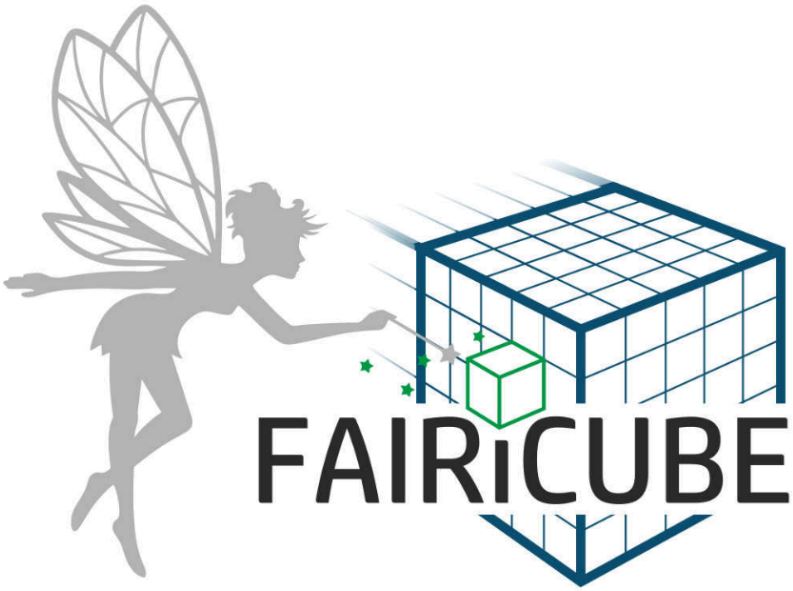


FAIRICUBE – F.A.I.R. INFORMATION CUBES

Project Number: 101059238

WP 3 Process
D3.1 UC exploratory data analysis

Deliverable Lead: NIL
Deliverable due date: 29/02/2024

Version: 2.1
2024-01-24

Document Control Page

Document Control Page	
Title	D3.1 UC exploratory data analysis
Creator	NIL
Description	D3.1 UC exploratory data analysis
Publisher	"FAIRICUBE – F.A.I.R. information cubes" Consortium
Contributors	NIL, WER, NHM, S4E, 4SF
Date of delivery	30/06/2023
Type	Text
Language	EN-GB
Rights	Copyright "FAIRICUBE – F.A.I.R. information cubes"
Audience	☒ Public ☐ Confidential ☐ Classified
Status	☐ In Progress ☐ For Review ☒ For Approval ☐ Approved

Revision History			
Version	Date	Modified by	Comments
0.1	16/05/2023	Stefan Jetschny, NIL	Draft setup, headings, and partner / contributor assignments
	27/05/2023	Rob Knapen, WER	Use case 2 contribution
0.2	11/06/2023	Stefan Jetschny	Ready for partial review, only UC4 contribution missing
1.0	21/06/2023	Stefan Jetschny	Ready for review, minor comments still open
1.1	21/06/2023	Jaume Targa, Stefan Jetschny	Review and minor modifications according to review comments.
2.0	07/11/2023	Stefan Jetschny	Extraordinary update to reflect the progress of the work, read-through version for assigning writing-tasks
2.1	11/01/2023	Jaume Targa	Review

Disclaimer

This document is issued within the frame and for the purpose of the FAIRICUBE project. This project has received funding from the European Union's Horizon research and innovation programme under grant agreement No. 101059238. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission.

Table of Contents

List of Figures

List of Tables

1 Introduction

WP3 aims to provide guidance, recommendations, technical expertise, and implementation support expertise to all use case efforts in terms of data analysis and processing. While the use case developers will execute the tasks, support will be given to assist in all data handling steps after ingestion and provision on both the Rasdaman- and EOxHub services as part of FAIRiCUBE's overall data and model services. Special emphasis is given to the data driven machine learning (ML) model generation.

This deliverable needs to be seen as one item of a classical and logical execution of a machine learning (ML) application. Given availability/ingestion of data, we first perform an exploratory data analysis to get familiar with the data, analyse statistical parameters and distribution, check for completeness, outliers and other characteristics which could be relevant for the choice of the machine learning. This in-depth data analysis is covered by this deliverable *D3.1 UC exploratory data analysis*.

Subsequently, the raw data might require conversion into features through a data engineering step. This could imply a combination of several input data sources or applying simple mathematical operations to enhance the meaningfulness of the raw data given the relationships that are to be revealed. The more domain knowledge, and a-priori information is available, the better the feature engineering process can be performed. Based on the findings from the exploratory data analysis, the formulation of the research question and the relationship between raw data sources / features, machine learning algorithms can be recommended to establish a baseline model if this is not provided by use case owners. Starting from the most efficient machine learning algorithm, more advanced ML methods can be identified to form a machine learning strategy. Several methods might also be tested to recommend a method based on computational demands and accuracy of the ML output. Typically, the testing of ML algorithms is performed on a subset of the original input data or on selected cases. The feature engineering process, testing of ML algorithms and the recommendation of a cascade to ML algorithms, as well as analysing the output of ML methods, is covered by this deliverable *D3.2 Machine learning strategy specific for each use case*.

As the FAIRiCUBE Hub ultimately aims to also provide resource estimations and guidance for ML applications, we aim to collect and share computational parameters, timings, requirements and give an outlook on the expected scalability of the ML problems defined by the use cases. For each ML algorithm identified and executed as described in D3.2 we collect information on e.g., disk storage, CPU runtime, main memory consumption, describe the hardware and environment where the ML algorithm is executed on and list essential libraries that are needed to exactly replicate the ML application. This technical documentation of the ML execution is covered in the deliverable *D3.3 Processing and ML applications*.

In summary, the exploratory data analysis (D3.1) can be seen as essential input to the development of a UC specific machine learning strategy (D3.2) whereas the technical description in D3.3 acts as a reference to follow up on the execution and serves as valuable input to estimate the demands for other ML applications. In the following, the exploratory data analysis is described for each use case.

2 Exploratory data analysis

UC1 Urban adaptation to climate change

The decision-making process regarding urban adaptation, to mitigate climate changes, requires some data to give a clear picture about the current situation. The aim of this use case is to provide such data. Some of the data need to be processed and analysed to make sense for decision makers (e.g., extract useful patterns instead of presenting the whole data).

i. Land use map for Functional Urban Areas

One data is the land use/land cover map for Functional Urban Areas (FUAs) in Europe, defined by the Copernicus Urban Atlas dataset[1] which presents the classes of each area regarding a level of detail. The data is under EPSG:3035 (ETRS89, LAEA) CRS[2] and with a resolution of 10m. It covers Europe, including Turkey (see Figure 1). Here, we consider the five level-1 classes (*1. Artificial surfaces; 2. Agricultural areas; 3. Natural and semi-natural areas; 4. Water; 5. Wetlands*). The classes (typologies) can be further detailed by considering level-2 classes (e.g., class 112 represents *Artificial surfaces* that in addition are *Discontinuous Urban Fabric*) and level-3 classes (class 11210 represents *Discontinuous Urban Fabric* that are *Dense*). However, in this study, we will only focus on the five classes of level 1. Finally, as mentioned before, the Urban Atlas data is presented at FUAs level. To get the coverage ratio of each class at cities' level, we also need cities' outlines. These are defined in the Urban Audit dataset.[3]

For example, for the city of Luxembourg in Figure 2 classes 1, 2, 3, 4 and 5 are covering 52%, 18%, 29%, 0.4% and 0% of the city, respectively. Using the data-cube constructed by 1- areas coverage (from the Urban Atlas dataset); and 2- cities outlines (from the Urban Audit dataset), the use case owners have generated a csv file for 1,042 cities, each line containing a city with its coverage ratios for the five classes. In addition, we have a 6th class representing the non-data areas. Finally, before conducting any data analysis, we filtered out cities with large non-class areas, i.e., the cities that have a high degree of incomplete data. We have considered cities with over 50% of non-class coverage as outliers and therefore removed a total of 56 cities which reduced the total of considered cities to 986 cities.

[1] https://land.copernicus.eu/local/urban-atlas/urban-atlas-2018?tab=metadata
[2] Coordinate Reference System.
[3] https://ec.europa.eu/eurostat/web/cities/data/database

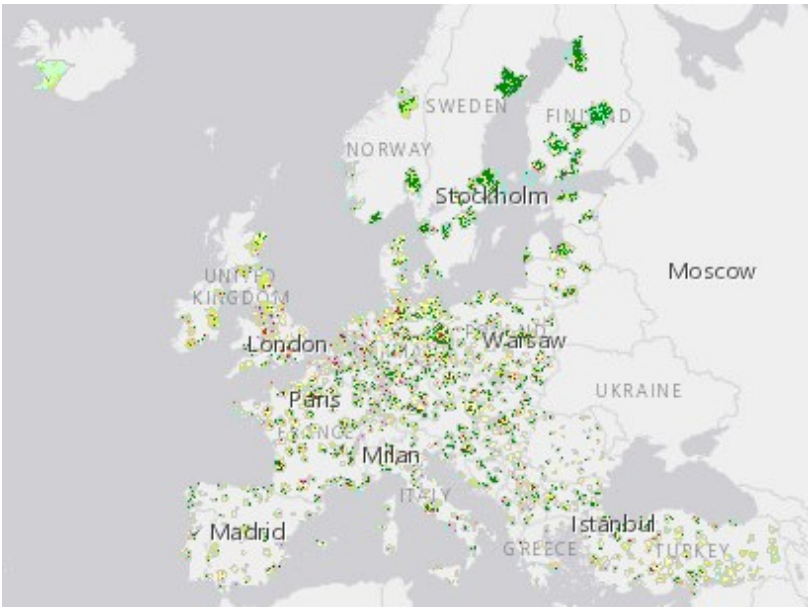


Figure 1: Urban Atlas data

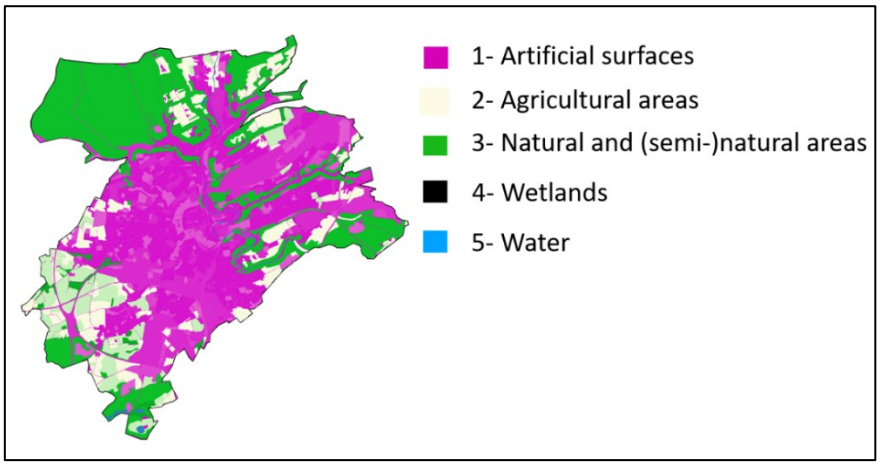


Figure 2: The level-1 coverage in Luxembourg

The class-ratios with respect to the 986 cities are presented in Figure 3. For each class (e.g., *Artificial surfaces*) we represent how many cities for which at least *r* of the city is covered by the class. For example, we have a total of 952 cities for class *Artificial surfaces* and *r*=0.1 i.e., 952 cities are covered by over 10% with *Artificial surfaces*. The red line represents the total number of cities (986). Finally, the blue bars in the right bottom plot represents the average ratio of each class where 1, 2, 3, 4 and 5 represent the classes *Artificial surfaces, Agricultural areas, Natural and semi-natural areas, Water and Wetlands*, respectively. We can clearly see that *Water* and *Wetlands* cover very few areas of the cities in average (only 2% and 0.3%, respectively). One reason is that the data available through Urban Atlas does not consider the sea to be part of the city and hence, coastal cities have less water coverage than they should have.

The data, as it is, does not give a complete picture of the cities w.r.t their coverage, or further, we cannot, manually, identify the cities that are similar in coverage. To achieve this, the data need further processing (e.g., using Machine Learning). Ratios of coverages classes with respect to the total area of a city can be input for an unsupervised classification which groups the selected cities according to dominant coverage types or a combination of coverage types.

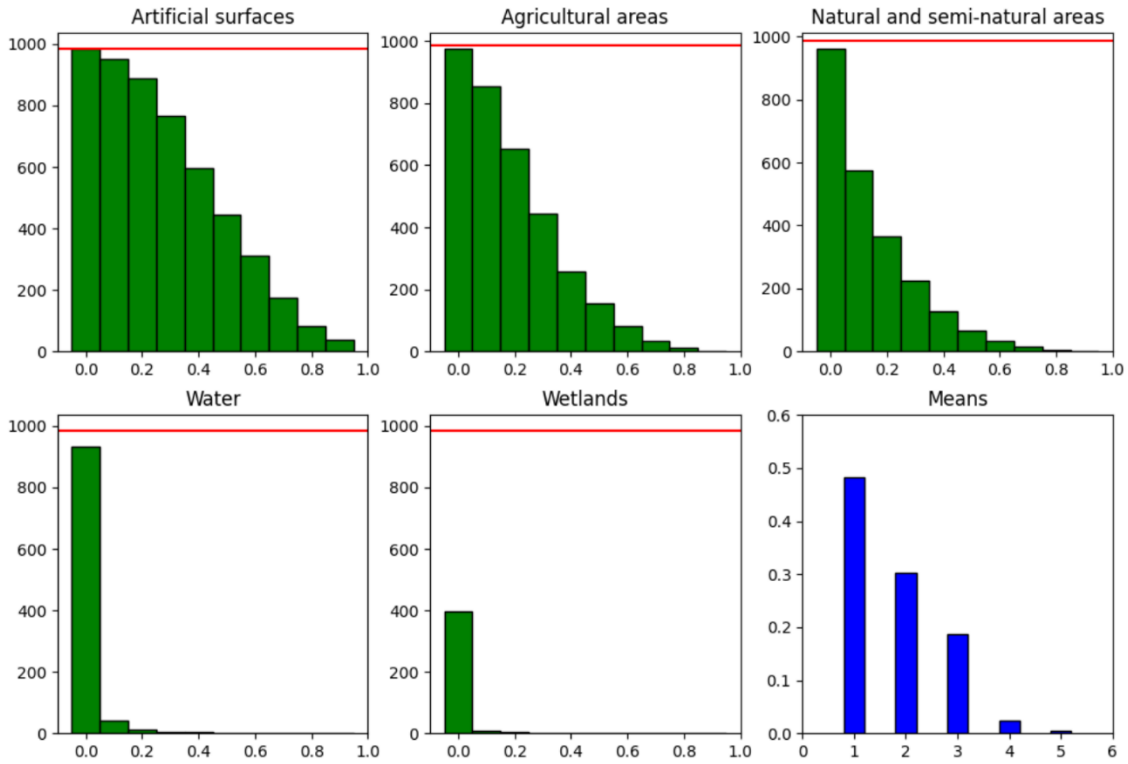


Figure 3: Number of cities w.r.t classes' ratios

ii. Socioeconomic data

Another valuable data source to be considered for clustering cities is socioeconomic data. Socioeconomic data reflect both the social and economic trend of a given population. In Europe, Eurostat and its Urban Audit database represents one of the most important and complete socioeconomic data sources for European cities.[4]

The data covers 82 variables and indicators, including, for example, *'Severely materially deprived persons'* (under code *EC3066V*) and *'Total Resident Population'* (under code *DE1001V*). As for now, Eurostat attempts to cover 910 administrative units between from countries to cities in Europe and a time range of 31 years, from 1991 to 2022. Given the above information, someone expects to have around 910 cities * 82 indicators * 31 years = 2,313,220 data points (indicator value for each year/city). Unfortunately, only 358,101 data points are available (that is 15.48% of the data). This leaves us with more than 84.52% of missing data. In what follows we give more insight into the data availability and its distribution.

Available data w.r.t cities (expected = 82 indicators * 31 years = 2,542 data points): The city with the most missing data is *'Mechelen'* in Belgium (under code *BE012C*), with only 27 data points. Figure 4 represents the number of available years w.r.t indicator/attribute index. On the other hand, *'Helsinki'* in Finland (under code *FI001C*) has the most available data with 1,117 data points. Figure 5 represents data availability for Helsinki w.r.t indicators and years.

[4] https://ec.europa.eu/eurostat/web/cities/database

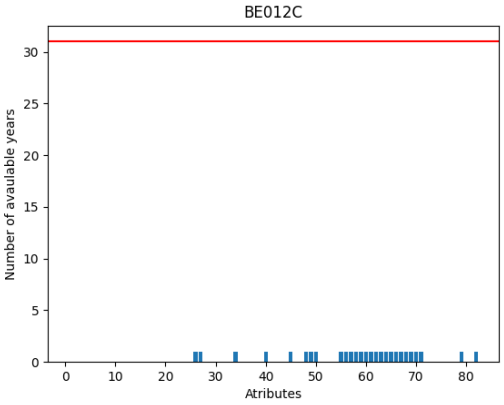


Figure 4: Data Availability for Mechelen

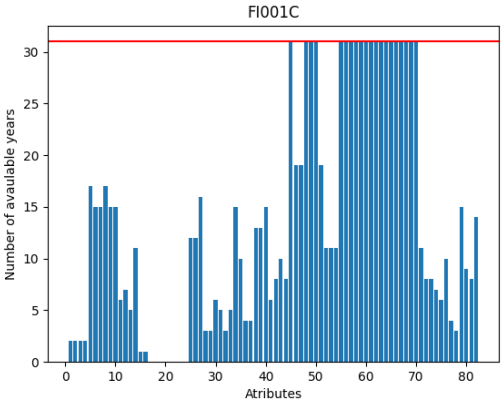


Figure 5: Data availability for Helsinki

Available data w.r.t years (expected = 910 cities * 82 indicators = 74,620 data points): The year with the most missing data is 1993 with 897 data points, the year 2011 has the most available data with 28,985 data points.
Available data w.r.t indicators (expected = 910 cities * 31 years= 28,210 data points): The indicator with most missing data is *'Severely materially deprived persons'* (under code *EC3066V*) with only 302 data points. On the other hand, the indicator *'Total Resident Population'* (under code *DE1001V*) has the most available data with around 14,559 data points.
Total Resident Population as example: In Figure 6 we present an example of the data distribution. Here, we represent the number of cities with available *'Total Resident Population'* values w.r.t each year index (1 representing the year 1991). We can clearly see that most years have less than 100 available cities. On the other hand, the year 2013 has more than 120 available cities. In Figure 7 and Figure 8 we present two cities with available *'Total Resident Population'* for all years (*Helsinki* and *Bari*). We can clearly see that the time series are very different and are city dependent. In *Helsinki* (i.e., *FI001C*), the total population change follows an increasing linear trend, while the city *Bari* (i.e., *IT008C*) follows a more irregular trend.

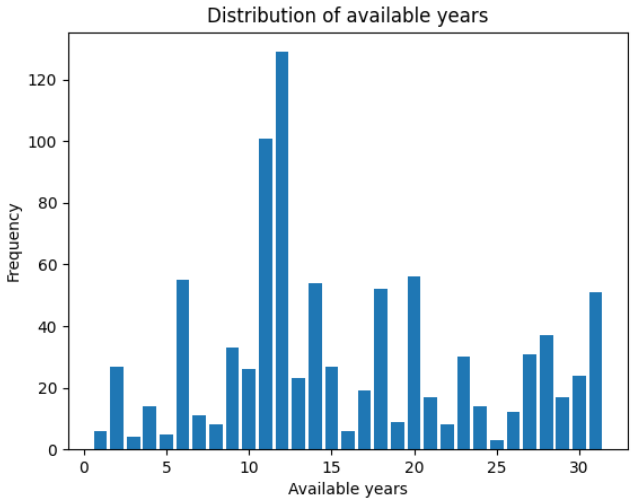


Figure 6: Total Population distribution

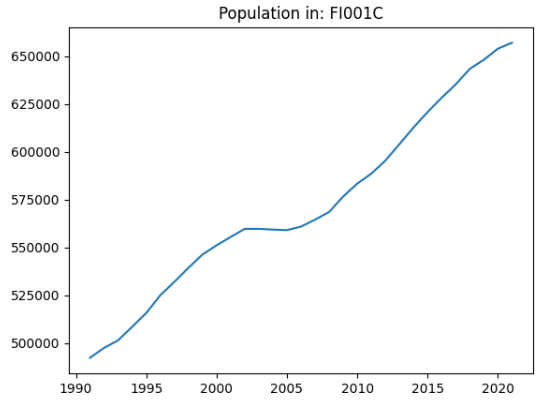


Figure 7: Population over the years in Helsinki

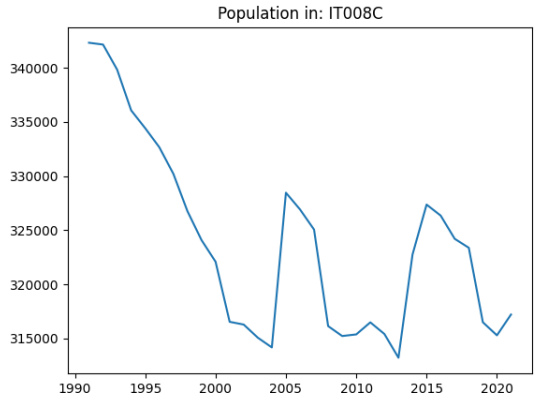


Figure 8: Population over the years in Bari

The data, as is, is limited and is not very useful to draw conclusions and cluster cities. Luckily, some Machine Learning approaches can be helpful in gap filling and recovering missing data.

UC2 Agriculture and Biodiversity Nexus

Use case 2 aims to study the effects of agriculture and farming activities on biodiversity, specifically in agricultural areas. While it is known that various activities can have different impacts on biodiversity, as illustrated by Figure 9, these associations are typically poorly understood, likely complex, and difficult to clarify. Related scientific studies are usually local in scope and lack sufficient standardization to allow comparison of results at different scales.

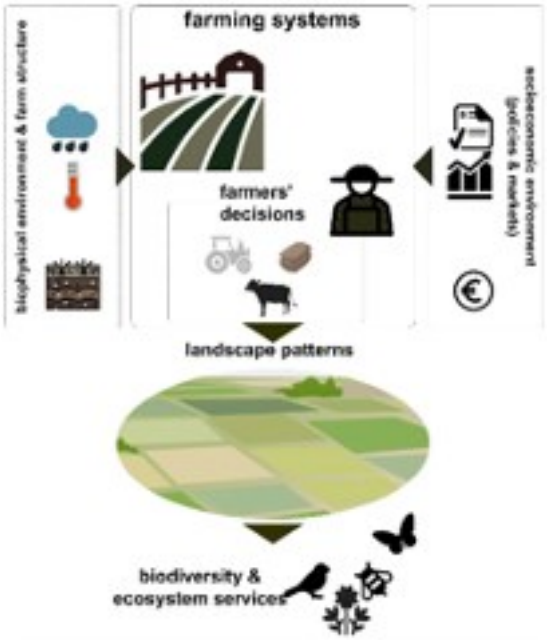


Figure 9: The agriculture - biodiversity nexus

Data collection and analysis for this use case is commencing but has proven to be challenging due to the nature of the data involved. On one hand biodiversity data (such as shown in Figure 10) is not easy to obtain, often requiring tracking down individual researchers involved in the studies and attempting to get consent for using the data for the project. Furthermore, this data usually is collected in-situ

following different protocols, e.g., species presence only, or both presence and absence, and the observations have various levels of inaccuracies (e.g., positional or in species abundance). Additionally, there has not been recognized one universal measure of biodiversity, it can be interpreted in many ways based on the available species observation data and the particular goal of a study. Fortunately, despite these challenges, there are also already established methods to handle biodiversity data and ongoing studies towards improvements such as the work on *species distribution modelling* (SDM) and the establishment of *Essential Biodiversity Variables*. Analysis of these topics and their relation to data cubes and machine learning (ML) is in progress.

FID	Shape	nl_name	sci_name	jaar	countsubje	orig_abund	straal
2639	Polygon	Graspieper	Anthus pratensis	2016	territorium	1	3
2647	Polygon	Bosrietzanger	Acrocephalus palustris	2016	territorium	1	3
2654	Polygon	Zanglijster	Turdus philomelos	2016	territorium	1	3
2657	Polygon	Kievit	Vanellus vanellus	2016	territorium	1	3
2663	Polygon	Rietgors	Emberiza schoeniclus	2016	territorium	1	3
2670	Polygon	Veldleeuwerik	Alauda arvensis	2016	territorium	1	3
2672	Polygon	Boomklever	Sitta europaea	2016	territorium	1	3
2680	Polygon	Grauwe vliegenvanger	Muscicapa striata	2016	territorium	1	3
2687	Polygon	Graspieper	Anthus pratensis	2016	territorium	1	3
2693	Polygon	Veldleeuwerik	Alauda arvensis	2016	territorium	1	3

Figure 10: An example of species observation data

The farming data, on the other hand, is somewhat easier to collect, but highly privacy sensitive due to the detailed scale and business character of the data. At a later stage this will have consequences in how much of the data can actually be obtained at various levels of detail, what can be made publicly available, or lead to an evolving need for data (cube) anonymization and/or restricted access capabilities.

Finally, environmental data will play a central role in the use case, as it not only contains information essential for SDM (or similar work), but expectedly also many confounding factors between agriculture and biodiversity. This type of data is most 'native' to the data cube technology as many of it originates from the Earth Observation (EO) domain.

For the initial data exploration, a small study area has been selected in the Netherlands. Figure 11 shows the location in the country, while Figure 12 shows the land use map (of 2018) of the area. The area has been selected because it contains a good mix of strongly agricultural region (to the left), some lakes (in the middle), and on the right side there is plenty of vegetation and urban areas.

Figure 11: Study area in NL

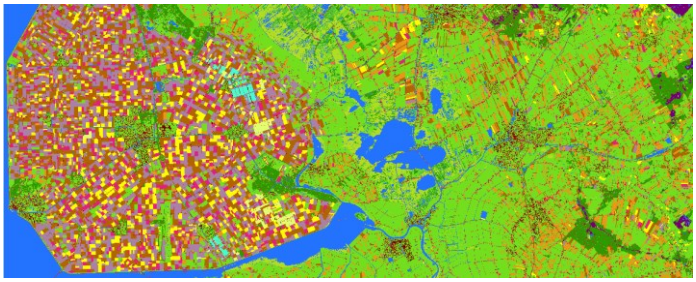


Figure 12: Landuse map (2018) of study area

i. Biodiversity Data Exploration

For the study area an initial three species distribution datasets have been acquired from the Dutch "Nationale Databank Flora en Fauna" (NDFF, www.ndff.nl). This is a large data warehouse that contains the distribution data of plants and animals in the Netherlands, providing data entry portals, a central archive, validation services and data export portals. In the Netherlands vast amounts of data on the whereabouts of species over the last century have been gathered and stored. However, it is scattered among different organizations, in different formats and not always digitally available. The NDFF has been built to make distribution data of (flora and fauna) species available through one National Data Warehouse. NDFF data can be requested for research purposes, and then might be provided without costs. In other cases, a subscription or per-request payment is required. In any case, license terms apply to the use of the data.

The three initial datasets chosen for purpose of data exploration and proof of concept for data interpretation (see Table 1) are a selection for the year 2016 of breeding birds (which do not have a large range of movement), 'other' species of interest for the use case (such as insects), and plants. These cover the whole country and have been provided as comma-separated values (CSV) files.

Table 1: Initial NDFF species distribution dataset properties

Filename	Content	File size	Record count
broedvogels_2016.csv	Nesting birds 2016	154 MB	454.453
overigesoorten_2016.csv	Other (selected) species 2016	133 MB	370.718
planten_2016.csv	Plants 2016	47 MB	128.968

It is good to note that, as a data warehouse, NDFF stores data from multiple sources (the various organizations involved) including crowdsourcing and that different collection protocols are being used. Confusingly, it appears that this can mean that attributes of data records with the same name don't necessarily have the same meaning. And, depending on the species and the collection protocol, available attributes might differ. Thorough understanding of the data and most likely expert advice is needed for a correct interpretation and meaningful processing. Also, since the received files already contain a selection of the data which has been validated and reaching certain confidence level, it is not precisely clarified what selection criteria (and possibly attribute renaming) have been applied.

The common structure in the provided csv files consists of the following attributes: (1) *sci_name*, the scientific name of the species observed; (2) *year,* the year of observation; (3) *countsubject,* the type of subject that has been counted (e.g. a nest, a living specimen, a territory, etc.); (4) *orig_abundance,* the amount observed (this can be a number, a '*', or a letter code); (5) *radius,* the range in which the amount has been observed (although this needs to be further clarified); and (6) *wkt,* a geometry (typically an octagon) in WKT representation [5] that indicates the location of the observer, with the size of the octagon representing the (GPS) accuracy of the positional information. The geometry is represented in the Dutch RD coordinate reference system (EPSG:28992). Some statistics are captured in Table 2.

[5] https://en.wikipedia.org/wiki/Well-known_text_representation_of_geometry

Table 2: Overview of some properties of the datasets

Attribute	Nesting Birds	Other Species	Plants
Total observations	454.452	370.717	128.967
Mean radius	221.8	161.6	94.1
Std radius	106.9	121.5	94.1
Subject: Territory	326.818	NA	NA
Subject: Specimen	115.299	321.733	103.536
Subject: Unknown	NA	39.663	9.450
Subject: Cover Herb Layer	NA	NA	9.396
Subject: Cover Vegetation	NA	NA	6.457

For all three datasets the *living specimen* is a well available *countsubject* category, for which the *orig_abundance* is most frequently 1 (one), and the *radius* often 283 meters (which seems odd and needs further investigation). A histogram of radius values for the living specimen records in the nesting birds' dataset is shown in Figure 13.

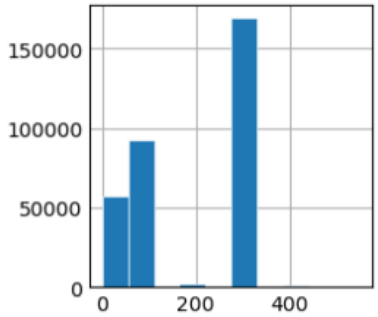


Figure 13: Distribution of nesting birds living specimen radius values

In the second phase of the data exploration, we got available extended dataset covering the same spatial extent of study area but including all available species records for the period of years 2014 – 2022. Dataset also include extended number of attributes (17 in total). For the further data exploration, we selected breeding birds for the year 2018.

Data engineering is needed to aggregate the species observation data recording with point/polygon geometries to show the location into grid cells that are suitable for ingesting into a data cube (i.e., to create an occurrences cube). This is further described in the UC2 section of deliverable D3.2. For an example result see Figure 14, which shows for a small area the calculated proportional (for each grid cell) abundance of nesting birds. Darker colours show higher abundance. The background, in grey, shows agricultural fields (arable and grasslands). Note that this is only based on positive abundance records, so the not coloured grid cells should not be interpreted as species absence, rather as unknowns.

Figure 14: Example of proportional abundance calculated for nesting birds at 100m x 100m grid cells

Another example of data exploration is focused on assignment of abundance values from observation polygons to grid cells based on different criteria. Initially were the abundance values assigned to the cell at the centre and at random position of observation polygon. Further, considering possible movement of animal species beyond observation polygon boundary was potential area of abundance extended for buffer of 100 and 500 meters. In these extended polygons were abundance values assigned to grid cells at random position. Such an extension of area allows map species abundance which are initially recorded out of agricultural areas but in proximity to boundaries within buffer polygons (Figure 15). Selection of buffer size is species specific and has to be further investigated.

Figure 15: Example of assignment of abundance value in random position (yellow dots) within observation buffer area (white circle with yellow hatching) within the boundaries of agricultural land (red lines). Original abundance locations are shown as blue dots

Allocation of abundance records was made to different grid size data layers. On the Figure 16 are values aggregated to 1 km grid using above mention allocation methods.

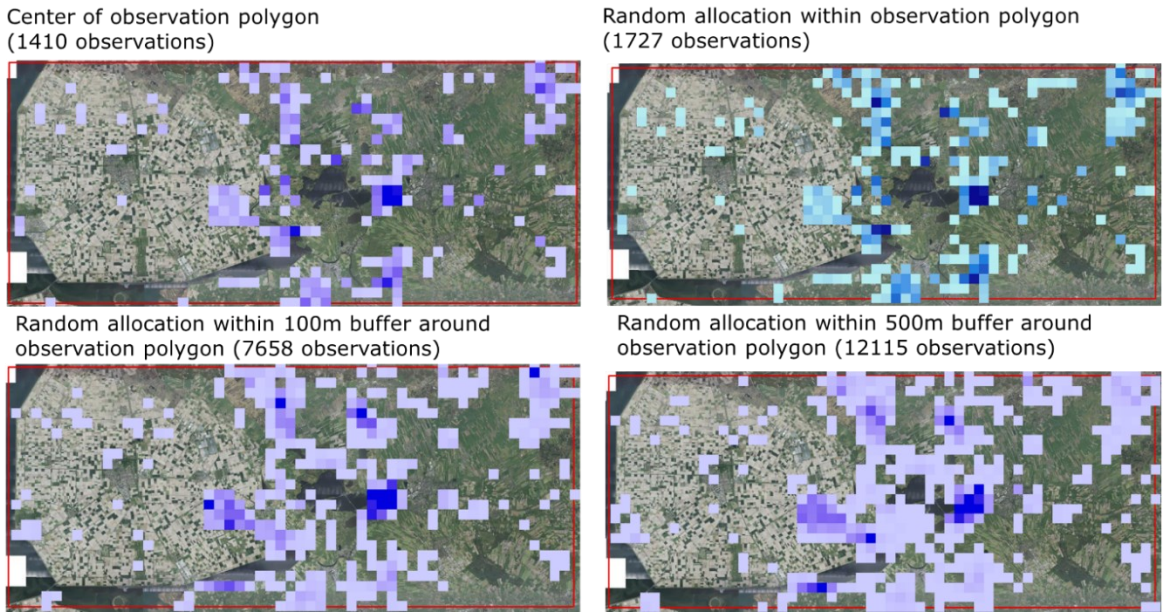


Figure 16: Example of nesting bird data in agricultural areas aggregated to 1000m grid cells using different criteria, darker colours indicate higher abundance values at given grid cell location

ii. Agricultural Data Exploration

Base Register Crop Fields (BRP) consists of the location of agricultural parcels with information on cultivated crops. The boundaries of the agricultural plots are based on the Agricultural Area of the Netherlands (AAN) dataset. The data are acquired from farmers when each owner of the plot must annually register his crop plots extent and indicate which crop is grown on the relevant plot. A dataset is generated for each year with a reference date of 15 May.

The Agricultural Nature and Landscape Management (ANLb) is a dataset indicating subsidy for agricultural collectives from provinces, water boards and the Common Agricultural Policy (CAP). With this grant: farmers protect and improve the environment of animals, working on water quality and contribute to climate goals.

Farmers do this by implementing practices such as reducing the fertilizer use on grasslands and delaying mowing to facilitate bird breeding. The collective ensures that the management of participants in different areas is consistent. Additionally, habitat areas of species can be located on the grounds of several companies.

iii. Environmental Data Exploration

Environment data are important to develop measures of biodiversity and to evaluate impact of agriculture activities. To translate individual observations to standardized biodiversity measures, we consider Essential Biodiversity Variables (EBVs) (Pereira et al. 2013). The concept of EBVs was established to advance the collection, sharing, and use of biodiversity information. To detect change, systematic biodiversity observations are collected using standard formats and methods, together with environmental monitoring. Ensuring that data are interoperable across databases will make efficient use of biodiversity information for guiding conservation and sustainable development strategies. EBVs providing a way to aggregate the many biodiversity observations collected through different methods such as in situ monitoring or remote sensing. EBVs can be visualised as biodiversity observations at one location over time, or in many locations, aggregated in a time series of maps.

UC3 Biodiversity occurrence cubes – *Drosophila* landscape genomics

UC3 aims to exploit the massive collection of DNA sequenced data of natural populations of the fruit fly *Drosophila melanogaster*. Apart from the challenge to adapt the data for storage as data cubes, the WP3 supports the processing and potential application of machine learning algorithm to enrich the data set and reveal further insights while making advantage of the scalability and accessibility of data storage and processing capabilities of the FAIRiCUBE Hub.

As a first demonstrator task, gap-filling of missing data in the genetic information of pools of individuals that were sequenced jointly (Pool-Seq) was identified and performed on a sample data set. The provided data set[6] comprises of 754 population-based samples of *Drosophila melanogaster* distributed over the globe (see Figure 17). In a first step, we focus on populations in North America, which are predominantly collected along the East Coast. Many of these samples are densely collected across multiple seasons and years.

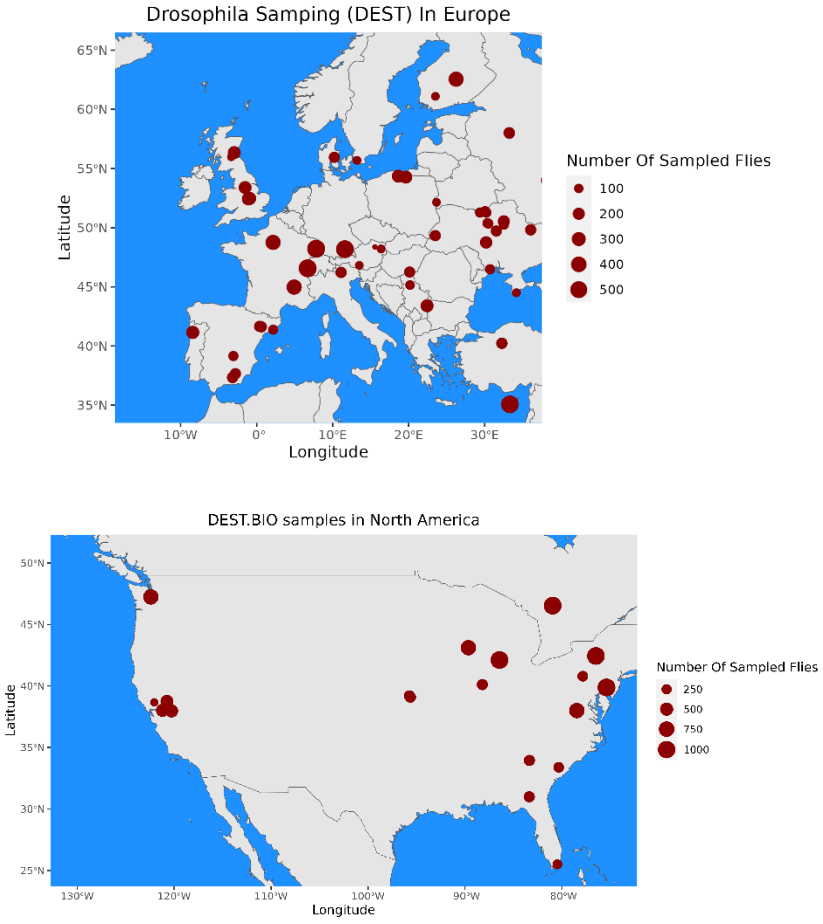


Figure 17: Location of DNA sequenced population of species "*Drosophila melanogaster*" as contained in the DESTv2 dataset

The available sequencing data was previously filtered for sequencing quality and afterwards aligned against a reference genome from *Drosophila melanogaster* to identify polymorphic genomic positions along chromosomes and their genes. Such positions are characterized by more than one allelic state, i.e., nucleotide variants (A, T, C or G). This means that individuals in a population carry different genetic variants which may be either neutral, i.e., they do not influence the phenotype of the fly, or they are

[6] https://dest.bio/

under selection because they have different influence the fitness of the carrier – potentially in the context of environmental variation.

The frequency of an alternative allele, which is a nucleotide that differs from the state in the reference genome, is ranging from 0 to 1. This means that an alternative allele is either not present or at 100% in a given sample. During the processing, the data volume has been reduced from containing the information of all genomic positions (180 x 10^6 base pairs) to only those polymorphic genomic sites that contain more than one allelic state in at least one population sample. Thus, the final dataset contains row-wise allele frequencies for every polymorphic position for every population (in columns). A sample dataset is shown in Figure 18.

Index	#CHROM	POS	AT_Kar_See_1_2014-08-17	AT_Kar_See_1_2016-08-01	AT_Nie_Mau_1_2014-07-20	AT_Nie_Mau_1_2014-10-19	AT_Nie_Mau_1_2015-07-20	AT_Nie_Mau
0	2L	1003…	0.04	0	0.07	0.03	0.13	0
1	2L	1003…	0.14	0.12	0.22	0.18	0.12	0.16
2	2L	1003…	0.58	0.44	0.72	0.69	0.69	0.57
3	2L	1003…	0	0	0	0	0	0
4	2L	1003…	0	0	0.04	0.07	0	0
5	2L	1004…	0.78	0.76	0.83	0.93	0.85	0.78
6	2L	1004…	0.15	0.22	0.12	0.13	0.18	0.12
7	2L	1004…	0	0	0	0	0	0
8	2L	1004…	0.24	0.11	0.11	0.11	0.11	0.07
9	2L	1004…	0.09	0.2	0.03	0.09	0.02	0.06
10	2L	1004…	0	0	0	0	0	0.02
11	2L	1004…	0	0	0	0	0.06	0
12	2L	1004…	0	0	0.02	0.05	0.01	0.01
13	2L	1004…	0	0.03	0.04	0	0	0
14	2L	1004…	0.49	0.57	0.51	0.76	0.49	0.46
15	2L	1004…	0	0	0	0	0	0
16	2L	1004…	0	0	0	0	0	0
17	2L	1004…	0.9	0.91	0.82	0.75	0.72	0.91
18	2L	1004…	0	0	0	0	0	0
19	2L	1004…	0	0	0	0	0	0

Figure 18: Sample dataset of Allele positions and frequencies for available North America

Errors during the DNA sequencing process, quality variation of the extracted DNA and the sequencing depth can result in gaps in the genomic data, where the allelic state and frequency cannot be estimated in a population sample at a given genomic position. In order to improve data completeness, which is pivotal and good statistical practice for unbiased analyses of genome-wide associations between allele frequencies and environmental data, we aim to apply a gap filling of data. That means we want to estimate allele frequencies based on the assumptions that the genetic information of populations shows similarities either through location or in the course of time.

The original DEST dataset (https://droseu.net), comprised of populations from North America show a diverse range of gaps. The data from 71 out of the 230 population consists of more than 10% gaps, is treated as incomplete and will be disregarded (see Figure 19, data with more than 10% gaps is shown in green the left image). For the remaining data with less than 10% gaps, we plot the length of gaps by means of subsequent single nucleotide polymorphism (SNP) positions without data (see Figure 19, right image). We can observe a logarithmic distribution of gap occurrence as function of the gap size, i.e. most of the gaps affect only a single SNP position. The larger the gaps, i.e. the more subsequent SNP positions without data, the lower the occurrence. On average 1% of the total amount of data consists of gaps of a length of 1 SNP position, while on average gaps larger than 10 subsequent SNP positions represent only 0.01 % of the total data.

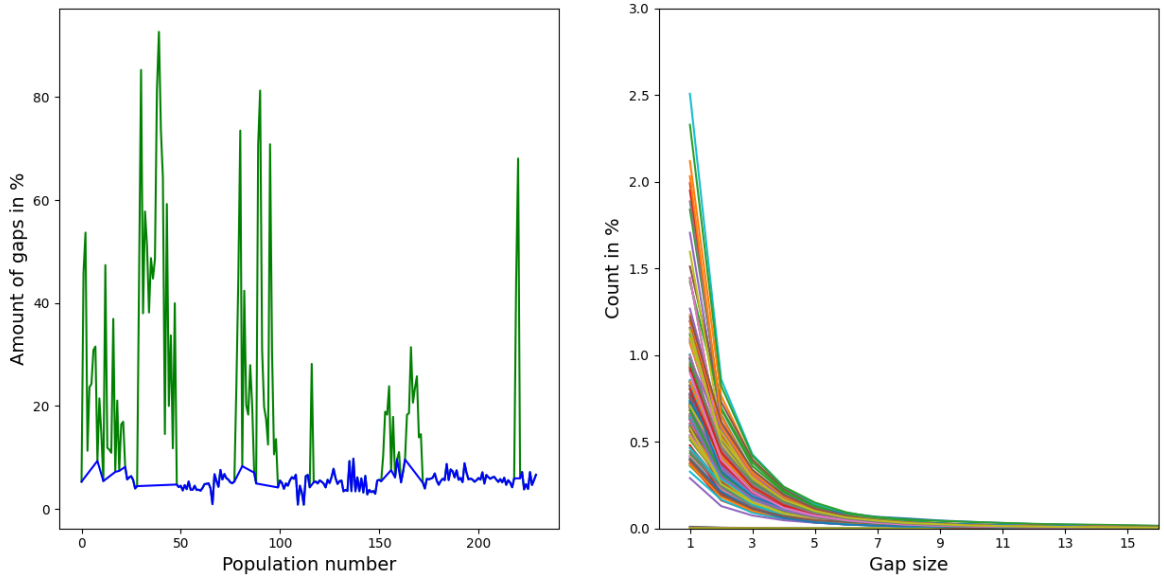


Figure 19: Gap size analysis of the North America populations of the DEST dataset.

For the purpose of developing a gap-filling method., we now focus on a subset of the DEST dataset, comprised of populations from North America, and randomly drew 50,024 polymorphic genomic positions without missing data in any population. After filtering by location, the test data set contains allele frequencies for 121 populations. Each sequenced population is referenced in time and space. During the exploratory data analysis, the data has been thoroughly analysed with a focus on the distribution of allele frequencies across all populations and within each population to gain insights about the uniqueness and meaningfulness of allele frequencies. This is the basis for further decisions on the machine learning algorithm that can be used to fill existing or artificially inserted data gaps.

Initially, we first get familiar with the distribution and variance of the allele frequencies across all populations. The minimum, maximum and mean value of allele frequencies of all the populations as a function of the single nucleotide polymorphism (SNP) position is shown in Figure 20, Figure 21 and Figure 22. As these statistics are derived across the population, the aspect of collection time and location of the sequenced individuals is removed and now gives information on the general variance of the allele frequencies. It is apparent, that a clear majority of SNP positions have a minimum of exactly 0 and the maximum is generally below 0.3 which indicates a low change towards the reference population. Only a fraction of SNP positions shows a significant difference to the reference. From Figure 21 we see that either the maximum frequency is below 0.3 or peaks at 1. The mean value of allele frequencies is close to 0, the corresponding distribution is very unbalanced (see Figure 22).

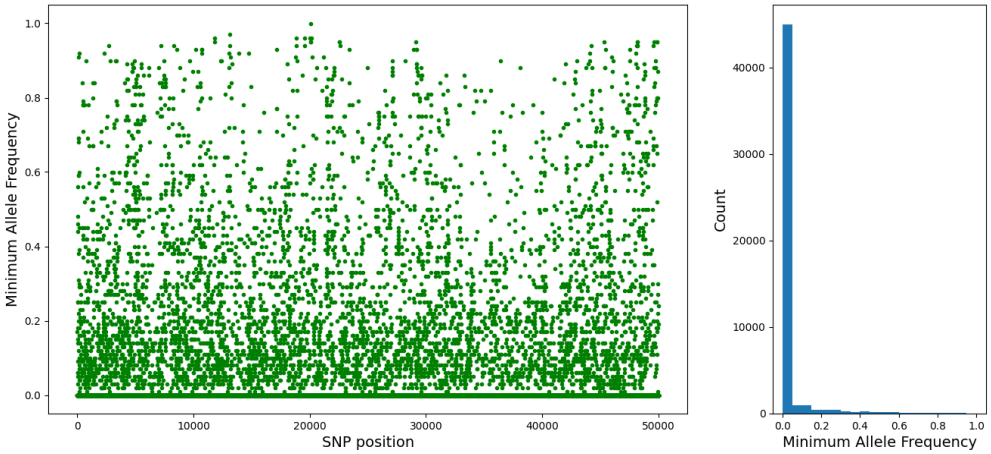


Figure 20: Minimum of allele frequencies across all populations and histogram of minimum allele frequencies in the North America data

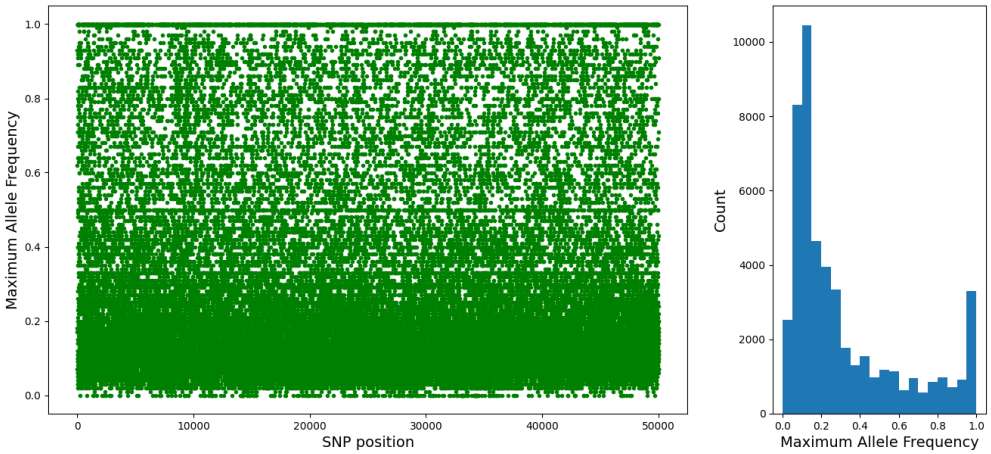


Figure 21: Maximum of allele frequencies across all populations and histogram of maximum allele frequencies in the North America data

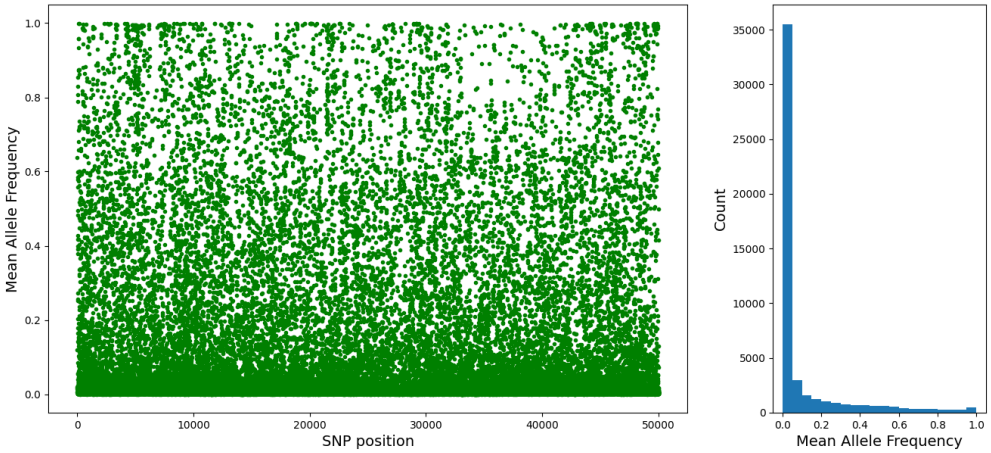


Figure 22: Mean of allele frequencies across all populations and histogram of mean allele frequencies in the North America populations

When performing the same analysis across the single nucleotide polymorphism (SNP) positions, we obtain some insights on how much individual populations differ from the reference population. As there is always an SNP position per population which is either 0 or 1, plotting the minimum or maximum across allele positions does not give any relevant information, only plotting the average shows variability of the populations (see Figure 23). According to the balanced distribution most of the populations have an average allele frequency around 0.108 but there are significant number of populations that have either significantly higher or lower average allele frequency which means a generally larger genetic difference to the reference allele, i.e. the allelic state in the originally sequenced genome. Further correlation of these "outliers" with higher or lower differences in allele frequencies and additional information on e.g., sampling location can reveal potential explanations.

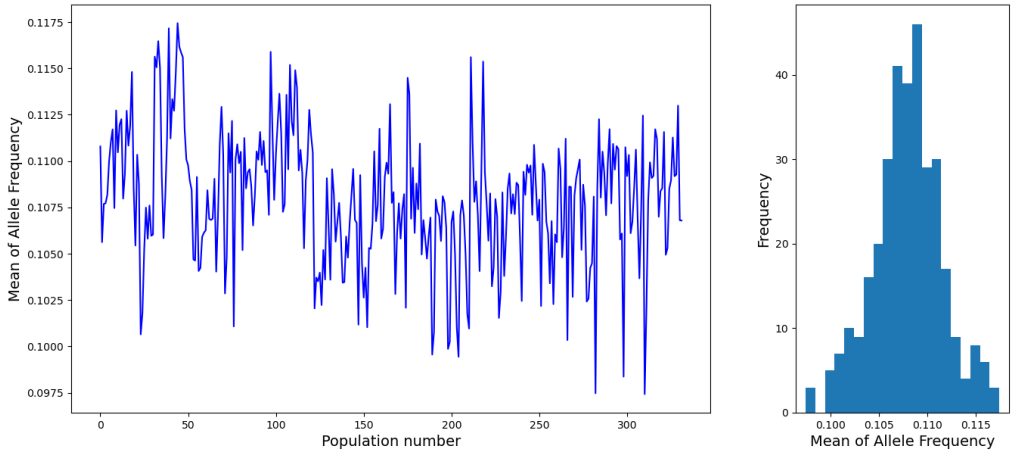


Figure 23: Mean of allele frequencies across all allele positions and histogram of the North America data

For the purposes of preparing the data for a potential gap filling ML application, we further analyse the standard deviation of allele frequencies across all populations to identify allele frequencies with the highest deviation from the average of all populations. As can be seen from Figure 24 the vast majority of allele positions shows low standard deviations, i.e., the allele frequencies do not change much across the populations at these positions. This data would therefore not exhibit any characteristic information and can be treated as redundant. By defining a threshold of selecting only 0.5 % of the highest standard deviations, we can identify the SNPs with the highest variance from the average. Figure 25 shows the filtered SNPs and the histogram which is basically a zoom of the descending flank of the local peak around standard deviation of 0.13 from Figure 24. After filtering, we select around 250 of the initial 50,024 SNPs which exhibits the most characteristic deviation from the average allele frequencies across all populations. These 250 positions can be seen as dimensions to span out a feature space which can be input for clustering. Reducing the dimensions of the clustering application has both numerical/computational as well as accuracy implications as we do not cluster on redundant but the most characteristic features (SNPs).

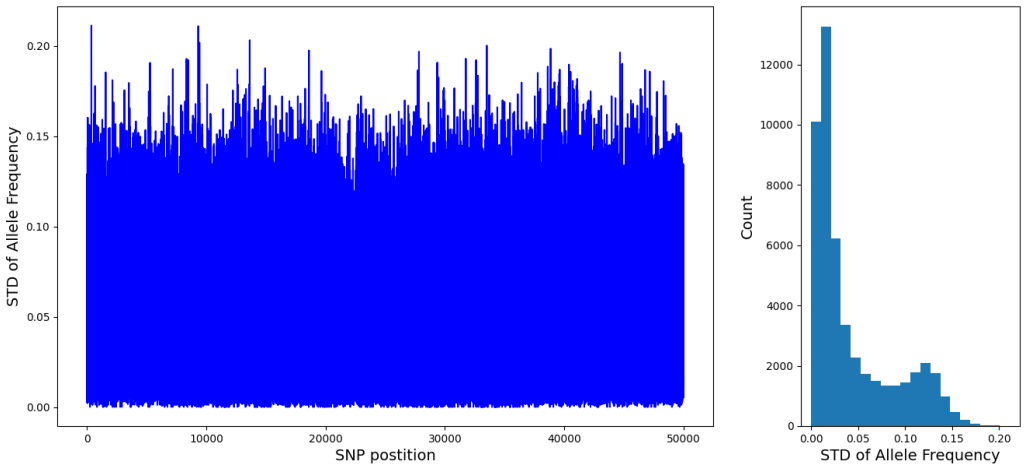


Figure 24: Standard deviation of Allele frequencies across all populations and histogram standard deviations of allele frequencies in the North America data

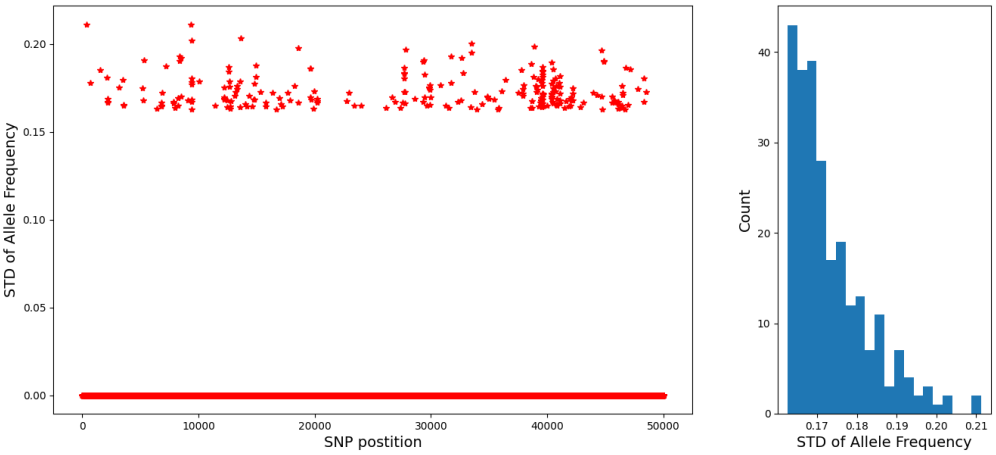


Figure 25: 0.5 % of the most significant standard deviation of allele frequencies across all populations and histogram of the North America data

UC4 Spatial and temporal assessment of neighbourhood building stock

UC4 has as aim to create a semantic model to later estimate the stock of materials and energy performance of primarily housing buildings and we take the join outcome of the IEE Project Episcope and Tabula[7] as a starting point (see also the final EPISCOPE report[8]). Country wise, a building classification is presented and a straightforward energy performance and building composition estimation is published there. Given the availability of the input parameters (construction year, building type and total floor area) we can thereby directly link public city data to our desired output parameters.

An essential part in the building's description stock is the building volume, which leads us to the need of setting a Level of Detail (LoD), a specification used with building data to describe the amount and degree of information used to abstract real world objects with. Other essential building parameters are for example construction year and a classification of residential housing types.

[7] https://episcope.eu
[8] https://episcope.eu/fileadmin/episcope/public/docs/reports/EPISCOPE_FinalReport.pdf

LoD are divided into 4 types (see Figure 26):

- LoD0 is a building footprint with no height attached just flat polygons aimed for two-dimensional analysis,
- LoD1 adds height information to LoD0.
- LoD2 include LoD1 and roof shapes, and last
- LoD3 covers detailed roof and façade shapes, as well as information on materials and textures Figure 26

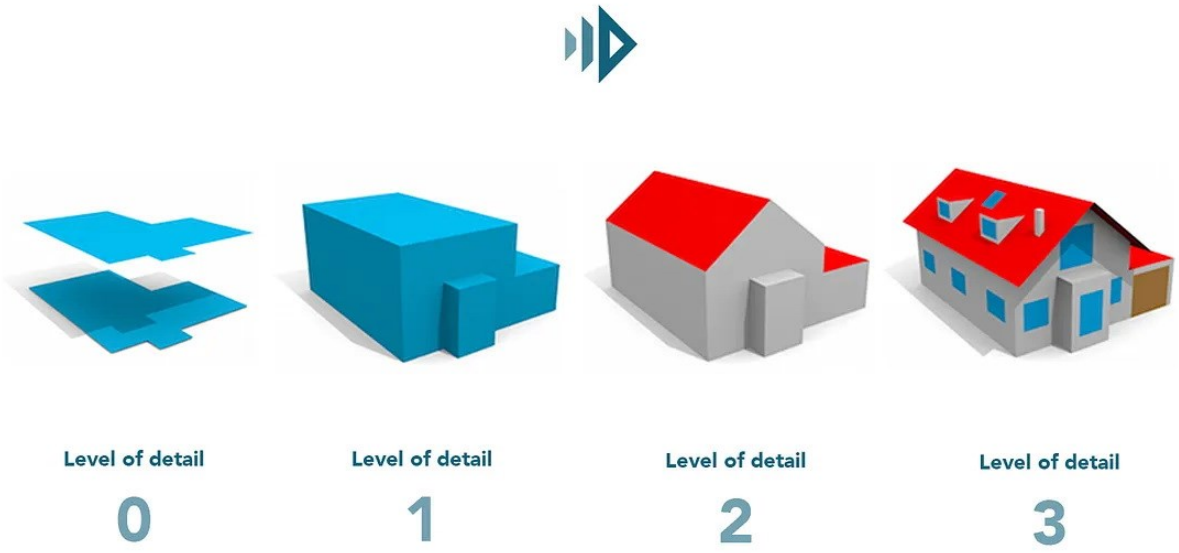


Figure 26: Illustration of level of datils (LoD) to describe building models, taken from 3DBuildings[9]

In the current context, we define building volume as building base area times building height. The base area is widely available (for example in Open Street Map and Google Maps) whereas the building height information is only partially available. Our efforts were therefore oriented firstly to get an overview of availability of building height data and secondly to evaluate different methods for gap filling when data is not available. Later, additional information such as construction year, building material composition or building classifications can be assigned to the building volume.

The exploratory data analysis was carried out for the city of Halle, Germany, mostly for reasons of data availability and applicability of the building height estimation methods that are described in deliverable 3.2. Once we conclude from the city of Halle test case, we will extend the building height estimation task to cover the original selected European test cities (e.g., Vienna, Austria; Oslo, Norway; Luxembourg; Barcelona, Spain) to allow for synergies with other FAIRiCUBE use case, e.g. UC 1. Table 3 summarises the data sources uses as input or reference to estimate building heights.

[9] https://3dbuildings.medium.com/how-building-data-works-level-of-detail-e9bad0b61baa

Table 3: Datasets used in the estimation of building heights.

Name	Reference / Link	Description
Building height ground truth	https://www.lvermgeo.sachsen-anhalt.de/de/download-lod1.html	Building height ground truth as calculated/derived from real estate, laser scan, aerial photo and terrain model data
Open Street Map (OSM)	http://overpass-api.de/	OpenStreetMap Overpass API. It provides building heights and other attributes using the tag "buildings" for a given bounding box. Using osmnx, a python package, the retrieved data is loaded into memory as a geopandas GeoDataFrame.
Copernicus building height dataset	https://land.copernicus.eu/local/urban-atlas/building-height-2012?tab=mapview	Copernicus land monitoring service, Provides heights and/or number of levels, the Copernicus building height data for the year 2012 (Urban Atlas 2012)
Digital Surface Model (DSM)	https://www.lvermgeo.sachsen-anhalt.de/de/dom2.html	Digital surface model with a grid spacing of 2 m (DOM2). Data download as a set of .xyz files.
Digital Terrain Model (DTM)	https://www.lvermgeo.sachsen-anhalt.de/de/dgm2.html	Digital terrain model with a grid spacing of 2m (DGM2). Data download as a set of .xyz files.
Urban Atlas land use/land cover classification	https://land.copernicus.eu/local/urban-atlas/urban-atlas-2018	Copernicus land monitoring service, high-resolution land use and land cover data for 788 Functional Urban Areas (FUA)

An administrative boundary of Halle in GeoJson format was created using the procedure described in https://peteris.rocks and used to crop all the building datasets.

The official building heights data for the city of Halle, Germany, was available as a set of gml files downloadable through the public Landesportal Sachsen-Anhalt. The dataset contained information for 11328 buildings. The distribution of building heights in the dataset is shown in Figure 27. Please note that data basis combines the input from:

- - Floor plan data of the buildings (house outlines) from the official digital real estate map,
- - building heights from laser scan data,
- - current aerial photo data,
- - terrain heights from the digital terrain model - ATKIS®-DGM,

which implies that this data must not be 100% accurate as it involves unknown processing of the input layers. We treat the data nevertheless as ground truth in absence of absolute correct alternative.

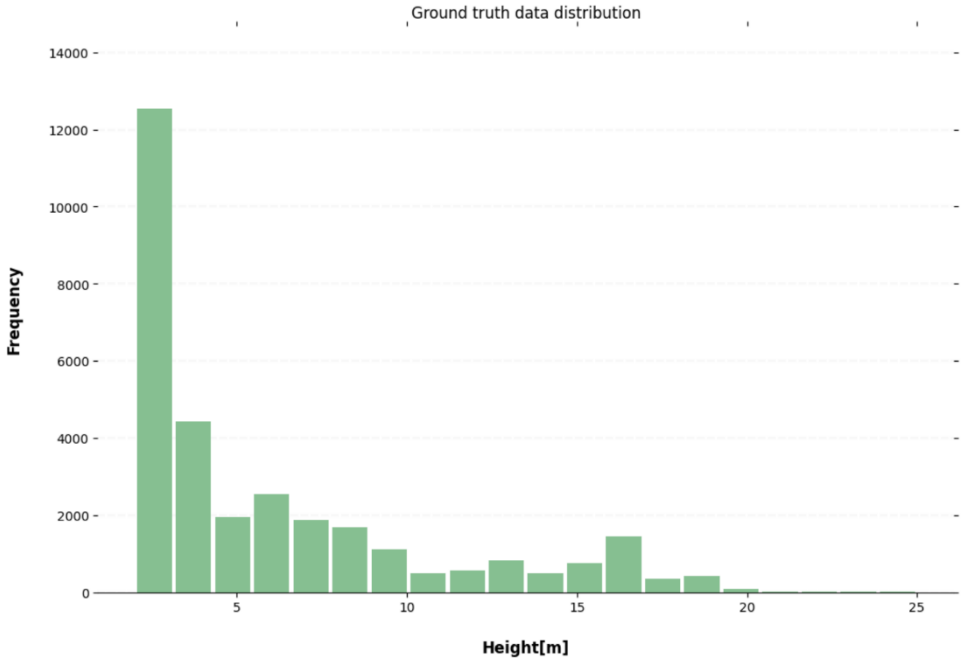


Figure 27: Distribution of heights in Ground truth data.

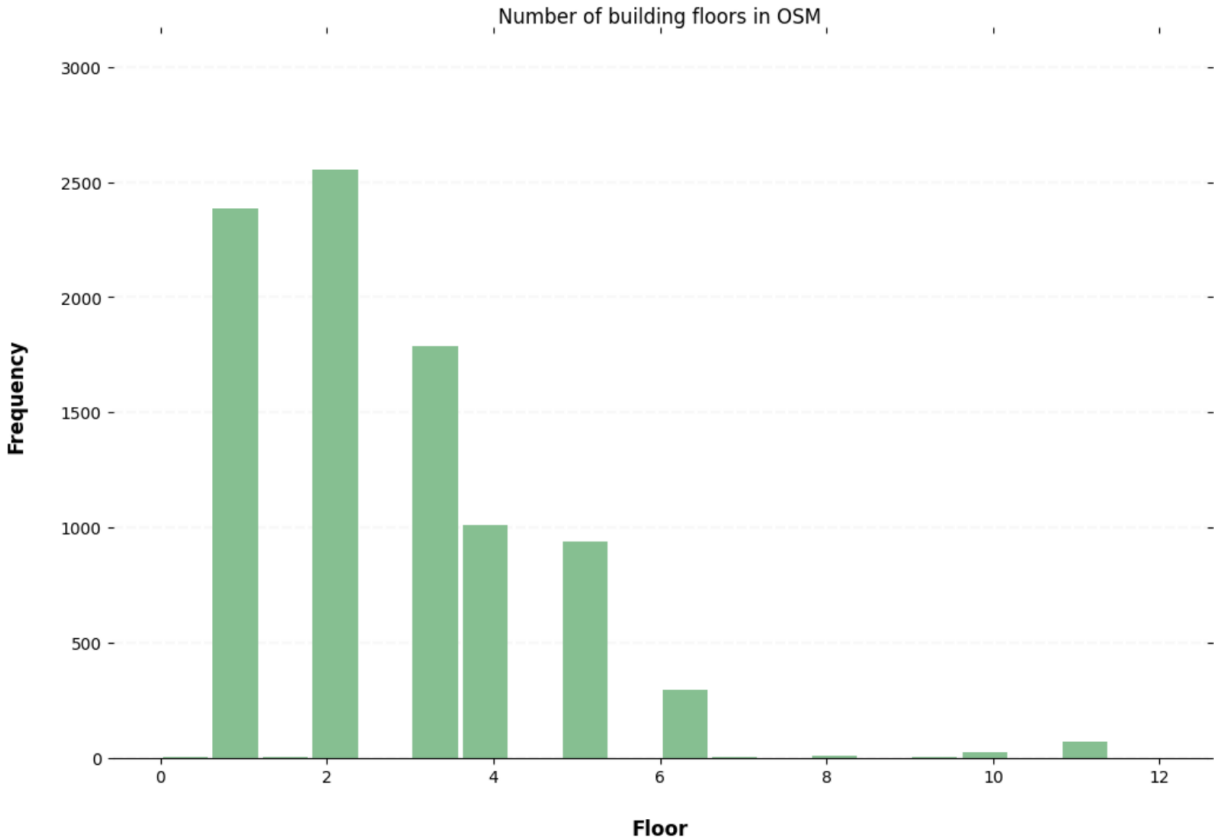


Figure 28: Distribution of number of stories in OSM data

The OSM building data contained a total of 40.608 buildings of which only 0.2% of them had building height (=81 buildings), 20.78% with number of stories/floors (=8438 buildings distributed as shown in Figure 28) and 20.85% number of stories or building height (=8467 buildings), as shown in Figure 29. This data source needs to be declared as significantly fragmented and incomplete.

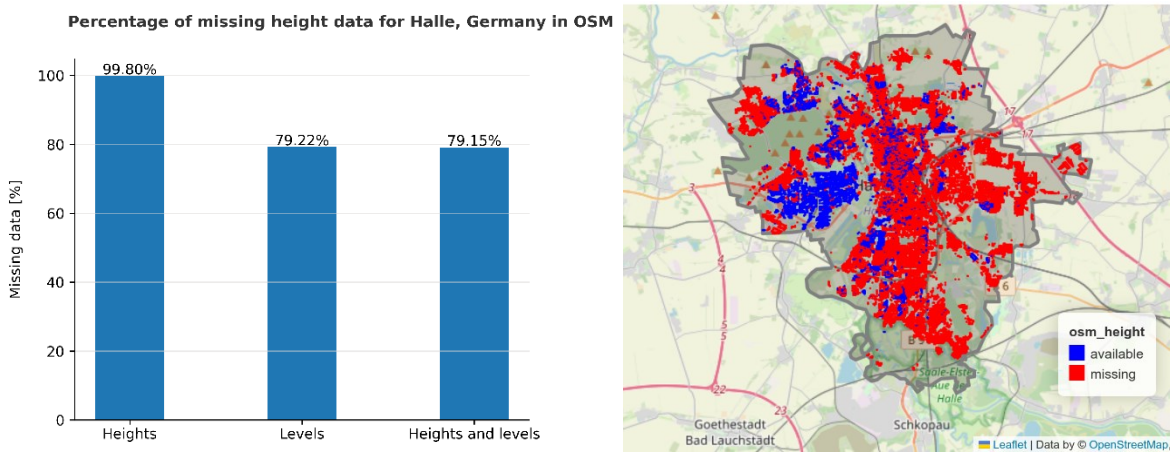


Figure 29: Percentage of missing height data (left). Area covered by missing height data(right)

The elevation model datasets (DSM and DTM) were retrieved in point cloud format (.xyz). Each dataset contained around 25x10e6 data points. The OSM layer with is vector definition of building outline was used to select DSM and DTM data points falling inside each building polygon. The mean of DSM and DTM on each polygon was used to estimate the building canopy height as the difference DSM – DTM, covering a total of 27805 buildings (table 4). The distribution is presented in Figure 30.

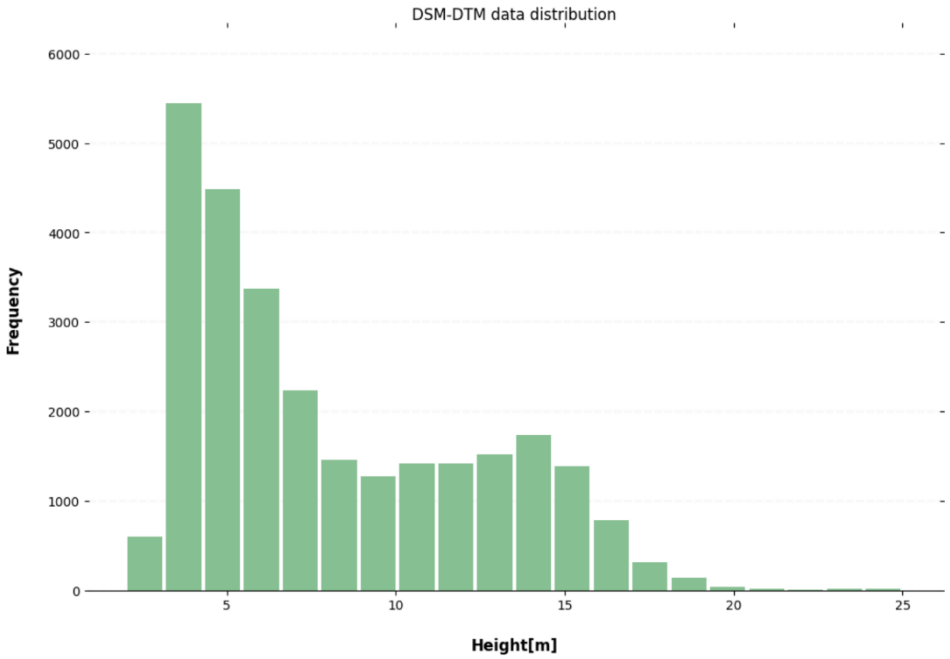


Figure 30: Distribution of canopy heights in the Digital Surface Model - Digital Surface Model.

As a validation step, we have used the Copernicus Urban Atlas land used/land cover (UA18)[10] to assess the OSM data and make sure if covers a sufficient number of buildings in Halle. First, we have filtered the UA18 data to keep only artificial areas (class 1). The UA18 after filtering, as well as the OSM are presented in Figure 31. As it can be observed, UA18 has significantly less polygons (2,524) than the OSM data (40,916). This is explained by the fact that polygons in OSM represent building, however in the UA18 data polygons represent huge parts of artificial areas (i.e., multiple building by a single polygon). To have a general overview of polygons present in UA18 and not in OSM, we have used the "Select within a distance" function of QGis (see Figure 32). Given two vector data layers A and B, the function "Select within a distance" filters polygons from layer A for which the closest polygon from layer

[10] https://land.copernicus.eu/local/urban-atlas/urban-atlas-2018?tab=metadata

B is distant with at-most d meters. The outcome with different distance values, d, is presented in Figure 33. We can clearly see that only 0.9% of polygons are absent from OSM with 100 m distance. Even with a very small distance (d = 1m), only 8.5% of polygons are absent from OSM. This confirms that OSM data covers a sufficient number of buildings in Halle.

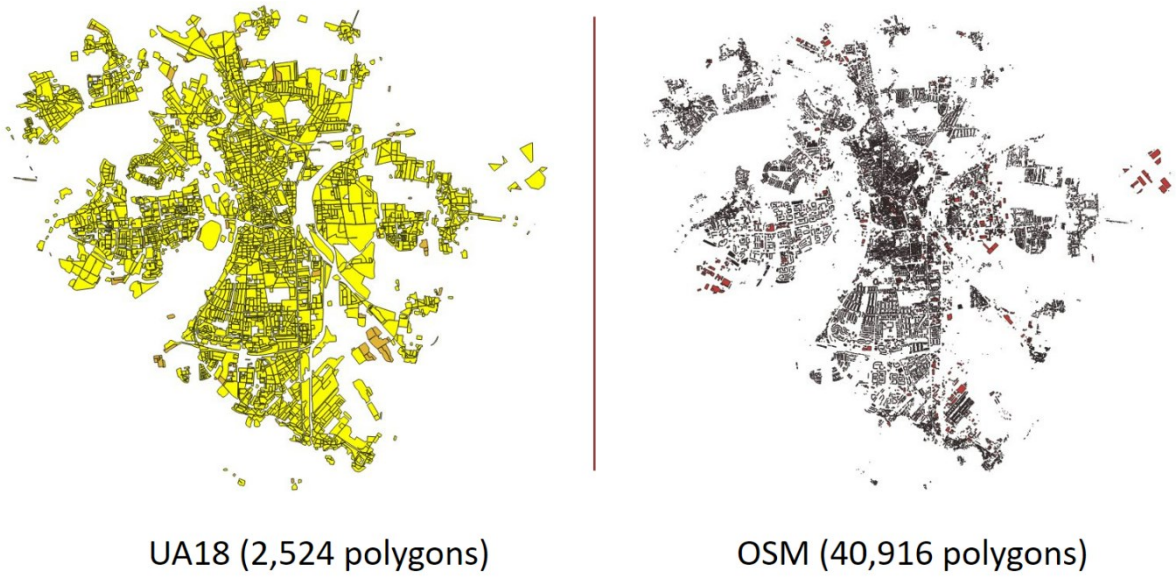


Figure 31: Polygons of both the UA and OSM plotted using QGis

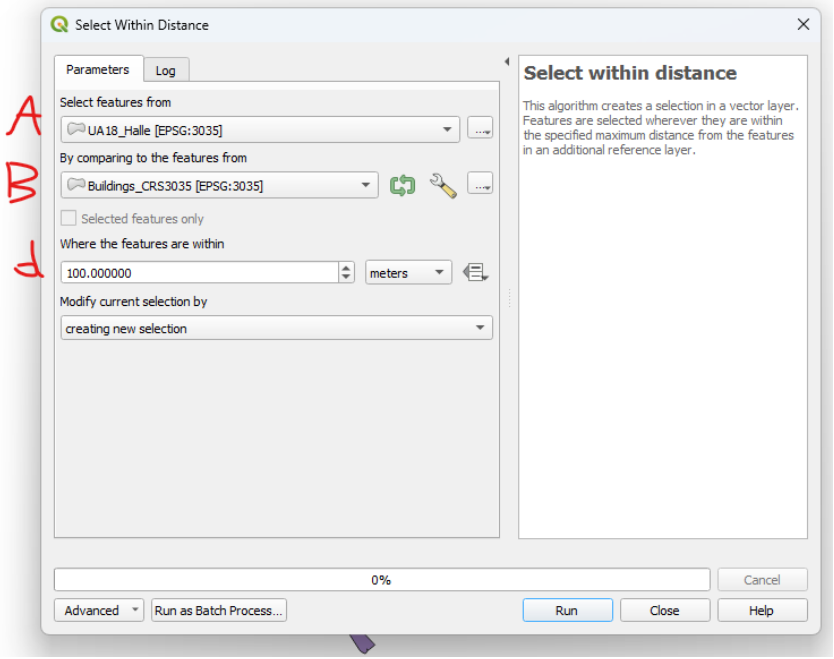
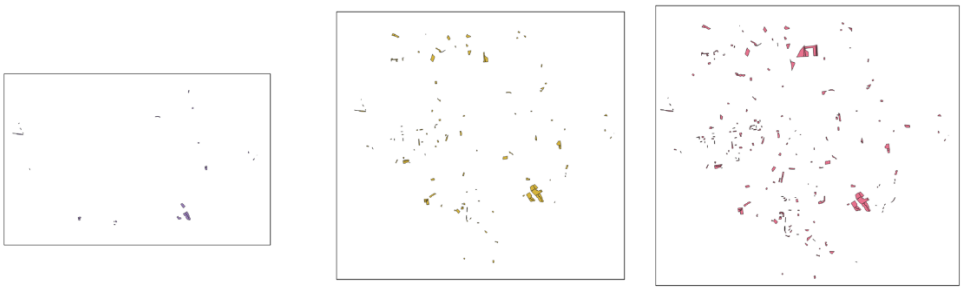


Figure 32: Overview of the QGis function *select within a distance*.

If not already available in 10 m spatial resolution, the output building height data was down sampled to 10 m resolution to make them comparable to the ground truth and rasterized as GeoTiff. Binary overlap layers between the GeoTiff layer of the ground truth and the results in all the three estimation methods were generated. The different overlap layers were used as a binary mask to extract data from the different estimation results and the ground truth. A description of the ground truth and the different estimation methods (D3.2) just before down sampling and rasterization is presented in Table 4.



d = 100m

23 out of 2,524 polygons (0.9%)

d = 10m

120 out of 2,524 polygons (4.7%)

d = 1m

215 out of 2,524 polygons (8.5%)

Figure 33: Outcome of "Select within a distance" between UA18 and OSM layers.

Table 4: Descriptive statistics of the different datasets (in GeoJson) just before down sampling and rasterization for comparison with the ground truth.

Parameter	Ground truth	DSM - DTM building height estimation	Copernicus Urban Atlas building heights
Max [m]	99.8	60.9	77.0
Mean [m]	5.05	8.2	8.5
Min [m]	0.005	3.0	3.0
Number of buildings	43960	27085	-

In preparation for the estimation of stock of materials and energy performance of primarily housing buildings, we have been reviewing and discovering data that can be used to provide the number of building floors / stories. The Copernicus urban atlas building heights data layer is available only for the year 2012 and is founded on the principle of subtracting the digital surface from the terrain data. This data is also available for our city of Halle test case. We have therefore also evaluated the data separately. Open street map (OSM) provides the number of stories and partially the heights of buildings but exhibits a reduced data completeness, for only a fraction of buildings, this data is available. In a further task, mainly described in deliverable D3.2, we will focus on improvement of coverage OSM number of building floors using a published ML gap filling technique. The Copernicus Urban Atlas building heights data and the subtracted data layer *DSM – DTM*, respectively, have better coverage and will be the input to the comparison of several building hight estimation methods. We aim to identify an optimal method or a strategy of applying methods based on data availability and completeness.

The outcomes of the carried-out investigation, presented in Table 4, shows the advantage of ground truth data compare to the other methods. However, the availability/accessibility of ground truth is very limited, making it hard to be generalized for cities across Europe. Since the aim of UC4 is to identify a flexible mode that can be easily applied to the four cities of choice (Oslo, Vienna, Luxemburg, and Barcelona), it is necessary to find a generalized approach that fits well with a range a variation. Hence, it is decided to choose the second-best method to estimate building height. However, it is worth noting that UC4 leaves the possibility of using ground true, as it contains.

The first city to estimate its residential energy demand and later calculate potential environmental affects and in-use stocks of building materials was the municipality of Oslo. In doing so, a list of primary data is extracted to carry out the estimations. Since the ground true data are not available (at least for free), the first two primary data to estimate canopy height of buildings are digital terrain model and digital surface model. Figure 34 shows these two data sources side-by-side. At the first glance, these

two data sources lookalike, but at pixel level their differences are clear. The calculated canopy height from these two data sources is presented in D3.2.

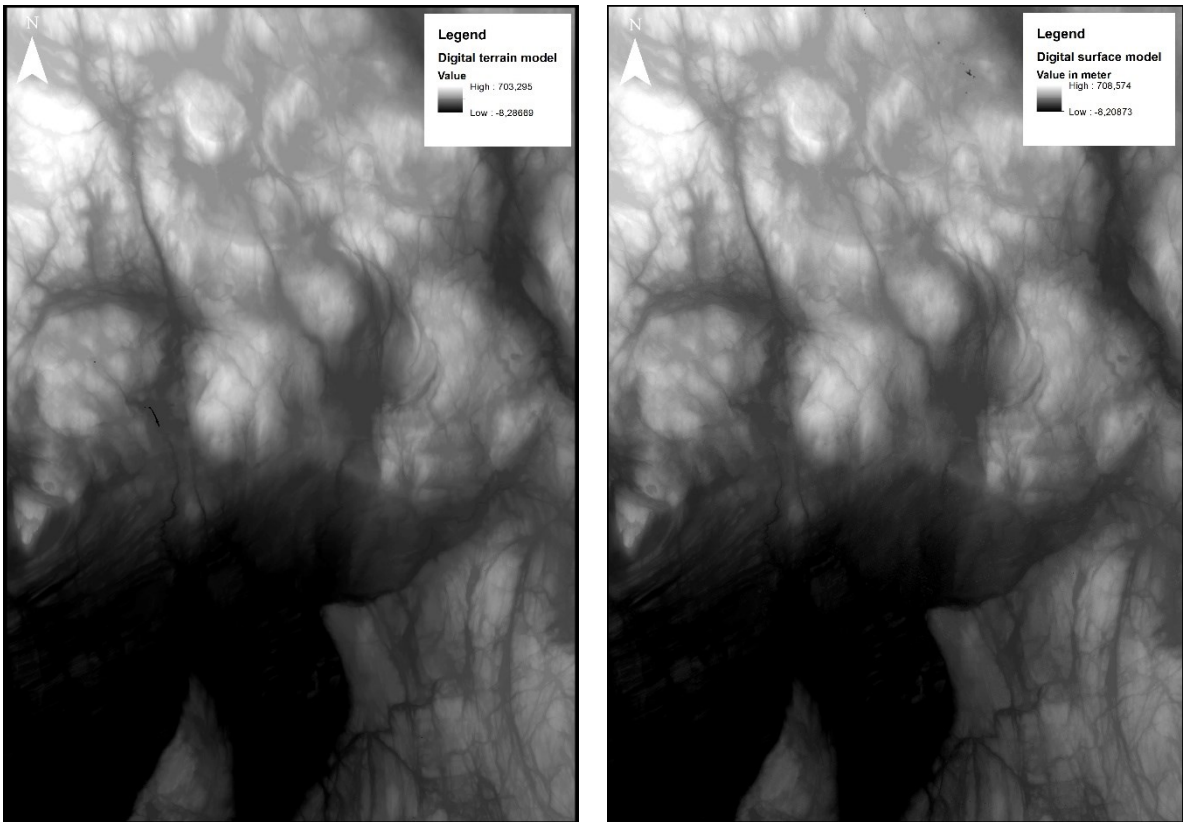


Figure 34: Digital terrain model (picture on the left) and digital surface model (picture on the right) of Oslo.

Besides DTM and DSM, local data are also used. Collection of local data was necessary as it was rather difficult to attain information about the construction year (i.e., the year of construction completion of a building where it is located) and building type by its functionality (i.e., for the time being only residential building are considered with annotations specifying whether a residential building is a single-family house, terraced house, multi-family house, or apartment). We requested these two additional data sources by contacting the municipality of Oslo and we received the data via email.

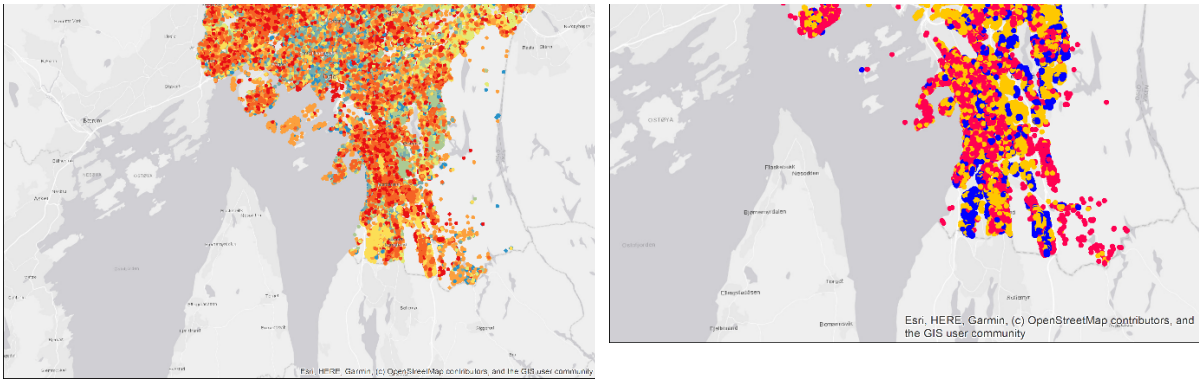


Figure 35: Construction year cluster (picture on the left) and building types (picture on the right) in Oslo.

Based on the presented data here for the city of Oslo, UC4 uses them to estimate building geometry (i.e., building height and shaded wall) to estimate energy demand for residential buildings. D3.2 explains the carried approach.

UC5 Validation of Phytosociological Methods through Occurrence Cubes

UC5 aims to validate the traditional methods applied in phytosociology to characterize and classify plant communities and to develop a new phytosociological approach to characterize and predict the presence of plant communities for yet unknown localities. This will be approached by linking distribution data of plant taxa and vegetation communities based on habitat types with EO environmental data.

2.5.1 Data sources

First and foremost, a list of habitats will be chosen from the EUNIS classification (European Nature Information System) of Habitat types. Occurrence data of the diagnostic taxa of the habitats will be gained based on records from human observations and collection samples from the Global Biodiversity Information Facility (GBIF, www.gbif.org) and an online collaboration platform for botanical collections (JACQ). The rest of the taxa comprised in the vegetation units of the Habitats chosen will be got from vegetation units present in Mucina et al. (1993) [11] and their occurrences gained from GBIF and JACQ.

When getting occurrence data of the taxa, which comprise coordinates and date of collection or observation, we will also check the taxonomy of the entities as they can be considered as taxonomic synonyms and/or additionally at infraspecific level. If this is the case, we will also request occurrence data for those listed as accepted taxa and all the infraspecific taxa.

Lastly, the vegetation communities will be based on Austrian communities from Mucina et al. (1993),[12] but the distribution of the taxa involved will be extended to all Europe when obtaining occurrence data from GBIF.

2.5.2 Data Cubes

A set of Data Cubes, based on occurrences of taxa, will be produced by combining biotic and abiotic data from Rasdaman services together with taxon occurrence data using the tool Wormpicker developed by UC3. The tool will retrieve EO point estimates from the Rasdaman interface based on point coordinates derived from taxa occurrences. Furthermore, a set of Community Cubes will be obtained from raster data of the vegetation communities chosen.

2.5.3 Sample data set

As a first sample data set, we started to collect the vegetation data for the EUNIS Habitat N1J4 (Mediterranean and Black Sea dune-slack grassland and heaths). The two diagnostic taxa for this habitat were *Salix rosmarinifolia* and *Salix arenaria*. When consulting GBIF database, *S. arenaria* was indicated to be the synonym of *S. arenaria subsp. argentea* and was therefore included in our dataset.

Once obtained the coordinates and corresponding dates of these three taxa from the GBIF database, we retrieved estimates for the environmental factor ´Air Temperature near surface´ from the Rasdaman interface using the Wormpicker software.

2.5.4 Data analysis and ML

[11] Mucina, L., Grabherr, G., & Ellmauer, T. (1993). Die Pflanzengesellschaften Österreichs-Teil 1: Anthropogene Vegetation, Teil 2: Natürliche waldfreie Vegetation, Teil 3: Wälder und Gebüsche. *Fischer, Stuttgart*.

[12] Mucina, L., Grabherr, G., & Ellmauer, T. (1993). Die Pflanzengesellschaften Österreichs-Teil 1: Anthropogene Vegetation, Teil 2: Natürliche waldfreie Vegetation, Teil 3: Wälder und Gebüsche. *Fischer, Stuttgart*.

In the UC5, data analysis will be carried out to investigate the distribution patterns of the taxa where plant communities have been identified. Furthermore, ML approaches will identify relations between identified communities and EO data, determine locations with favourable environmental conditions and predict possible presences at these locations of plant communities corresponding to the ones investigated. However, to date, analysis of the distribution patterns of taxa and vegetation communities and Machine Learning strategies still need to be addressed.

3 Summary and conclusion

Given the current status of the data ingestion and the UC owners progress to identify and describe scientific research questions in terms of data relationships that are to be discovered and exploited, a data exploratory analysis has been performed for each use case.

UC1 successfully tested the collaboration on the EOX Hub across partners and with a focus on how to upload and register own data. The clustering exercise was a demonstrator and a first step towards finding similarities of European cities according to the Urban Atlas land classification.

UC2 constructed a comprehensive architecture how to relate environmental and agricultural data to describe biodiversity and will discover available data sources while making a start on biodiversity observations. The main focus will be to harmonize, increase data completeness and regularize the input data to create data cubes ready for ML applications.

In UC3, the genetic allele frequency data was studied and analysed to prepare for a gap filling exercise. Due to the sparsity of sequenced populations, the goal is to increase data completeness. First interesting patterns were identified that can already be a starting point for correlating high genetic variance with other external [environmental] data.

UC4 discovered and described several data sources that are input to the estimation of building height which is a first crucial parameter for the estimation of stock of materials and energy performance of buildings.

Finally, UC5 started late, and the data exploration phase is not complete yet.