LOCAL CLIMATE ZONE CLASSIFICATION SYSTEM USING WEB GIS APPROACH

Bruno Ricardo Jorge Martins Marques

Project Work

presented as partial requirement for obtaining a Master's Degree in Geographical Information Systems and Science

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

LOCAL CLIMATE ZONE CLASSIFICATION SYSTEM USING WEB GIS APPROACH

by

Bruno Ricardo Jorge Martins Marques

Project Work presented as partial requirement for obtaining the Master's degree in Geographical Information Systems and Science

Supervised by

Pedro da Costa Brito Cabral, PhD, NOVA Information Management School

Ana Oliveira, PhD, +ATLANTIC CoLAB

February, 2025

STATEMENT OF INTEGRITY

1. Digital signature

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or not published) are referenced.

This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Assinado por: **Bruno Ricardo Jorge Martins Marques**
Num. de Identificação: 12976896
Data: 2025.02.15 15:39:08+00'00'

USE OF GENERATIVE ARTIFICIAL INTELLIGENCE

Tasks	NO	YES	Generative Artificial Intelligence tools
Better understand issues related to the research		X	ChatGPT
Summarizing text from bibliography / resources	X		
Summarizing the method(s) used	X		
Translating text	X		
Grammar check	X		
Paraphrase or rewriting text from other people / resources	X		
Coding in R, Python, etc.		X	GitHub Copilot
Get help on a software		X	ChatGPT
Creating and editing images, maps, videos, etc.	X		
Data analysis	X		

DEDICATION

I dedicate to my wife, Susana Martins Marques, for the support and understanding that free time was lacking during the creation of this thesis, and patience as sometimes I wondered if I'd ever finish it. It was a long journey.

To my mom, thank you for the life lessons, I wouldn't be who I am without you.

ACKNOWLEDGMENTS

I'd like to thank my wife Susana Martins Marques for pushing me to believe that Geographic Information Systems were interesting to learn, and to motivate me during the time I needed to study for this master's course, and to be there to listen through my difficulties. This thesis is the culmination.

I want to thank my co-supervisor, Ana Oliveira, who also works with me at +ATLANTIC CoLAB for giving me the initial idea and general purpose and for all the time supporting me during the development of this thesis. This also extends to +ATLANTIC CoLAB for giving me the conditions, time and technologically speaking, to be able to work on this.

To my supervisor, Pedro Cabral, thank you for accepting me to supervise me, despite the distance, you were always available to help whenever I needed.

ABSTRACT

Climate change along with population growth and urbanization trends across the coastal areas are pressuring our cities to become more and more resilient to the ''new normal''. Our resilience is a function of hazards' severity and probability, people's and assets' exposure and societies' sensitivity, and our ability in responding not only to a warmer climate, but also and foremost to extreme climate events will determine our resilience, in the future. Midlatitude countries such as Portugal have been experiencing warmer summers, in which heatwaves are becoming more frequent and severe, exposing an increasingly ageing population to correlated health and energy poverty issues. Part of the challenge moving forward is ensuring our ability to monitor and predict extreme heatwaves in space and time, with enough level of detail as to allow to highlight hot spots within cities and larger metropolitan areas which should be prioritized. The temporal aspect, and the spatial accuracy are something that the national and international weather services already take well care of - but there is a gap in what concerns managing cities, especially in more complex and coastal environments, where temperatures asymmetries are greater due to the complex land-sea interactions. To overcome this issue, remote sensing now offers spatially complete time series of observational data from which to derive insights into what concerns to the way we occupy these territories, as well as the thermal footprint of that land use land cover. When processed using machine learning and artificial intelligence, time series of geospatial data obtained from satellite imagery hence allows us to downscale weather and climate up to the neighborhood scale. But to do that we need to ensure consistency, precision, and standardization of processes and benchmarking of climate-relevant land use land cover data. Local Climate Zones (LCZ) is now the gold standard scheme for Land Use/Land Cover (LULC) classification; nevertheless, while the LCZ's overall criteria are well-established, the methods employed for classification are very diverse, often local-specific. Alternatively, some highly scalable satellite imagery-based classification approaches have been attempted, based on open-source data – but these lack sufficient local detail, and may lead to misleading urban climate modelling results. This thesis aims to build

upon pre-existent attempts at transforming geospatial data products into accurate and sub-kilometric resolution LCZ classification maps, to optimize them in terms of precision and scalability using open-source tools and libraries and a WEB GIS interface. The GIS-based classification is implemented using Portuguese and Denmark cities for benchmark with previous works. The objective is to attain the best-performing algorithm without overfitting it to a specific location, while its usability to tackle urban climate shall be tested, particularly the predictability of the Urban Heat Island (UHI) effect using such LCZ as a predictor, but open for scientific community to use.

KEYWORDS

Land Use/Land Cover; Urban Heat Island; Local Climate Zone; Web; GIS; Climate Adaptation

Sustainable Development Goals (SGD):

INDEX

INDEX OF TABLES

INDEX OF FIGURES

ACRONYMS

API – Application Programming Interface

BH – Building Height

CLMS – Copernicus Land Monitoring Service

DLT – Dominant Leaf Type

GIS – Geographic Information System

GRA – Grassland

IMD – Imperviousness Density

LCZ – Local Climate Zone

ML – Machine Learning

OSM – Open Street Map

TCD – Tree Cover Density

UA – Urban Atlas

1. INTRODUCTION

Cities (here referred to in broader terms, interchangeably as municipalities, metropolitan areas or conurbations) are central to our current way of living. The industrial revolution of the 19th century placed immense pressure on cities to develop at a faster pace to accommodate new land uses and the many rural migrants which lead to more extensive urban plans which were not always responsive to local climate conditions (Oliveira et al., 2020b). While some evolutions have happened such as moving the industries outside the city centers, urban planning and architectural design during the 20th century still frequently ignored climatic performance, although collective awareness regarding climate change, sustainability and environmental protection has been growing which is driving regulatory entities to develop climate-related risk assessment frameworks for urban planning (IPCC, 2014).

As such, international standards for climate-relevant Land Use/Land Cover (LULC) classification schemes, such as the Local Climate Zones, also known as LCZ, were developed to help local climate experts to convey information to urban planners and architects, for which several works and studies appeared. (Alexander et al., 2016; Mills et al., 2010)

LCZ's help to define the land cover classes of a city with finer granularity than simple classifications such as Urban or Rural, reducing uncertainty about the actual exposure and land cover of the cities, and offering a more reliable alternative for weather and climate modelling LULC inputs. The LCZ framework as proposed by (Stewart & Oke, 2012) normally comprises 17 zone types at the local scale (10^2 to 10^4m). Each type is unique in its combination of surface structure, cover and human activity – in turn, these reflect classes of urban morphology, materials and their thermal properties, which is the physics foundation of climate modelling, namely Bowen-ratio, roughness length or thermal resistance, which influence the calculation of the urban energy balance components and resulting air and surface temperatures.

Tools such as World Urban Database and Access Portal Tools (WUDAPT) were created to acquire and make accessible coherent and consistent descriptions and information on form and function of urban morphology relevant to climate weather and environmental studies on a worldwide basis, as well as providing a portal with tools that extract relevant urban parameters for models and for model applications. It primarily uses open-source satellite data,

it works well for cities which are regularly gridded, with a high percentage of certainty, but for cities which are intricate and with complex morphological characteristics, it needs improvement, as this is a limitation arising from the spatial resolution of the used data, but also of the difficulty in distinguishing certain LULC classes that have similar spectral signatures (e.g., bare soil versus concrete) or the compromise between the generalization of the algorithms employed globally versus local-specific urban features (Hidalgo et al., 2019). To overcome these limitations, (Oliveira et al., 2020b) developed a Geographic Information Systems (GIS)-based method which uses ArcGIS's Model Builder and data from Copernicus Land Monitoring Service (CLMS), such as Imperviousness Density, Tree Cover Density, Grasslands, Urban Atlas which cover 785 Pan-European and local datasets, and CORINE Land Cover. It also uses OpenStreetMap for additional information on industrial built types. This method, according to the paper, reaches a higher certainty in European cities such as Athens or Barcelona, up to 81% accuracy versus 50% of WUDAPT (Oliveira et al., 2020b) – the underlying reason is that such datasets are produced through a complex merging of many satellite data sources, including very high resolution missions (i.e, pixel sizes below 5m) from proprietary private operators, thus providing a baseline for the classification that goes beyond the level of detail of open imagery (typically, 20m or more).

The issue with this approach is that, so far, it requires a non-open-source tool, which is ArcGIS, to generate results, and it is a long manual approach, taking from 2 to 3 days of work, and it requires manual changing input data for each zone that needs to be analyzed. Hence, it is a laborious task not suitable to produce at scale.

This project tackles these issues by giving it a more similar feeling to WUDAPT, but using (Oliveira et al., 2020a) algorithm, which is open and attached to her publication, using the latest available data from CLMS. In the end, it results in a Web GIS tool, in which the user can select the zone he wants to obtain an LCZ classification and after some processing time, the LCZ is generated and made available for all users, while notifying the user that the data is ready to be viewed via email. In this tool, one can check other users generated LCZ's, as it may be that among the already generated LCZ's, by searching for the city's name, the user finds one that fits his study area.

This tool is also able to run pre-determined Functional Urban Areas (FUAs) such as Denmark's major cities like Copenhagen, Odense, Aarhus, and Aalborg, as is necessary for CLIM4Cities,

an European Space Agency (ESA)-funded project (Contract No. 4000143628/24/I-DT, under the AI TRUSTWORTHY APPLICATIONS FOR CLIMATE call of the Future-EO programme) in which this tool was developed. CLIM4Cities is a consortium made up of experts from the Danish Meteorological Institute (DMI) and the +ATLANTIC CoLAB.

As such, the LCZ's generated by this tool will be used scientifically, as a predictor for a bigger machine learning (ML) prediction system that is able to downscale urban climate data (forecasts, reanalysis or climate projections), and the algorithm developed in this project will be generic, so that it can be used scaled to novel projects in the pipeline that will involve European Cities.

2. DATA & METHODS

2.1. STUDY AREA

The tool is capable of processing LCZ's for any bounding box drawn in the map of the web platform developed in this project, although the focus of this thesis is on Lisbon (Portugal) and Aarhus (Denmark).

The criteria for the selection of these cities specify that they need to be regionally relevant metropolitan areas, and they depict coastal urban settlements. Lisbon is one of the cities selected in the work of (Oliveira et al., 2020b) and Aarhus is one of the cities to be studied in the CLIM4Cities project in which this tool was built for.

Population statistics are estimated as follows:

- Lisbon Municipality: According to (Statistics Portugal, 2024), it has 567,131 inhabitants, with a surface area of 100.05 km^2 with a population density of 5,445.7 per km^2
- Aarhus Municipality: According to (Statistics Denmark, 2024), it has 367,095 inhabitants, with a surface area of 468 Km^2 and a population density of 784 inhabitants per km^2

Given that the tool uses the concept of bounding boxes to process, these are the coordinates in WGS 84 that will be analyzed for each of the cities, in the format of (Southwest Longitude, Southwest Latitude, Northeast Longitude, Northeast Latitude):

- Lisbon: (-9,34020663ºE, 38,68953237ºN, -9,05103795ºE, 38,84845948ºN) as seen on Figure 1 - Lisbon Study Area
- Aarhus: (10,06240749ºE, 56,06537071ºN, 10,35157617ºE, 56,22429782ºN) as seen on Figure 2 - Aahrus Study Area

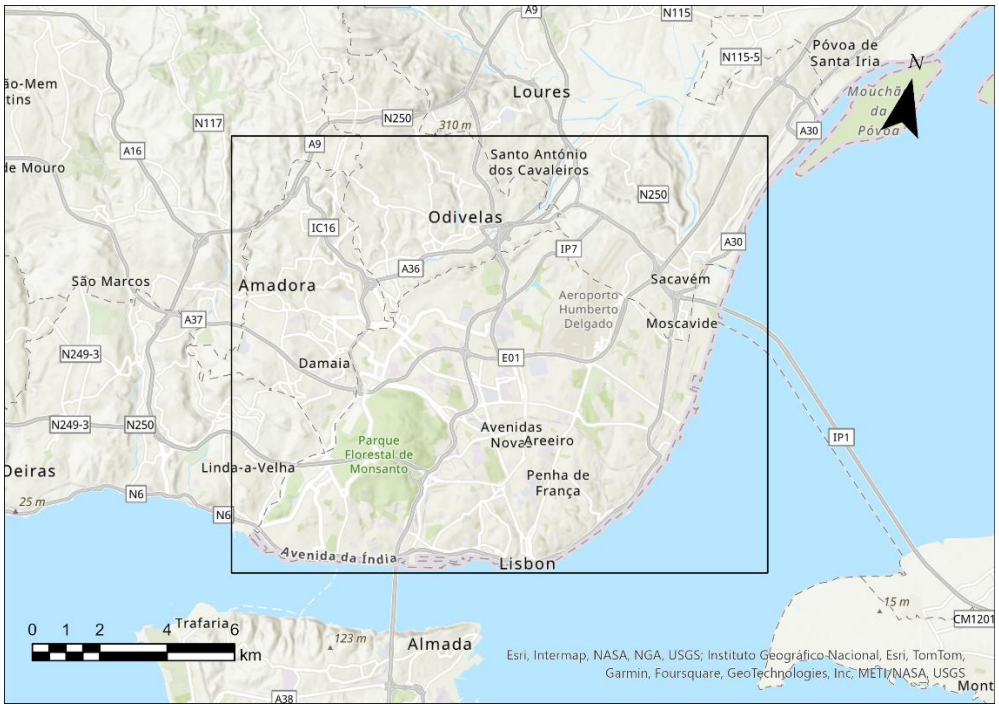


Figure 1 - Lisbon Study Area

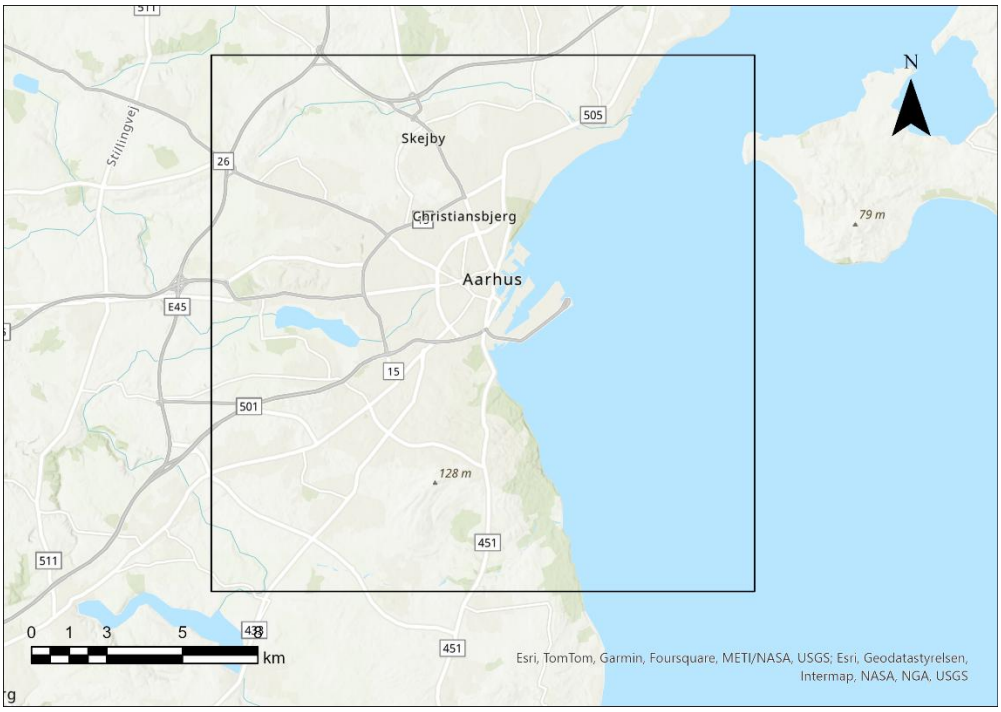


Figure 2 - Aahrus Study Area

Climatic characteristics are contrasting between both locations (Figure 3 - Present-day World Map of Koppen-Geiger Climate Classification. (Kottek et al., 2006)):

- Lisbon is located within the Koppen-Grieiser 'Csa' class (Kottek et al., 2006), corresponding to the 'warm temperature', 'summer dry' and 'hot summer' sub-classes

that characterize the southern European Mediterranean climate, in which thermal discomfort from air temperature extremes during the summer tends to be related to very dry conditions arising from high-pressure blocking systems stationed over the Iberian Peninsula, often influenced by southerly air mass flows from Africa, which occasionally also transport Saharan dust.

- Aahrus is located within the 'Cfb' class, corresponding to 'warm temperature', 'fully humid', and 'warm summer' subclasses, albeit closer to the 'Dfb' and 'Dfc' classes which characterize Scandinavian countries, thus being in the transition between the European Continental and Boreal climates. Compared to Lisbon, it is much wetter and cooler, and thermal discomfort from air temperature extremes is also influenced by the presence of humidity, which may reduce the human tolerance to heat.

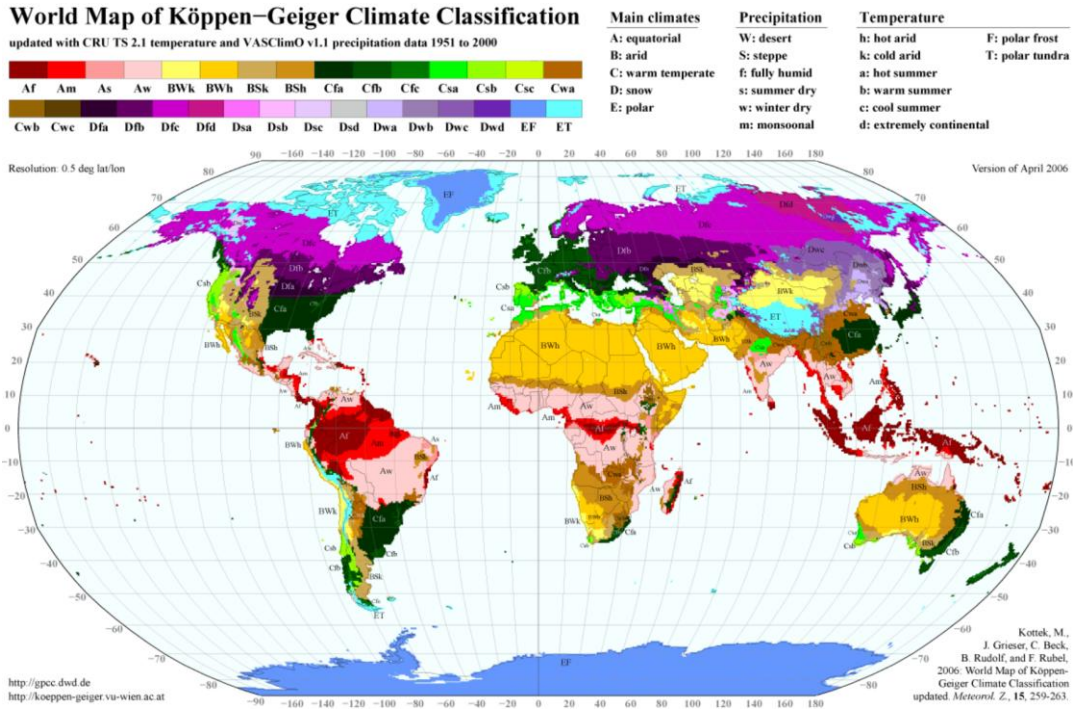


Figure 3 - Present-day World Map of Koppen-Geiger Climate Classification. (Kottek et al., 2006)

2.2. DATA

This project uses Urban Atlas and Corine Land Cover data from Copernicus Land Monitoring Service (CLMS) dataset catalogue, provided by European Environmental Agency (EEA). For specific parts which help to determine the LCZ 8 and 10 classes, OpenStreetMap was also used. Table 1 – Data used in this study contains all the details on the data used in this study.

Table 1 – Data used in this study

Dataset	Version	Release Date	Spatial resolution	Format	Reference	Accuracy
UA	2018	2021-07-16	10m	GPKG	(EEA, 2021)	>= 80%
UA BH	2018	2022-10-06	10m	GeoTIFF	(EEA, 2022b)	n/a
CLC	2018	2020-05-13	100m	GPKG	(EEA, 2019)	>= 85%
IMD	2018	2020-08-18	10m	GeoTIFF	(EEA, 2020b)	>= 90%
TCD	2018	2020-09-18	10m	GeoTIFF	(EEA, 2020d)	>= 90%
DLT	2018	2020-09-18	10m	GeoTIFF	(EEA, 2020c)	>= 90%
GRA	2018	2020-08-18	10m	GeoTIFF	(EEA, 2020a)	>= 85%
OSM	n/a	2024-10-03	n/a	GPKG	(OSM, 2024)	n/a

Unlike the original ArcGIS tool, which required manual data acquisition, this project does not require the local download of the pre-built datasets that CLMS provides, but rather it is requesting the required data via the CLMS API (EEA, 2022a) using the bounding-box approach, using the area drawn from the user in the web platform. This request is sent to CLMS's processor, which puts the request in a queue. The advantage is that it minimizes the need for any manual tasks; nevertheless, as with the manual download options, our request may take an uncertain amount of time to start processing, depending on the CLMS service speed.

The last dataset is OpenStreetMap data, to retrieve more detailed built-up and industrial data. The tool uses pre-downloaded data from OpenStreetMap, which is divided by countries. To find out the country the user is requesting data from, the tool uses geocoding data from Nominatim, which uses OpenStreetMap (OSM) data, to retrieve the location selected by the user in the web platform. Future developments will see the use of OSM's API to automatically retrieve the required data instead of needing to have pre-downloaded data.

2.3. METHODS

2.3.1. REQUIREMENTS

To allow the scalability of the solution, this project needs to be able to classify LCZs using the same algorithm implemented in (Oliveira et al., 2020a) Toolbox, but using open-source tools and in an automated approach. It's also necessary to have a friendly user interface, preferably using a website rather than a desktop application.

This led to the opportunity to implement a Participatory Web GIS in which not only +ATLANTIC CoLAB will be able to use it, but other users as well, such as from Academia, from other Scientific users and Urban Planners. The Participatory Web GIS's requirement is to offer users the possibility to report inaccuracies in the generated LCZ. This will allow us to gather data for posterior improvement using ML with a supervised learning approach, namely by gathering a set of pre-labelled ground truth points which can then be used to train alternative LCZs.

Regarding non-functional requirements, this project needs to be implemented using open-source libraries and tools, preferably the same ones used already in +ATLANTIC CoLAB. For the LCZ classifier algorithm, we chose to use Python and geospatial libraries such as GeoPandas for vector operations and Rasterio for raster operations. Regarding the Web GIS platform development, HTML, CSS and JavaScript will be the programming languages used and PostGIS will be used as the database server to gather input data from users.

2.3.2. IMPLEMENTATION

The project contains multiple systems, each one with its own responsibility so that it allows the user to view the map, request a specific area to be classified, and request data from CLMS and Nominatim. Once the data is available for download from CLMS, the LCZ classifier retrieves the data and starts processing. Once it's processed, it informs the user, via email, that the classification map can be viewed on the web platform. How the systems interact with each other can be viewed in Figure 4 - System architecture.

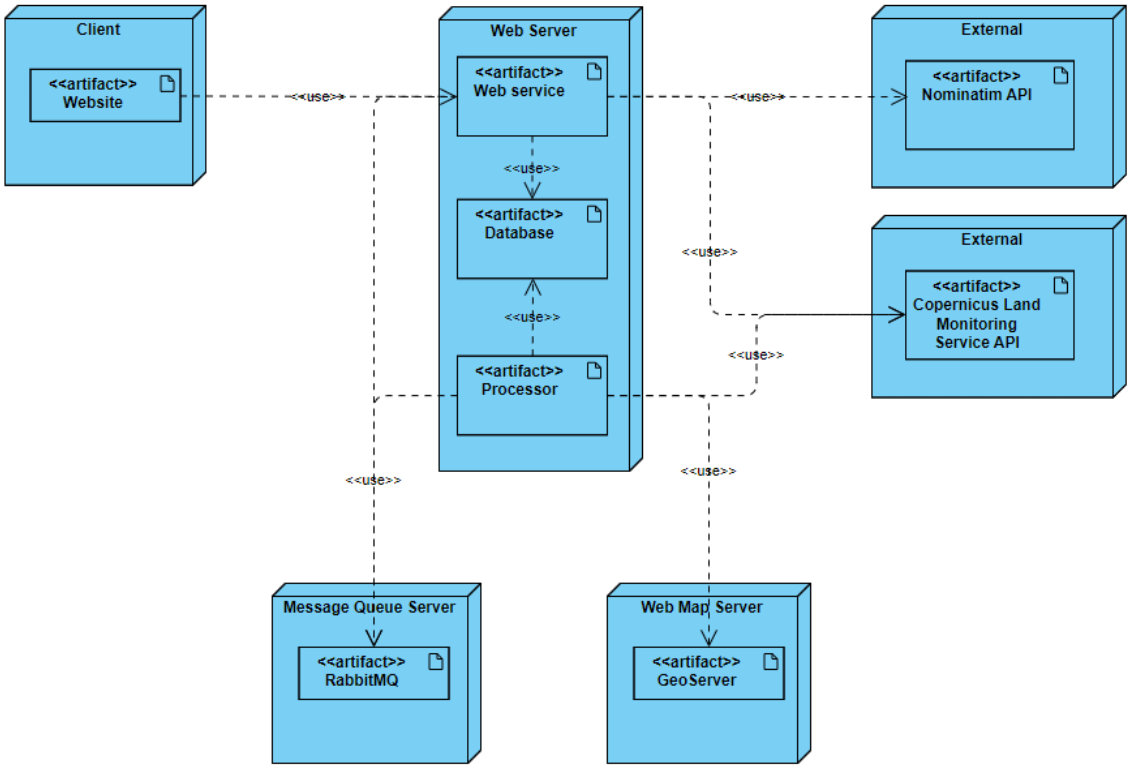


Figure 4 - System architecture

The Participatory Web GIS platform is the main point of interaction of the user, which is a website developed with Python, HTML, CSS and JavaScript. It displays a Leaflet map in which the user may select the area of interest where an LCZ classification should be created. This is useful for Scientists where they can select the study area, but also for normal users where they can select area of interest to be classified. The main use case for this is to generate an LCZ layer for a given city to be used in climate modelling (regardless of ML or physics-based). All climate and weather modelling starts with the land (and ocean) scheme, and the coverage type is one of the main constraints that determines land surface and atmospheric energy balance fluxes. Operationally, weather services tend to use simpler urban-rural-forest classification schemes, at the national or pan-European scales. For urban climate, however, one requires a much higher level of detail to infer the energy fluxes across many more classes of land cover types according to surface materials, morphology and vegetation typologies. Another possible use case is someone looking for an area to live in and wants to know the types of land cover of the area, knowing that certain types are typically related to increased heat exposure (e.g., high-density neighborhoods).

The core of the project is the classifier (also called processor), which inspired heavily on the classification algorithm from (Oliveira et al., 2020a) but using open-source software rather than

proprietary tools such as ArcGIS. It is an automated classifier, there is no need for a GIS Specialist to spend time on manual tasks which leads to increased processing time that can take up to 3 days, because this specialist would need to look for input data that suits the study area, configure Model Builder's various steps to run and in case something goes wrong, such as ArcGIS crashing, restart the process.

The classification algorithm was implemented differently from (Oliveira et al., 2020a) because this project, for performance reasons, uses pre-processed input raster data, while (Oliveira et al., 2020a) uses pre-processed input vector data, converted from the raster data into polygons, which slows down the processing time in every step, although it is slightly more accurate when analyzing on the boundaries of the polygons. Nevertheless, the accuracy loss is only noticeable when analyzing results at a resolution closer to the minimum mapping unit of the CLMS layers (i.e., 10 m), which is rarely the target resolution for LCZs usage (typically, on the order of 100 m). In addition, accuracy assessment will be carried out both in rasterized (this project's output) and vectorized form (ArcGIS's output) to assess the difference between both approaches.

Some geospatial operations which were used by (Oliveira et al., 2020a) ArcGIS's toolbox is not readily available to be used with open-source tools, such as spatial join one-to-one join operation and the polygon to raster feature, which uses a specific "Maximum combined area" option which can't be found anywhere else (i.e., it is a proprietary algorithm not openly documented). These had to be custom implemented for this project, but without a detailed description of how it works, it may be difficult to replicate (in the case of the spatial join) or use a slower method (in the case of the polygon to raster).

As mentioned earlier, the LCZs classification system provides a research framework for urban heat island studies and standardizes the worldwide exchange of urban temperature observations.

Urban Heat Island refers to the atmospheric warmth of a city compared to the countryside, and it occurs in almost every urban area. The main causes of heat islands relate to structural and land cover differences of urban and rural areas, as cities are rough with buildings extending above ground level and are dry and impervious with construction materials extending across natural soils and vegetation.

These characteristics alter the natural surface energy and radiation, making cities warmer places (Oke, 1982), which in cases where the cities are already in a warm climate, as it increases discomfort, potentially raises the threat of heat stress and mortality, heighten the cost of air conditioning and demand for energy.

(Stewart & Oke, 2012) found that many studies were not accurate enough by only separating urban and rural areas, failing to give quantitative metadata of site exposure or land cover. Hence the Local Climate Zone classification, which is a climate-based classification for urban and rural sites that applies universally and relatively easily.

The article says that the aim of the LCZ classification is twofold, to facilitate consistent documentation of site metadata and thereby improve the basis of intersite comparisons and provide an objective protocol for measuring the magnitude of the urban heat island effect in any city.

The LCZ Classification is normally made up of 17 classes as shown in Figure 5 – LCZ classification types. As (Stewart & Oke, 2012) refers, 15 classes are defined by surface structure and cover and 2 classes by construction materials and anthropogenic heat emissions. The standard set is divided into built types from 1 to 10 and land cover types from A to G. Built types are composed of constructed features on a predominant land cover, which is paved for compact zones and low/plants scattered trees for open zones. Land cover types can be classified into seasonal or ephemeral properties (i.e. bare trees, snow-covered ground, dry/wet ground). This serves as the baseline for most studies, in cases where not all data is readily available, they may add more classes that may better fit their study area. In this project, due to lack of data on building heights from CLMS, two additional types were defined. These are LCZ 123 and LCZ 456, as the name implies, it's a mix of LCZ 1, 2, 3 for LCZ 123 and LCZ 4, 5, 6 for LCZ 456. LCZ 123 represents areas which are compact, irrespective of building height, and LCZ 456 represents areas which are open with buildings but irrespective of their heights.

Normally the LCZ classification is based on the physical properties and analysis in-situ of the study area, while other approaches such as WUDAPT analyze Landsat satellite data. Since this project is based on the work of (Oliveira et al., 2020a) which uses high quality datasets of imperviousness, tree cover density, dominant leaf types and grassland, the approach is

different, it's done by analyzing CLMS pre-processed raster data and apply a reclassification algorithm with multiple steps which result in a LCZ classification.

Figure 5 – LCZ classification types

The algorithm of (Oliveira et al., 2020a)'s implementation is as follows:

1. Pre-processing:
    a. GRA (Figure 6 – Grasslands input) , IMD (Figure 7 – Imperviousness input), TCD (Figure 8 – Tree Cover Density Input) and DLT (Figure 9 – Dominant Leaf Type) layers are clipped to the Region of Interest and converted into polygon shapefile format.
    b. The outputs are then reclassified according to equivalent LCZ classes, for example, GRA is reclassified into LCZ D.
2. LCZ Baseline:
    a. UA, CLC and OSM input vector layers are subject to class selection and merged into an LCZ baseline polygon shapefile layer.

    b. The baseline is merged with IMD-based LCZ classes and the corresponding area per feature is calculated.
    c. The baseline is then merged with TCD, DLT and GRA-based LCZ classes from step 1b and the corresponding area per feature is calculated.
3. LCZ classification:
    a. Reclassify each feature from the LCZ baseline, according to the dominant LCZ class in each feature.
4. LCZ with Building Height:
    a. Merges the DHM raster layer with step 3a. shapefile classification output and reclassify each feature from the LCZ baseline. This step mainly will convert LCZ 123 to LCZ 1, 2, 3 or LCZ 456 to LCZ 4, 5, 6, depending on the building heights.
5. LCZ conversion to raster:
    a. Convert the shapefile LCZ classification into raster, according to user's pixel size specification.

As mentioned previously, this project closely follows this algorithm, with a few modifications as will be explained now.

2.3.2.1. PRE-PROCESSING

The pre-processing step is required to accurately access the land cover type from LCZ A to D, as these are about trees, bushes and low plants.

Once a request is made to create an LCZ classification, it is necessary to retrieve the required input data from Copernicus Land Monitoring Service API (CLMS API), which is freely available, although a free account is required. The API enables data retrieval for a user-selected bounding box. The area of the bounding box influences the process duration, with smaller bounding boxes resulting in shorter processing times.

Requests made to CLMS API will be queued for an uncertain duration, then once the queue is cleared, it starts processing on their side, and it takes around 20 to 30 minutes to process. Since it's not certain when the data from CLMS will be ready, we need to ask the API every 30 minutes for the status of the request until it's ready, making it effectively a polling service. Once the data is ready, we download the data to our file server, so that our classifier can start.

The first step is to pre-process the downloaded Grasslands (GRA), Imperviousness (IMD), Tree Cover Density (TCD) and Dominant Leaf Tree (DLT) inputs.

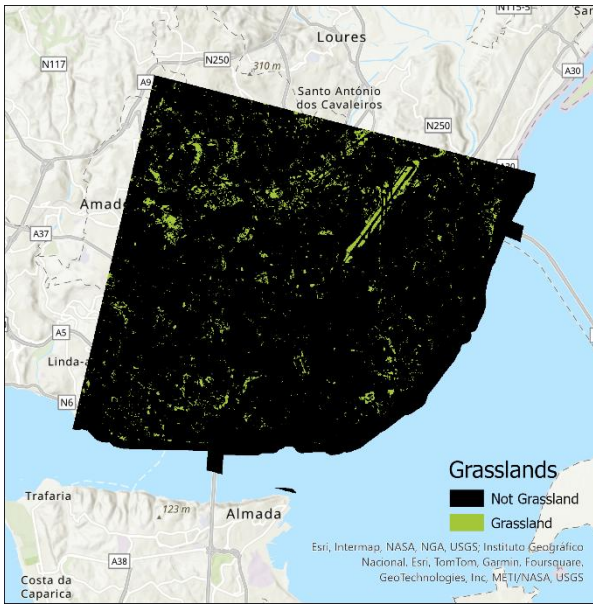


Figure 6 – Grasslands input

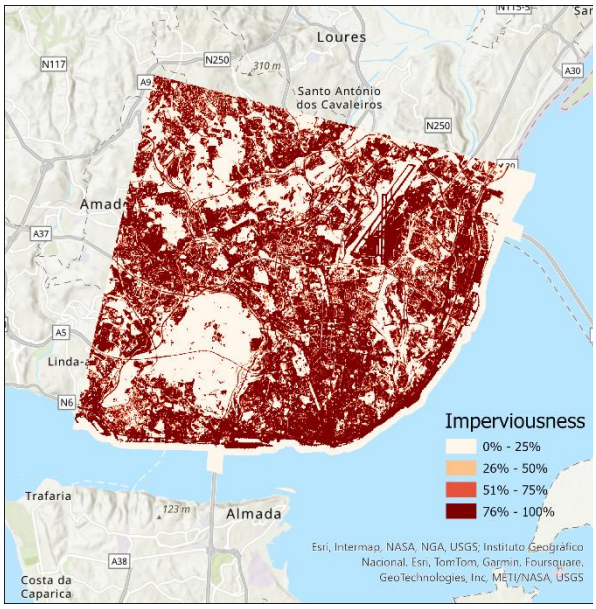


Figure 7 – Imperviousness input

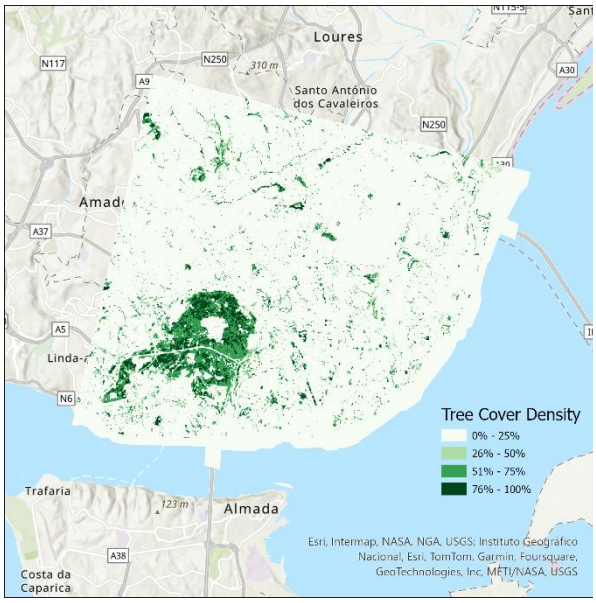


Figure 8 – Tree Cover Density Input

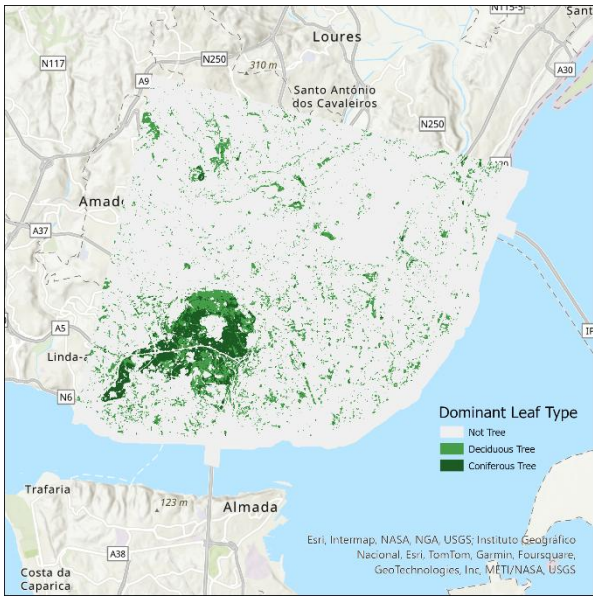


Figure 9 – Dominant Leaf Type

The pre-processing reclassifies the input data using the following logic:

- Every pixel identified as GRA is classified as LCZ D, otherwise it's classified with a NoData value.
- The DLT is either 1 or 2, which can be reclassified as Deciduous or Coniferous respectively.
- The IMD contains continuous percentage values, and it's not directly reclassified into a specific LCZ class due to the nature of the data as it needs to be reclassified together with the other raster data to produce a definitive LCZ class:

a. If pixel has less than 20% imperviousness, it's reclassified as Nature
b. If pixel has more than 20% imperviousness, but less than 40% imperviousness, it's reclassified as Sparsely
c. If pixel has more than 40% imperviousness, but less than 70% imperviousness, it's reclassified as Open
d. If pixel has more than 70% imperviousness, then it's reclassified as Compact

- The TCD has continuous percentage values, but they can be reclassified into a specific LCZ land cover type classification. Despite this, it's not yet a final LCZ class type as it will need to be reclassified later together with other values:
    a. If the density is less than 100%, it's classified as LCZ A
    b. If the density is less than 50%, it's classified as LCZ B
    c. If the density is less than 20%, it's classified as LCZ C
    d. If the density is less than 10%, it's classified as LCZ D

Once each pixel has been calculated according to the previous reclassification logic, it's necessary to do yet another raster calculation for each pixel, using GRA, DLT and TCD together to create a more accurate LCZ land cover (LCZ A to D) classification, based on the following logic.

- If the TCD pixel is LCZ D and is also GRA, then it's reclassified as LCZ D
- If the TCD pixel is LCZ A and is also Deciduous DLT, it's reclassified as LCZ A
- If the TCD pixel is LCZ A and is also Coniferous DLT, it's reclassified as LCZ A
- If the TCD pixel is LCZ B and is also Deciduous DLT, it's reclassified as LCZ B
- If the TCD pixel is LCZ B and is also Coniferous DLT, it's reclassified as LCZ B
- If it's only a TCD Pixel, then it stays classified as is.

2.3.2.2. CREATING A BASELINE

A baseline is the feature class, in ArcGIS terms, that will contain all the polygons and data required to classify the LCZ. This feature class is based on the FUAs data of the study area, for example Lisbon or Aarhus, but clipped to the extent, or bounding box, which the user drew on the map. The land cover types contained in the Urban Atlas (UA) are not detailed enough outside of the urban core, deeming it necessary to complement with CLC data on those areas. For that, it is necessary to overlay UA data with CORINE Land Cover (CLC) data, using an

Identity operation (ESRI, 2024a). This means that new polygons will be created based on UA and CLC intersected features.

The UA land cover types which require more detailed information are:

- Arable Land
- Permanent Crops
- Complex and mixed cultivation patterns
- Herbaceous vegetation associations
- Wetlands

All the areas with these types will overlay with CLC, resulting in a more complex geometry.

Figure 10 – Lisbon Functional Urban Area unchanged

Figure 11 – Lisbon Functional Urban Area with CLC

As shown in Figure 10 – Lisbon Functional Urban Area unchanged and Figure 11 – Lisbon Functional Urban Area with CLC, after doing the Identity Overlay, this area has been enriched with CLC data. In this scenario, the FUA is identified with code 32000, which means Herbaceous vegetation associations, while CLC identifies this area as being discontinuous urban fabric. As mentioned by (Oliveira et al., 2020a), UA features are chosen to represent urban classes, while CLC features are preferred in non-urban land cover typologies.

As explained by (Oliveira et al., 2020a), additional inputs from OSM are necessary to enrich the LCZ's 8 and 10 classes, hence the need to do a spatial join(ESRI, 2024c) of the OSM data with the baseline.

The baseline classification is in line with the logic applied by previous authors (Oliveira et al., 2020a), with the differences being, a more streamlined approach which requires no manual work from GIS specialist, faster processing with very similar results. The results may differ in some cases due to differences in implementation of specific spatial operations. The ArcGIS toolbox uses a very efficient Identity operation, while the GeoPandas identity operation may take an hour to process which is not suitable, therefore it was necessary to make a custom implementation based on the explanation found on (ESRI, 2024a). After careful testing, it was found to deliver the same results at a fraction of processing time - in a couple of minutes.

The classifier also has a custom-made Spatial Join operation, as the ArcGIS toolbox uses a very specific type of Spatial Join, a is one to one join, which according to ArcGIS documentation (ESRI, 2024c), if multiple join features are found that have the same spatial relationship with a single target feature, the attributes from the multiple join features will be aggregated using a field map merge rule. There is nothing similar implemented open source and the documentation is not detailed nor specific enough to program a similar implementation, as such the current spatial join operation implemented in the classifier is not delivering the same results and will need to be improved further.

Once this initial operation is completed, the geometry of the baseline will remain unchanged, only the data associated with each feature will be enriched in each step.

2.3.2.3. CLASSIFICATION

The LCZ classifier requires the usage of Local Statistics to know how many 10x10m cells of each type of LCZ are contained within each feature of the LCZ Baseline. This is the base information the classifier will use to assess the correct LCZ type to assign to the feature. Once the calculation of the Local Statistics is done, the classifier starts. As a reminder, this step does not yet contain building height data, which means that this first classification will be simplified, using LCZ 123 instead of separate LCZ 1, LCZ 2 and LCZ 3, and LCZ 456 instead of LCZ 4, LCZ 5 and LCZ 6.

LCZ 1 through 6 depends not only on the built-types, but also on the amount of greenery. For that, we calculate the percentage of LCZ A, B, C, D per feature and store that data on the feature. This is output as stored on the baseline with the field TREES_SUM.

Afterwards, the classifier proceeds to convert the Urban Area code of each feature in the Baseline into an LCZ type, so that it's used in the final reclassification together with the other classifications done so far. This results in an additional field in the LCZ Baseline, named LCZ_UA

Table 2 – LCZ classes from Urban Area code

Urban Area	OpenStreetMap	LCZ_UA
Continuous urban fabric (11100)		123
Discontinuous dense urban fabric (11210)		456
Discontinuous medium-density urban fabric (11220)		456
Discontinuous low-density urban fabric (11230)		9
Discontinuous very low-density urban fabric (11240)		9
Isolated Structures (11300)		9
Industrial commercial, public, military- and private units (12100)	Industrial	10
Industrial commercial, public, military- and private units (12100)	not industrial	8
Fast transit roads and associated land (12210)		E
Other roads and associated land (12220)		E
Railways and associated land (12230)		E
Port areas (12300)		E
Airports (12400)		E
Mineral extraction sites (13100)		F
Dump sites (13200)		F
Construction sites (13300)		F
Land without current use (13400)		F
Green urban areas (14100)		ABCD
Sport and leisure facilities (14200)		ABCD
Arable land (annual crops) (21000)		D
Permanent crops (22000)		ABCD
Pastures (23000)		D
Complex and mixed cultivation patterns (24000)		ABCD
Orchards (25000)		B
Forests (31000)		A
Herbaceous vegetation associations (32000)		ABCD
Open spaces with little or no vegetations (33000)		F
Wetlands (40000)		ABCD

Water (50000)	G

Similar logic is applied to convert CLC to LCZ types, each corresponding CLC type corresponds to a LCZ type, and a new field called LCZ_CLC is created in the feature.

Table 3 - LCZ classes from CORINE Land Cover codes

CLC Code	Corine Land Cover Type	LCZ_CLC
211	Non-irrigated arable land	D
212	Permanently irrigated crops	D
213	Rice fields	D
221	Vineyards	C
222	Fruit trees and berry plantations	B
223	Olive groves	B
231	Pastures	D
241	Annual crops associated with permanent crop	D
242	Complex cultivation patterns	D
243	Land principally occupied by agriculture, with significant areas of natural vegetation	D
244	Agro-forestry areas	B
311	Broad-leaved forest	A
312	Coniferous forest	A
313	Mixed forest	A
321	Natural grassland	D
322	Moors and heathland	C
323	Sclerophyllous vegetation	C
324	Transitional woodland/shrub	C
331	Beaches, dunes, sands	F
332	Bare rock	F
333	Sparsely vegetated areas	F
334	Burnt areas	F
335	Glaciers and perpetual snow	G
411	Inland water bodies	D
412	Peatbogs	D
421	Salt marshes	G
422	Salines	G
423	Intertidal flats	G
511	Water courses	D
512	Water bodies	D
521	Coastal lagoons	G
522	Estuaries	G
523	Sea and ocean	G

Afterwards, for each feature, the classifier compares which Land Cover Type from A to D has the biggest area and sets the respective LCZ for the TCD, creating a new field in the LCZ Baseline called LCZ_TCD. For example, for a determined feature, if LCZ A has more covered area than LCZ B, LCZ C or LCZ D, then the classifier defines LCZ_TCD field as being A. LCZ_TCD can also be empty if a feature has no Land Cover type from A to D.

The final LCZ reclassification is done next, it uses all the reclassifications applied so far to process a new reclassification.

As a recap to simplify the understanding of the algorithm, Table 4 - LCZ Baseline fields summary contains the fields used for the LCZ reclassification and a brief explanation of what it represents.

Table 4 - LCZ Baseline fields summary

LCZ Baseline Field	Description
LCZ_UA	Corresponding LCZ type for the urban area. Verify Table 2 – LCZ classes from Urban Area code to know which values this field may contain.
LCZ_CLC	Corresponding LCZ type for the CORINE Land Cover. Verify Table 3 - LCZ classes from CORINE Land Cover codes to know which values this field may contain.
LCZ_TCD	This field represents Tree Cover Density, which may contain values from A to D.
TREES_SUM	Despite the name, it stores the percentage of green area per feature.

The LCZ reclassification algorithm is as follows:

- If LCZ_UA is not ABCD and LCZ_UA is not A and LCZ_UA is not B, then it should use the current LCZ_UA value

- If LCZ_UA is not ABCD and LCZ_TCD is empty and LCZ_CLC is empty, it should use the LCZ_UA values
- If LCZ_UA is ABCD and LCZ_TCD is empty and CLC is empty, then it should be LCZ 5
- If LCZ_UA is ABCD and LCZ_TCD is not empty and TREES_SUM is more or equal than 25 or LCZ_CLC is empty, then it should use the LCZ_TCD
- Otherwise use the LCZ_CLC

2.3.2.4. EXPAND WITH BUILDING HEIGHT

The building height is fundamental to define the LCZ 1 to 6 classification, as the height of the buildings impacts the Urban Heat Island intensity. The classifier uses the Urban Atlas Building Height dataset from CLMS to further improve the LCZ classification and change the simplified LCZ 123 and LCZ 456 into specific LCZ 1 to 6 types. Still, some features lack building height information, and these will be classified as LCZ 123 and LCZ 456. This is because the dataset only provides height information for core urban areas of selected cities (capitals) in EEA 38 (EEA, 2022b).

To recap, LCZ 1 and LCZ 4 are High-rise, LCZ 2 and LCZ 5 are Mid-rise, LCZ 3 and LCZ 6 are Low-rise.

This project's classifier defines the following:

- Low-rise: From 1 to 10m
- Mid-rise: From 11m to 30m
- High-rise: From 31m

This is different from what other authors (Oliveira et al., 2020a)(Stewart & Oke, 2012) define, as they define that High-rise starts at 25m. Although the prepackaged datasets from CLMS contain specific building height information as a continuous variable (i.e., data of type 'float), this project automates the downloading procedure by using the CLMS's API to retrieve the datasets, which, in the case of Building Height, returns a raster dataset containing values as a classification, in interval, such as 20 to 30m, then 30-60m. Regardless of this, according to the original LCZs authors (Stewart & Oke, 2012) a high-rise building is tens of stories. A story is typically defined as approximately 3 to 3.3 meters in height. Based on this definition, a High-

rise building is more than 30 meters tall. Consequently, this project adopts this definition, differing from the previous approach (Oliveira et al., 2020a), and establishes that a High-rise building exceeds 30 meters in height. This will result in a slightly different classification between LCZ 1 and LCZ 2, LCZ 4 and 5, which will be verified in the accuracy assessment.

Figure 12 – Lisbon Local Climate Zones before rasterization shows the vector based LCZ classification, a quick analysis shows that the south-center of Lisbon, which corresponds to the city's inner core, is primarily contained classified as LCZ 1 to 10, with small green areas containing tree coverage (LCZs from A to B). On the north-west and easternmost areas, which correspond to the city's limits, low-rise buildings and lower vegetation prevails (LCZs C to D), signaling the transition between Lisbon and the adjacent to suburban municipalities.

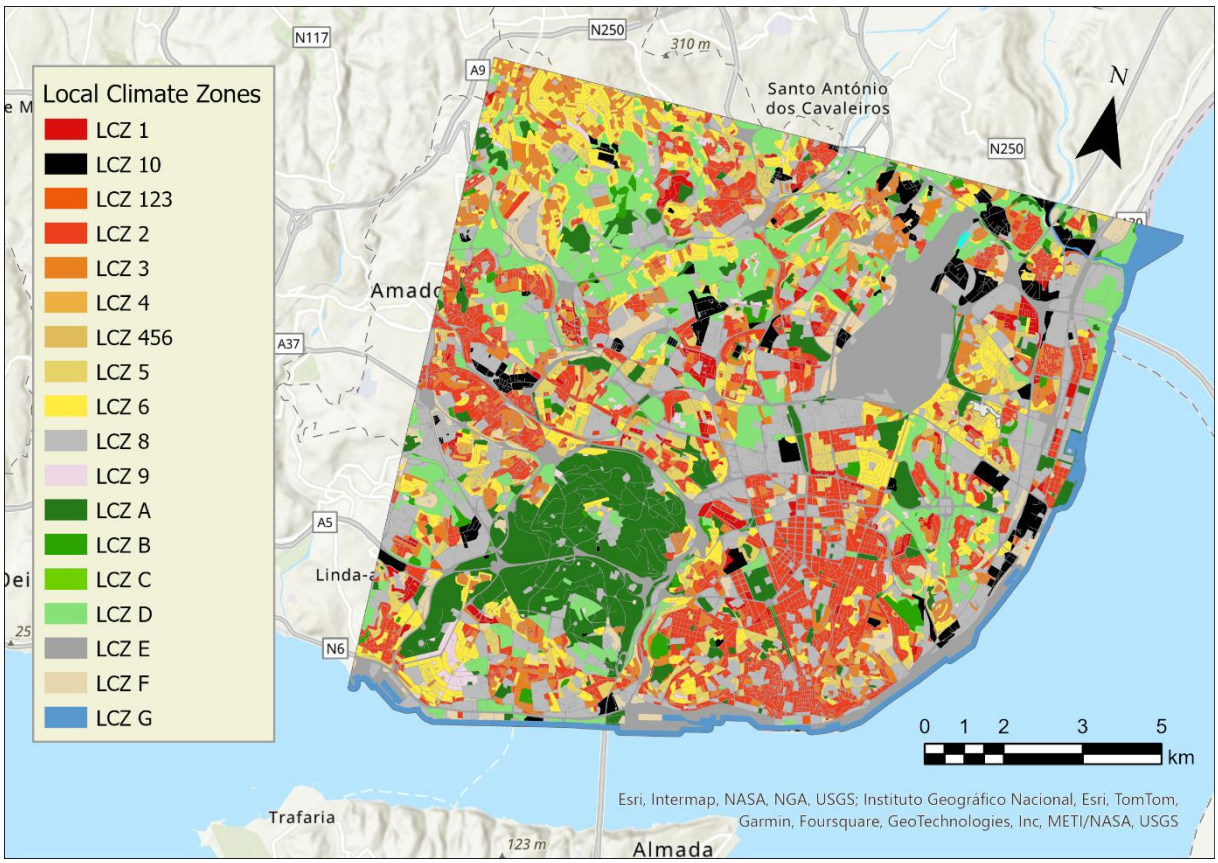


Figure 12 – Lisbon Local Climate Zones before rasterization

2.3.2.5. RASTERIZATION

The rasterization process converts the resulting vector LCZ baseline into a raster. The spatial resolution adopted is the same as (Oliveira et al., 2020a), which is 50x50m cell size. Previous studies (Oliveira et al., 2020a) used the ArcGIS's Polygon to Raster conversion geoprocessing

tool with a Cell assignment type of Maximum Combined Area (ESRI, 2024b), which combines features with the same value, and the largest area within the cell will determine the value to be assigned to the cell. There is currently no open-source tool or library that can be used to do this, therefore a custom implementation had to be programmed. The custom implementation delivered the same results, although it's, currently, a slower algorithm.

The algorithm uses a combination of techniques:

- Dissolve features with the same LCZ type. This is an optimization step to reduce the number of features to intersect.
- Create a fishnet of 50x50m, this fishnet should be the extent of the LCZ Baseline.
- Intersect the output of step 1 with the fishnet. This will be the baseline.
- Within each cell, dissolve features with the same LCZ type.
- Calculate the area of each dissolved feature within the cell.
- Create a new grid with the same size and spatial resolution as the fishnet previously created. For each cell in the grid, assign the biggest feature calculated in step 4 in the corresponding fishnet cell, assigning no-data in cells where there is no LCZ type.
- Rasterize the output of step 6.

There are plans to open source this custom implementation as it was implemented in a generic approach, it should be possible to be used in any Python code base, its only dependencies are Rasterio and GeoPandas libraries. It is still necessary to do some performance optimizations before sourcing the code available, as it can take up to 30 minutes to complete in bigger areas.

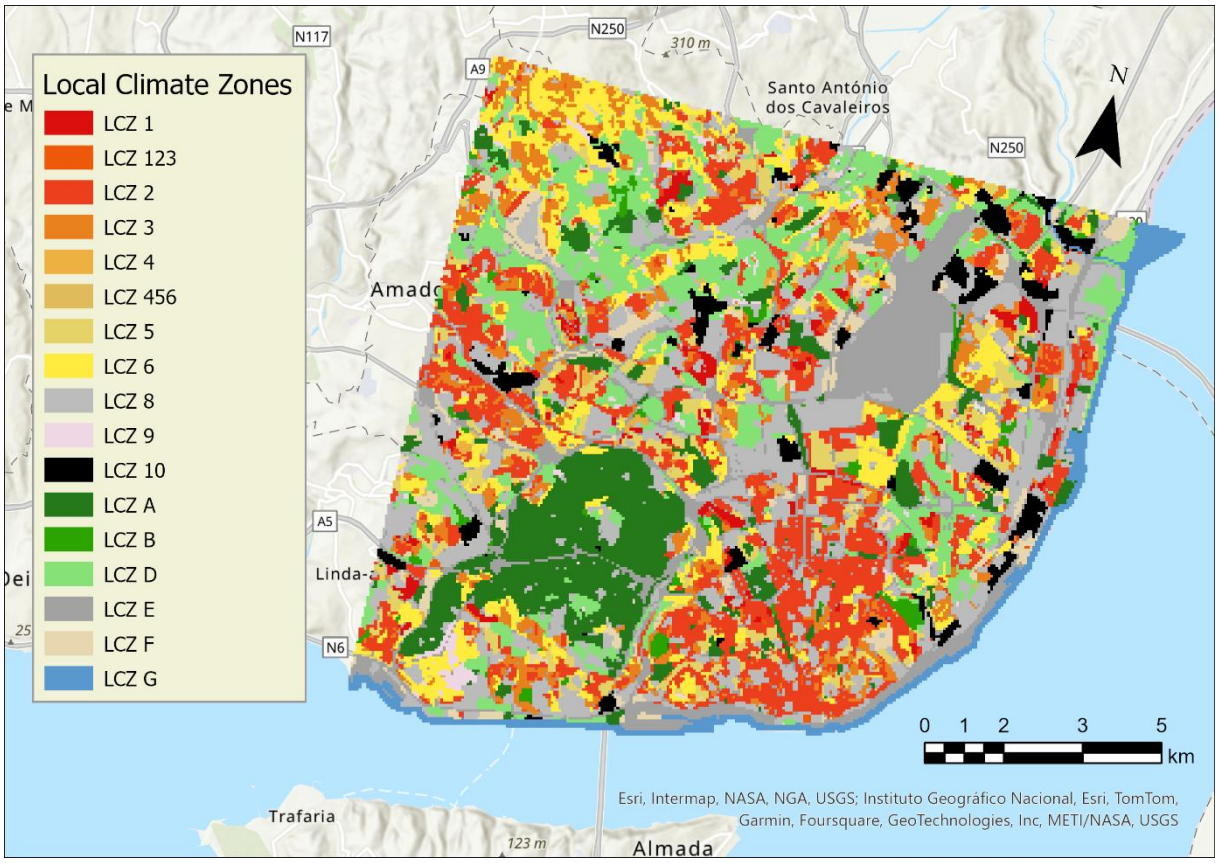


Figure 13 – Raster of Local Climate Zone

As shown in Figure 13 – Raster of Local Climate Zone, the raster map is less detailed than the vector map shown in Figure 12 – Lisbon Local Climate Zones before rasterization. Due to the downscaling, LCZ C is no longer present in the map because it was not the predominant LCZ type in any of the 50x50m squares used in the rasterization. This is a characteristic of prevailing vegetation types in the Lisbon area, which correspond mostly to contrasting tree covered urban parks and vacant land occupied by natural grasslands.

2.3.2.6. UPLOAD TO GEOSERVER

Once the rasterization step finishes, the algorithm then uploads the LCZ raster classification output to GeoServer using its REST API. Before the algorithm can do its job, GeoServer needs to be configured by a system administrator and initially set up with a Workspace named lczcgen, and a configured Style using a pre-created Styled Layered Descriptor (SLD) file, which according to best-practices (OGC, 2007) defines an encoding that extends the WMS standard to allow user-defined symbolization and coloring of the geographic feature and coverage data.

The Web GIS platform will be able to retrieve the raster and display it on the map, using Leaflet's WMS capabilities, applying the correct legend based on the SLD style.

Regarding the user's participation and usage of the platform, they will receive an email informing them that the data is ready with a link they can visit.

They will be able to interact with the map by zooming in and out, also pan left or right. They will be able to click on the map to check what the LCZ type is on the clicked cell, using WMS's GetFeatureInfo operation. If the user feels the data is incorrect, they may submit their suggestion by adding a label in that location, referring to a different LCZ type. Nevertheless, the self-labelling functionality serves only as a feedback tool and their suggestions will not reflect in a direct raster data change. This data is saved on a PostGIS database, as it will be required to know exactly the location which the user clicked on the map, which represents a point geometry, with the properties being the LCZ classified by the algorithm and the correction.

This platform is a Participatory GIS, the users can see LCZ's generated from other users and give feedback to the platform. Once enough data is gathered for 2 to 3 years, it will be used as input for a machine learning algorithm, it will need to be pre-processed, some exploratory spatial data analysis as well, outliers need to be analyzed, among other operations that may be required. Once the data is cleaned up, it will be possible to use a supervised machine learning algorithm to improve the output of the LCZ classification, namely in areas outside of the CLMS layers domain.

2.3.3. Accuracy Assessment

The chosen methodology involved creating confusion matrices using a set of points selected based on a specific strategy. This approach yields metrics such as Overall Accuracy, User's Accuracy, Producer's Accuracy, and the Kappa coefficient. In this project, the ArcGIS Accuracy Assessment tools were used, due to the easiness of implementation, and 'one-time-only' nature of the task. The first step was to create Accuracy Assessment points, choosing a minimum of 50 per class. Since there are 19 possible classes of LCZ, 50 times 19 is 950. This will generate a minimum of 950 points with empty Ground Truth and Classified fields. The strategy chosen for the distribution was Stratified Random, which according to ArcGIS's documentation, creates points that are randomly distributed within each class, and each class

has several points proportional to its relative area. The reason for choosing this sampling strategy is that some areas of the functional urban areas are much bigger than others and therefore need more accuracy points for validation.

The ground truth is the output generated by the toolbox created by previous work (Oliveira et al., 2020a). The classified field is the output generated by algorithm developed in this project.

The last step is to run the Compute Confusion Matrices tool which uses the Accuracy Assessment points previously generated and will produce a table with user's and producer's accuracy, and a Kappa result. A positive result was considered to have a Kappa higher than 0.8.

The result will be important to understand which LCZ type is more or less accurate and is an important tool to understand which part of the algorithm needs to be improved.

3. RESULTS AND DISCUSSION

3.1. WEB GIS PLATFORM

The Web GIS platform is online, it can be viewed on the website page http://lczcgen.atlanticsense.com, it's ready to accept requests from users to generate a LCZ classification of a given area selected by the user as shown in Figure 14 - LCZC Generator website. The full processing time takes around 1 hour for a 140 km^2 area, from requesting data to CLMS until the classifier outputs a raster file to GeoServer, although the waiting time from when the user requests the data until it is processed depends on the queue at CLMS's side. If the queue is too big, it may take up to a day to process. The duration of the classification varies depending on the amount of data to process, so the bigger the FUAs, the longer it takes to finish. The classifier will output a GeoTIFF raster file, representing a grid of 50x50m, in which each pixel represents the LCZ type with the greatest area in it.

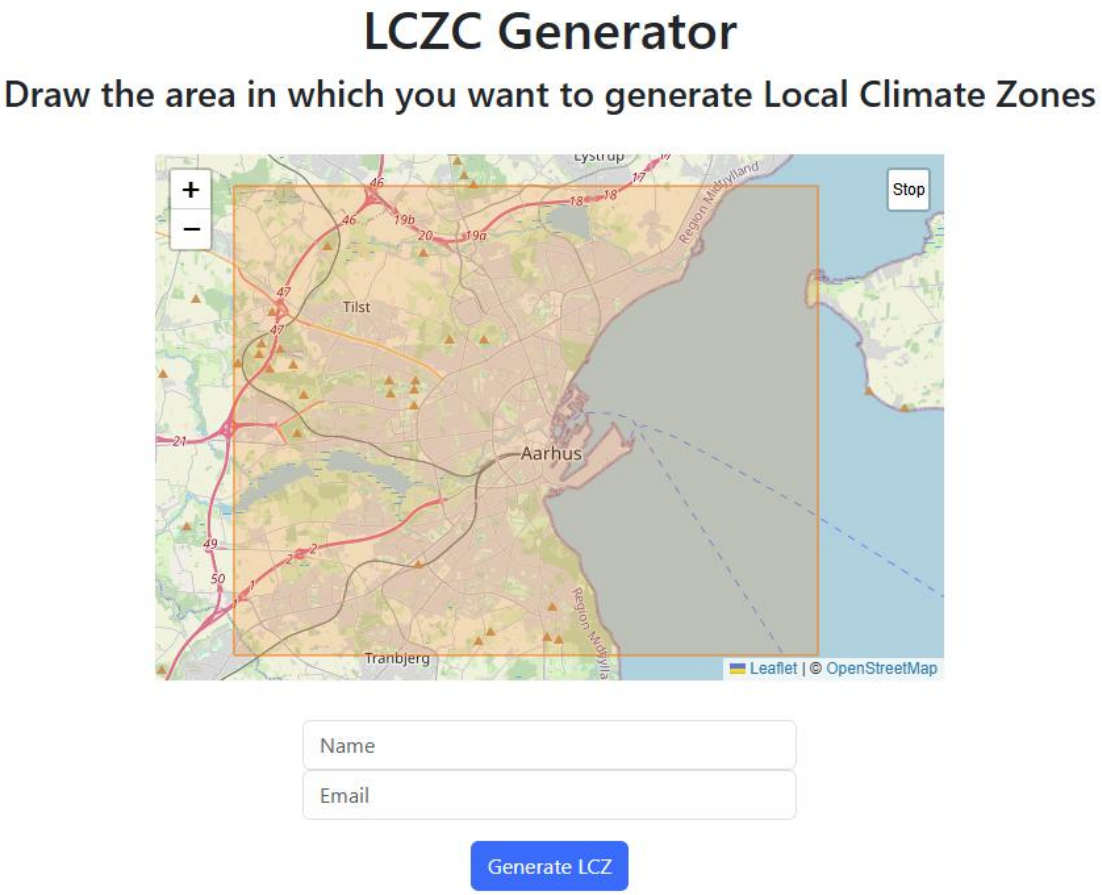


Figure 14 - LCZC Generator website

When the data reaches GeoServer, the user will be informed via email that the visualization is ready, with a link in the email so that he can directly view it on the platform. He will also be informed if something wrong happened during the process, as according to testing, CLMS sometimes fail silently, failing to start processing and not informing the clients. Normally if the process is not done within 24 hours, the background job that is waiting for new data from CLMS will send an email to the user to inform them that the process didn't happen and cancel the request.

Once the data is ready and the email is sent, it contains a direct link to the platform to verify the LCZ classification. This map, as seen in Figure 15 - View map in the Web GIS platform, will allow for panning, zooming in/out and downloading the data so that it can be analyzed more carefully in a GIS tool such as QGIS.

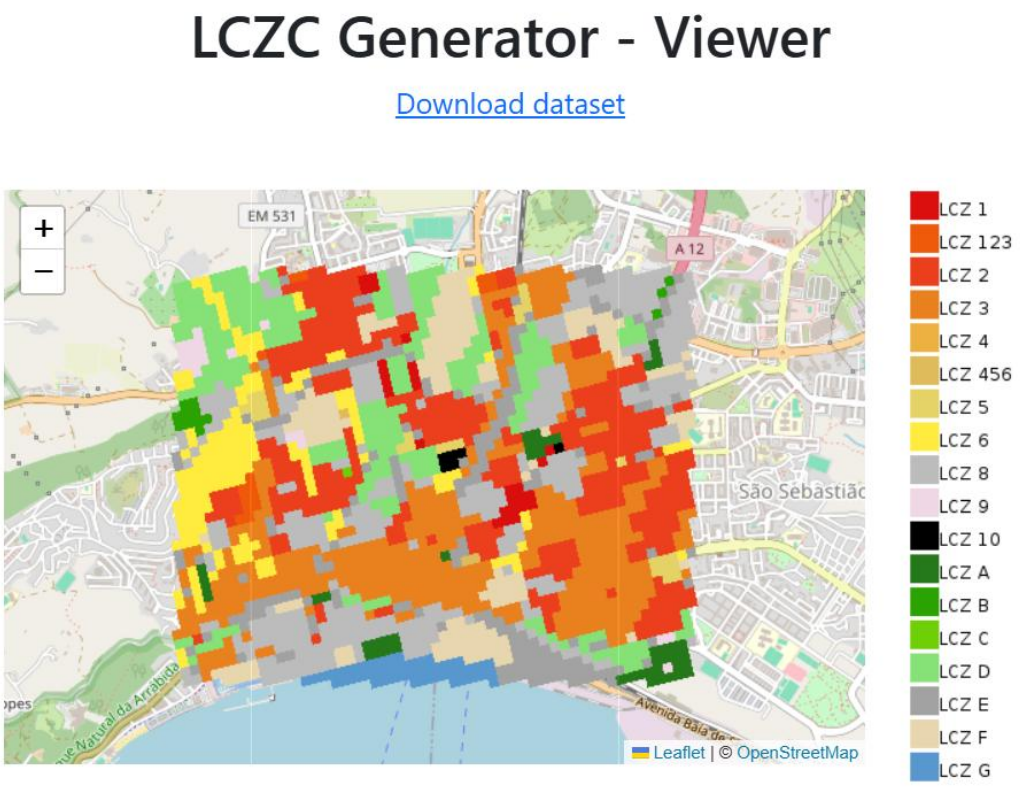


Figure 15 - View map in the Web GIS platform

The user can also report on the accuracy of the map by clicking on the map, as see in Figure 16 - Labeling, which will show a pop-up with the LCZ where to add an alternative label, for documentation purposes.

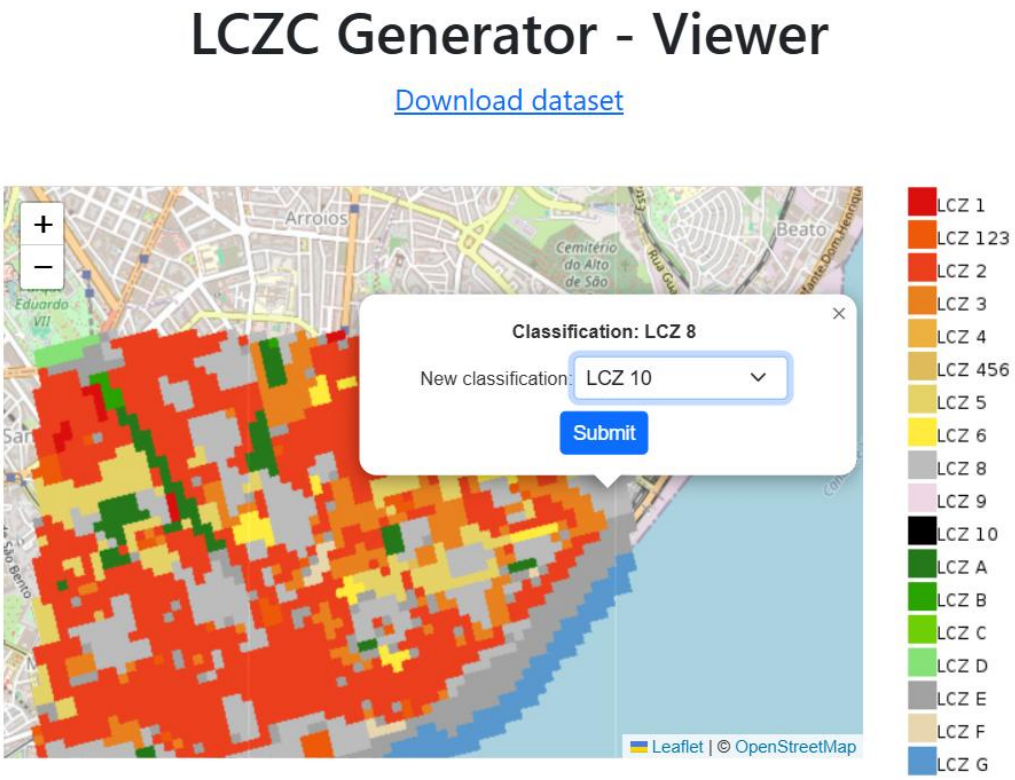
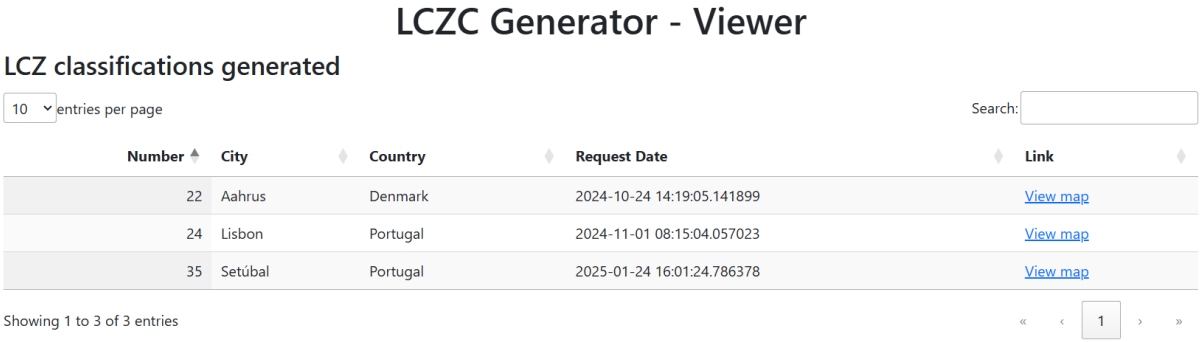


Figure 16 - Labeling

As this is a Participatory GIS, the user can see a list containing every generated LCZ, search for specific LCZ by filtering by country. This can be useful in case there is already a generated LCZ for his study area as he will be able to download it in GeoTIFF format.



3.2. Accuracy Assessment

The accuracy assessment realized for Lisbon can be found in Appendix A (Table 5 - Lisbon Accuracy Assessment), and Appendix B (Figure 17 – Lisbon's LCZ classification with ArcGIS toolbox, Figure 18 – Lisbon's LCZ Classification with this project's classifier) contains a side-by-

side visual comparison of the LCZ classification realized with the ArcGIS toolbox and a LCZ classification realized with this project's classifier.

The results show a Kappa of approximately 91%, most types resulted in equal or similar results, except for LCZ 123, LCZ 2, LCZ 5 and LCZ 10. LCZ 5 has an accuracy of 65%, it represents mid-rise buildings with open spaces, the algorithm has some difficulties due to the presence of low-rise buildings near the limits of the boundaries of the area, which get excluded from the zonal statistics calculation. On (Oliveira et al., 2020a) these are normally classified as LCZ 6, which are open low-rise buildings. There are also some cases which have a very low density of buildings, as such in those scenarios the algorithm uses the generalized LCZ 456 classification, whereas (Oliveira et al., 2020a) consider the low density due to the different approach of vector vs raster, so that 0.1% of density impacts the classification.

Regarding LCZ 10, heavy industry, it has 67% accuracy due to the difference in spatial join algorithm between this project and (Oliveira et al., 2020a). As will be seen in Aarhus's accuracy assessment, it's a weak point that needs to be worked on.

As for LCZ 123 differences, it has 71% accuracy. After some analysis, it is because many LCZ 123 areas don't have a full cell of $10m^2$ inside, and as such, the Local Statistics algorithm is excluding them.

Finally, LCZ 2 has 78% accuracy, it's mostly confusing LCZ 2 with LCZ 1 due to the difference in the high-rise and mid-rise classification between this project and (Oliveira et al., 2020a), as explained previously, in which (Oliveira et al., 2020a) considers high-rise to start at 25m, but this project considers starting at 30m.

As for the accuracy assessment of Aarhus, as it shows in Appendix A (Table 6 – Aahrus Accuracy Assessment), the Kappa of the classification is approximately 97%, Appendix B (Figure 19 – Aahrus's LCZ classification with ArcGIS toolbox, Figure 20 – Aahrus's LCZ Classification with this project's classifier) shows the visual comparison between both LCZ classifications. The accuracy assessment shows significantly different results in LCZ 3, LCZ 4 and LCZ 10. All other types show a high percentage of accuracy with a minimum of 94% accuracy.

LCZ 3, compact low-rise, has 55% accuracy. After careful analysis, it was identified that it is due to how the zonal statistics algorithm is culling cells touching the limits of the feature in which

it is contained. This is because many areas in Aarhus have mid-rise buildings near the boundary of the feature and low-rise buildings in the center.

LCZ 4, open high-rise, has 50% accuracy, due to the difference of classification regarding high-rise buildings and mid-rise buildings. As mentioned previously, this project classifies high-rise buildings having more than 30 meters, while (Oliveira et al., 2020a) classifies them as having more than 25 meters. In several locations of the study area, many buildings are 26 to 29 meters tall, which influences the difference in accuracy.

As for LCZ 10, heavy industry, it has 69% accuracy, it has a low percentage as found in Lisbon for the same reasons.

Despite these differences, as the purpose of the project is to support urban climate modelling in a 100-200m pixel size grid, they were deemed as not having a significant negative impact on the results at this scale.

4. CONCLUSIONS

LCZs have been adopted worldwide as the standard by the research community in urban climate-related studies (Stewart & Oke, 2012) and they have been implemented in multiple studies, although they are normally adapted to the local datasets, so they are normally not generic enough to be used elsewhere. Previous authors have proposed a GIS approach (Oliveira et al., 2020b) that leverages the data from CLMS so that it can be used, with a high degree of certainty, in Europe, but it depends on ArcGIS and requires manual intervention to interact with the tool. The main requirement for this project was to make an open-source tool that can provide the same algorithm and same results, but in an automated way and whenever required.

As previous authors (Oliveira et al., 2020a) used some ArcGIS's geospatial operations that are not readily available in Rasterio or GeoPandas or any other open-source tool, such as spatial join one to one approach and the rasterization process, these needed to be custom implemented in this project, with the spatial join one to one operation needing some improvement, as the accuracy assessment shows.

In general, these results were positive, as they have a Kappa coefficient over 0.9 which gives it enough confidence to be used in future projects like the ESA-funded CLIM4Cities project which use LCZ as a predictor for machine learning weather predictions downscaling.

For future research, the model needs to improve some aspects already identified such as better spatial join to improve LCZ 10 accuracy and general improvement to the rasterization process which was identified to be slowest part of the classifier. The algorithm also needs to improve the Zonal Statistics calculation to account for the presence of raster cells near the functional urban area boundaries, as these affected the identification of building heights.

This model will need to be adapted to run outside of the areas covered by CLMS, it will require substantial changes to work with different input data, because the current algorithm heavily depends on the FUAs structure to create the baseline for classification.

Furthermore, to have more up to date datasets and to avoid issues such as building heights being from 2012, or using functional urban area data from 2018, work can be carried out to start doing a satellite-based data classification on demand, using new algorithms such as

Segment Anything. This is important as cities are growing fast, with cities projected to hold 68% of the population by 2050, according to (UN Department of Economic and Social Affairs, 2018).

Usage of Computer Vision to analyze street-level images, such as using Mapillary platforms should improve the fidelity of the data by improving the understanding of the area being analyzed, their reflectance (albedo), the sky-view factor, and other aspects mentioned on the (Stewart & Oke, 2012) publication.

The last identified action point for future improvement is that it's expected that users will report on inaccuracies of the map, and we may run a machine learning algorithm to improve the classification before rasterization. This is projected to be done in 2 to 3 years, depending on the amount of gathered data.

BIBLIOGRAPHICAL REFERENCES

Alexander, P. J., Fealy, R., & Mills, G. M. (2016). Simulating the impact of urban development pathways on the local climate: A scenario-based analysis in the greater Dublin region, Ireland. *Landscape and Urban Planning*, *152*, 72–89. https://doi.org/10.1016/J.LANDURBPLAN.2016.02.006

EEA. (2019, June 14). *CORINE Land Cover 2018*. https://land.copernicus.eu/en/products/corine-land-cover/clc2018

EEA. (2020a, August 18). *Grassland 2018*. https://land.copernicus.eu/en/products/high-resolution-layer-grassland/grassland-2018

EEA. (2020b, August 18). *Imperviousness Density 2018*. https://land.copernicus.eu/en/products/high-resolution-layer-imperviousness/imperviousness-density-2018

EEA. (2020c, September 18). *Dominant Leaf Type 2018*. https://land.copernicus.eu/en/products/high-resolution-layer-dominant-leaf-type/dominant-leaf-type-2018

EEA. (2020d, September 18). *Tree Cover Density 2018*. https://land.copernicus.eu/en/products/high-resolution-layer-tree-cover-density/tree-cover-density-2018

EEA. (2021, July 16). *Urban Atlas Land Cover/Land Use 2018*. https://land.copernicus.eu/en/products/urban-atlas/urban-atlas-2018

EEA. (2022a). *CLMS API*. https://eea.github.io/clms-api-docs/introduction.html

EEA. (2022b, October 6). *Building Height 2012*. https://land.copernicus.eu/en/products/urban-atlas/building-height-2012

ESRI. (2024a). *Identity (Analysis)*. https://pro.arcgis.com/en/pro-app/latest/tool-reference/analysis/identity.htm

ESRI. (2024b). *Polygon to Raster (Conversion)*. https://pro.arcgis.com/en/pro-app/latest/tool-reference/conversion/polygon-to-raster.htm

ESRI. (2024c). *Spatial Join (Analysis)*. https://pro.arcgis.com/en/pro-app/latest/tool-reference/analysis/spatial-join.htm

Hidalgo, J., Dumas, G., Masson, V., Petit, G., Bechtel, B., Bocher, E., Foley, M., Schoetter, R., & Mills, G. (2019). Comparison between local climate zones maps derived from administrative datasets and satellite observations. *Urban Climate*, *27*, 64–89. https://doi.org/10.1016/J.UCLIM.2018.10.004

IPCC. (2014). *Climate Change 2014: Impacts, Adaptation, and Vulnerability - Part A*.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, *15*(3). https://doi.org/10.1127/0941-2948/2006/0130

Mills, G., Cleugh, H., Emmanuel, R., Endlicher, W., Erell, E., McGranahan, G., Ng, E., Nickson, A., Rosenthal, J., & Steemer, K. (2010). Climate Information for Improved Planning and Management of Mega Cities (Needs Perspective). *Procedia Environmental Sciences*, *1*(1), 228–246. https://doi.org/10.1016/J.PROENV.2010.09.015

OGC. (2007). *Styled Layer Descriptor* . https://www.ogc.org/publications/standard/sld/

Oke, T. R. (1982). The energetic basis of the urban heat island. *Quarterly Journal of the Royal Meteorological Society*, *108*(455). https://doi.org/10.1002/qj.49710845502

Oliveira, A., Lopes, A., & Niza, S. (2020a). Local climate zones classification method from Copernicus land monitoring service datasets: An ArcGIS-based toolbox. *MethodsX*, *7*. https://doi.org/10.1016/j.mex.2020.101150

Oliveira, A., Lopes, A., & Niza, S. (2020b). Local climate zones in five southern European cities: An improved GIS-based classification method based on Copernicus data. *Urban Climate*, *33*. https://doi.org/10.1016/j.uclim.2020.100631

OSM. (2024). *OpenStreetMap Data Extracts*. https://download.geofabrik.de/

Statistics Denmark. (2024, January). *Statistics Denmark*. https://www.statbank.dk/BY1

Statistics Portugal. (2024, January). *Statistics Portugal*. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0008272&contexto=bd&selTab=tab2

Stewart, I. D., & Oke, T. R. (2012). Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, *93*(12). https://doi.org/10.1175/BAMS-D-11-00019.1

UN Department of Economic and Social Affairs. (2018). 68% of the world population projected to live in urban areas by 2050, says UN. *United Nations News*.

APPENDIX A

The following two figures are the results of the accuracy assessments done for Lisbon and Aahrus

Table 5 - Lisbon Accuracy Assessment

LCZ	1	123	2	3	4	456	5	6	8	9	10	A	B	D	E	F	G	Total	User Accuracy	Kappa
1	15	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0,94	0
123	0	10	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0,71	0
2	11	0	107	20	0	0	0	0	0	0	0	0	0	0	0	0	0	138	0,78	0
3	0	0	0	64	0	0	0	0	0	0	0	0	0	0	0	0	0	64	1	0
4	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0
456	0	0	0	0	1	10	0	1	0	0	0	0	0	0	0	0	0	12	0,83	0
5	0	0	0	0	5	0	32	12	0	0	0	0	0	0	0	0	0	49	0,65	0
6	0	0	0	0	2	0	0	85	0	0	0	0	0	0	0	0	0	87	0,98	0
8	0	0	0	0	0	0	0	0	137	0	7	0	0	0	0	0	0	144	0,95	0
9	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	10	1	0
10	0	0	0	0	0	0	0	0	12	0	24	0	0	0	0	0	0	36	0,67	0
A	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	100	1	0
B	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	10	1	0
D	0	0	0	0	0	0	0	0	0	0	0	0	0	112	0	0	0	112	1	0
E	0	0	0	1	0	0	0	0	0	0	0	0	0	0	118	0	0	119	0,99	0
F	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	33	0	34	0,97	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	22	1	0
Total	26	10	109	88	10	10	32	98	149	10	31	100	11	112	118	33	22	969	0	0
Producer Accuracy	0,58	1	0,98	0,73	0,2	1	1	0,87	0,92	1	0,77	1	0,91	1	1	1	1	0	0,92	0
Kappa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,91

Table 6 – Aahrus Accuracy Assessment

LCZ	1	123	2	3	4	456	5	6	8	9	10	A	B	C	D	E	F	G	Total	User Accuracy	Kappa
1	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	1	0
123	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	1	0
2	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	16	0,94	0
3	0	1	3	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0,56	0
4	0	0	0	0	3	0	3	0	0	0	0	0	0	0	0	0	0	0	6	0,5	0
456	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	5	1	0
5	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	17	1	0
6	0	0	0	0	0	1	5	198	0	0	0	0	0	0	0	0	0	0	204	0,97	0
8	0	0	0	0	0	0	0	0	71	0	1	1	0	0	0	0	0	0	73	0,97	0
9	0	0	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0	28	1	0
10	0	0	0	0	0	0	0	0	8	0	18	0	0	0	0	0	0	0	26	0,69	0
A	0	0	0	0	0	0	0	0	0	0	0	98	0	0	0	0	0	0	98	1	0
B	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	10	1	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	8	1	0
D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	413	1	0	0	414	0,998	0
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50	0	0	50	1	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	11	1	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	29	1	0
Total	9	6	18	5	3	6	25	198	79	28	19	99	10	8	413	52	11	29	1018	0	0
Producer Accuracy	1	0,83	0,83	1	1	0,83	0,68	1	0,90	1	0,95	0,99	1	1	1	0,96	1	1	0	0,98	0
Kappa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,97

APPENDIX B

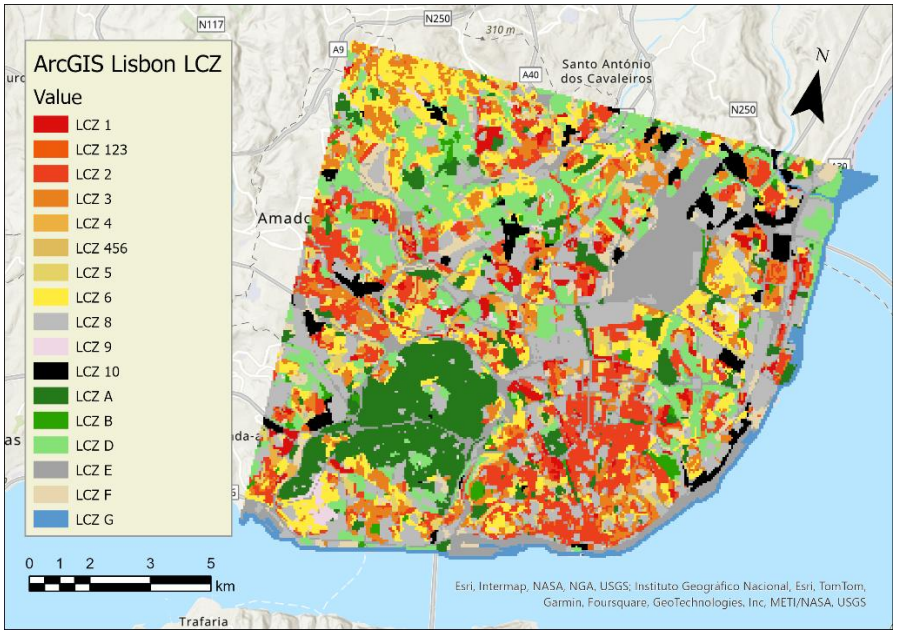


Figure 17 – Lisbon's LCZ classification with ArcGIS toolbox

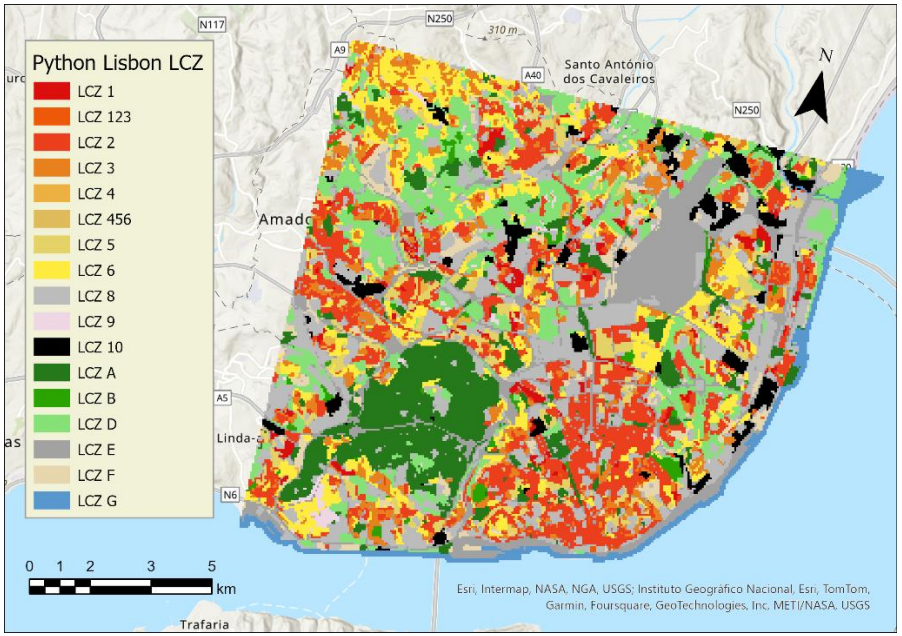


Figure 18 – Lisbon's LCZ Classification with this project's classifier

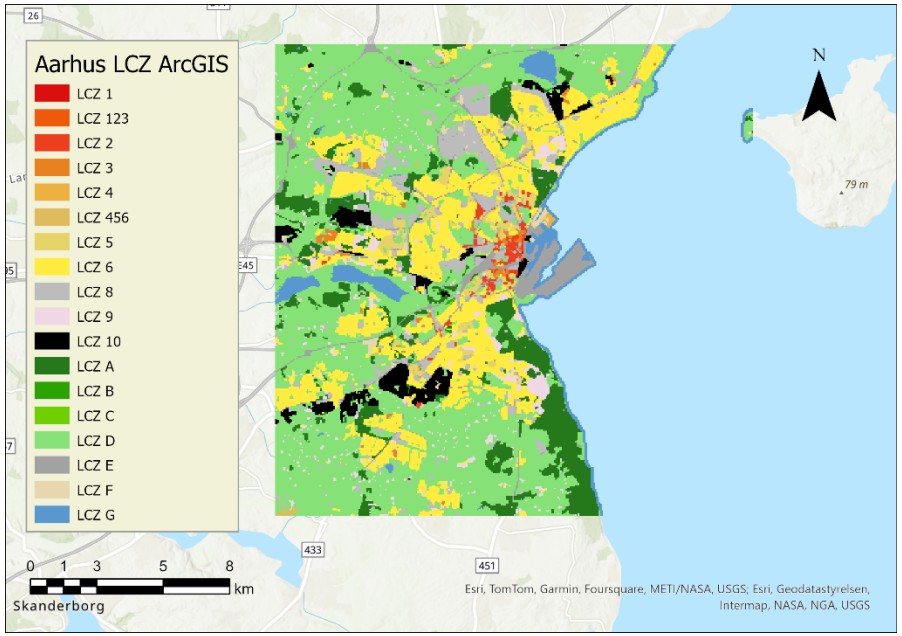


Figure 19 – Aahrus's LCZ classification with ArcGIS toolbox

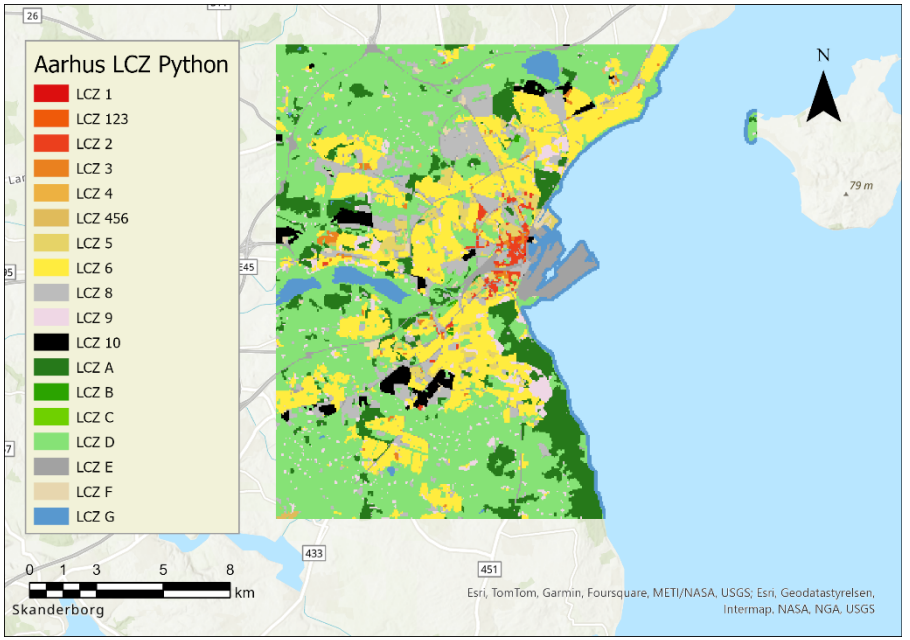


Figure 20 – Aahrus's LCZ Classification with this project's classifier

ANNEX A

Dear Bruno Marques,

Dear Professor Pedro Cabral,

Thank you for filling in the Research Ethics Checklist. After reviewing your request, you can proceed with the study we do not foresee any major ethical concerns with the project.

Project No.: **GEO2024-10-285677**

Project Title: **LOCAL CLIMATE ZONE CLASSIFICATION SYSTEM USING WEB GIS APPROACH**

Principal Researcher: **Bruno Ricardo Jorge Martins Marques**

according to the regulations of the Ethics Committee of NOVA IMS and MagIC Research Center this project was considered to meet the requirements of the NOVA IMS Internal Review Board, being considered **APPROVED** on 20/11/2024.

It is the Principal Researcher's responsibility to ensure that all researchers and stakeholders associated with this project are aware of the conditions of approval and which documents have been approved.

The Principal Researcher is required to notify the Ethics Committee, via amendment or progress report, of

- Any significant change to the project and the reason for that change;

- Any unforeseen events or unexpected developments that merit notification;

- The inability of the Principal Researcher to continue in that role or any other change in research personnel involved in the project.

Lisbon, 20/11/2024

NOVA IMS Ethics Committee

ethicscommittee@novaims.unl.pt

UNIGIS PT

NOVA IMS Information Management School

Grelha de revisão da formatação de DPE

Secção do trabalho	Conforme	Não conforme
Capa		
Contracapa		
1ª página do trabalho (cópia da capa)		
2ª página		
Declaração de originalidade		
Agradecimentos		
Resumo		
Palavras-chave		
Índice do texto		
Índice de tabelas		
Índice de Figuras		
Acrónimos		
Margens e limite de palavras		
Paginação		
Figuras, tabelas, gráficos e cartogramas		
Referências bibliográficas		
Anexos		