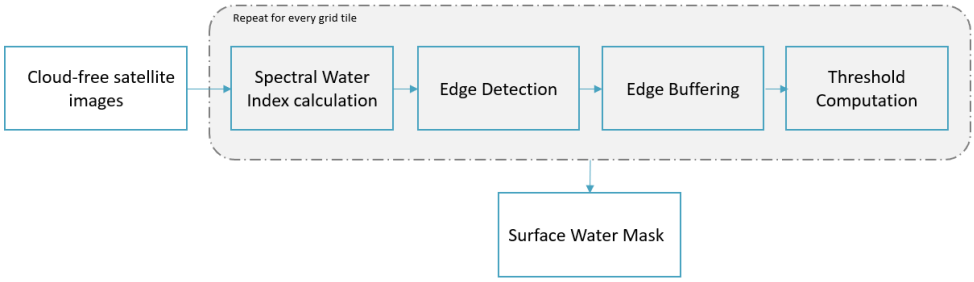


Figure 4.2: Water detection pipeline

4.1.2 Canny edge detector

The Canny edge detector is a widely used method for accurate edge detection in images [Duan et al., 2005]. The edge filter can assist in the detection of boundaries between water (black pixels) and non water (white) pixels (Figure 4.3), which helps further to reduce the extend of the area where the Otsu thresholding is applied (see Section 4.1.3).

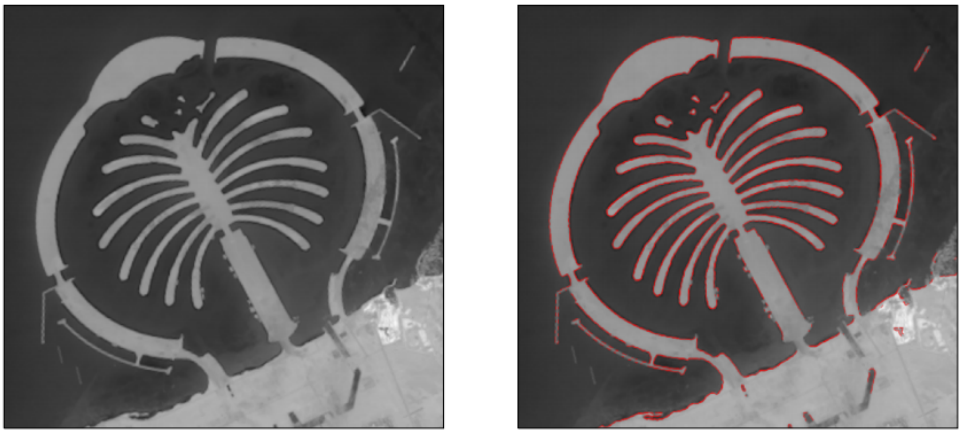


Figure 4.3: Canny Edge Detection in Satellite Image [Donchyts, 2018]

The algorithm consists of the following stages:

1. Image Smoothing: Edge detectors are prone to noise and therefore they are firstly smoothed with a square-sized Gaussian structural kernel usually of size 5×5 Li et al. [2019].

2. Gradient intensity calculation: The gradient direction defines the orientation of an edge, whereas the gradient magnitude indicates the intensity of a change in the reflectance values. High gradient magnitudes reveal the detection of an edge.

$$G = \sqrt{{G_x}^2 + {G_y}^2} \tag{4.1}$$

$$\theta = tan^{-1}\frac{G_x}{G_y} \tag{4.2}$$

where G_x and G_y the x,y derivatives of the current pixel. θ is rounded to 0 (horizontally),45 (diagonally), 90 (vertically) or 130 (diagonally) degrees.

3. Non-maximum suppression: All pixels are checked to see if they are a local maximum in certain neighborhood. If they are not, they are suppressed, resulting in very thin edges.

4. Double thresholding: Removal of small pixel noises, based on two threshold values of the intensity gradient. It is applied to detect strong edges only.

5. Hysteresis thresholding: Pixels below a certain threshold are discarded. This way edges that are weak are suppressed.

The detected sharp edges between water and land will be expanded with a buffer zone to make sure that all probable water and land pixels around the boundary are captured. This way, we it will be possible to obtain a bimodal distribution which will assist the distinction of the two classes in the derived histogram of MNDWI values (Figure 4.4 [5]).

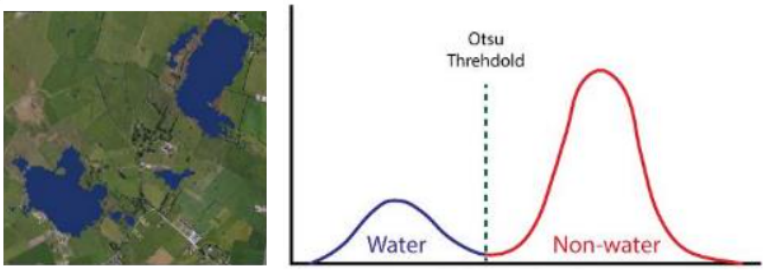


Figure 4.4: Otsu's Thresholding in Water Indexed Image

4.1.3 Otsu Thresholding

In order to separate water from non water features, a threshold value for MNDWI has to be estimated. Dynamic local thresholding will help avoiding errors in the surface water extraction procedure. Otsu thresholding is based on a histogram of all MNDWI values in a certain area Otsu [1979]. The goal of this method is to create a binary image [0,1] of two different classes, no water (object pixels) and water (background) pixels Figure 4.5.

5 Reference of image otsuexample https://www.gsi.ie/en-ie/programmes-and-projects/groundwater/activities/groundwater-flooding/gwflood-project-2016-2019/Pages/Mapping-methodologies.aspx

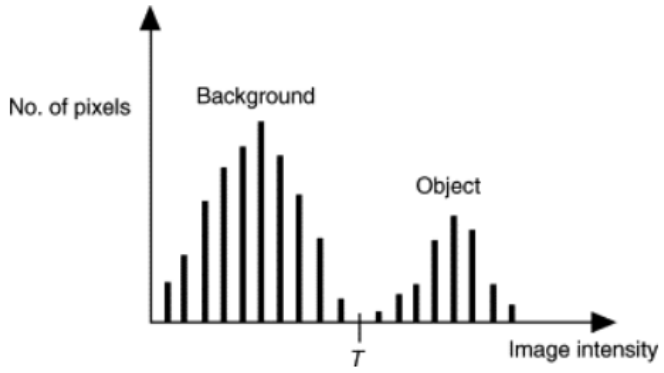


Figure 4.5: Bimodal histogram and selected threshold value (T) [Rogowska, 2009]

The general idea of the method is to find the threshold that minimizes the weighted within-class variance which is equal to the weighted sum of variances of the two classes (Equation 4.3). This is done by exploring all possible threshold values and calculating the variance of all pixels on each side of the threshold.

$$\sigma_w{}^2(t) = \omega_0{}^2(t) * \sigma_0{}^2(t) + \omega_1{}^2(t) * \sigma_1{}^2(t) \tag{4.3}$$

where σ_w is the intra-class variance, ω_0 and ω_1 are the probabilities of the two classes separated by the threshold t and σ_0 and σ_1 are the variances of the two classes. The class probability is computed from the number (L) of the bins of the histogram.

$$\omega_0 = \sum_{i=0}^{t-1} P(i) \tag{4.4}$$

$$\omega_1 = \sum_{i=t}^{L} P(i) \tag{4.5}$$

The class means μ_0, are given by the following equations:

$$\mu_0(t) = \frac{\sum_{i=0}^{t-1} iP(i)}{\omega_0(t)} \tag{4.6}$$

$$\mu_1(t) = \frac{\sum_{t}^{L-1} iP(i)}{\omega_1(t)} \tag{4.7}$$

Yousefi [2011] proved that maximizing the between-class variance, instead of minimizing the within-class variance has a higher performance. Therefore,

$$\sigma_b{}^2(t) = \sigma^2 - 2\sigma_w = \omega_0(t) * \omega_1(t) * (\mu_0(t) - \mu_1(t))^2 \tag{4.8}$$

4.2 COMPARISON OF ACCURACCY OF DATASETS

The comparison of accuracy analysis is performed in terms of the completeness of water area of the datasets (Section 4.2.1) and their positional accuracy (Section 4.2.2,

Section 4.2.4). The purpose of a positional accuracy analysis is to provide quantitative information about the positional difference between two objects, in other words how they are placed in Euclidean space relatively to each other (Section 4.2.3). The assessment of this quality indicator, is performed in terms of tho measures:

1. Percentage of overlap between two objects
2. Hausdorff distance between two objects

4.2.1 Completeness Analysis

The completeness of registered features is an important measure to assess the quality of OSM. To estimate the completeness of the water reservoir features in the OSM database, the total overlap between the datasets is calculated. In order to do this in an aggregated form for the whole extend of the study area, a 40×40 km regular grid is generated Figure 4.6. Thereafter, for every grid cell, the total covered water area of each dataset, the common water area between datasets, and the missing area of each dataset are computed. The choice of the size of the grid has no effect on the computation of the completeness for the whole study area, as it serves only visualization purposes for the results.

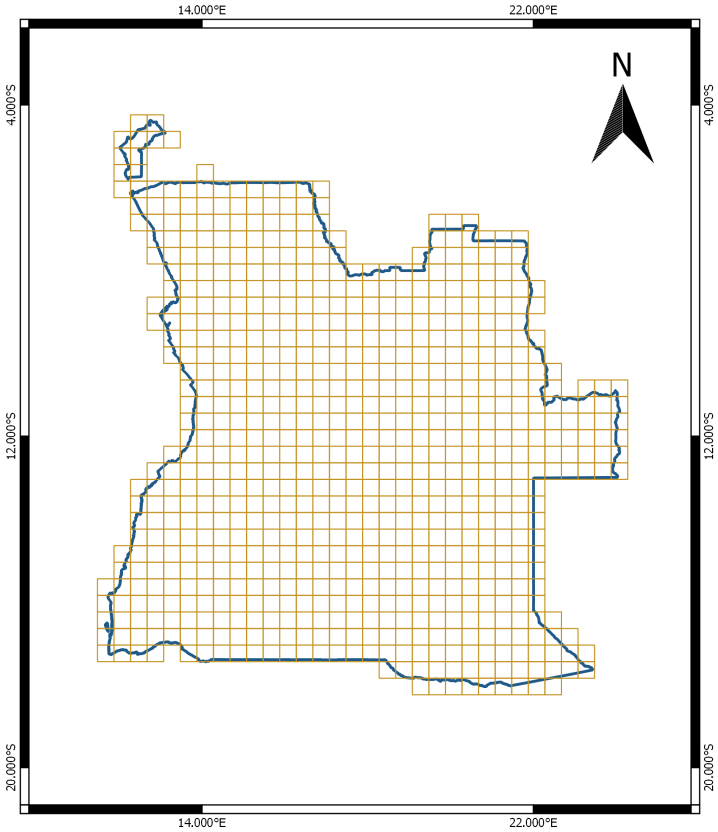


Figure 4.6: Grid of 40×40 km over the extend of the study area

4.2.2 Percentage of overlap

The percentage of overlap is a method introduced by Goodchild and Hunter [1997] for the assessment of the positional accuracy of geometric primitives. The method calculates the percentage of area of one dataset that is within a specified distance

of another dataset (see Figure 4.7). To do this, one of the two datasets is considered the reference (higher accuracy) feature and the percentage of overlap of the tested feature is calculated for a number of buffers of different widths for the reference source. The concept of the increasing buffers, is based on the idea that by calculating the percentage of overlap for different buffer zones, we can identify the accuracy of the tested dataset, i.e the closeness (or offset) to the true location of the object.

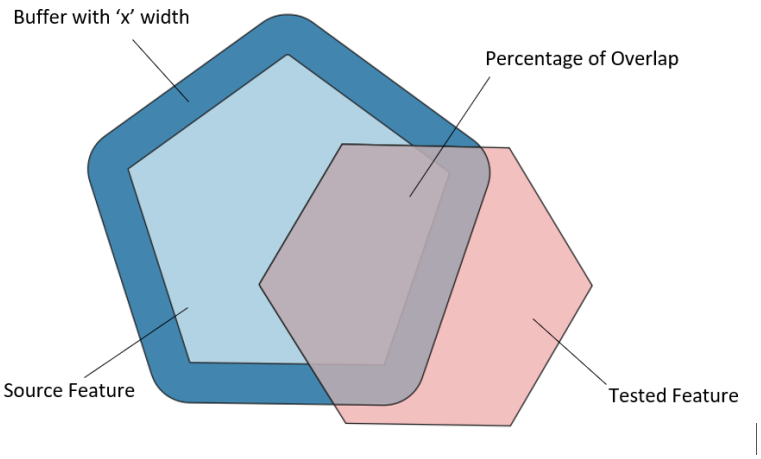


Figure 4.7: Goodchild's Percentage of overlap

The advantage of this quality metrics is that it can be implemented, for both raster and vector representations. Moreover, a) it is statistically based, b) is relatively insensitive to extreme outliers and, c) does not require matching of points between representations. Moreover, it is performed with a simple overlay process (see Section 2.2.4).

4.2.3 Euclidean Distance

The Euclidean Distance is defined as the minimum distance between two objects, i.e. the length of the shortest line that connects them. In raster discrete space (see Section 2.2.2) the euclidean distance is the distance from each cell to the nearest source cell location, i.e. the cells with value 1. It is calculated from the center of the source cell to the center of each of the surrounding cells (see Figure 4.8).

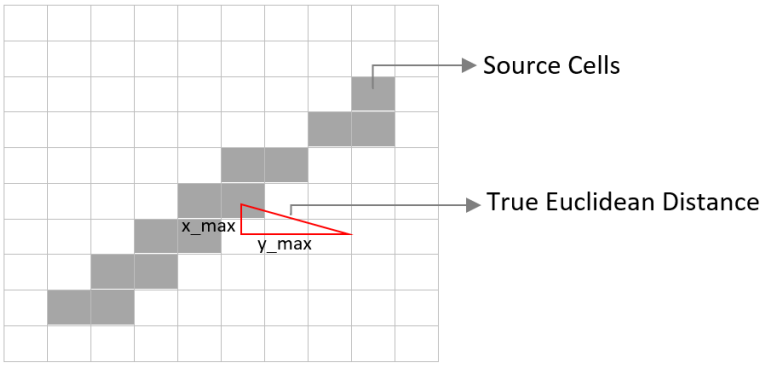


Figure 4.8: Euclidean Distance

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	1	1	0
0	0	0	0	0	1	1	0	0	0
0	0	0	0	1	1	0	0	0	0
0	0	0	1	1	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0
0	1	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

(a) Binary Raster Image

6.40	5.66	5.00	4.48	4.12	3.61	2.83	2.24	2.00	2.24
5.66	5.0	4.24	3.61	3.16	2.83	2.24	1.41	1.00	1.41
5.00	4.24	3.61	2.83	2.24	2.00	1.41	1.00	0.00	1.00
4.24	3.61	2.83	2.24	1.41	1.00	1.00	0.00	0.00	1.00
3.61	2.83	2.24	1.41	1.0	0.00	0.00	1.00	1.00	1.41
2.83	2.24	1.41	1.00	0.00	0.00	1.00	1.41	2.00	2.24
2.24	1.41	1.00	0.00	0.00	1.00	1.41	2.24	2.83	3.16
1.41	1.00	0.00	0.00	1.00	1.41	2.24	2.83	3.61	4.12
1.00	0.00	0.00	1.00	1.41	2.24	2.83	3.61	4.24	5.00
1.41	1.00	1.0	1.41	2.24	2.83	3.61	4.24	5.00	5.66

(b) Euclidean Distance Grid

Figure 4.9: Euclidean Distance Grid for Rasterized Linear feature

4.2.4 Hausdorff Distance

The Hausdorff distance (h) measures the maximum deviation between two sets. More specifically, it is the maximum from all minimum distances from one set to the nearest point in the other set, as given by the Equation 4.9. It is calculated by finding the minimum distance of a point $a \in A$ to any point in B, and then by selecting the maximum distance from all the shortest distance values computed for all the points in set A.

$$h(A,B) = max\Big\{min\ d(a,b)\Big\} \tag{4.9}$$

where d is the Euclidean distance between point $a \in A$ and $b \in B$.

However, as the forward distance (from set A to set B), is not always equal to the backward distance (from set B to set A), the bidirectional Hausdorff Distance is defined as a symmetric distance from the Equation 4.10 [Huttenlocher et al., 1993; Thirusittampalam et al., 2013]:

$$H(A,B) = max\Big\{h(A,B), h(B,A)\Big\} \tag{4.10}$$

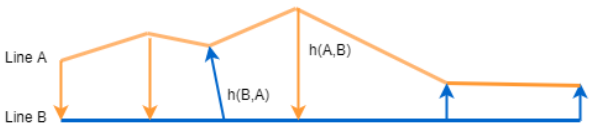


Figure 4.10: Hausdorff Distance between two polylines

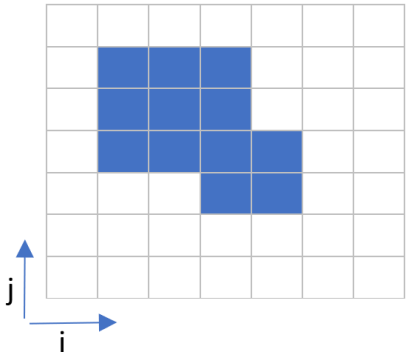
Hossain et al. [2012] proposes a method for computing the Hausdorff Distance between raster representations, by making use of a Distance Transform , also known as Distance Map. This map is generated by computing the distance of each pixel to the closest boundary point based on the Euclidean Distance [Meijster, 2004]:

$$EDT(x,y) = min(x-i)^2 + G(i,y) \tag{4.11}$$

$$G(i,y) = min(y-j)^2\ where\ F(i,j) = 0 \tag{4.12}$$

where F(i,j) the input image, i,j the rows and columns of the image array respectively.

The algorithm works as follows: The image is stored in an array of columns and rows. Afterwards, the algorithm iterates through all pixels from top to bottom and then bottom to top to, to compute the minimum distances in this dimension G(x,y) to the closest boundary pixel. Then the array G is scanned from left to right and right to left to calculate again the minimum distance to the closest boundary point Thissen [2019].



(a) Raster Image B

0	0	0	0	0	0	0
0	1	1	1	0	0	0
0	1	1	1	0	0	0
0	1	1	1	1	0	0
0	0	0	1	1	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

(b) Binary Image

3.6	2.8	2.2	2.0	2.4	2.8	3.2
3.2	2.2	1.4	1.0	1.4	2.0	2.2
3.0	2.0	1.0	0.0	1.0	1.0	1.4
3.0	2.0	1.0	0.0	0.0	0.0	1.0
3.0	2.0	1.0	0.0	0.0	0.0	1.0
3.0	2.0	1.0	0.0	0.0	0.0	1.0
3.2	2.2	1.4	1.0	1.0	1.0	1.4

(c) Euclidean Distance Transform Map of Object A

3.6	2.8	2.2	2.0	2.4	2.8	3.2
3.2	2.2	1.4	1.0	1.4	2.0	2.2
3.0	2.0	1.0	0.0	1.0	1.0	1.4
3.0	2.0	1.0	0.0	0.0	0.0	1.0
3.0	2.0	1.0	0.0	0.0	0.0	1.0
3.0	2.0	1.0	0.0	0.0	0.0	1.0
3.2	2.2	1.4	1.0	1.0	1.0	1.4

(d) Hausdorff Distance h(B,A)

Figure 4.11: Computation of backwards Hausdorff Distance h(B,A)

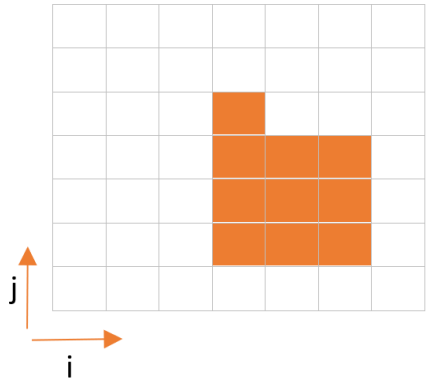
There are two different methods to compute the Hausdorff Distance from the Distance Transform Map. The first method utilizes the Equation 4.10, and the second is a modified version as shown in Equation 4.15.

$$h(A,B) = avg\Big\{\dot{min}(a,b)\Big\} \tag{4.13}$$

$$h(B,A) = avg\Big\{\dot{min}(b,a)\Big\} \tag{4.14}$$

Finally, the Hausdorff Distance is calculated by Equation 4.15:

$$H(A,B) = h(A,B) + h(B,A) \tag{4.15}$$



(a) Raster Image A

0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	0	0	1	1	1	0
0	0	0	1	1	1	0
0	0	0	1	1	1	0
0	0	0	0	0	0	0

(b) Binary Image

1.4	1.0	1.0	1.0	1.4	2.2	3.2
1.0	0.0	0.0	0.0	1.0	2.0	2.8
1.0	0.0	0.0	0.0	1.0	1.4	2.2
1.0	0.0	0.0	0.0	0.0	1.0	2.0
1.4	1.0	1.0	0.0	0.0	1.0	2.0
2.2	2.0	1.4	1.0	1.0	1.4	2.2
3.2	2.8	2.2	2.0	2.0	2.2	2.8

(c) Euclidean Distance Transform Map B

1.4	1.0	1.0	1.0	1.4	2.2	3.2
1.0	0.0	0.0	0.0	1.0	2.0	2.8
1.0	0.0	0.0	0.0	1.0	1.4	2.2
1.0	0.0	0.0	0.0	0.0	1.0	2.0
1.4	1.0	1.0	0.0	0.0	1.0	2.0
2.2	2.0	1.4	1.0	1.0	1.4	2.2
3.2	2.8	2.2	2.0	2.0	2.2	2.8

(d) Hausdorff Distance h(B,A)

Figure 4.12: Computation of forwards Hausdorff Distance h(A,B)

This metric can also be used as an indication of the level of similarity/dissimilarity between two raster images of size m. All pixels with value 1 represent an object whereas pixels with 0 values are background pixels. The Hausdorff Distance can be computed by:

1. Method 1: First the h(B,A) is computed by assigning to it the maximum value of the Euclidean Distance Transform (EDT) of image A of the cells where $B[i][j] = 1$ edt such that the image B is in the object A (Figure 4.11). Secondly, $h(A, B)$ is computed similarly from the edt of B and then by taking tha maximum of all values in edt in the cells where there is an object of A (A[i,j]=1) (Figure 4.12). Finally, the Hausdorff Distance is the max ($h(A, b)$,$h(B, A)$).

2. Method 2: To compute the Hausdorff distance h(B,A) and h(A,B) are assigned with the average, instead of the maximum, values of EDT of A, where $B[i][j] = 1$ and EDT of B where $A[i][j] = 1$.

Lastly, the Hausdorff Distance algorithm as described above is slightly adjusted to work also with vector data while following the same logic. More specifically, instead of estimating the Hausdorff Distance by overlapping two raster representations A and B, only the feature for which the Euclidean Distance Map (EDM) is be computed will be rasterized (e.g. object A). Afterwards, this generated EDM of object A will be overlapped with the vector feature B (Figure 4.14). Each of the points forming the vector B will be overlapped with the values of the EDM of the rasterized feature A. Same process will also be applied reversely, resulting in the estimation of the Euclidean distance values for the points of vector A based on its overlap with the EDM of rasterized object B (Figure 4.13).

(a) Vector Polygon Object A

3.6	2.8	2.2	2.0	2.4	2.8	3.2
3.2	2.2	1.4	1.0	1.4	2.0	2.2
3.0	2.0	1.0	0.0	1.0	1.0	1.4
3.0	2.0	1.0	0.0	0.0	0.0	1.0
3.0	2.0	1.0	0.0	0.0	0.0	1.0
3.0	2.0	1.0	0.0	0.0	0.0	1.0
3.2	2.2	1.4	1.0	1.0	1.0	1.4

j
i

(b) Intersection of EDM of B with object A

Figure 4.13: Euclidean Distances of Object A to B

(a) Vector Polygon Object B

1.4	1.0	1.0	1.0	1.4	2.2	3.2
1.0	0.0	0.0	0.0	1.0	2.0	2.8
1.0	0.0	0.0	0.0	1.0	1.4	2.2
1.0	0.0	0.0	0.0	0.0	1.0	2.0
1.4	1.0	1.0	0.0	0.0	1.0	2.0
2.2	2.0	1.4	1.0	1.0	1.4	2.2
3.2	2.8	2.2	2.0	2.0	2.2	2.8

j
i

(b) Intersection of EDM of A with object B

Figure 4.14: Euclidean Distances of Object B to A

5 DATASETS AND TOOLS

5.1 DATASETS

5.1.1 Sentinel-2

The Sentinel-2 mission was launched by the European Space Agency (ESA) in June 2015 as part of the Copernicus program. The constellation is equipped with two identical polar-orbiting multispectral-imaging satellites, Sentinel-2A (2015) and Sentinel-2B (2017). They are placed in the same sun-synchronous orbit, phased at 180° to each other [6]. The purpose of this mission is to monitor the land environment (the vegetation, inland waterways, soil, water and coastal areas). Its territorial coverage extents globally, between latitudes 56 ° south and 84° north.

The wide swath width along with the high revisit time (10 days) allows the monitoring of temporal changes of the Earth's surface (Figure 5.1[7]). With the combined satellites, the overlap between swaths from adjacent orbits increases the frequency of observations, as all areas at the Equator are revisited every 5 days under the same viewing conditions.

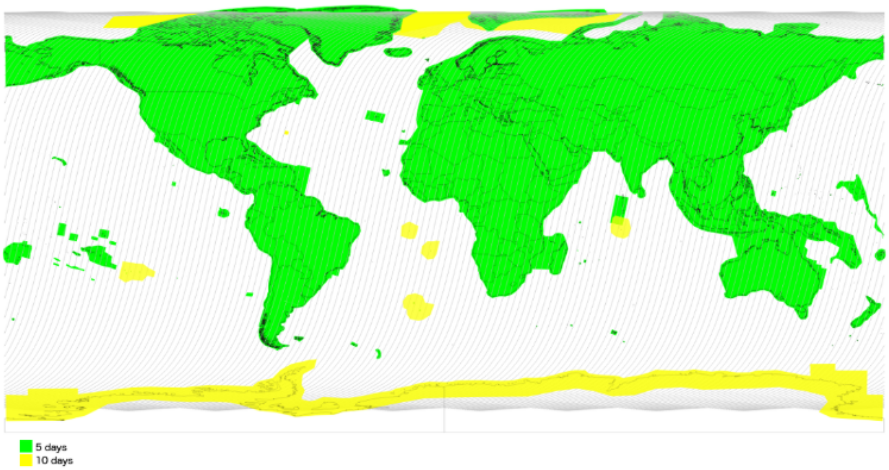


Figure 5.1: Sentinel 2 revisit temporal resolution

For this research, the dataset used is the Sentinel-2 MSI (MultiSpectral Instrument Level-1C). This product is a map projection of the acquired image referenced in the WGS84 global reference system, using a DEM which offers geometric correction for ground distortions. It contains 13 spectral channels representing TOA reflectance values, which are provided together with the parameters to transform them into pixel based radiances. This offers a radiometric correction to the raw pixel values of the images. Lastly a bitmask band with cloud mask information (QA60) is included. (Table 6.8[8]). The spatial resolution of this dataset varies from 10 to 60 offering dif-

6 https://sentinel.esa.int/web/sentinel/missions/sentinel-2
7 https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/revisit-coverage
8 https://www.usgs.gov/centers/eros/science/usgs-eros-archive-sentinel-2?qt-science_center_objects=0qt-science_center_objects

ferent Ground Sample Distance (GSD), i.e. the ground area captured in a single pixel.

Band	Resolution (m)	Wavelength (nm)	Description
B1	60	443	Aerosols
B2	10	490	Blue
B3	10	560	Green
B4	10	665	Red
B5	20	705	Red Edge 1
B6	20	740	Red Edge 2
B7	20	783	Red Edge 3
B8	10	842	NIR
B8A	20	865	Red Edge 4
B9	60	945	Water Vapor
B10	60	1375	Sirus
B11	20	1610	SWIR 1
B12	20	2190	SWIR 2
QA60	60	-	Cloud mask

Table 5.1: Sentinel-2 Spectral and QA Bands

5.1.2 JRC Global Surface Water Mapping

The Global Surface Water Mapping dataset was created by the Joint Research Center in collaboration with Google Earth Engine. It is a map generated based on 3 million Landsat satellite images of 30 m resolution collected between 1984 and 2018. It offers information about the location and the changes of the extend of the surface water all over the globe in the past three decades (Table 5.2[9]).

Band	Range (%)	Description
occurrence	[0,100]	Frequency of of presence of surface water
change_abs	[-100,100]	Absolute change in occurrence between 1984-1999 vs 2000-2018.
change_norm	[-100,100]	Normalized change in occurrence
seasonality	[0,12]	Number of months where water occured
recurrence	[0,100]	The frequency with which water returns from year to year.
transition	-	Categorical classification of change between first and last year.
max_extent	-	Binary image containing 1 anywhere water has ever been detected.

Table 5.2: JRC Global Surface Water Bands

5.1.3 OpenStreetMap

OpenStreetMap, as a croudsourcing platform, is built upon the contribution of volunteers that manually share the location of geographical information on an editable web map. The acquired information is stored in a database organised according to a specific data model, which is based on expressing the geometric features as three objects: nodes, ways and relations. According to OpenStreetMap Wiki contributors [10] these three elements are described as follows:

- Nodes: Are used to represent point features and they are defined by their coordinates (latitude, longitude). They can be presented as standalone components or as part of a group of nodes.

9 https://developers.google.com/earth-engine/datasets/catalog/JRC_GSW1_1_GlobalSurfaceWater
10 http://wiki.openstreetmap.org/wiki/Main_Page

- Ways: They represent linear features that connect two or more nodes. Closed ways (first and last node are the same) specifically, describe area features (e.g. polygons).
- Relations: Are used to describe the relationship between multiple nodes or ways and to organize them into groups.

All types of data objects are saved as tagged geometric primitives. These tags are used as description of the map features and consist of two text fields in the form of a dictionary (key and value). The keys add meaning to the features by describing a category or type whereas the value provides the detailed information of this key-feature (e.g. name). There is no fixed and predefined tag list, as every user is able to create a new tag and add it on existing or new map objects. However, there is a tag info website [11] that contains all the existing tags in the database, together with statistics about the frequency and type of their use.

The OSM data are distributed in different digital formats. The data used for this research were download in vector format from the recent files in the Geofabrik website [12] for the country of Angola. In order to acquire only the water reservoir features from the original file, a filtering and data format conversion intermediate step was performed. The specific tags used to extract the water features are shown in Table5.3 [13]. The degree of accuracy of the OSM data varies depending on location of data and also the node spacing in the OSM is not standard.

Finally, from the filtered features, only polygons and multipolygons were considered. Specifically, multipolygons had to be broken down to simple polygons.

Key	Value	Description
Natural	Water	Areas of water
Natural	Spring	Area to which water is discharged from an underground source
Waterway	Dam	A wall built across a river to stop the flow of water
Waterway	weir	Low-rise dam
Landuse	Reservoir	Man made body to store water
Landuse	Saltpond	Area where salt is extracted from sea water by humans
Landuse	Basin	Area of land artificially created to hold water by lowering its level
Barrier	Ditch	Man-made barrier dugged in the ground

Table 5.3: OSM filtering tags

5.1.4 Global Reservoir Datasets

- **HydroLAKES**: Dataset that provides the boundary polygons of all global lakes with a surface area of at least 10 ha.
- **GRanD**: Contains information about existing dam and reservoir features in vector format at a planetary scale.

11 https://wiki.openstreetmap.org/wiki/Elements
12 https://download.geofabrik.de/africa/angola.html
13 https://taginfo.openstreetmap.org/tags

The number of the water features included in each of the input datasets are given in Table 5.4. In the case of GSW and Sentinel 2, the stated number refers to all possible features including rivers.

Dataset	Number of Water Features
OSM	1189
HydroLAKES	433
GRaND	10
GSW	47952
Sentinel 2	44079

Table 5.4: Number of registered water features in the source datasets

5.2 TOOLS

Google Earth Engine is a cloud-based platform that is offering access to multi-petabyte satellite imagery and other geospatial datasets at a global scale. It provides the possibility of analysis and visualization of space and time changes and trends over large Image and Feature Collections, as it utilizes parallel processing. This enables the simultaneous execution of processes and calculations, a concept used in big data analysis, where instead of using for loops the functions are being "mapped" and executed for all data at the same time. The integrated code editor enables the computation of statistics and other image processing functions [Gorelick et al., 2017] by using the JavaScript programming language, the outcome of which can be directly displayed at the built-in output console. GEE is also available through Python and JavaScript Application Programming Interface (API). Moreover, the user has the option of uploading and exporting datasets.

Quantum GIS or QGIS is a free, open-source software application for analysing, editing and visualising geographic information. In this research it was used initially for separating the OSM GeoPackage files into individual lines, points and polygon shapefiles (SHP) and dissolving multipolygons. Moreover, it was employed for the visualization of the datasets and results of the implemented algorithms.

Python for computng statistics and visualising results. The following packages were used:

- **matplotlib**: Library for creating visualizations.
- **NumPy**: organizes and stores the data in arrays and matrices and offers various mathematical functions for manipulation of the data of this format.
- **pyproject**: Library for cartographic projections and coordinate transformations to check the coordinate system of the OSM data.
- **Shapely**: Module to process planar vector data and perform geometric operations.

The acquire the OSM information about reservoirs a filtering and format conversion pre-processing step was performed by using the following tools:

- **Osmfilter**: Command line Java application for filtering of OSM data with specifically desired tags
- **ogr2ogr**: It is part of the Geospatial Data Abstraction Library (GDAL) which is used for vector and raster data manipulation. In particular, ogr2ogr converts simple features (standardized geographic data model) between different formats. During this research it was utilized to convert OSM files from .osm format to GeoPackage (.gpkg) files

5.3 IMPLEMENTED ALGORITHMS

This section provides the link to the generated scripts for the pre-processing and main analysis of the data. Moreover, the resulting datasets of the implemented methods are included. All scripts are available in the following Google Earth Engine Repository or Github Repository .

6 IMPLEMENTATION AND RESULTS

6.1 STUDY AREA

The study area of this research is the country of Angola in Africa, located in the tropical zone of the Earth. Its climate is characterized by alternating rainy and dry seasons. However, millions of litres of rainwater are being lost from evaporation or ground absorption. Especially the southern Angola, where the climate is more arid, is facing a significant lack of available water. Reservoirs and lakes can support the storage and management of surface and atmospheric freshwater that can be collected throughout the year to assist the water supply in periods of drought. Both natural and man-made water features of Angola have been registered in different datasets (OSM, HydroLAKES, GRaND). However, these databases are still incomplete, so new information or the assessment of the accuracy of existing data on water reservoirs in Angola is essential. Another reason this area was selected is the fact that it is a 1.247 million km^2 region, which is completely in line with the objective of this study to create an accuracy control tool that can perform large scale analysis.

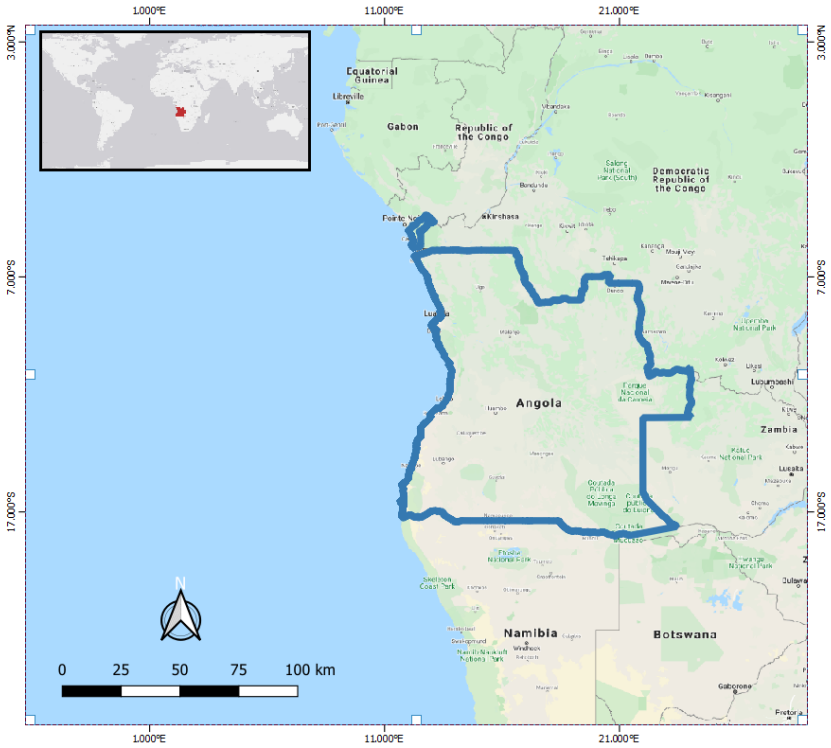


Figure 6.1: Overview of study area (Angola)

6.2 PREPROCESSING

6.2.1 Sentinel 2

The method for surface water detection with Sentinel 2 as described in Section 4.1.1, required huge computational resources, causing the GEE platform to fail when processing the area of Angola as a whole. Therefore, the process had to be decomposed in intermediate steps and be applied in smaller sub areas. A grid of $1km \times 1km$ was generated over the extend of the research area. This way the steps of the described methodology could be implemented for each tile individually.

As satellite images are often disturbed by cloud obstructions, the most clean images over the sampled time period (2016-2019) for each tile were chosen. Cloudy pixels appear very bright in all visible bands. To assess the cloudiness of each image, their reflectance values in the Green visible band were computed at a pixel level. This way the same pixels were sampled over the years and for each image a histogram of the reflectance values was created. The distribution of these values for every image was then assessed, so that they can be sorted depending on their cloudiness (Figure 6.3a[14]). The acquired histogram is basically a joint distribution of all possible types of noise (clouds, shadows etc). Therefore, a 75% percentile was utilized, to find for each image what is the value of reflectance for which 75% of pixels lie within. Once this characteristic quality score, i.e their reflection in Green band was acquired for each image, the images were sorted. Then, to distinguish between cloudy and cloudless images the mean annual cloud frequency by MODIS satellite imagery was used as a threshold for the calculated cloud-scores (Figure 6.3b). The reasoning behind the 75% percentile comes from the fact that we were interested in finding the top reflectance values for most pixels in each image, and therefore a high percentile was logical to be chosen. Finding the best performing percentile heuristically was considered out of the scope of this thesis, due to time constraints.

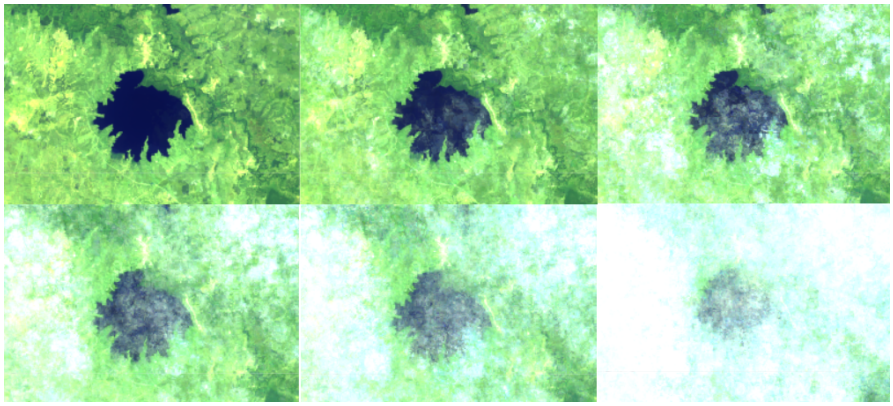
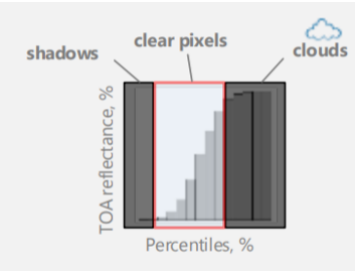
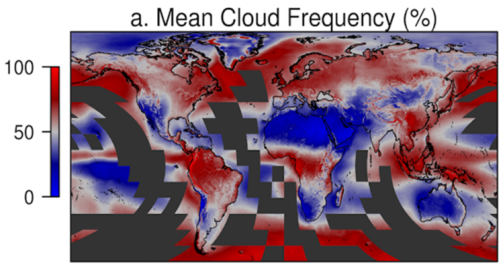


Figure 6.2: Cloudy and Cloud-free Satellite Images

14 Planetary-scale geospatial analysis with Google Earth engine, Gennadii Donchyts & Josh Friedman, Deltares, https://www.slideshare.net/Delft_Software_Days/dsd-int-2015-planetaryscale-geospatial-analysis-with-google-earth-engine-gennadii-donchyts-amp-josh-friedman-deltares



(a) Sorted images depending on cloudyness

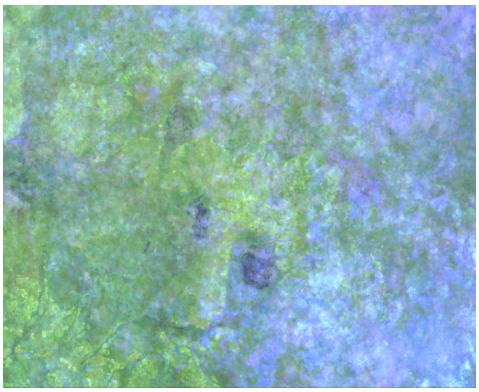


(b) Mean Cloud Frequency MODIS

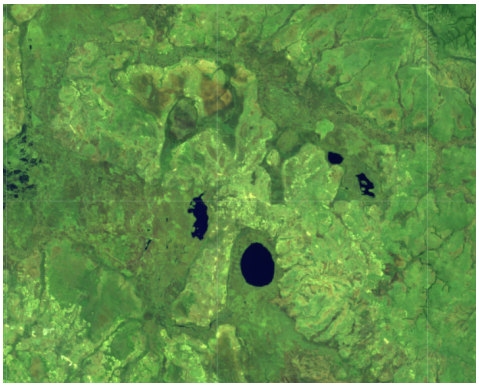
Figure 6.3: Classification of satellite images to cloudy and cloud-free images

Once, the most clean images were acquired, to detect the existing water areas the MNDWI water index for every pixel in each image was calculated by employing the Green and SWIR1 bands. Next, the median reflectance value for each pixel amongst all images was computed to create a single image. The problem however, with this part of the workflow is that the amount of clean images was so big that the process eventually failed when computing the median image. Therefore this part of the algorithm had to be adjusted in a way that it requires less computational effort. The surface water mask was delineated by employing the 20^{th} percentile, of the reflectance values for each pixel in the two selected bands, instead of the median, and then computing the MNDWI. This would result in the value of Green and SWIR1 reflectance for which 80% of the observations at pixel level are above. The 20^{th} percentile was chosen, apart from visual inspection, based on the logic that when acquiring the most clean images, a small percent of the pixels remain cloudy. The idea was to choose a low percentile which corresponds to higher water land, but not too low to acquire shadows, so practically the lower the percentile the bigger the water area that is detected.

This process could not be fully automatized, because of the ruinous required computational resources for analyzing thousands of cloud-free satellite images. Therefore, the resulting surface water masks were firstly exported and imported again to be used as input for the next steps. The reason for this was to avoid the crashing of the algorithm when functioning for all the steps consecutively. The detection of the surface water masks was followed by the Canny edge detector. Ones the edges of each waterbody were dilated, they were buffered in order to focus on the sharp changes between water and land that take place within this buffer zone. Then, a bimodal distribution of the MNDWI values of the water and land pixels was obtained and thereafter separated with Otsu's Thresholding method.

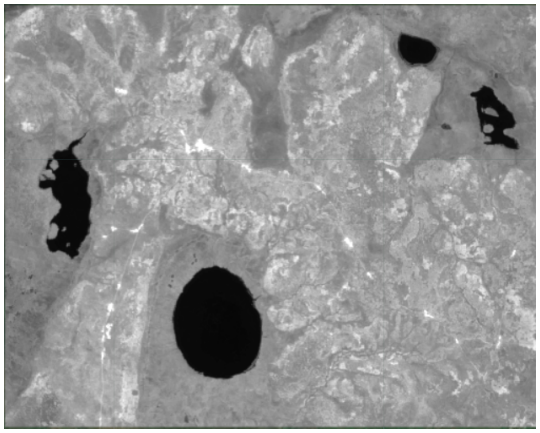


(a) Cloudy Image

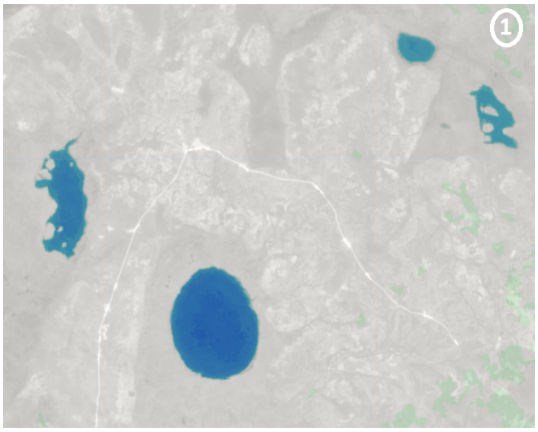


(b) Cloud-free Image

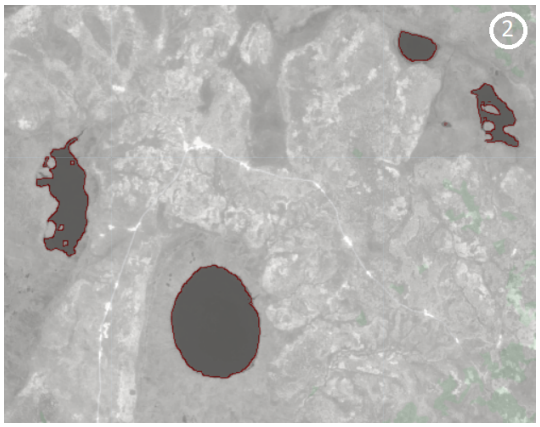
Figure 6.4: Cloudy and cloud-free image of sampled area



(a) Sentinel 2 Satellite Image

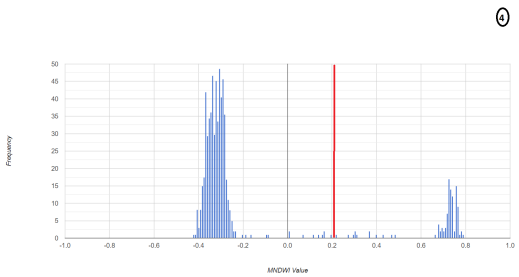


(b) MNDWI Calculation

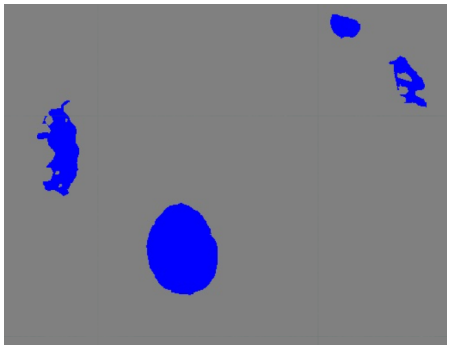


(c) Edge Detection

(d) Buffered Edges



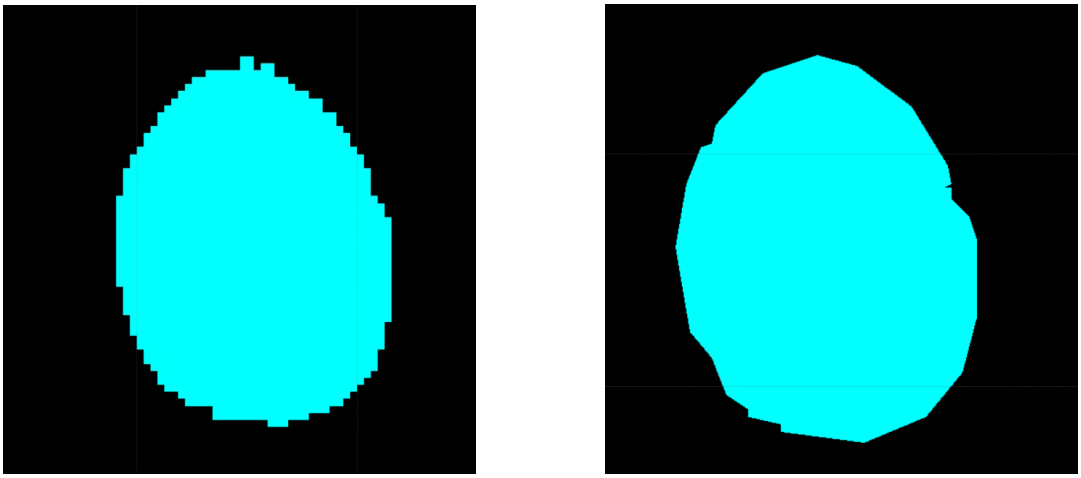
(e) Threshold computation with Otsu's dynamic method



(f) Final Water Mask

Figure 6.5: Sentinel 2 surface water dataset workflow

After acquiring the final water mask for the whole extend of Angola, a vectorization process was performed. To avoid bulky polygons as shown in Figure 6.6, the boundaries of the polygons were simplified according to Deuglas Peucker Line Algorithm (see Section 2.2.5). This was a necessary step also for exporting the final vector features of Sentinel 2, as without it, huge computational efforts were required and the algorithm ultimately failed.



(a) Water feature before simplification

(b) Water feature after simplification

Figure 6.6: Line Simplification

6.2.2 Global Surface Water

The Global Surface Water (GSW) dataset in raw format contains raster images of 30 m resolution where each pixel contains information about seasonal and year changes of the water extend. More specifically its values ([0,1]) show how often water appeared in each pixel throughout the years. To use this dataset in the main comparison analysis, a pre-processing step was performed (Figure 6.7). First of all, converting this raster data directly to vector polygons would result in acquiring bulky boundaries, because of their 30×30 m resolution. To avoid this and obtain a more detailed water mask and thereafter a vector, the GSW raster images were resampled with bicubic interpolation to 10 m resolution (see Section 2.2.3). This way more pixels were generated, although it is to be mentioned that the accuracy of the contained pixels remained at 30 m.

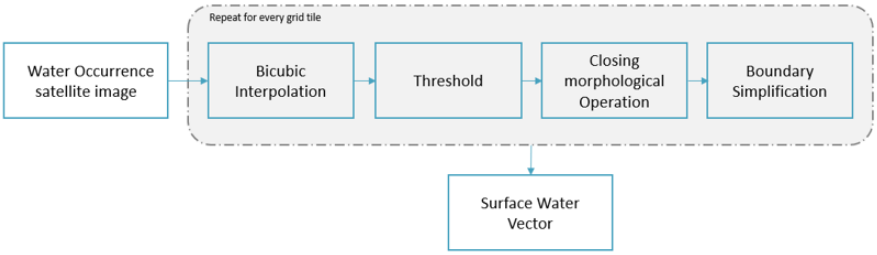


Figure 6.7: Global Surface Water pre-processing workflow

Another issue with the raw GSW data is that sometimes they contain gaps or no data values. To remove this gaps and delineate better the water masks, a closing morphological operation was performed, to eliminate holes while keeping the general shape of the original water mask (see Section 2.2.1). Finally, the water occurrence was thresholded to 0.1 (90%) occurrence to obtain the final watermask, which was subsequently vectorized and simplified. As the amount of processed data was enormous, and the creation of one vector water dataset for the entire Angola at once was not possible, a grid of $1km \times 1km$ was used to divide the study area in smaller regions. Thereafter, the generated vectors were initially exported per tile and then merged into one file that contained the GSW water vectors of the entire Angola.

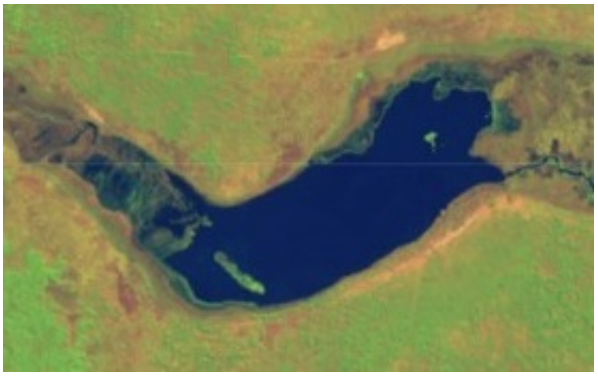
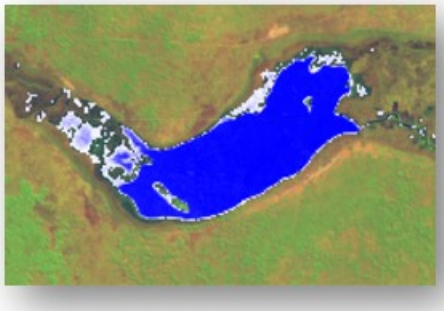
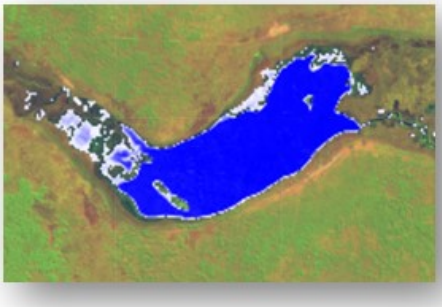


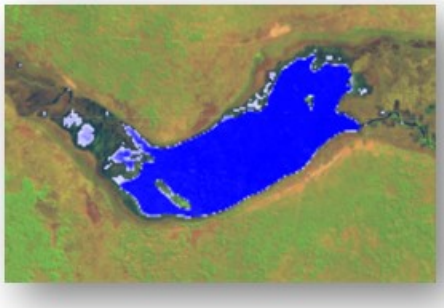
Figure 6.8: Satellite Image



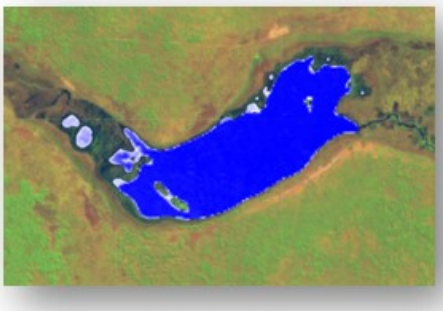
(a) Water Occurrence



(b) Bicubic Interpolation



(c) Thresholding of water mask



(d) Morphological closing operation

(e) Vectorized surface water mask

(f) GSW feature after simplification

Figure 6.9: GSW surface water dataset workflow

6.2.3 OpenStreetMap

The OSM data were initially filtered to make sure they would contain information only relative to water reservoirs and lakes. Nevertheless, the initial filtering was not completely successful, due to the fact the OSM features might have multiple tags. For example the tag "natural = water" was contained also in data that were not water areas. More specifically, island features, rivers, land covered with shrub and bushes or stunted trees amongst others, were falsely included in the OSM reservoir dataset. Therefore, throughout this research the OSM data were multiple times refined whenever a false entry was located.

The data in the OSM database are represented as Simple Features (SF), meaning they have standardized representation each simple feature is a sequence of spatial coordinates. For OSM specifically only XY coordinates are stored as strings. According to SF there are 7 standardized classes [15]:

1. POINT
2. POINT
3. MULTIPOINT
4. LINESTRING
5. MULTILINESTRING
6. POLYGON
7. MULTIPOLYGON
8. GEOMETRYCOLLECTION

Even though only the Polygon and Multipolygon geometries were chosen during the initial filtering, the OSM data still GeometryCollections which include several geometries in one. During this research this SF class created problems in the implementation of the methodology. Therefore all GeometryCollections had to be broken down in single geometries and then only the Polygons were choosen for further processing. Same operation was performed also in the case of Multipolygons, which were converted into single Polygons.

6.3 COMPARISON OF ACCURACCY OF DATASETS

The assesment of the water area completeness and the positional accuracy computation of the datasets, was performed in pairwise comparison. More specifically, all input datasets (GSW, Sentinel 2, HydroLAKES, GRaND) were compared against the OSM dataset. The reason for selecting the OSM as part of each individual comparison analysis, is the fact that it contains both large scale information, highly detailed geometrical primitives and features that refer specifically to water reservoirs. Nonetheless, it is not considered of higher accuracy compared to the other datasets, since it has its own limitations.

The completeness analysis was performed relatively to the water area, and indicates apart from the common area between datasets and total water area of each, also the lack or overestimation of water in the datasets. The percentage of overlap indicates the distance between two overlapping features, i.e. how much one feature is spatially shifted relatively to the other. Lastly, the Hausdorff distance measures the degree of mismatch between two corresponding features from different datasets.

15 https://cran.r-project.org/web/packages/osmdata/vignettes/osm-sf-translation.html

6.3.1 Completeness

To calculate the completeness, the OSM and the rest of the input datasets were clipped into the 40×40 km grid squares. Within each cell, the total water area of each dataset was computed but also the difference of total covered area, indicating the places where the two compared datasets do not agree. The completeness quality metric is assessing the total common area of OSM with each of the other datasets. However, as the Sentinel 2 and GSW datasets contain also information about rivers, their total water area was overestimated. The thematic differences presented in Figure 6.10, Figure 6.12, Figure 6.14 and Figure 6.16, indicate the cell regions in which each dataset detected water.

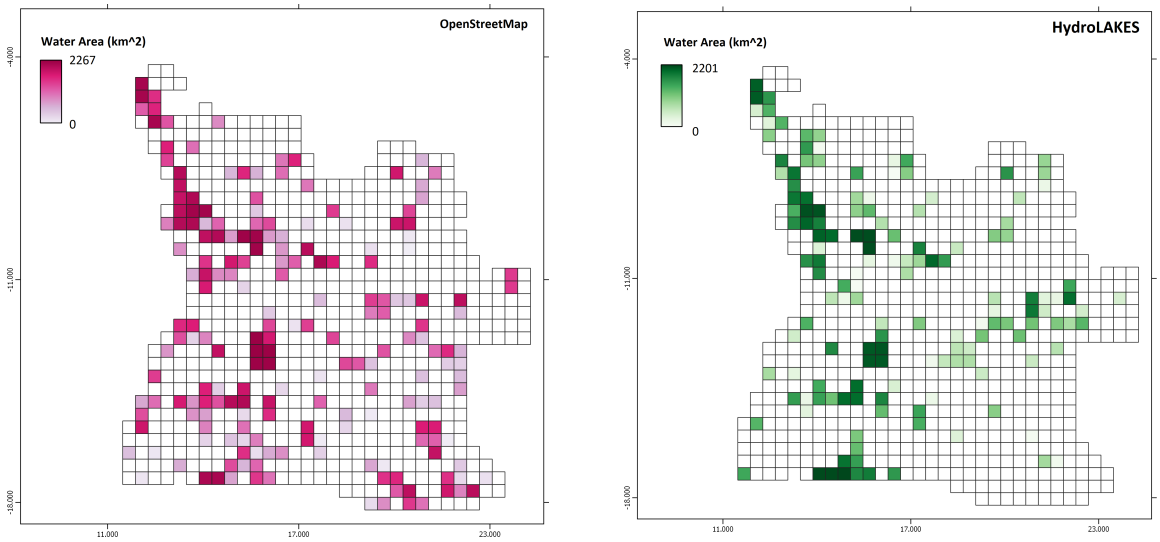


Figure 6.10: Thematic differences between water surface of OSM and HydroLAKES

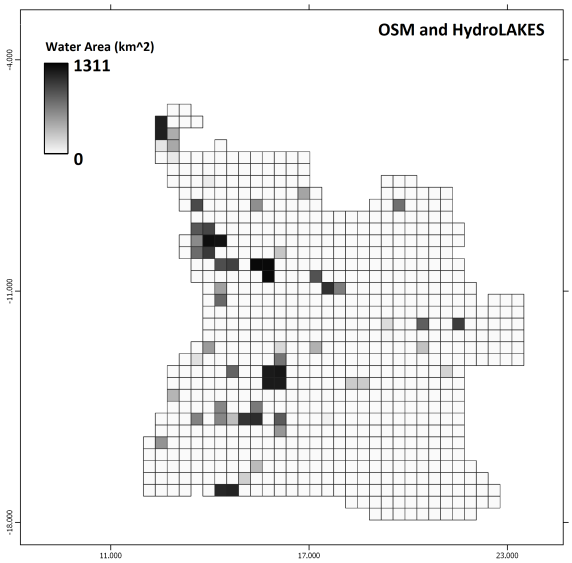


Figure 6.11: Area of intersection of OSM and Hydrolakes

Dataset	Area (km^2)	Ratio(%)
Total Water	5437	100 %
HydroLAKES	2201	41%
OSM	2267	42%
OSM and HydroLAKES	1311	24%
OSM,no HydroLAKES	956	18%
HydroLAKES, no OSM	3171	58%

Table 6.1: Overlap between surface water of OSM and HydroLAKES

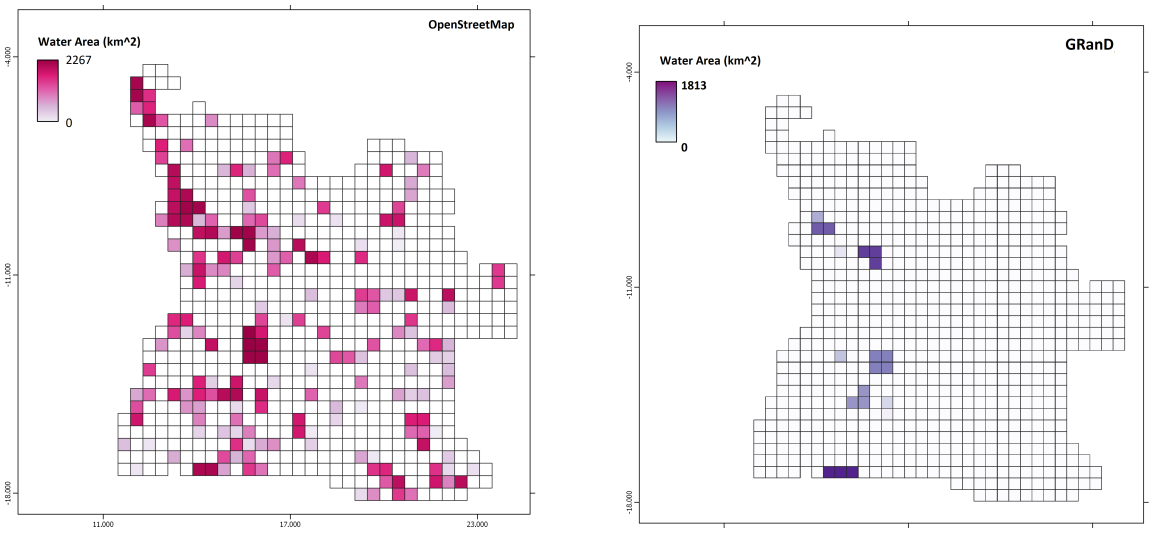


Figure 6.12: Thematic differences between water surface of OSM and GRaND

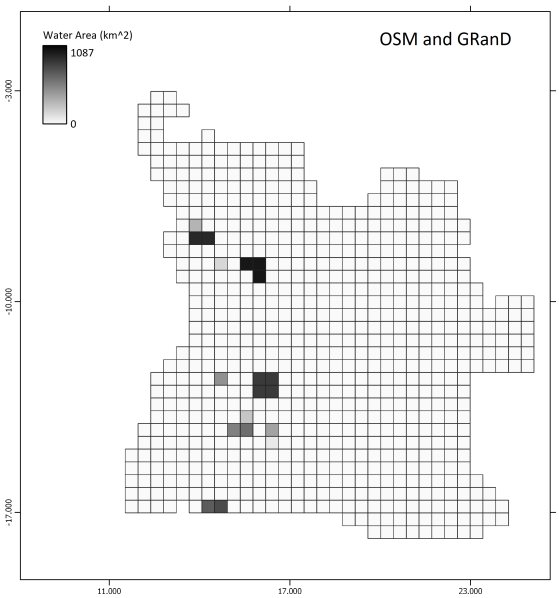


Figure 6.13: Area of intersection of OSM and GRaND

Dataset	Area (km^2)	Ratio (%)
Total Water	5030	100 %
OSM	2267	45%
GRanD	1813	36%
OSM and GRanD	1087	22%
OSM, no GRanD	1180	24%
GRaND, no OSM	2763	54%

Table 6.2: Overlap between surface water of OSM and GRanD

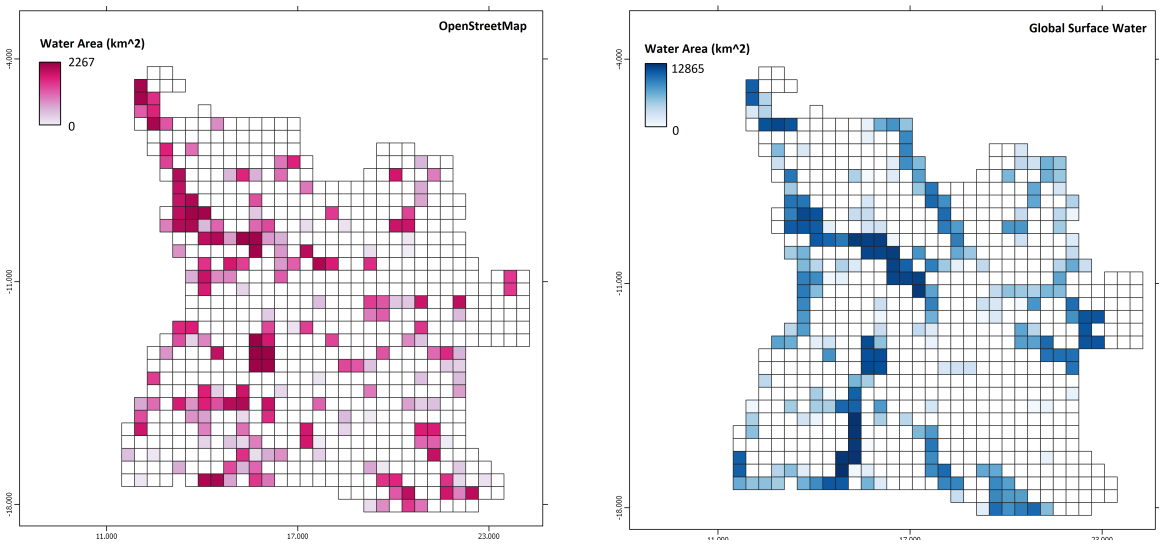


Figure 6.14: Thematic differences between water surface of OSM and GSW

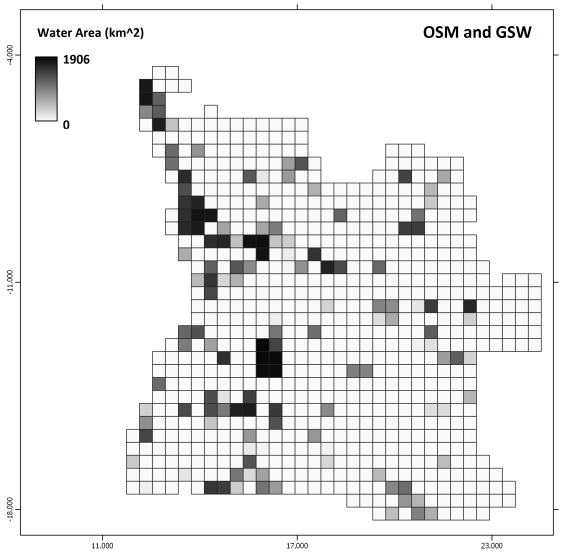


Figure 6.15: Area of intersection of OSM and GSW

Dataset	Area (km^2)	Ration(%)
Total Water	21759	100 %
OSM	2267	10%
GSW	12865	59%
OSM and GSW	1906	9%
OSM, no GSW	8894	41%
GSW, no OSM	10959	50%

Table 6.3: Overlap between surface water of OSM and GSW

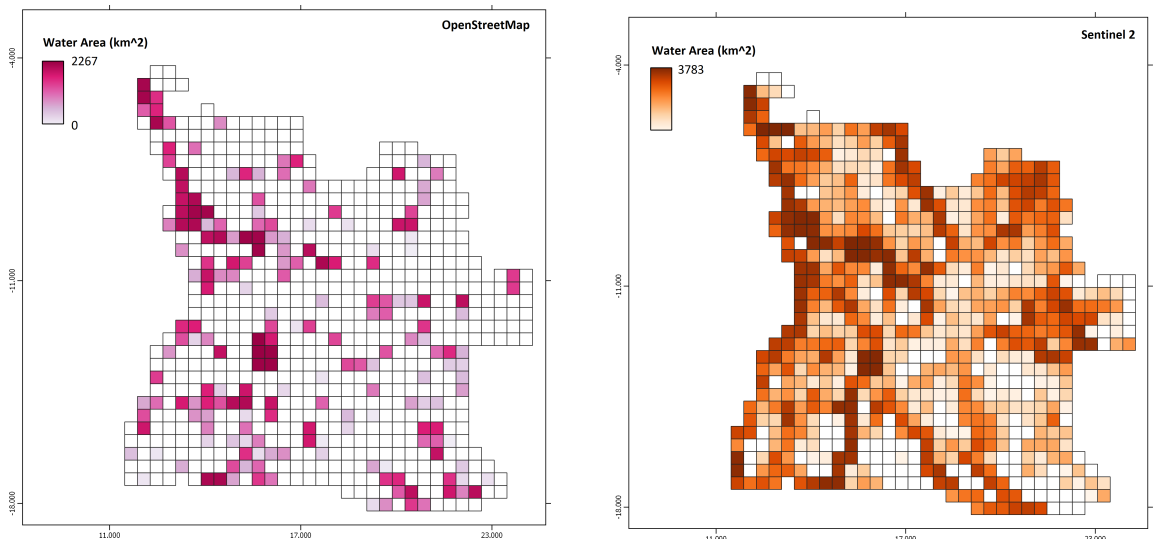


Figure 6.16: Thematic differences between water surface of OSM and Sentinel 2

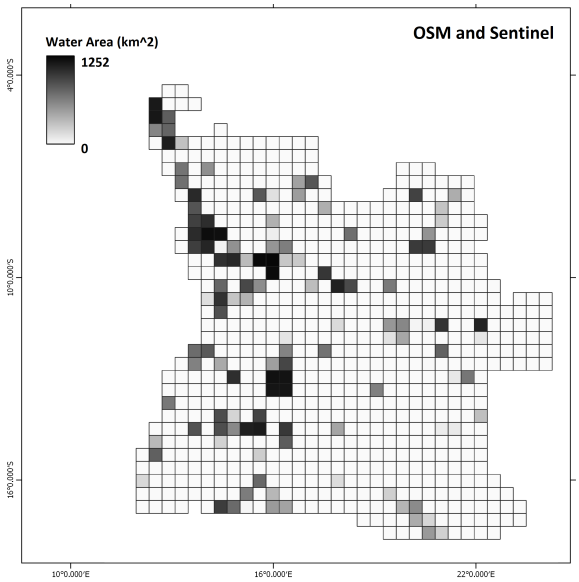


Figure 6.17: Area of intersection of OSM and Sentinel 2

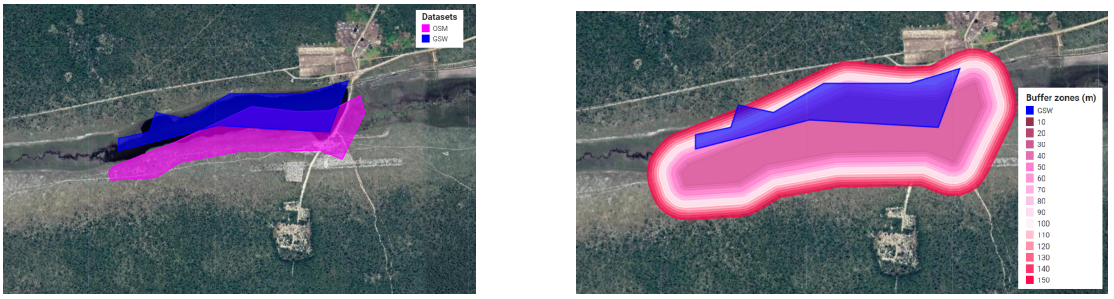
Dataset	Area (km^2)	Ration(%)
Total Water	6933	100%
OSM	2267	33%
Sentinel	3783	55%
OSM and Sentinel	1252	19%
OSM,no Sentinel	3151	45%
Sentinel,no OSM	2530	36%

Table 6.4: Overlap between surface water of OSM and Sentinel 2

6.4 POSITIONAL ACCURACCY

6.4.1 Percentage of overlap

The percentage of overlap was computed for 15 different buffer zones from 0 to 150 m, created around the OSM features (see Figure 6.18). Afterwards, to determine the distance between each pair of features, the distribution of the percentages was examined (see Figure 6.19). The peak of the histogram is identified as the distance that quantifies how far away one feature is from the other. This conclusion was based on the fact that as the buffer size increases, the compared features have higher overlap until a certain buffer zone, after which the overlap reduces, i.e. the features are spatially further apart from each other (e.g Figure 6.18).



(a) OSM and GSW features

(b) Buffered OSM and GSW

Figure 6.18: Example of Percentage of Overlap

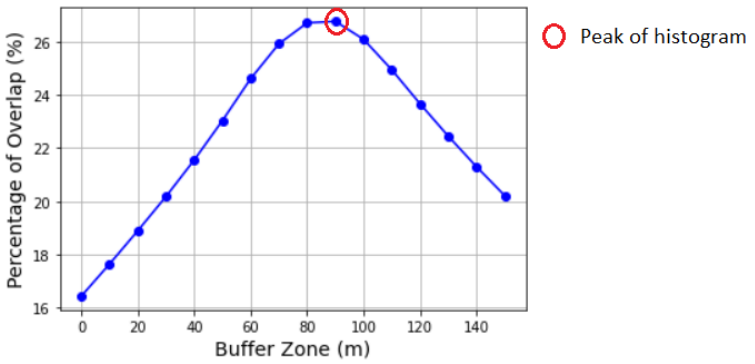
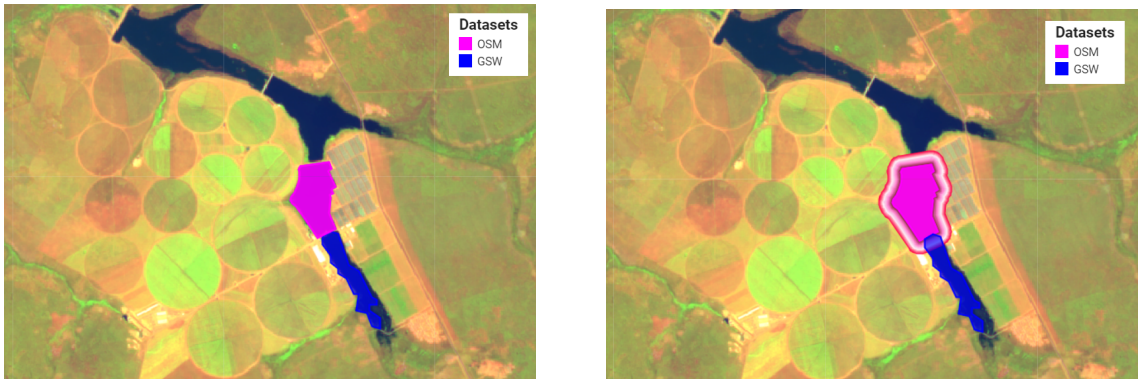


Figure 6.19: Peak of overlap percentage and and estimated positional difference in meters

The number of tested buffer zones used in this method were limited, because an infinite number of buffer zones would not be supported by GEE, due to memory limitations. Therefore, the 150 buffer zone implies a 150 m at the minimum, distance between two features (see Figure 6.20).



(a) OSM and GSW features **(b)** Buffered OSM and GSW

Figure 6.20: Example of Percentage of Overlap

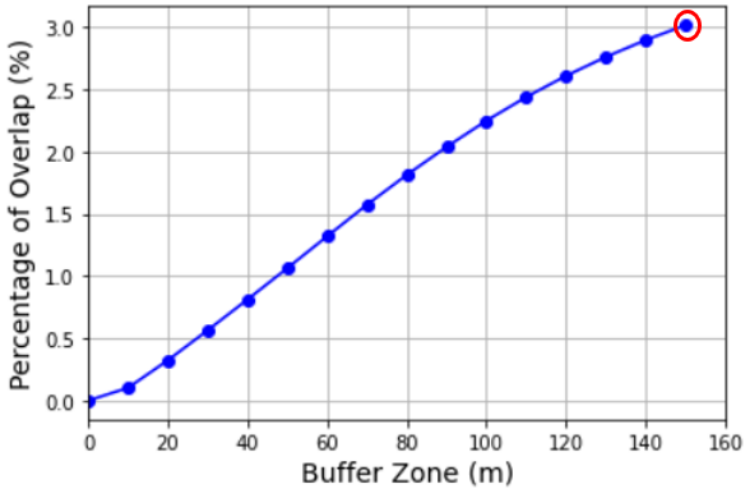
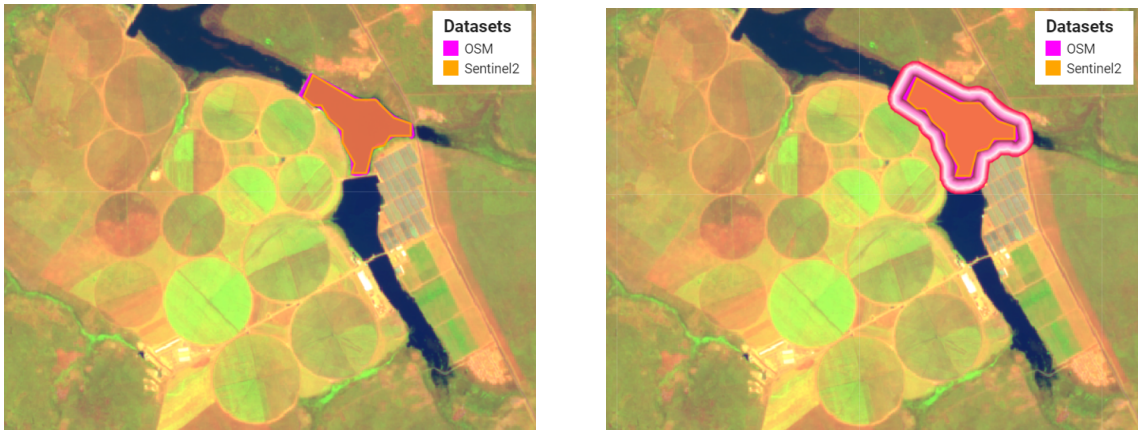


Figure 6.21: Peak of overlap percentage and and estimated positional difference in meters

In Figure 6.22 the geometries of the Sentinel 2 and OSM features are very similar. Therefore, as expected, the distance between the two features corresponds to the 0 m buffer zone.



(a) OSM and SEntinel 2 features **(b)** Buffered OSM and Sentinel 2

Figure 6.22: Example of Percentage of Overlap

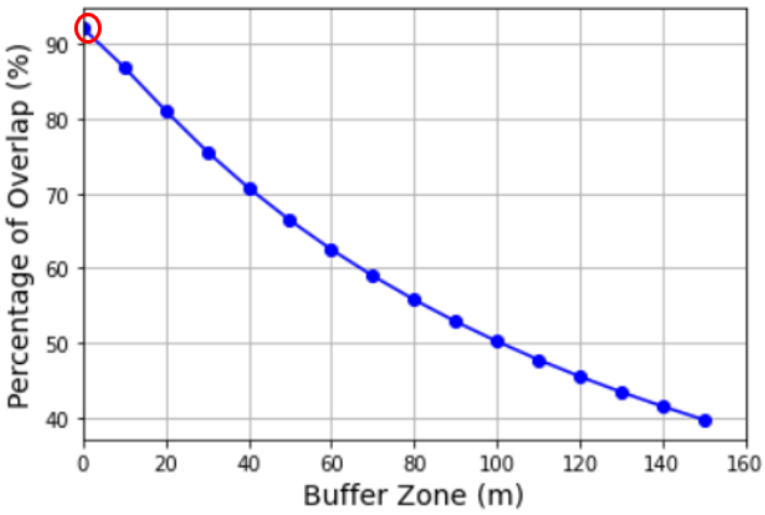


Figure 6.23: Peak of overlap percentage and and estimated positional difference in meters

The positional accuracy results in terms of percentage of overlap for the entire country of Angola, are presented in the following figures.

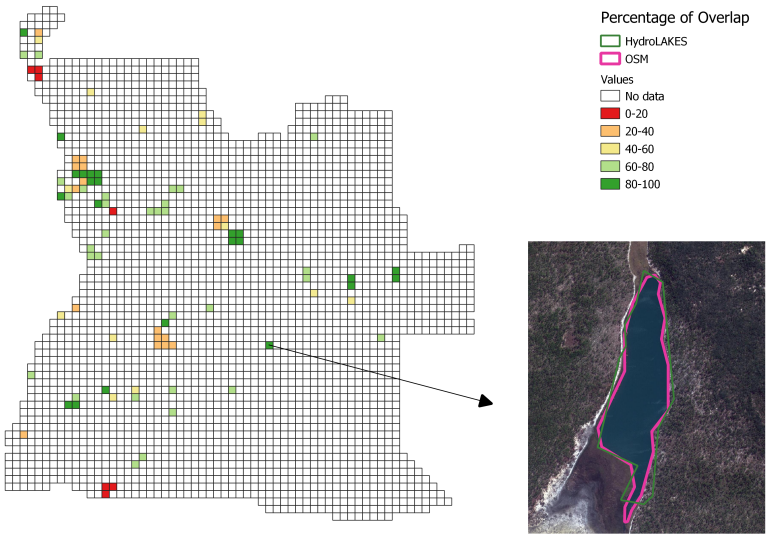


Figure 6.24: Percentage of overlap between OSM and HydroLAKES

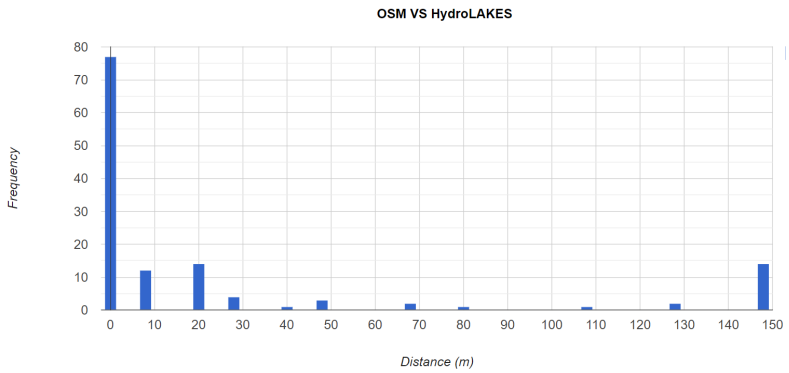


Figure 6.25: Distances between OSM and HydroLAKES

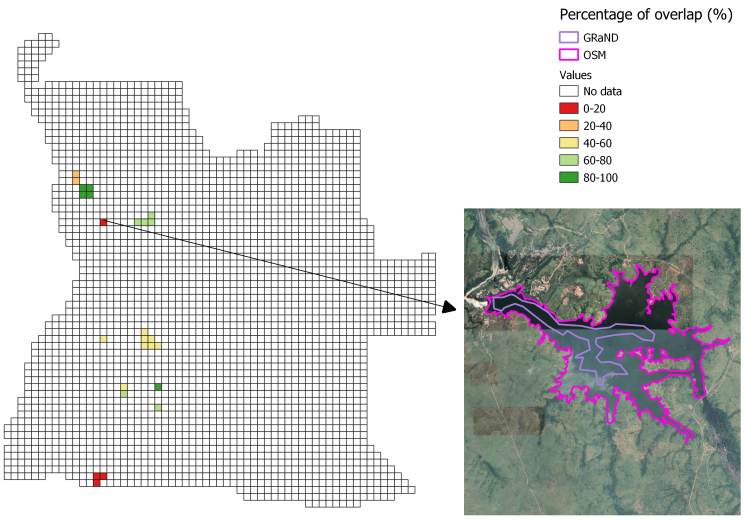


Figure 6.26: Percentage of overlap between OSM and GRanD

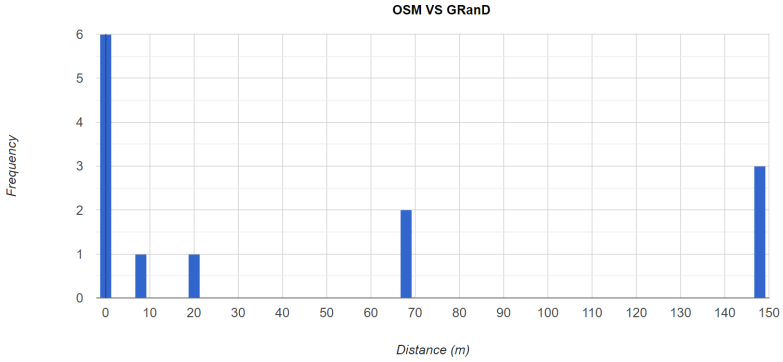


Figure 6.27: Distances between OSM and GRanD

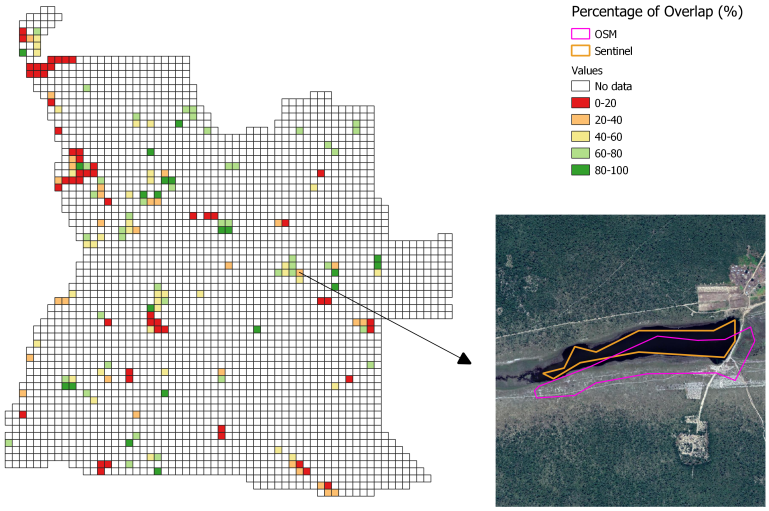


Figure 6.28: Percentage of overlap between OSM and Sentinel 2

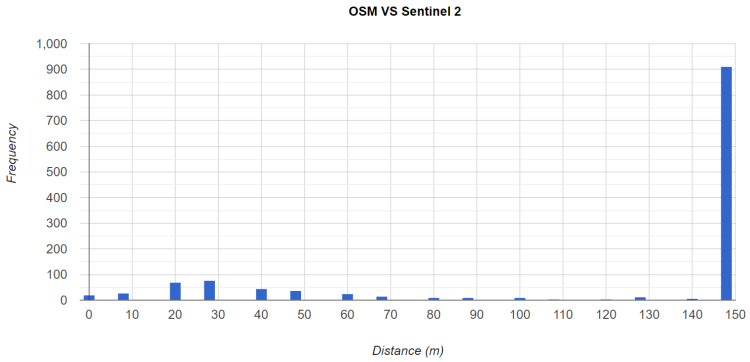


Figure 6.29: Distances between OSM and Sentinel

The Goodchild's quality metric in the case of Sentinel 2 versus OSM, shows a high concentration of features with a distance of 150 m (Figure 6.29). This is expected, as one peculiarity of the generated Sentinel 2 water dataset is that, in many cases it contains clusters of smaller water features that actually represent one bigger feature (see Figure 6.30). Although Sentinel 2 is considered a valuable source of information for this and future research, the creation of an optimal Sentinel 2 dataset was considered scope of this thesis as it was not feasible within the given available time frame.

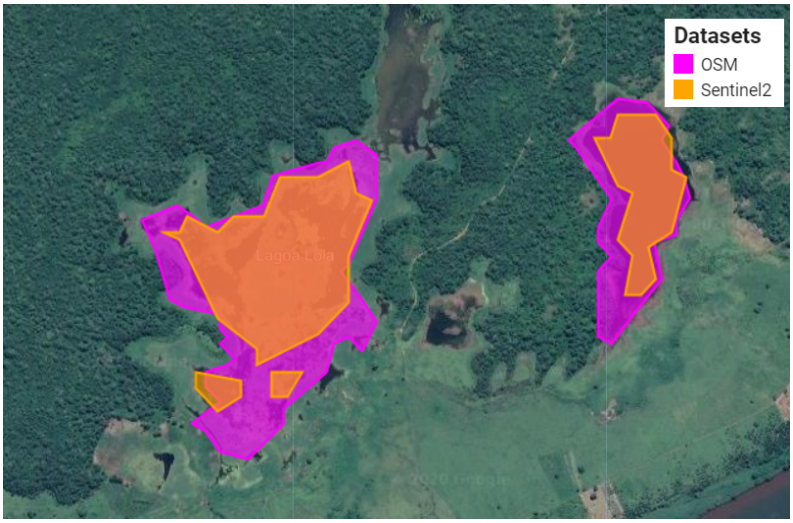


Figure 6.30: Example of cluster of Sentinel 2 water features

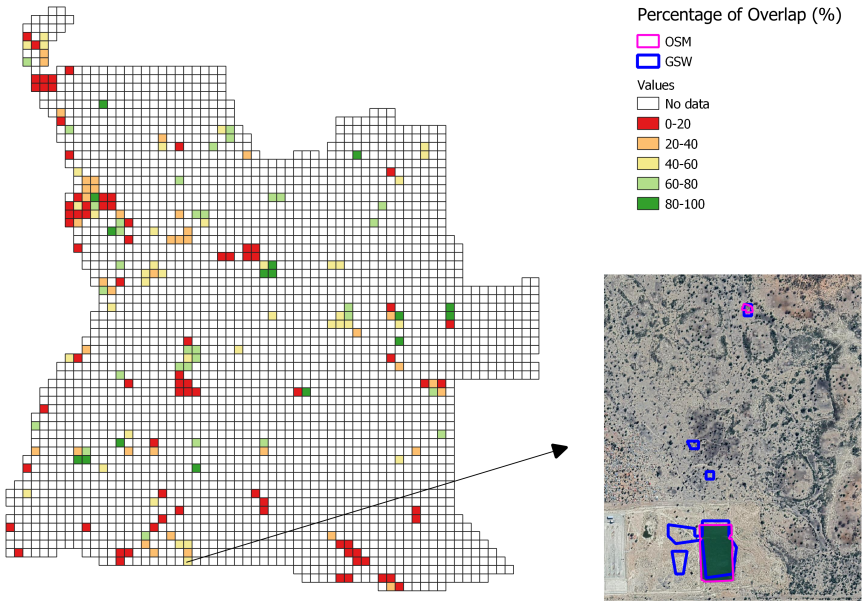


Figure 6.31: Percentage of overlap between OSM and GSW

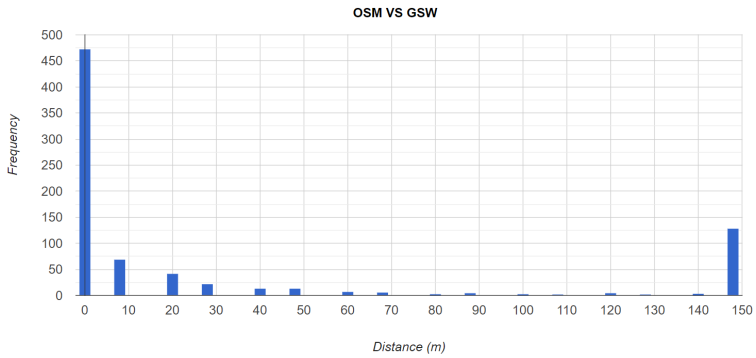


Figure 6.32: Distances between OSM and GSW

6.4.2 Euclidean Distance

The Euclidean Distance as defined in Section 4.2.3, is the minimum distance between two points in Euclidean 2D space. Each polygon feature consists of a set of point primitives. Therefore, to estimate how far away the OSM polygons are to the corresponding features from the other datasets, the euclidean distance per point was computed. However, the amount of points and the distance between them, varies amongst datasets and polygons. To avoid any potential bias due to the lack of equidistance of the point primitives, new points were sampled along the boundaries of the polygons to increase the point density with a stepsize of 10 m (Figure 6.34). More details about the selection of the sampling step size are given in Chapter 7.

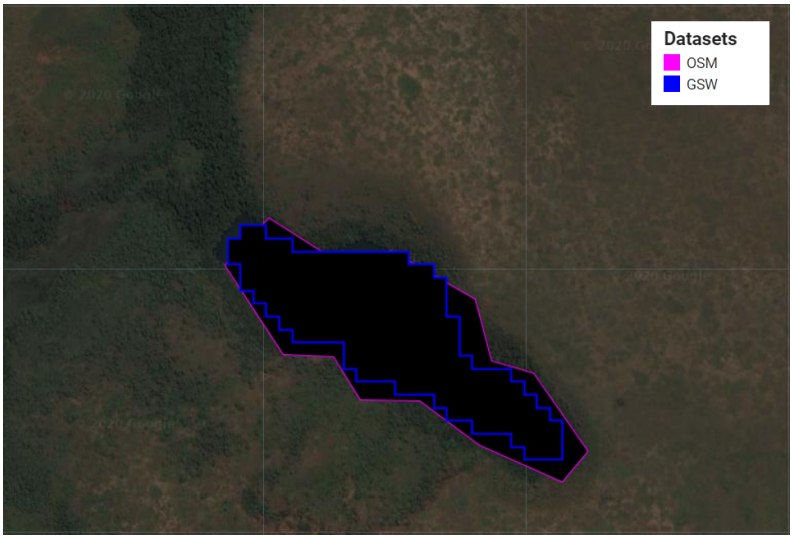
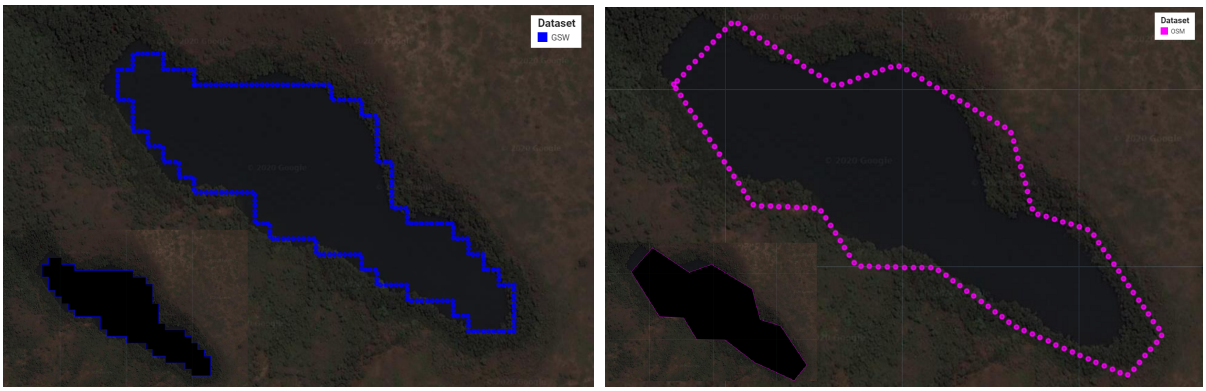


Figure 6.33: GSW and OSM water features

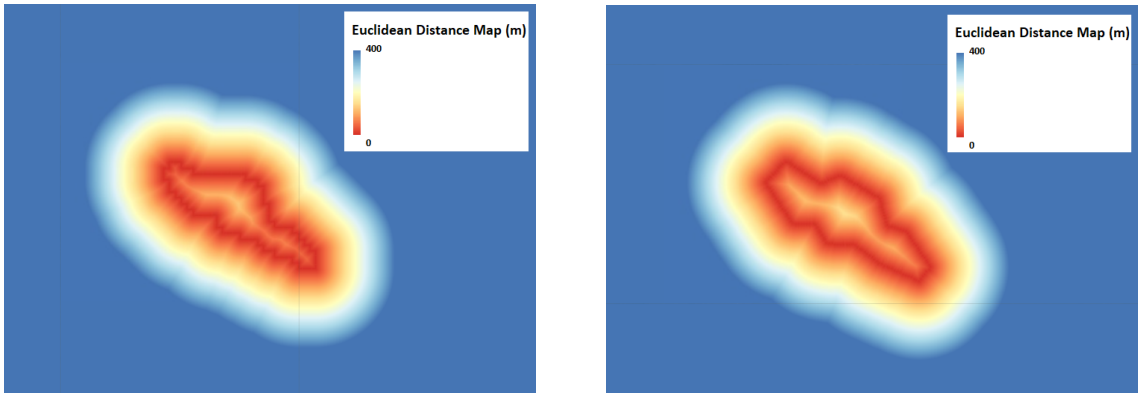


(a) Sampling of GSW points

(b) Sampling of OSM points

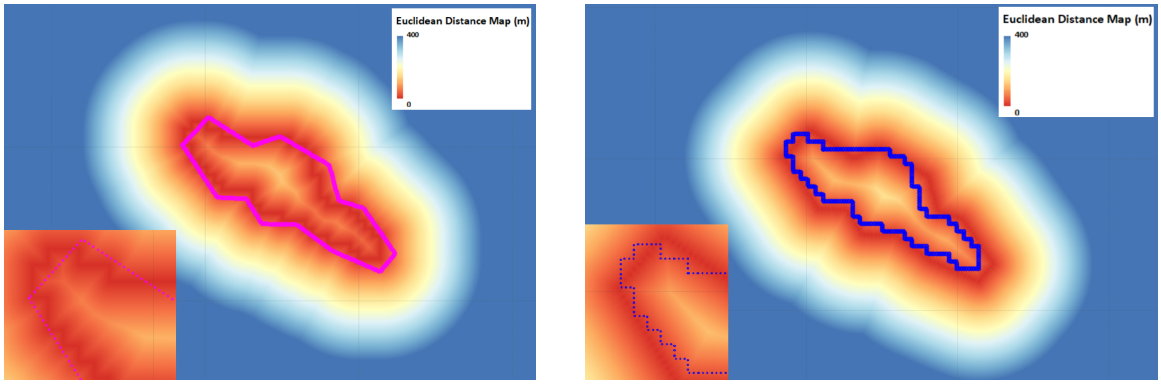
Figure 6.34: Example of point sampling

Next, a EDM was obtained based on the new generated boundary points. The fact that more densified points were acquired along the polygon boundaries, resulted in the construction of a more detailed and accurate EDM. Each pixel of this raster distance map contains the distance value to the closest boundary point (Figure 6.35). To acquire the values of euclidean distance for the points of a polygon primitive, each polygon was overlapped with the EDM of the other dataset. The location of the overlap between points and EDM, are the places where the euclidean distance values for each feature were collected (see Figure 6.36). From this the average euclidean distances between the boundaries of the various features were computed (see Figure 6.38, Figure 6.37 and Figure 6.39).



(a) Euclidean Distance Map of GSW **(b)** Euclidean Distance Map of OSM

Figure 6.35: Euclidean Distance Maps



(a) Overlap of OSM points with GSW Euclidean Distance Map **(b)** Overlap of GSW points with OSM Euclidean Distance Map

Figure 6.36: Euclidean Distances Calculation

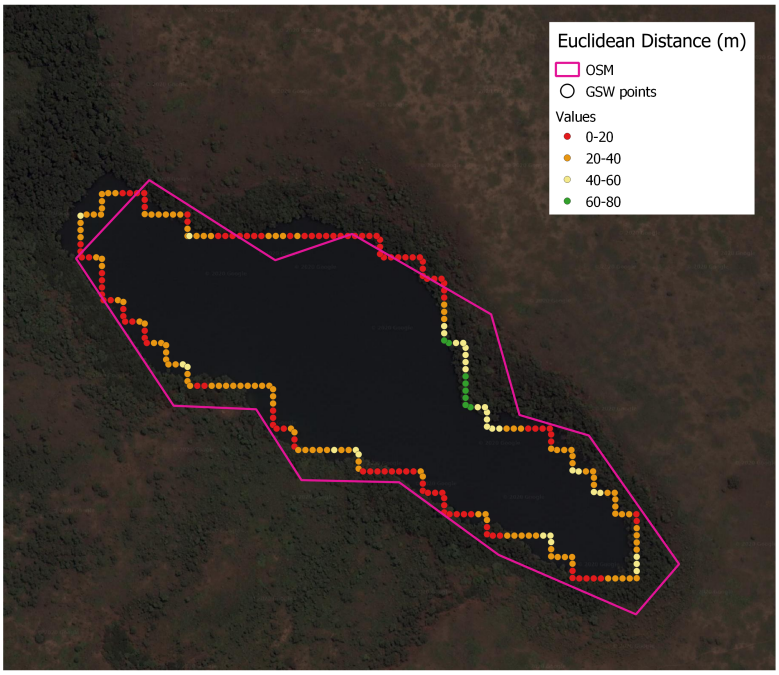


Figure 6.37: OSM Point Distance Values

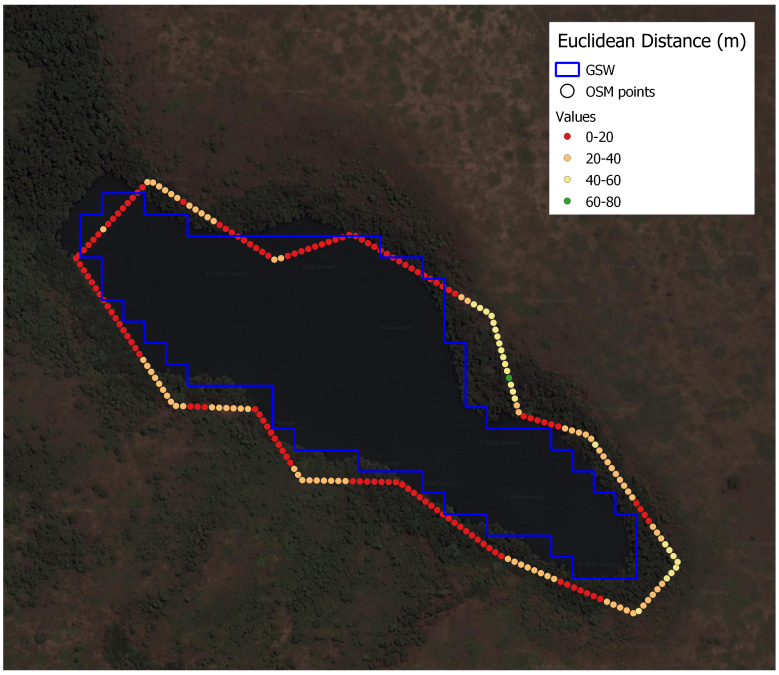
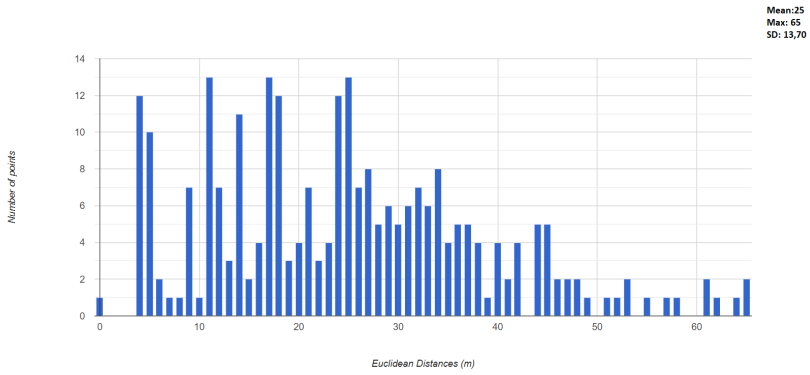
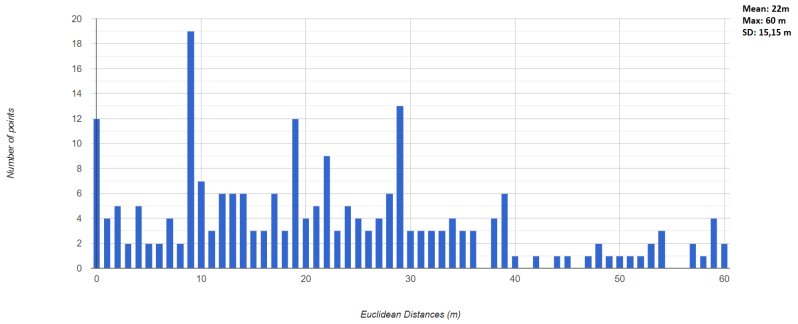


Figure 6.38: GSW Point Distance Values



(a) Euclidean Distance Values GSW to OSM



(b) Euclidean Distance Values OSM to GSW

Figure 6.39: Euclidean Distance Values

6.4.3 Hausdorff Distance

The Hausdorff distance is a measure of the shape similarity between polygons. It shows the maximum deviation between the boundaries of two polygons. However, it suited for detecting significant dissimilarities and not for giving an estimation

about the average similarity between them. Therefore, even if most of the euclidean distance values of a polygon primitive are below a certain value, the Hausdorff distance will converge to the value of the highest euclidean distance amongst all distances.

For this research, as described in Section 4.2.4, to compute the Hausdorff distance between two features , the sum of the mean Euclidean values of both EDM was used. The results for all overlapping features between OSM and the other input datasets over the extend of the entire Angola, are presented in the Figures and Tables below.

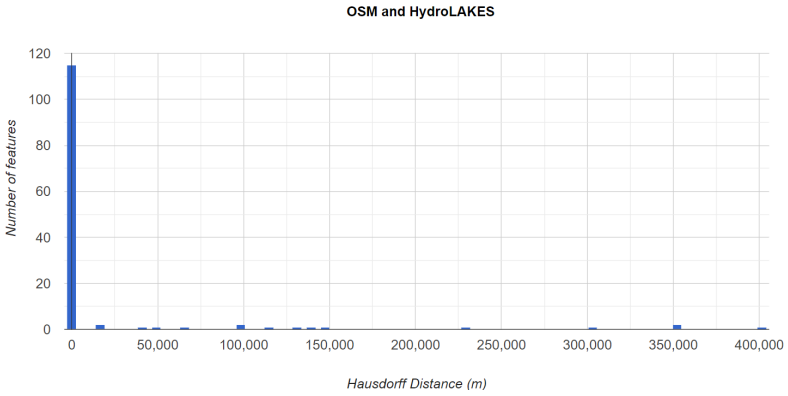


Figure 6.40: Hausdorff Distance Values for OSM and HydroLAKES

Statistical Value	Distance (m)
Mean	20017.22
Median	206.19
Standard Deviation	68701.56

Table 6.5: Hausdorff Distance Statistics for OSM and HydroLAKES

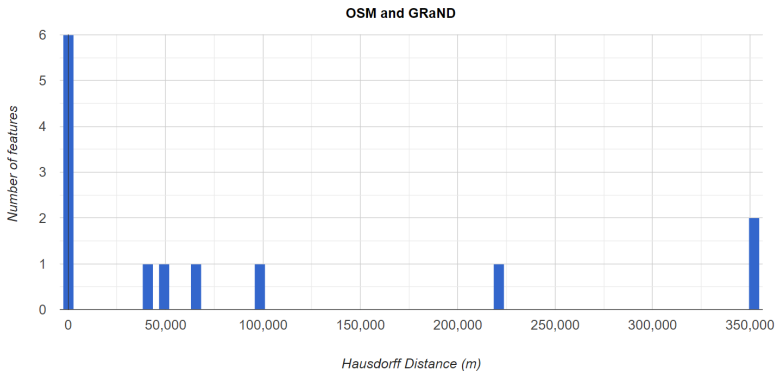


Figure 6.41: Hausdorff Distance Values for OSM and GRaND

Statistical Value	Distance (m)
Mean	93129.97
Median	41761.69
Standard Deviation	128483.62

Table 6.6: Hausdorff Distance Statisticsfor OSM and GRaND

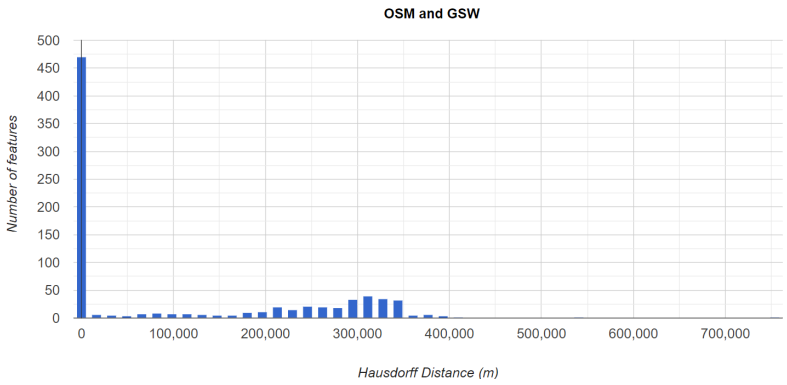


Figure 6.42: Hausdorff Distance Values for OSM and GSW

Statistical Value	Distance (m)
Mean	109133.59
Median	236.07
Standard Deviation	143024.91

Table 6.7: Hausdorff Distance Statistics OSM and GSW

Due to some technical issues, a subset of the OSM and Sentinel 2 comparison analysis was exported. From the total of 1228 common features between the two datasets, only the 600 could be exported and visualized, as the processing time for the entire dataset was extremely high without giving any specific notification for an error (Figure 6.43). This possibly has to do with the of utilization of point resampling, when dealing with relatively high number number of vertices, in the case of very big geometrical primitives or high number of features. In particular, there is a chance that the number of sampling points exceeded the computational capacity of GEE. Due to time limitations of this research, this issue remains unsolved for now. However, some optimization will be necessary in future research to increase the scalability of the Hausdorff Distance algorithm when dealing with bigger datasets.

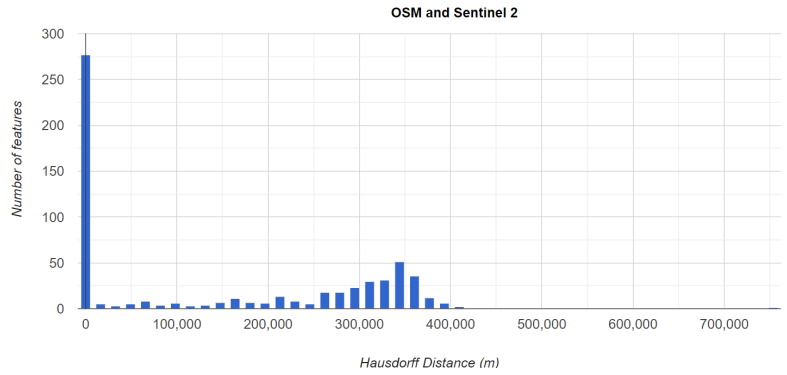


Figure 6.43: Hausdorff Distance Values for OSM and Sentinel 2

Statistical Value	Distance (m)
Mean	152535.25
Median	92301.64
Standard Deviation	158434.63

Table 6.8: Hausdorff Distance Statistics for OSM and Sentinel 2

7 SENSITIVITY ANALYSIS

In this chapter a sensitivity analysis is performed to explore the influence of a certain variable of the algorithm to the estimated Hausdorff Distance. The sensitivity is analyzed based on sampled features.

As mentioned in Section 6.4.3, the number of points that form the water polygons and the distance between them varies amongst features and input datasets. The number of vertices used to create the EDM of the water polygons, affect the level of detail and quality of the EDM. More specifically the distance between those boundary points is very important for narrow geometries, where a high number of vertices is necessary to compute an accurate EDM and therefore an accurate Hausdorff Distance.

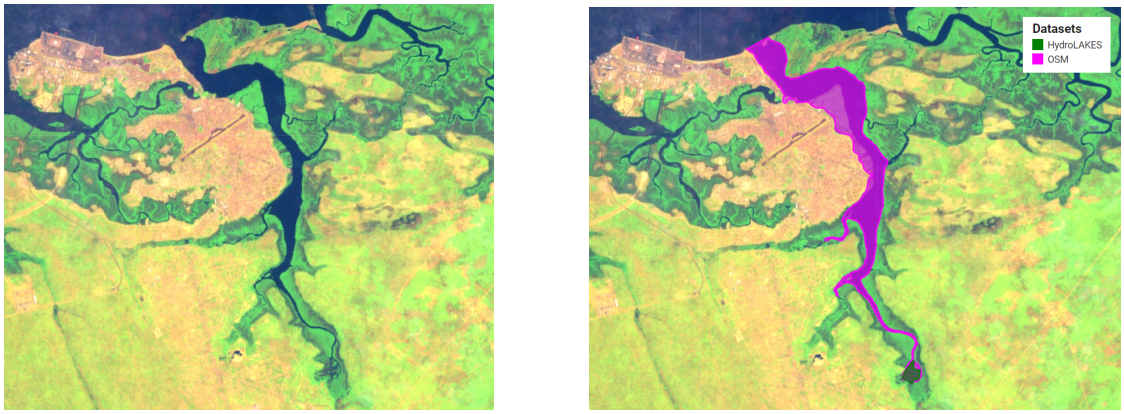
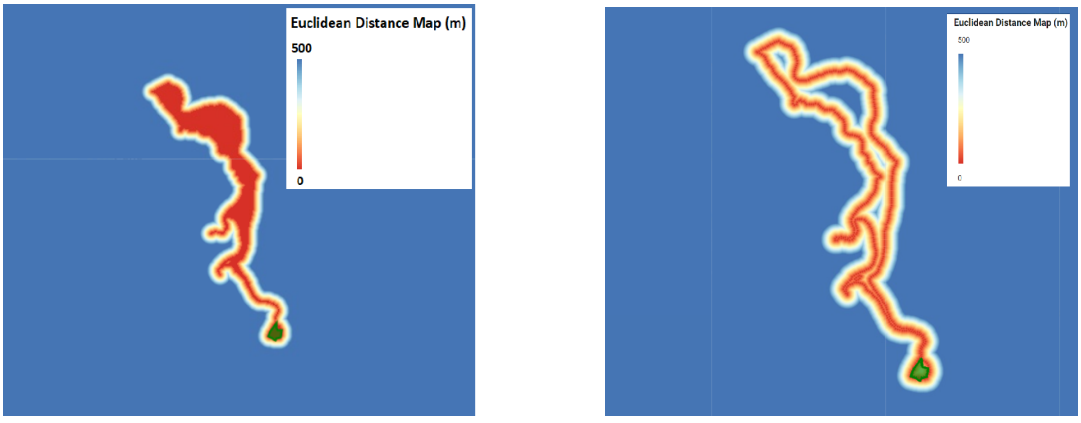


Figure 7.1: Example of OSM and HydroLAKES features

The differences in the estimated EDM for the features presented in Figure 7.1 can be seen in Figure 7.2. The red zone in the EDM of the OSM feature represents the area with zero distance from the boundary. With the exclusion of the resampling along the water polygons, this zone has a wide extend causing all the boundary points of the HydroLAKES feature that are located inside of the OSM polygon to be falsely estimated with a zero Euclidean Distance (see Figure 7.3).



(a) EDM without resampling of vertices

(b) EDM with resampling of vertices

Figure 7.2: The impact of the number of vertices towards the quality of the EDM

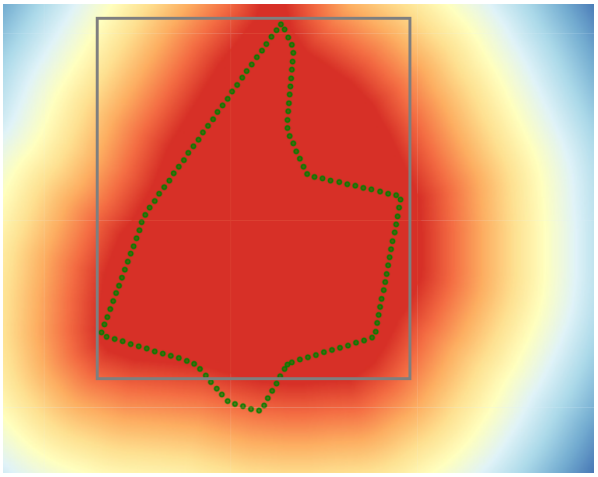


Figure 7.3: Vertices with zero Euclidean Distance values

The impact of this less accurate EDM on the computed Euclidean Distance can be seen in Table 7.1. Considering the fact that the Hausdorff Distance is estimated based on the sum of the mean Euclidean Distances from HydroLAKES to OSM and OSM to HydroLAKES (Table 7.2) as explained in Section 4.2.4, we observe that this false estimation of the mean Euclidean distance of 3.57 m instead of 47.04 m will affect the result the Hausdorff Distance between the two features. The mean Euclidean Distances with and without point sampling are 360.06 km and 360.02 km respectively, resulting in a difference in the estimation of the Hausdorff Distance up to 40 m.

Statistical Unit	Euclidean Distance (m)	
	Without Sampling	With Sampling
Max	50.67	152.64
Mean	3.57	47.04
Median	0	39.80
Mode	0	43.65
Standard Deviation	11.55	31.80

Table 7.1: Euclidean Distances from HydroLAKES to OSM

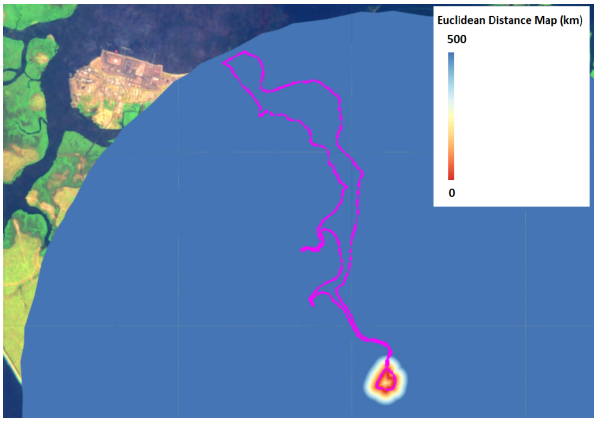
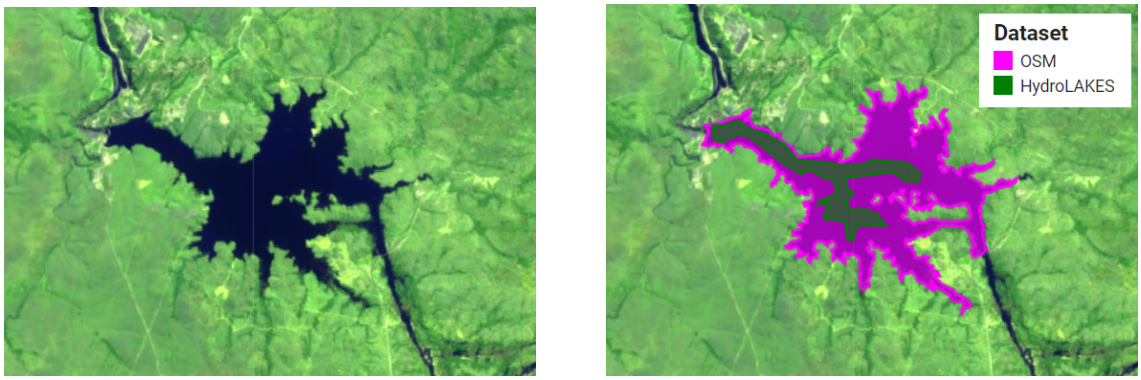


Figure 7.4: Euclidean Distances between OSM points to HydroLAKES features

Statistical Unit	Euclidean Distance (km)
Max	460.57
Mean	306.02
Median	298.70
Mode	406.33
Standard Deviation	216.93

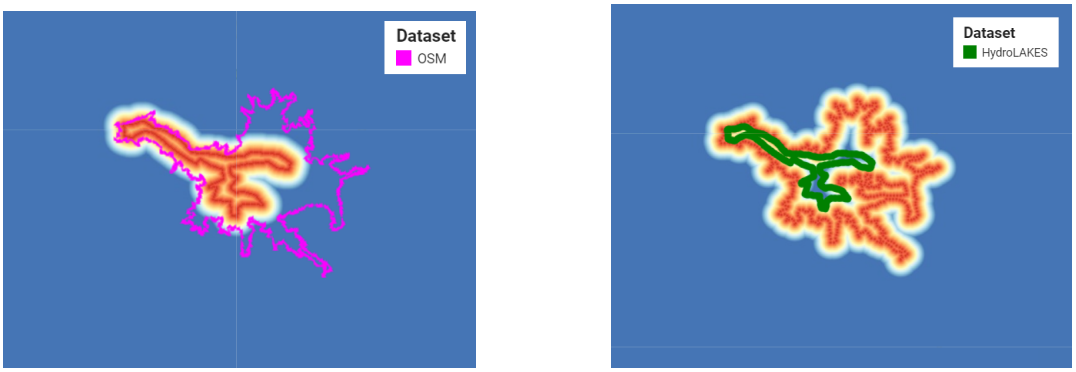
Table 7.2: Euclidean Distances from OSM to HydroLAKES

The implementation of the Hausdorff Distance algorithm was performed with a resampling of the vertices of the water polygons with a step size of 10 m . This decision was based on two factors: to choose a point density that is enough to create a detailed EDM with as few points as possible, so that the necessary computational resources are operational from the GEE platform. To investigate the sensitivity of this particular variable, two sample features were selected from the OSM and HydroLAKES input datasets. The estimated Euclidean and thereafter Hausdorff Distances were tested for ten different step sizes (3, 5, 8, 10, 20, 50, 100, 200, 500, 1000). The results demonstrated in Table 7.4 and Figure 7.7, show that the values of the Hausdorff Distance for less than 20 m step size converge, but at the same time the amount of sampled points increases largely. To avoid an enormous processing time or failure of the algorithm, the 10 m step size was chosen as the most suitable for being both computationally less intensive and frequent enough to provide accurate results for the Hausdorff Distance computations.



(a) Satellite Image **(b)** OSM and HydroLAKES features

Figure 7.5: Example of OSM and HydroLAKES features



(a) Euclidean Distances of OSM points to HydroLAKES **(b)** Euclidean Distances of HydroLAKES points to OSM

Figure 7.6: Euclidean Distances between OSM and HydroLAKES features

Step Sizes (m)	Mean ED (Forward)	Mean ED (Backward)
3	663.95	212.72
5	663.94	212.69
8	663.86	212.68
10	663.88	212.65
20	663.90	212.65
50	663.73	212.38
100	663.17	211.67
200	661.30	211.47
500	655.64	210.60
1000	657.46	226.05

Table 7.3: Mean Euclidean Distances (ED) between OSM and HydroLAKES features for various stepsizes

Step Sizes (m)	Amount of Points	Hausdorff Distance (m)	Relative Difference [%]
3	14.357	876.67	-
5	8615	876.63	- 0.004
8	5385	876.54	- 0.010
10	4308	876.53	- 0.001
20	2154	876.55	+ 0.002
50	862	876.11	- 0.050
100	432	874.84	+ 0.145
200	217	872.77	- 0.237
500	88	866.20	- 0.753
1000	45	883.51	+ 1.998

Table 7.4: Hausdorff Distance between OSM and HydroLAKES features for various stepsizes

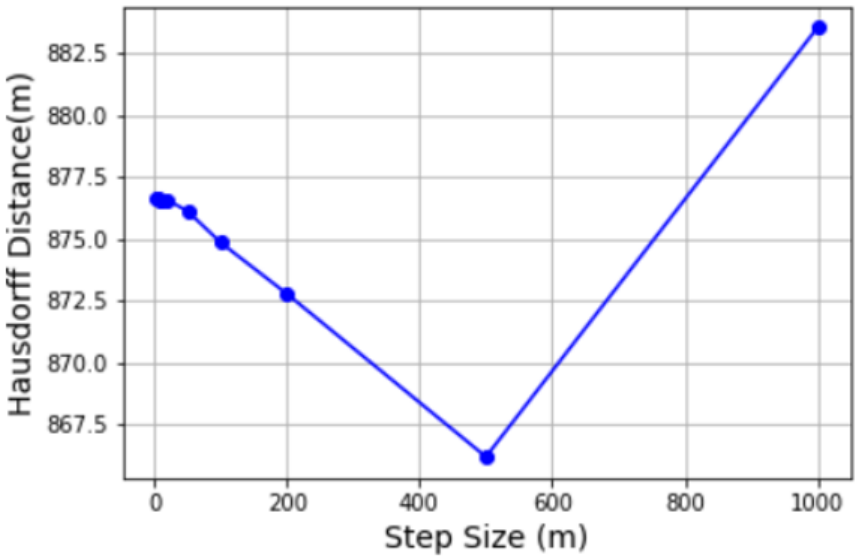


Figure 7.7: Sensitivity of the Hausdorff Distance towards icreasing step sizes

8 CONCLUSIONS AND FUTURE WORK

This chapter presents the conclusions of this research as well as the answers to the main and sub research questions. Moreover, recommendations for the continuation of this study are given.

8.1 CONCLUSIONS

The aim of this study was to describe a workflow for comparing the accuracy of some of the publicly available water reservoir datasets. This comparison was made, based on the assumption that none of these datasets was considered ground truth data. Therefore, the results of this research do not indicate or prove that one dataset is better or worse than the other. Nevertheless, it focuses on describing a technical methodology to facilitate the automatic comparison of remotely sensed and other vector water reservoir datasets in one environment where both raster and vector data can be processed, and most importantly in a large scale. This quality analysis was conducted relatively to the geometry of the datasets that are compared.

This research shows promising results that large scale analysis is possible for the comparison of reservoir waterbodies. However, due to time constraints, parts of the developed algorithm are in semi-automated form. The main limitation arises from the fact that two out of five input water reservoir datasets for this research, had to be generated in order to perform the quality comparison. More specifically, the generation of the Sentinel 2 and GSW remotely sensed datasets were performed in a discontinuous workflow of intermediate steps, as the large amount of the available data could not be processed successfully at once. As a result, the pre-processing of these data was implemented with some manual export and import operations to reduce the processing time for the generation of their vector water features and be able to proceed to the main quality analysis, which is fully automated.

During the implementation of the research methodology, some interesting observations were made relatively to the data. The OSM data, although they are a valuable source of information, they were quite challenging to filter properly so that only geometrical information related to water reservoirs were included. Due to the semantics (tagging) in the OSM data, it is possible that water features have been overlooked, resulting in possible underestimation of the water area when calculating the completeness quality metric, or overestimation when land features were falsely registered as water primitives.

In the case of Sentinel 2, the acquired most clean images were not comletely cloudless, since the resulting samples still contained a small percent of cloudy pixels. Apart from the remained cloudy parts in the satellite images, the surface water detection was also challenging, because of cloud and terrain shadows that lead to missclassification of water pixels. Moreover, the fact that a low percentile was used instead of the median of the water index values for all images, in some cases, resulted in worse representation of the water mask. This was identified with visual inspection of the satellite imagery. Consequently, the quantative results of the performed quality analysis are as accurate as the input data.

To answer the main reasearch question, two subquestions were investigated:

1. **Question 2: What are the differences in terms of spatial coverage?**

   When analyzing the actual intersecting surface water area of OSM with HydroLAKES, GRaND, GSW and Sentinel, a match of 24%, 36%, 9%, 19% respectively was found.

2. **Question 2: What are the differences in terms of positional accuracy?**

   The positional accuracy of the OSM data was evaluated in terms of percentage of overlap and Hausdorff Distance only for the features that intersect amongst the OSM and the other input datasets. More specifically, OSM compared to HydroLAKES and GRaND shows a good agreement concerning the Goodchild's percentage of overlap with 59.6% and 54.69% respectively. Moreover, this metric showed that there is a 35.93% overlap for the intersecting features of OSM and GSW and a 43.73 % for OSM and Sentinel 2. With regards to the histograms of the distances between the datasets based on the percentage of overlap, it was found that 57% of the HydroLAKES and 46% of GRaND features lie within a 10 m buffer zone from the OSM data. For GSW and Sentinel 2 the scores for this metric are significantly lower, as only 15% and 1.7 % accordingly, are found in a distance of 10 m from the OSM data.

   When analyzing the Hausdorff Distance which indicates the shape similarity between water reservoir features, in many cases a wide range of values in the order of hundreds of kilometers was found. Therefore, although this metric is feature based and the use of one statistical value for the entire Angola is not representative, the median of all Hausdorff Distance values was chosen to give a general insight in the variability of the shapes of the water features. In particular, the median Hausdorff Distance when comparing OSM to HydroLAKES is 206.19 m, whereas the median for GRaND is 41761.69 m. In the case of GSW and the used subset of Sentinel 2, the median Hausdorff Distances are equal to 236.07m and 92301.64 respectively.

The main research question :

"What are the spatial differences between Earth Observation based and Volunteered Geographic Information for water reservoirs and how can they be addressed in an automated way at a large scale?"

is answered as follows:

Based on existing concepts, a method was proposed to compute the spatial differences between water reservoir datasets. The comparison was perfomed in pairs of datasets. The approach for performing this analysis was based on three quality metrics: the completeness of water area, Goodchild's percentage of overlap for different buffer sizes and lastly the Hausdorff Distance.

As none of the data was considered of higher accuracy due to their limitations, the results of this approach have not been validated. Instead, the outcome of this research indicates only the level of agreement in terms of spatial coverage and positional accuracy between OSM and the other input datasets.

To assess the overall large scale and possibly global applicability of the approach, the methods have been tested over the extend of the country of Angola. This was possible by exploiting the planetary-scale analysis capabilities of GEE. The quantitative results from the accuracy comparison of the datasets were successfully estimated for the entire Angola in most cases, proving that cloud based methods are scalable. Even though the Hausdorff Distance metric when comparing Sentinel 2 and OSM provided results only for a subset of its features, possibly due to a scalability error, the overall performance of the algorithm showed promising results.

With an optimization of the Hausdorff Distance algorithm, the computational speed could be increased, making even a global scale analysis for all three quality metrics possible.

8.2 CONTRIBUTIONS

The main contributions of this thesis are given below.

- To develop an automated tool in an environment that enables firstly the use of both big volumes of available satellite data for surface water detection and vector datasets, and secondly is able to perform large scale analysis that compares the accuracy of these data.
- The cross validation of datasets coming from different origins related specifically to water reservoirs.

8.3 FUTURE WORK

This section gives a few recommendations that may be helpful for future research:

- Additional processing of the input datasets is suggested, to refine the Sentinel 2 and GSW datasets. More specifically, in the case of Sentinel 2, possible improvements to the method of water detection might include the selection of the ideal percentiles in a heuristic way instead of visual inspection. Further research for the selection of an optimal threshold value of the GSW water masks is suggested, to distinguish permanent water from short-term flooded areas.
- Exclusion of rivers from the GSW and Sentinel 2 datasets so that the results for the completeness and positional accuracy metrics are more accurate. Moreover, this will reduce the amount of processed Sentinel 2 and GSW data, so the algorithm will need less computational effort and therefore the pre-processing of these datasets could be fully automated.
- Merging the clusters of Sentinel 2 smaller waterbodies into single homogeneous features for more accurate water reservoir delineation.
- Apart from the identification of dis/similarities between the datasets, the research can be further extended by classifying into bad and good reservoirs. As due to the time limitation of this research, the mismatches amongst the datasets have not been resolved, a recommendation for future research would be the development of a data fusion algorithm that combines the strengths of all five input datasets. The process for resolving the mismatches and harmonizing existing vector and raster water mask datasets would require the introduction of objective criteria (e.g. topographic conditions) regarding confidence of every water mask. By doing this, all input datasets could be integrated into one that is more accurate regarding the true geometries of the water reservoir features, by consulting the accuracy metrics and making the most out of the value of these datasets. This could lead to the generation of water reservoir feature dataset with a better overall quality.
- Classifying the registered water features into types of water features (e.g. lakes, agricultural water reservoirs, valley-dammed reservoirs etc.).

BIBLIOGRAPHY

Avisse, N., Tilmant, A., François Müller, M., and Zhang, H. (2017). Monitoring small reservoirs' storage with satellite remote sensing in inaccessible areas. *Hydrology and Earth System Sciences*, 21(12):6445–6459.

Barron, C., Neis, P., and Zipf, A. (2014). A comprehensive framework for intrinsic openstreetmap quality analysis. *Transactions in GIS*, 18(6):877–895.

Bhatia, G. and Goel, D. (2011). An enhanced approach to improve the contrast of images having bad light by detecting and extracting their background. *International Journal of Computer Science and Management Studies*, 11.

Bhattacharya, P. (2012). Quality assessment and object matching of OpenStreetMap in combination with the Dutch topographic map TOP10NL. pages 23–88.

Brovelli, M. A., Minghini, M., Molinari, M., and Mooney, P. (2017). Towards an Automated Comparison of OpenStreetMap with Authoritative Road Datasets. *Transactions in GIS*, 21(2):191–206.

Brovelli, M. A. and Zamboni, G. (2018). A new method for the assessment of spatial accuracy and completeness of OpenStreetMap building footprints. *ISPRS International Journal of Geo-Information*, 7(8).

Campbell, J. B. (2002). *Introduction to Remote Sensing*.

Cardinal, J. (2019). Sets, graphs, and things we can see: A formal combinatorial ontology for empirical intra-site analysis. *Journal of Computer Applications in Archaeology*, 2:56–78.

Domenikiotis, C., Loukas, A., and Dalezios, N. R. (2003). The use of noaa/avhrr satellite data for monitoring and assessment of forest fires and floods. *Natural Hazards and Earth System Sciences*, 3(1/2):115–128.

Donchyts, G. (2018). *Planetary-scale surface water detection from space*. PhD thesis.

Donchyts, G., Schellekens, J., Winsemius, H., Eisemann, E., and Van de Giesen, N. (2016). A 30 m resolution surface water mask including estimation of positional and thematic differences using landsat 8, srtm and openstreetmap: A case study in the murray-darling basin, australia. *Remote Sensing*, 8(5).

Douglas, D. and Peucker, T. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 10:112–122.

Duan, R.-l., Li, Q.-x., and Li, Y.-h. (2005). Summary of image edge detection [j]. *Optical Technique*, 3(3):415–419.

Fairbairn, D. and Al-Bakri, M. (2013). Using geometric properties to evaluate possible integration of authoritative and volunteered geographic information. *ISPRS International Journal of Geo-Information*, 2(2):349–370.

Fan, H., Zipf, A., Fu, Q., and Neis, P. (2014). Quality assessment for building footprints data on openstreetmap. *International Journal of Geographical Information Science*, 28(4):700–719.

Feng, M. and Bai, Y. (2019). A global land cover map produced through integrating multi-source datasets. *Big Earth Data*, 3(3):191–219.

Feyisa, G. L., Meilby, H., Fensholt, R., and Proud, S. R. (2014). Automated water extraction index: A new technique for surface water mapping using landsat imagery. *Remote Sensing of Environment*, 140:23 – 35.

Frazier, P. and Page, K. (2000). Water body detection and delineation with landsat tm data.

Girres, J. F. and Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, 14(4):435–459.

Goetz, M. and Zipf, A. (2013). *The Evolution of Geo-Crowdsourcing: Bringing Volunteered Geographic Information to the Third Dimension*, pages 139–159.

Goodchild, M. F. and Hunter, G. J. (1997). A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, 11(3):299–306.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18 – 27. Big Remotely Sensed Data: tools, applications and experiences.

Haklay, M. (2010). How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, 37(4):682–703.

Hansen, M., Potapov, P., Moore, R., Hancher, M., Turubanova, S., Tyukavina, A., Thau, D., Stehman, S., Goetz, S., Loveland, T., Kommareddy, A., Egorov, A., Chini, L., Justice, C., and Townshend, J. (2013a). High-resolution global maps of 21st-century forest cover change. *Science (New York, N.Y.)*, 342:850–853.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., and Townshend, J. R. G. (2013b). High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160):850–853.

Hossain, M. J., Dewan, M., Ahn, K., and Chae, O. (2012). A linear time algorithm of computing hausdorff distance for content-based image analysis. *Circuits Systems and Signal Processing - CIRC SYST SIGNAL PROCESS*, 31.

Huang, C., Chen, Y., Wu, J., Li, L., and Liu, R. (2015). An evaluation of suomi npp-viirs data for surface water detection. *Remote Sensing Letters*, 6(2):155–164.

Huang, C., Chen, Y., Zhang, S., and Wu, J. (2018). Detecting, Extracting, and Monitoring Surface Water From Space Using Optical Sensors: A Review. *Reviews of Geophysics*, 56(2):333–360.

Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. A. (1993). Comparing images using the hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(9):850–863.

Ivanov, A., Krylov, S., and Tatarnikov, A. (2000). Automation of map generalization. *Geodesy and cartography*, (May):1 – 12.

Jakovljević, G., Govedarica, M., and Álvarez-Taboada, F. (2019). Waterbody mapping: a comparison of remotely sensed and GIS open data sources. *International Journal of Remote Sensing*, 40(8):2936–2964.

Kato, L. V. (2018). Integrating Openstreetmap Data in Object Based Landcover and Landuse Classification for Disaster Recovery.

Khodaei, K. and Nassery, H. R. (2008). Groundwater exploration using remote sensing and geographic information systems in a semi-arid area (southwest of urmieh , northwest of iran). (June).

Koukoletsos, T., Haklay, M., and Ellul, C. (2012). Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS*, 16(4):477–498.

Lehner, B., Liermann, C. R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., and Frenken, K. (2011). Global Reservoir and Dam (GRanD) database. *Technical Documentation*, (March):12.

Li, Yang, and Wu (2019). A Method of Watershed Delineation for Flat Terrain using Sentinel-2A Imagery and DEM: A Case Study of the Taihu Basin. *ISPRS International Journal of Geo-Information*, 8(12):528.

Manavalan, P., Sathyanath, P., and Rajegowda, G. L. (1993). Digital image analysis techniques to estimate waterspread for capacity evaluations of reservoirs.

McFeeters, S. K. (1996). The use of the normalized difference water index (ndwi) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7):1425–1432.

Meijster, A. (2004). *Efficient sequential and parallel algorithms for morphological image processing*. PhD thesis. Relation: https://www.rug.nl/ $date_submitted$: $2004Rights : Universityof Groningen$.

Mooney, P., Corcoran, P., and Winstanley, A. C. (2010a). Towards quality metrics for openstreetmap. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 514–517. ACM.

Mooney, P., Corcoran, P., and Winstanley, A. C. (2010b). Towards quality metrics for openstreetmap. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, page 514–517, New York, NY, USA. Association for Computing Machinery.

Mueller, N., Lewis, A., Roberts, D., Ring, S., Melrose, R., Sixsmith, J., Lymburner, L., Mcintyre, A., Tan, P., Curnow, S., and Ip, A. (2016). Remote Sensing of Environment Water observations from space : Mapping surface water from 25 years of Landsat imagery across Australia. 174:341–352.

Müller, F., Iosifescu, I., and Hurni, L. (2015). Assessment and Visualization of OSM Building Footprint Quality. *International Cartographic Conference*.

Ogilvie, A., Belaud, G., Massuel, S., Mulligan, M., Le Goulven, P., and Calvez, R. (2018). Surface water monitoring in small water bodies: Potential and limits of multi-sensor Landsat time series. *Hydrology and Earth System Sciences*, 22(8):4349–4380.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66.

Ozesmi, S. L. and Bauer, M. E. (2002). Satellite remote sensing of wetlands. *Wetlands ecology and management*, 10(5):381–402.

Rogowska, J. (2009). Chapter 5 - overview and fundamentals of medical image segmentation. In BANKMAN, I. N., editor, *Handbook of Medical Image Processing and Analysis (Second Edition)*, pages 73 – 90. Academic Press, Burlington, second edition edition.

Sachs, J. (2001). Image Resampling. pages 1–14.

Santoro, M., Wegmüller, U., Lamarche, C., Bontemps, S., Defourny, P., and Arino, O. (2015). Strengths and weaknesses of multi-year envisat asar backscatter measurements to map permanent open water bodies at global scale. *Remote Sensing of Environment*, 171:185 – 201.

Srisha, R. and Khan, A. (2013). Morphological operations for image processing : Understanding and its applications.

Tan, B., Masek, J. G., Wolfe, R., Gao, F., Huang, C., Vermote, E. F., Sexton, J. O., and Ederer, G. (2013). Improved forest change detection with terrain illumination corrected landsat images. *Remote Sensing of Environment*, 136:469–483.

Thirusittampalam, K., Hossain, M. J., Ghita, O., and Whelan, P. (2013). A novel framework for cellular tracking and mitosis detection in dense phase contrast microscopy images. *IEEE Journal of Biomedical and Health Informatics*, 17:642–53.

Thissen, J. (2019). Automating surface water detection for rivers : the estimation of the geometry of rivers based on optical earth observation sensors.

Xu, H. (2006). Modification of normalised difference water index (ndwi) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14):3025–3033.

Yamazaki, D., Trigg, M. A., and Ikeshima, D. (2015). Remote Sensing of Environment Development of a global ˜ 90 m water body map using multi-temporal Landsat images. *Remote Sensing of Environment*, 171:337–351.

Yousefi, J. (2011). Image binarization using otsu thresholding algorithm. *University of Guelph, Ontario, Canada*.

Zhu, L., Suomalainen, J., Liu, J., Hyyppä, J., Kaartinen, H., and Haggrén, H. (2018). *A Review: Remote Sensing Sensors*.

Zhu, Z. and Woodcock, C. E. (2014). Continuous change detection and classification of land cover using all available landsat data. *Remote Sensing of Environment*, 144:152 – 171.

A REPRODUCIBILITY SELF-ASSESSMENT

A.1 MARKS FOR EACH OF THE CRITERIA

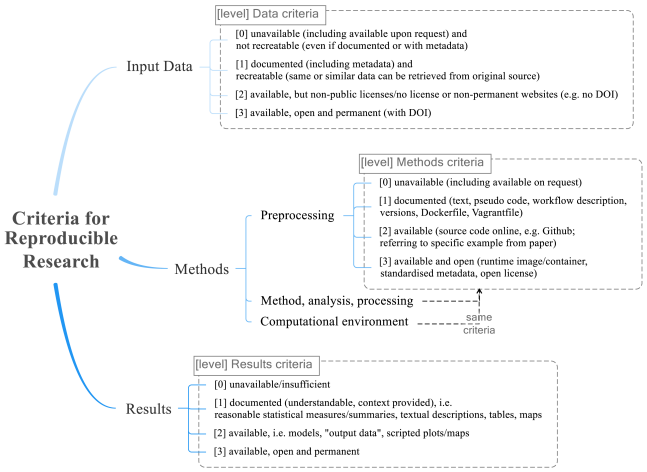


Figure A.1: Reproducibility criteria to be assessed.

A.2 SELF-REFLECTION

This study were performed in collaboration with Deltares [15]; which is an independent institute for applied research in the field of water and subsurface. All the information used is publicly available. More specifically:

- Input Data [3]: Five Datasets have been used during this thesis. These datasets are provided freely available in the Google Earth Engine platform, the OpenStreetMap platform, and lastly the HydroLAKES and GRaND databases. The information they provide is limited in the extend of the country of Angola.
- Preprocessing [1/2]: All the pre processing procedures are documented with text. In the case of Sentinel 2 and Global Surface Water a workflow and the scripts for the generation of these datasets in the Google Earth Engine Platforms is included. For OpenStreetMap, apart from the preprocessing steps, the list of the manipulation open source command line tools that were used is presented.
- Methods [2]: The source code for the comparison analysis used in this research is available online in Google Earth Engine repository. The algorithms are documented with text.
- Computational environment [3]: This research is conducted by using the Google Earth Engine platform which has a commercial license program, and the open source QGIS software for the manipulation of data and visualization of the results.

15 https://www.deltares.nl/en/

- Results [2]: The are documented through text, graphs, maps, and they can also be reproduced by using the freely available code.

COLOPHON

This document was typeset using LaTeX. The document layout was generated using the arsclassica package by Lorenzo Pantieri, which is an adaption of the original classicthesis package from André Miede.