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

Work Package 2: Use
Deliverable 2.2: Use Cases Analysis Plans

Deliverable Lead: 4sfera
Deliverable due date: 30/06/2024

Version: 3.4

Document Control Page

Document Control Page	
Title	Use Cases Analysis Plans
Creator	Jaume Targa and Cristina Carnerero and María Colina
Description	Deliverable D2.2 Use Case Analysis Plans
Publisher	"FAIRICUBE – F.A.I.R. information cubes" Consortium
Contributors	Use Cases leads
Date of delivery	30/06/2024
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
0.0	20/11/2022	Jaume Targa and Cristina Carnerero	Internal draft
0.1	13/12/2022	Jaume Targa and Cristina Carnerero	First draft
0.2	21/12/2022	Jaume Targa and Cristina Carnerero	Incorporate reviewers' comments
1.0	22/12/2022	Jaume Targa and Cristina Carnerero	Final Month 6 delivery
2.0	29/05/2023	Cristina Carnerero	Internal draft for M12 update
2.1	27/06/2023	Stefan Jetschny and Jaume Targa	Review and adding required input to UC
3.0	19/12/2023	Several UC authors	Update information on UC
3.1	20/12/2023	Jaume Targa & María Colina	Ready for review
		Mirko Gregor	Review
3.2	15/01/2024	Jaume Targa	Final
3.3	01/06/2024	Jaume Targa	Addressing reviewers comments, use cases 5 added in more detailed, UC5 processing and analysis workflow included
3.4	18/06/2024	Stefan Jetschny	Internal review, format checking, minor edits of UC allocation to FAIRiCUBE Hub services

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.

This document and its content are the property of the FAIRICUBE Consortium. All rights relevant to this document are determined by the applicable laws. Access to this document does not grant any right or license on the document or its contents. This document or its contents are not to be used or treated in any manner inconsistent with the rights or interests of the FAIRICUBE Consortium or the Partners detriment and are not to be disclosed externally without prior written consent from the FAIRICUBE Partners. Each FAIRICUBE Partner may use this document in conformity with the FAIRICUBE Consortium Grant Agreement provisions.

Table of Contents

List of Figures

1 Contextualisation of this deliverable

1.1 Overall objective of WP2

The overall objective of Work Package 2: Use (WP2) is to ensure efficient execution of the use cases, assuring those potential synergies pertaining to both data and processing are identified and leveraged.

1.2 Description of WP2 work

WP2 focuses on the execution and cross-coordination of the different use cases (UC) on the FAIRiCUBE. Acting with an "outsider" role, the supervision crosscuts through all UC activities ensuring harmonisation with both upstream (data sources, ingestion, and processes) and downstream (results, promotion, and distribution of outputs) activities.

1.3 Description of Task 2.3

Task 2.3 (Use Case Analysis) focuses on detailing and documenting the objectives for each UC, as well as analysing and describing similarities on UC analysis approaches across UCs.

2 Analysis of Use Cases

The Use Case Analysis focuses on detailing and documenting the objectives for each UC. This information is being generated as each UC makes progress throughout the project. Part of the work under WP2 is to analyse and describe similarities on UC analysis approaches across UCs.

Before the Use Cases are analysed to find similarities and synergies, these are briefly described in terms of objectives, data analysis plan and outcome expectation.

2.1 Brief description of Use Cases

UC 1: Urban adaptation to climate change

2.1.1.1 Objectives

As reported by the most recent EEA report on urban adaptation to climate change, cities face a lot of challenges combatting the impacts of climate change, such as (i) mitigating the Urban Heat Island effect; (ii) providing shading and cooling through urban green spaces and trees; or (iii) adapting to changing precipitation patterns and preparing for heavy rains and associated flash flood events. Climate change also causes pressures on (urban) biodiversity by the changes in temperature and precipitation patterns (heat waves, drought, wildfires, torrential rains, flash floods) and on agricultural surfaces and the entire agricultural system, also due to the changing patterns of temperature and rainfall. Other land use activities do also have an impact or lead to an exacerbation of the risks, such as land take, sealing of surfaces or the removal of green spaces and trees/forest. Thus, climate change together with human activities exert a lot of pressure on ecosystems, one of which are cities (urban ecosystem).

This situation triggers activities on two spatial levels. On the European level it is required to gather information on the different impacts of climate change across the continent and if and how other factors apart from climate change influence the adaptation capacity of cities. This analysis can be realised, for instance, by collecting a large amount of data from the land cover/land use, climate or socio-economic domains and subject them to a cluster analysis. This results in several groups of cities that share similar characteristics. We can then determine which of the factors have the largest impacts and provide this information to the relevant institutions that have the mandate to address local or national governments and inform them. By providing higher weights to certain parameters, we could also steer the cluster analysis to be more relevant for defined policy targets (e.g., urban green spaces). The cluster analysis can also help to identify positive examples, best practice cases if you will, which could serve as an inspiration for other, less favoured cities. We're currently continuing to find other applications that would help to create meaningful information from the vast data that we have at hand already.

The second spatial level that we aim to serve is the local level, i.e., the cities themselves. It is indispensable that they put in place concepts and measures that identify and set up clear objectives and concrete actions to mitigate the impacts and adapt to the future situation. Following the management principle "If you can't measure it, you can't manage it", the basis for all actions are reliable and accessible data and information of high quality. Even more than on the European level, data comes from different sources, are of different quality and often lack metadata or information on their sources and processing. Likewise, they come in different formats which makes it difficult to combine and integrate them to derive more specific and customised information.

For all these issues, data cubes and the integration of data therein can be a powerful tool to receive the information they need. This information can be provided in a tailor-made format, depending on the knowledge and interest of the respective stakeholder. While policymakers will prefer non-technical and concise information (short factsheets and visualisations) in an easy language, data engineers or researchers will be interested in the underlying database and more specific maps and visualisations and might even support further development of the analytical process.

Eventually, the FAIRiCUBE Hub will provide a unique access point to data and tools that can be used by experts to process various data sets with the aim of receiving customised results for their immediate use.

2.1.1.2 General research questions (formulated at the beginning of the project)

The main research questions of UC1 can be summarised as follows:

- Can big data (historical, real-time, and modelled forecast spatial data) and Machine Learning approaches help European cities to prepare for the impacts of climate change and take adaptive measures/make informed decisions?
- Would it be possible for cities to simulate the impacts of the implementation of certain policies and decisions, e.g., which impact would the 10% urban tree cover target of the Nature Restoration Law (NRL) have on the other critical parameters of cities that do not yet comply with this target (e.g., urban heat island effect, air pollution, noise, or human health)?
- In how far can data cubes enable local, regional, national, and European decision-makers to achieve the goals of the European Green Deal?
- Does the European Green Deal data space provide the best possible means to collect, store and provide European data on climate change impacts on cities?

2.1.1.3 More specific research questions

- What are commonalities of and differences between cities in terms of their situation with regard to climate adaptation taking into account comparable European data sets from various sources (Copernicus, ESTAT)?
- Can a cluster analysis identify groups of cities with similar conditions so that "good" and "bad" cities become apparent and can serve as "good" and "bad" examples for the analysis?
- Can we develop a cluster analysis that allows to set weights to defined parameters to make them more "important" than others?
- Can we identify the impacts of certain measures in cities? E.g., what does it mean to increase the urban tree cover by x% as requested by European policies?
- Can we help cities in identifying the most suitable areas for a retention area or a green space so that these are located in the most ideal place? For instance, where does a city need to allocate areas as not suitable for building activities because this area is of high risk of flooding and could serve as retention area; or where would a city need to create new urban green spaces so that more inhabitants have a shorter connection to their closest green space?

2.1.1.4 Required input data

Below we present a non-exhaustive list of potential input data sets. Many of them come from Copernicus services, such as the Copernicus Land Monitoring Service (CLMS) or the Copernicus Climate Change Service (C3S). Some of the data sets are ready to be used immediately (e.g., most of the CLMS data), while others need to be accessed via an API or downloaded to be used in our system (C3S climate data). We've encountered several problems with the climate data, in particular the processing of NetCDF files is very slow and, thus, expensive, which necessitates a workaround by converting them first into a cloud-optimised file format (zarr) before processing them in the AWS S3/EOXHub system.

In addition, other data sets and indices are also readily available to be included in the use case on demand. In their case, some pre-processing (such as rasterisation) will be necessary.

The following list provides a high-level overview of relevant data sets:

- Copernicus Land Monitoring Service (CLMS) products:
  - CLMS Pan-European component, e.g., CORINE Land Cover, High-Resolution Layers Imperviousness, Forest, Grassland, Water and Wetness, Small Woody Features (https://land.copernicus.eu/pan-european)
  - CLMS Local component, e.g., Urban Atlas (including land use/land cover, street tree layer and building heights), Riparian Zones, Natura 2000 (https://land.copernicus.eu/local)
- Copernicus Climate Change Service (C3S) Climate data:
  - Climate Data Store (https://cds.climate.copernicus.eu/#!/home).

- Climate-ADAPT platform (https://climate-adapt.eea.europa.eu/).
  - European Climate Data Explorer (https://climate-adapt.eea.europa.eu/knowledge/european-climate-data-explorer) providing data produced in the context of the Copernicus Climate Change Service (https://climate.copernicus.eu/)
  - European Climate and Health Observatory (https://climate-adapt.eea.europa.eu/observatory)
- Other data:
  - Data collected and accessible via EEA Spatial Data Infrastructure (https://sdi.eea.europa.eu)
  - OpenStreetMap
  - Eurostat Urban Audit, socio-economic data

2.1.1.5 Data Analysis

This use case covers "cities" as spatial entities of analysis. The term "city" can be understood differently by different people. It can be an administrative unit (delimited by the administrative border), a morphological unit (characterised by the urban fabric) or a functional unit (adding the commuting zone to the core city from where people travel into the city for work). On the European level, the reference units for cities are the "City" which corresponds to the core city (an administrative unit), the "Commuting Zone" (also delimited by administrative borders but identified according to their function) and the "Functional Urban Area (FUA)" which is the sum of the two (i.e., aggregating the city and commuting zone). We will use this definition in our use case.

The following activities are foreseen or already being implemented in this use case:

- Collection of climate-related data and recent land use/land cover data with the highest spatial resolution possible; creation of a cube (using a dedicated AWS S3 bucket and EOXHub, a cloud-based platform with a direct link to several DIASes that is also supposed to be the basis of the FAIRiCUBE Hub) that integrates such data together with thematic data, such as information on green urban areas, flood-prone areas, location of residential and commercial/industrial areas or socio-economic data like total population or population density.
- Using ML approaches for the calculation of various indicators and information products for cities (covering FUAs, core city and commuting zone) for different stakeholders, ranging from policy- and decision-makers over scientists and experts to the wider public, thereby combining the land use/land cover data with the climate data that are of relevance at the city scale (e.g., Urban Heat Island data). We aim to work on a subset of the entire Urban Atlas data layer. The main output should be the cube and several visualisations of the main information products, e.g., by using business intelligence software. This process includes gap filling, in particular for the socio-economic data which are patchy.
- Assess the applicability, usefulness and quality of the new data cube and the calculated products (e.g., indicators, visualisation tools) compared to the existing information; assess the compliance of the data cube with the European Green Deal data space requirements.
- Collection of local data with a much higher resolution and analysis of their usefulness and the potential to be combined with the European data. Attempt to integrate these data from different levels into data cubes to be able to develop and implement simulations of the decision-making process (e.g., what happens if I change parameter x?).

Figure 1 presents the expected processing workflow in UC1. The flowchart describes the data flow and processing steps for the first part of the use case, namely the analysis of cities across EU. In this part, a preparatory step is the calculation of descriptive indicators for a large number of cities across EU and for several years. These indicators are partly derived from EU-wide spatial datasets (land-based and climate data) and partly are already available as socio-economic indicators. All the indicators are eventually collected into a data cube where the spatial coordinates are not the traditional geographic coordinates, but rather the city identifiers, which can in turn be linked back to geographic coordinates by using either the city boundaries or the city centre point coordinates.

There are three different types of data sources:

- Land based data, mostly derived from the Copernicus Land Monitoring Service (CLMS).
- Climate data, mostly originating from the Copernicus Climate Change Service (C3S); and
- Socio-economic data from the Urban Audit database hosted by Eurostat.

Additional spatial datasets already publicly available in the EDC Hub will also be used.

The following processing steps are being implemented:
Land and climate data are harmonised and ingested in a spatial data cube (spatial coordinates are lat/lon). Harmonisation consists for example in ensuring that the dimensions all share the same projection and geographical extent.

- The spatial data cube is used to compute various city indicators. One indicator has one value per city/timestamp.
- The spatial data cube will also be used in subsequent spatially-aware analysis (e.g. green distribution with cities) and for visualisation purposes.
- These land- and climate-based indicators are then fed into a city data cube (technically implemented as a postgres database for the time being). The city data cube differs from the spatial data cube because the spatial dimension (coordinates) are not the traditional geographic coordinates (e.g. lat/lon) but rather the city identifiers. The spatial dimension in the traditional sense can be at any time recovered by linking the identifiers to the city boundaries or centre point coordinates. This data representation is less memory intense and is more practical for further analysis.
- At the same time, socio-economic data is ingested. Socio-economic data do not have spatially explicit coordinates, but rather they are indexed by the city identifier. Therefore, indicators are directly fed into the city data cube, and derived indicators are computed.
- Attempts at ML-based gap filling have been made on the socioeconomic data, but with limited success due to the large extent of the gaps. Rather the original dataset has been heavily filtered down to the indicators that have sufficient data.
- Cluster analysis is then carried out on the city data cube. The goal is to generate different clustering scenarios driven by different themes/questions (e.g. urban heat, flood retention, green infrastructure). This can be achieved by weighted clustering, where a weight is assigned to each feature (indicator) to control its influence on the clustering. An advantage of weighted clustering is also that it can be easily tuned to different needs, thereby generating multiple scenarios relatively quickly.
- Ultimately, this framework can be used to run simulations, by tuning the indicators values and measuring their influence on the outcome.
- The city data cube can be linked to visualisation tools to create dashboards and city fact sheets.

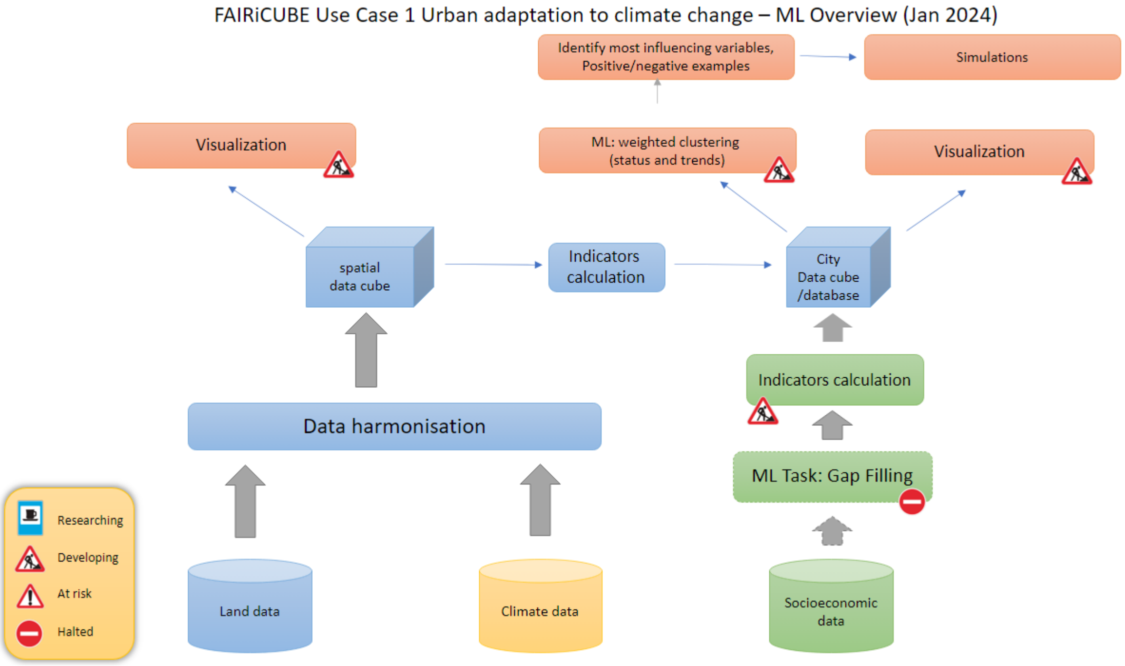


Figure 1: UC1 data, analysis, and processing workflow

2.1.1.6 Status and expected results

In the past months, we have been working on several strands of work in parallel:

- Collection and ingestion of data into the AWS S3 bucket and the EOXHub.
- Setting up of a PostgreSQL database (with PostGIS extension) to store our data and make them available for Tableau as visualisation software.
- Developing several scripts for data ingestion, data exploration, and data conversion.
- Working on gap filling of data, in particular the socio-economic data for which we find gaps in the time series.
- Setting up a Tableau dashboard to show the ingested data sets and explore them visually (https://public.tableau.com/app/profile/stefan.kleeschulte/viz/FAIRICUBECITY-Eurostat/Story1?publish=yes).
- Preparing the concept for a cluster analysis using the largest number of data sets with sufficient coverage; identification of policy-relevant priority parameters that will get a higher weight in the analysis.

The expected results of this use case consist of a cubified database system containing all available data at 10m and 100m spatial resolution. This system will incorporate dashboards with indicators and indices describing the status and trends of the different parameters and the ability to run simulations allowing the user to simulate changes in the overall system when changing single parameters. This will allow the users to see which parameters have the strongest influence on mitigating and adapting their urban ecosystems to climate change. Business intelligence software may be used to build the visualisation of the main products.

UC 2: Agriculture and Biodiversity Nexus

2.1.1.7 Objectives

Focusing on biodiversity, one of the European Green Deal priority actions, this use case considers the agricultural landscape as the main environment to investigate the impact that farming activities have on biodiversity.

To describe a basic conceptual design of biodiversity assessment, this use case uses the Dutch Biodiversity Monitor (DBM), which measures the effect on biodiversity resulting from the impact that farming has on the physical conditions of the environment expressed by Key Performance Indicators (KPIs). This makes it possible to monitor the role of agricultural activities in the preservation of the landscape and the environment in a standardised approach. KPI data are complemented by other relevant agricultural data which enable additional analysis at agricultural parcel and landscape level. This use case focuses on finding correlations and causal relationship between the obtained farm activity indicators and biodiversity estimations. Interpretable AI and Causal machine learning will be used to attribute detected changes in biodiversity to specific agricultural activities. This study will be implemented on the FAIRiCUBE hub where all relevant data will be collected and made accessible, evaluated, and further analysed by using machine learning algorithms.

The main objective of this use case is to improve the knowledge about the relations between different agricultural practices and biodiversity, using a machine learning approach which is consistent at large scales. It also aims at increasing awareness about data cubes and AI in domain stakeholders involved in the smart agriculture and biodiversity fields.

This use case also aims at investigating how a data cube-based data infrastructure can improve the access to information related to biodiversity, for stakeholders interested in linking biodiversity and human activities in agricultural areas with changes in the physical conditions of the region (e.g., soil, groundwater, emissions etc.). This would provide a step forward in making more precise estimates of biodiversity in a spatial context, as well as allowing stakeholders to make better-informed decisions on selecting more nature-inclusive practices promoting biodiversity.

The analysis tools used to extract causal relationship between different interventions at farm and landscape level, specific biodiversity indicators and further derived measures of biodiversity will be provided within FAIRiCUBE for reuse both in different locations as well as pertaining to related questions.

2.1.1.8 Research questions

The main research questions of UC2 can be summarised as follows:

- Can the integration and ML-based analysis of currently available biodiversity, agriculture, environmental, and remote sensing data provide comprehensive, verifiable, and actionable insights for different regions?
- Can data cube functionality and ML help in finding causal relationships between effects of farm level measures, indicators of physical conditions, and direct measures of biodiversity?
- Can the insights obtained in the study region be extended to other regions by reusing learned patterns applying transfer learning?

2.1.1.9 Data Analysis

In the use case, three main categories are considered: biodiversity related data, environmental data, and agricultural data (Figure 2). Each of these data categories are being handled primarily within their individual processing flow where distinctive data cubes are generated. The processing flows are then ultimately merged using causal machine learning. These methods support causal inference and discovery, which can provide insights into the underlying mechanisms describing the impact of agricultural practices on biodiversity. They do not only statistically predict the correlations but also provide meaningful explanations for those predictions, enhancing the overall interpretability of the results.

As a first step, the relevant datasets on species occurrences, (bio)physical environment and agricultural land management, were acquired and ingested for the study region. The use case will mainly use datasets with local and national (Netherlands) coverage, further complemented with European or global datasets. Examples of relevant datasets considered in this use case within a study area include local scale datasets as KPI data related to farm activities (e.g. rotation index, crop diversity, annual and winter greenness…); national scale datasets as species occurrence data, land use, soil, weather, interannual farm crop field data and landscape management; and European/global scale datasets as Sentinel 1,2 and derived indices.

One essential data source to build biodiversity indicators consist of species occurrence records. For this purpose, primarily the National Databank Flora and Fauna (NDFF) will be used. However, consistent mapping of biodiversity over space is not a straightforward task. Methods of gap filling for the missing species occurrences using species distribution modelling and the concept of so-called occurrence cubes must be applied. To link spatial data of individual species occurrences to biodiversity estimates, environmental covariates will be contained in a common high dimensional feature space contributing to create a biodiversity index (BI). Ultimately as a result, the estimates will be performed in several time steps to evaluate the effect of farmers' interventions on biodiversity.

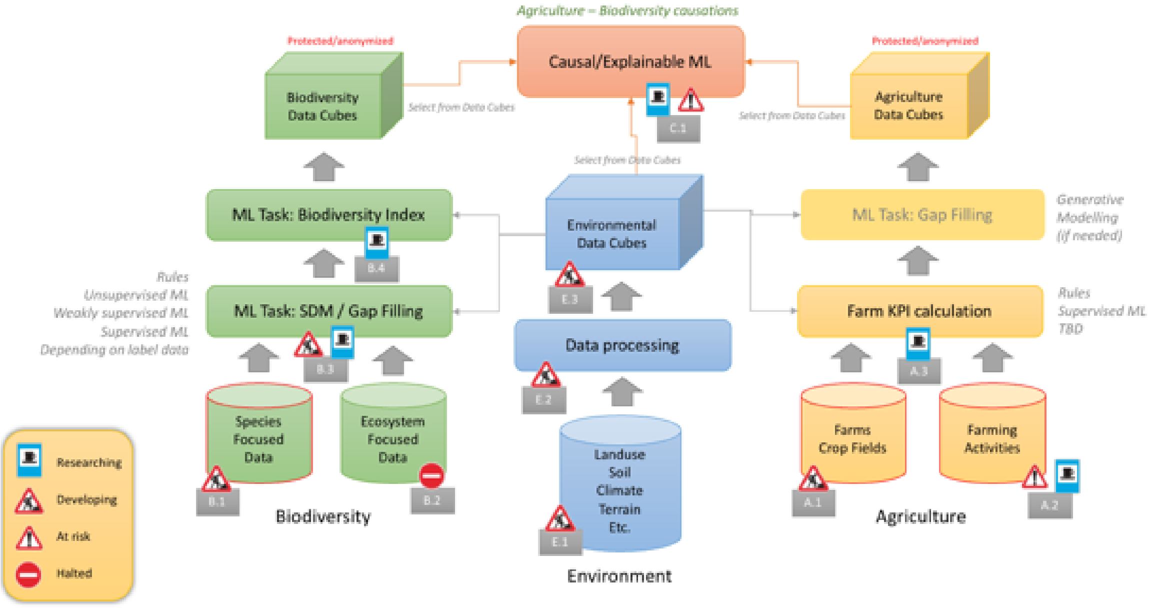


Figure 2: UC2 data, analysis, and processing workflow

2.1.1.10 Expected results

The expected results of this use case consist of a standardised procedure to monitor agricultural activities in relation to biodiversity using KPIs as variables. While KPIs are presented as integrated sets, the individual biodiversity indicators not only target biodiversity goals, but also contribute to soil, climate, air, and landscape objectives. The use case will provide a series of F.A.I.R. analysis-ready datasets for reuse by stakeholders. These will be openly accessible, when possible, but may be restricted if needed. The data processing pipelines, trustworthy AI models and deployed concept applications will be shared via FAIRiCUBE HUB and other channels for stakeholders to use.

UC 3: Biodiversity occurrence cubes – *Drosophila* landscape genomics

2.1.1.11 Objectives

The main objective of this use case is to intersect quantitative environmental and genomic datasets to elucidate how climatic and anthropogenic variation affects genetic diversity in *Drosophila melanogaster*.

The fruit fly *D. melanogaster* is one of the best-studied genetic model organisms, which is characterised by a world-wide distribution and thus very well-suited to quantitatively study the impact of the environment on genetic variation.

Combining genomic data from densely sampled populations from diverse environments with high-resolution geospatial data will allow studying ecological factors that affect local adaptation and to identify genomic targets of selection in *Drosophila melanogaster*. Using already well-established population genetics theory and approaches shall allow drawing conclusions from the evolutionary history of *Drosophila* populations for predicting future evolutionary responses to changing environments.

2.1.1.12 Research questions

Specifically, we want to address the following research questions:

1) To which extent do environmental factors influence the distribution and amount of genetic variation in natural *D. melanogaster* populations in Europe?

2) Which environmental factors contribute to adaptation in response to climate change in European *D. melanogaster*?

3) Which genes are under strong spatially varying selection and what is their function?

4) Can insights into environmental adaptation of a model organism be used to make predictions for other, non-model organisms?

2.1.1.13 Data Analysis

The analyses as shown in Figure 3, will be based on already available genome-wide sequencing data from spatiotemporally sampled populations of *Drosophila melanogaster* on a global scale (754 natural populations world-wide) generated by the DrosEU consortium (https://droseu.net/). These data were collected from over 100 locations in more than 20 countries on four continents. Several of these locations have been recurrently collected at different seasons and years, which will allow us to assess spatial variation as well as temporal changes and seasonal patterns. In our analyses, we will specifically investigate populations collected in North America (*n*=221) and Europe (*n*=343).

An existing analysis pipeline (Kapun et al. 2021) will be adapted to process the raw data which will facilitate incorporating upcoming datasets and to extend these analyses beyond *Drosophila*. The focus will be on positions along the whole-genome DNA-sequence that differs among individuals. Such polymorphic positions at single nucleotides (the basic building blocks of the DNA) are usually composed of different variants - so called alleles - that vary in frequency among populations. The allele frequency data of millions of genome-wide single nucleotide polymorphisms (SNPs) estimated by this bioinformatic pipeline will then be used for downstream analyses.

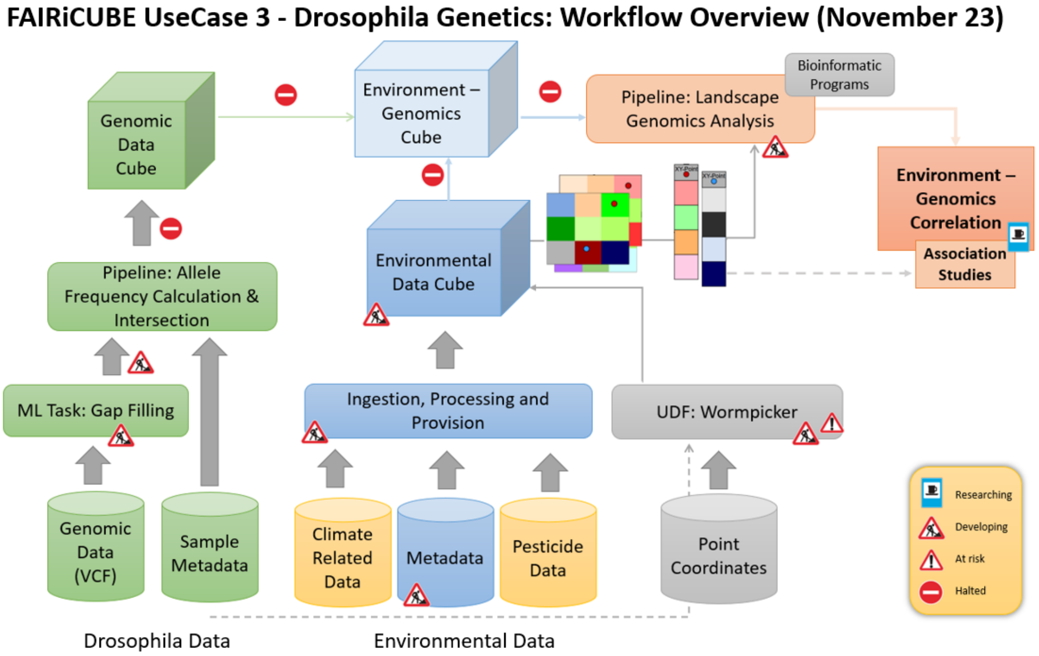


Figure 3: UC3 data, analysis, and processing workflow

An API will be built to make common data formats for genomic data (variant calling format; VCF) compatible with data structures used by the Rasdaman interface, which will facilitate intersecting variation data with gridded climate-related data (temperature, precipitation, seasonality etc.) and gridded agriculture and socio-economic data (land-use, traffic, human demography, pesticide application on crops) following FAIRiCUBE principles.

A software to obtain point data estimates from rasterised environmental data for specific coordinates has been developed, which allows to intersect point data from different domains to test for correlations among genetic and environmental data. In our approach, we will employ two statistical methods, BayPass (Gautier 2015) and LFMM (Caye, et al. 2019), which both test for correlations between allele frequencies at every genomic position and environmental factors while controlling for the evolutionary history of the populations, which may confound the detection of causative links between environmental effects and allele frequency changes due to natural selection.

Missing allele frequency data in the genomic datasets have an impact on the statistical analyses presented above and may bias the outcomes of genome-wide tests for signals of selection. To account for this problem, we developed a ML supported method to impute missing allele frequency information that takes advantage of available genetic information in neighbouring populations. More specifically, populations will be clustered according to genetic similarity (more explanations in D3.2). The missing allelic information at a specific population and genomic position will be obtained from averaging allelic data at this position in highly similar samples with available information. To test the accuracy of this new approach, we generated test datasets by focusing on a subset of the data without missing information. We then artificially removed known allelic information from each of the population samples and used our imputation approach to estimate the missing allelic information. By comparing imputed and expected allele frequencies, this method thus allows to obtain estimates of accuracy and robustness of our approach.

We will further investigate demographic factors that influence *isolation by distance* (IBD). The concept of IBD assumes that geographically proximate populations are more closely related and share genetic variation due to a limited dispersal ability. It will be assessed whether multidimensional geographic distance (elevation, geographic obstacles, human routes, etc.) better fits genetic differentiation than plane geographical distance (latitude and longitude only), which will allow to test the hypothesis that human transportation is the major source of dispersal in human-associated fruit fly *D. melanogaster*. Testing different distance measures that are either based on Euclidean distance or geographic and anthropogenic factors will thus reveal which factors contribute most to population structure in European *Drosophila*.

Additionally, possible new drivers of spatially and temporally varying selection will be identified by extending established population genetics strategies (GWAS, PCA, regression analysis using environmental data e.g., WorldClim; Fick & Hijmans 2017). These advances will be based both on the availability of extended metadata and new high-resolution environmental data as well as on using innovative machine learning approaches as novel tools to unravel hidden interactions between genomic and environmental variation.

2.1.1.14 Expected results

The expected results of this use case include the development of User Defined Functions (UDF), that can be integrated into the Rasdaman Web Coverage Processing Service (WCPS), which will simplify the intersection of quantitative genomic and environmental datasets for *Drosophila* and beyond and allow to conduct specific genetics analyses directly in the web interface.

Moreover, the analysis of correlations between genetic and environmental variation will reveal previously unknown ecological factors driving local adaptation, uncover genomic targets of spatially varying selection along the *Drosophila* genome and provide important insights into the effects of climate change on biodiversity loss. Additionally, the identification of targets of selection in response to insecticide applications may provide helpful insights into pesticide resistance and support targeted measures against related pest species, such as *Drosophila suzukii*.

Novel results and analysis methods derived from this use case, including the imputation of missing data and the transformation of genomic data to data cubes, will be disseminated in the form of scientific publications, reports to stakeholders, oral presentations, and practical courses at workshops.

2.1.1.15 2.1.3.5 References

Caye, K., Jumentier, B., Lepeule, J. & François, O. 2019. LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. *Mol Biol Evol* **36**: 852–860. Oxford Academic.

Gautier, M. 2015. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics* **201**: 1555–1579.

Fick, S.E. & Hijmans, R.J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol* **37**: 4302–4315.

Kapun, M., Nunez, J.C.B., Bogaerts-Márquez, M., Murga-Moreno, J., Paris, M., Outten, J., *et al.* 2021. Drosophila Evolution over Space and Time (DEST): A New Population Genomics Resource. *Molecular Biology and Evolution* **38**: 5782–5805.

UC 4: Spatial and temporal assessment of neighbourhood building stock

2.1.1.16 Objectives

Buildings are responsible for about 40% and 36% of energy demand and greenhouse gas (GHG) emissions, respectively, in the EU. Reducing the energy demand and environmental impacts associated with buildings have a crucial role for achieving the EU's energy and climate goals. "Renovation wave strategy" and "fit for 55" are a set of policies aiming to pave the wave to make the European building stocks energy efficient and less carbon intensive. While enhancing energy efficiency and having environmentally sound buildings assist the EU's climate neutrality objectives, the need for having a circular use of building materials become more prominent. Huge amounts of materials are consumed and stocked during the entire lifetimes of buildings, which makes buildings a mining asset for future supply of materials. Adopting a circular economy principle may assist in both reducing environmental impacts associated with production of building materials and increasing the resilience to supply chain disruption as well as avoiding increased prices of raw materials.

Ensuring that the EU's climate and energy objectives are achievable and the inventory of in-use building materials are correctly mapped, it requires a sufficient knowledge of the properties of our built environment. However, what one can often find is a mix of detailed and generic data. For newly built buildings, due to current intensive regulative codes (particularly in EU), a prominent level of clarity is demanded giving the possibility to understand the properties of assets. However, for the older buildings a detailed level of clarity is often scarce until assets are (deeply) refurbished, rehabilitated or inspected. This mixed level of insights on the properties of buildings constrains the possibility to make decisions at the national level to prioritise dwellings with the highest return on investments and promotion of circular and local materials.

The main objective of UC4 is to develop models based on data compliant with F.A.I.R-data definition to estimate (a) material use intensity and energy performance; and (b) associated greenhouse gas emissions of building stocks (LOD1 or above).

2.1.1.17 Research questions

The following research questions strive to reaching the UC4 objective:

- How can the inconsistency of data availability across European cities be enhanced by machine learning models? And what independent variables are rudimentary for training the models?
- To what extent can data cube infrastructure support actors, researcher, stakeholders, etc. to tackle the Green Deal priority action plans related to climate change and circular economy?
- How efficiently and effectively can data cube infrastructure be scaled up to cover building stock at regional or national level?
- Can data cube infrastructure knowledge be easily transferred and reproduced for other types of infrastructure (i.e., green, blue, and grey infrastructure)?

2.1.1.18 Data Analysis

UC4 will create building 3D models to estimate energy and environmental performance, as well as potential stocks of building materials. The estimations will be narrowed down to four European cities. Oslo, Barcelona, Luxemburg, and Vienna are considered for the time being to assure the diversity of the cities from the architectural and climate zone viewpoints. At the same time, one representative of each city is available in FAIRiCUBE's consortium. The availability of each city representative assists in tapping into local data faster due to circumventing language barriers.

Vectorised and rasterised data from publicly available platforms (e.g., OpenStreetMap and INSPIRE geoportal) will be used in the first place to create the 3D models. Such public data contains some necessary information to create the 3D models like buildings footprints and digital elevation model. However, in case of missing to attain necessary data from the public platforms, local data from public bodies will be considered. Since the cities of choice in this use case have local representatives in this consortium, there is a higher likelihood to bridge the data gap. However, if contacting local bodies doesn't yield a suitable outcome, the work will expand its scope by means of other data fulfilling F.A.I.R-data definition, like satellite data. In addition to the identification and collection of data, suitable machine learning techniques will be used for the gap filling of missing information.

Figure 4 presents the expected processing work in UC4. So far, three types of data sources are considered. The differences of these data are based on their spatial resolution. Ground-based data (presented in blue), remote sensing data (presented in yellow), and tabular data (presented in green) are these three data types. At this stage, ground-based data are attained from governmental data portal or local authorities, remote sensing data are attained from Sentinel hub to access satellite data, and tabular data are attained from different sources like Tabula/Episcope, literature, and statistics.

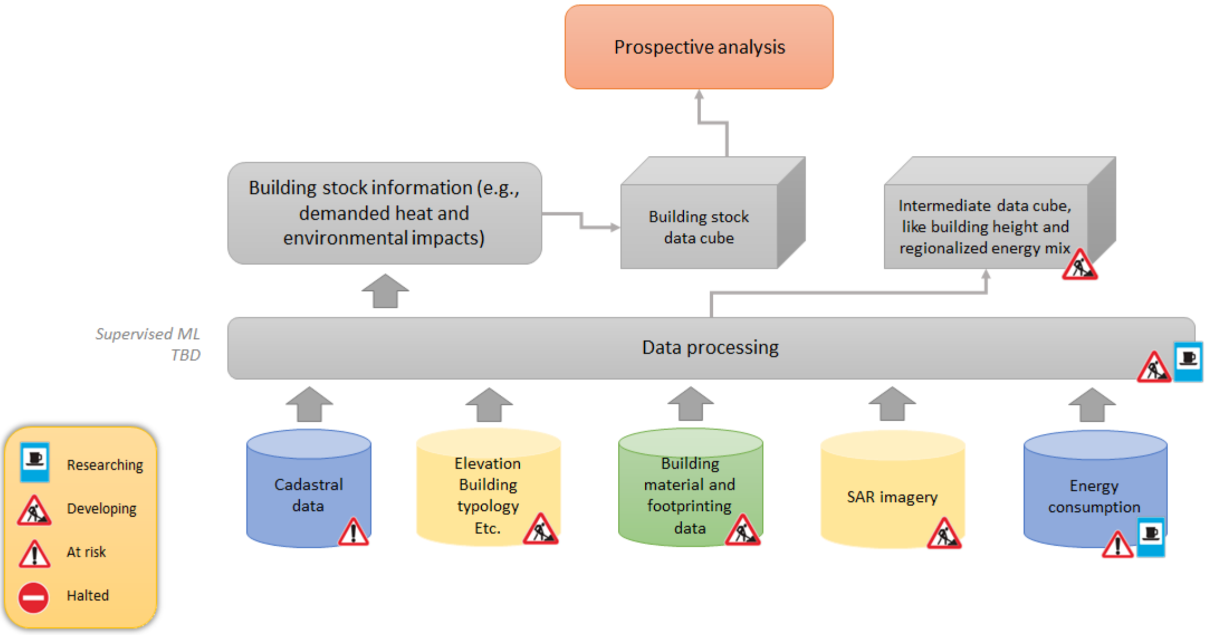

Prospective analysis
Building stock information (e.g., demanded heat and environmental impacts)
Building stock data cube
Intermediate data cube, like building height and regionalized energy mix
Supervised ML
TBD
Data processing
Researching
Developing
At risk
Halted
Cadastral data
Elevation Building typology Etc.
Building material and footprinting data
SAR imagery
Energy consumption

Figure 4: UC4 data, analysis, and processing workflow

These data types are later fed to data processing section which carries a cascade of modelling to provide estimates for the investigated stocks of buildings. Building stock information will later be stored as data cube for further analysis, like analysis of energy retrofitting of existing buildings while trying to keep environmental impacts associated with the energy retrofitting as low as possible. At the same time, some intermediate results are generated during the data processing which are grided and stored in FAIRiCUBE hub. These intermediate data might potentially be of interest for some of UCs in this project.

Since UC4 deals with both vector and grided data, different conversion methods will be tested to achieve the vector to grid data conversion. In addition, the suitability of using ML for certain processes will be evaluated and documented. This will be done to maximise the possibility of covering all options before carrying out an ML.

D3.2 provides detail explanations about these data sources and the carried-out processes.

2.1.1.19 Expected results

One of the interim results that UC4 can generate at this stage is the gridded building height which is generated by means of a digital elevation model. Also, UC4 can estimate as-built energy demand for space heating. At this stage, the estimated energy demand is in vector format, but it is intended to find a suitable way to store vector data.

For the energy demand estimate, UC4 requires some explanatory variables defining a building type by its functionality. For instance, if a building is a residential building with two important properties: construction year (e.g., Building x was built in 1976), and building type (e.g., single family house and apartment). UC4 has identified the need for ML in estimating construction year since such cadastral data are not easy to access, UC4 tests developed ML models (based on scientific research) to predict building age. The outcome of this investigation will be a generic model and results that can be scaled up for different cities.

Another expected outcome from UC4 is the map of energy mix at regional scale. At this stage, UC4 estimates total energy demand for space heating. However, it is unable to specify what energy mix might be delivered to a dwelling. UC4 will investigate the possibility of using Eurostat data in combination with some energy models to create final energy mix at household level.

UC 5: Validation of Phytosociological Methods through Occurrence Cubes

2.1.1.20 Objectives

Phytosociology is a branch of vegetation science that studies vegetation communities and classifies the species occurring in such communities in hierarchical vegetation units. A classic phytosociological method is based on the cover that the species have on ground and translates this pattern into units of species combinations forming the community.

However, this approach does not explain the reason why they occur on the spot and form such communities since the environmental conditions are only partially or not at all considered.

In this context, FAIRiCUBE provides the ideal framework for ground-truthing, as we can compare known environmental conditions with the areas where these plant communities occur.

In particular, the main objective of this use case is to validate the traditional methods applied in phytosociology to characterise and classify plant communities. This will be approached by linking distribution data of plant species 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).

Moreover, this use case aims also to develop a new phytosociological approach to characterise and predict the presence of plant communities in unknown localities, based on satellite and occurrence data of corresponding known communities.

2.1.1.21 Research questions

For this use case, the research questions addressed are as follows:

- Which abiotic or biotic factors contribute to the distribution of taxa to form vegetation communities?
- Do occurrences of taxa vary along environmental gradients and what are the driving forces behind the observed distribution patterns.
- Is it possible to predict the presence of vegetation communities in unknown locations based on known occurrences of corresponding communities and contributing environmental factors?

2.1.1.22 Data Analysis

First and foremost, a list of habitats will be chosen from the EUNIS classification (European Nature Information System) of Habitat types. The diagnostic taxa related to the habitat types will be obtained together with their occurrences from the Global Biodiversity Information Facility (GBIF) and the Virtual Herbaria System (JACQ).

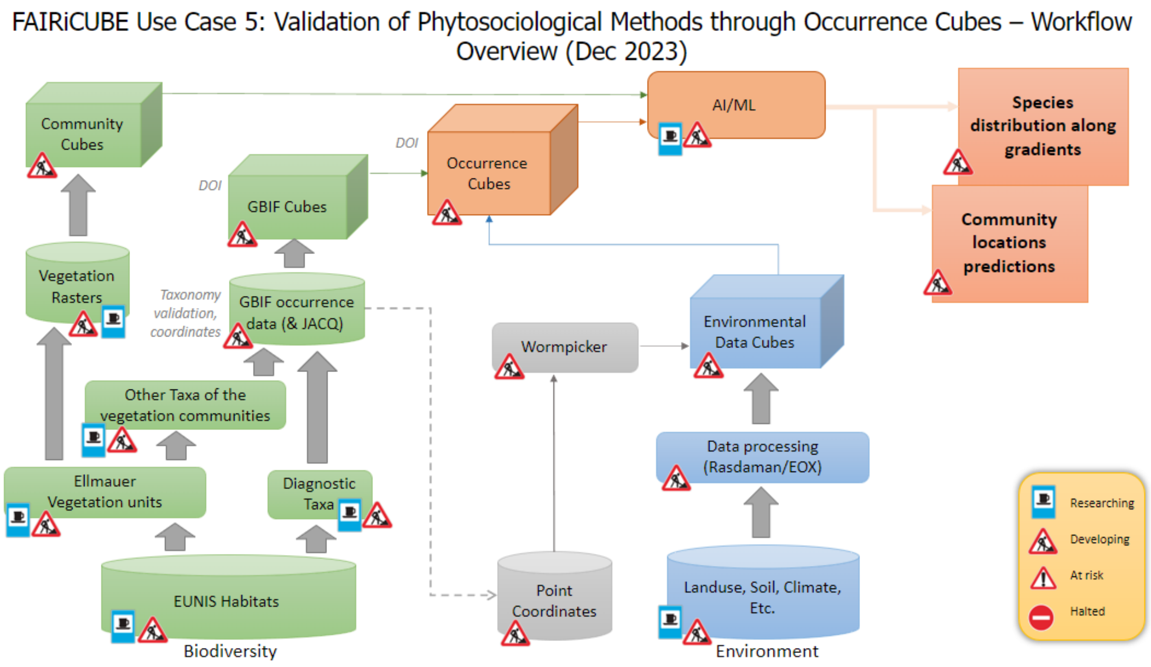
The rest of the taxa comprised in the vegetation units of the Habitats chosen will be obtained from vegetation units present in Mucina et al. (1993)[1].

Furthermore, known distributions of plant communities will be obtained as raster data based on the vegetation units (Mucina et al., 1993)[1] to produce Community Cubes.

A second 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.

Once we obtained the Occurrence Cubes and Community Cubes, we will investigate the distribution patterns of the taxa where plant communities have been identified.

Lastly, we will use ML and AI approaches to identify relations between identified communities and EO data, determine locations with favorable environmental conditions and predict possible presences at these locations of plant communities corresponding to the ones investigated. Figure 5 presents the data sources and the expected processing work in UC5.


FAIRiCUBE Use Case 5: Validation of Phytosociological Methods through Occurrence Cubes – Workflow Overview (Dec 2023)

Community Cubes
AI/ML
Species distribution along gradients
DOI
Occurrence Cubes
DOI
GBIF Cubes
Community locations predictions
Vegetation Rasters
Taxonomy validation, coordinates
GBIF occurrence data (& JACQ)
Wormpicker
Environmental Data Cubes
Other Taxa of the vegetation communities
Data processing (Rasdaman/EOX)
Ellmauer Vegetation units
Diagnostic Taxa
EUNIS Habitats
Biodiversity
Point Coordinates
Landuse, Soil, Climate, Etc.
Environment
Researching
Developing
At risk
Halted

Figure 5: UC5 data, analysis, and processing workflow

2.1.1.23 Expected results

The expected results of this use case include the creation of Data Cubes on the basis of habitat and vegetation types, taxon occurrences and earth observation data, the investigation of the distribution patterns of vegetation communities based on environmental factors and predictions of vegetation communities' presence in unknown localities.

The results of this use case will be in line with F.A.I.R. principles and available for use of a broader scientific audience. In particular, the data cubes produced will be disseminated and identified with Digital Object Identifiers (DOIs). Moreover, API specifications will be provided for data delivery from JACQ and GBIF together with linked data concepts for collection data integrated in data cube processes and analysis. Taxon distribution patterns in relation to biotic and abiotic factors will be published. Lastly, a guidance for legacy data preparation will be provided.

2.1.1.24 References

[1] 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*.

2.2 Data analysis

The detailed data analysis description of all 5 UCs is work in progress under WP3 and WP4. More information on this can be found in deliverable *D3.1 – UC exploratory data analysis*. This section will be expanded in the updated report expected in month 30 as WP3 and WP4 make progress.

The different use cases set up within FAIRiCUBE address different topic areas as well as spatial scales. Although all have their particularities in data sources, data analysis and processes, due to the nature of each UCs, there are potential common areas on the data analysis side.

The data analysis for **UC1 Urban Adaptation to Climate Change** and **UC4 Spatial and Temporal Assessment of Neighbourhood Building Stock** will be at a local spatial scale, hence, covering "hotspots" and not "wall-to-wall". Both will use "cities" as spatial entities of analysis, in the case of UC1 represented by the Urban Atlas/Urban Audit reference units "Functional Urban Area" and "Core City". Both will focus their analysis on specific urban areas, even though UC1 will as a first step produce outcomes for all European cities that are part of the Urban Atlas/Urban Audit. The analysis will require specific local data for each UCs. Local data will require harmonisation with other data across spatio-temporal and thematic content. European data sets will be used in the data analysis. At the moment, the two UCs are discussing about organising a joint seminar to inform and engage stakeholders that could be interested in the work at the city level.

UC2 Agriculture and Biodiversity Nexus will do the data analysis at farm and regional/national level. The analysis will aim at investigating links between biodiversity and agriculture. This UC will analyse private data pertaining to farming practices and productivity on specific farms. This data, per se, may be restricted. However, the data analysis will be fully documented and described. Other local data not private will not be restricted.

UC3 Biodiversity Occurrence Cubes – Drosophila landscape genomics and **UC5 Validation of Phytosociological Methods through Occurrence Cubes** aim to garner data from diverse sources to improve their understanding of various biological processes. The data analysis will be based on transformed point-based and vector based to gridded data. This initial transformation will allow carrying out data analysis together with other gridded resources.**C3** in Drosophila data analysis is expected to be different. However, it will also use site specific data which will also be gridded for analysis.

3 Definition of AI/ML algorithms

The task will assist in the definition of AI/ML algorithms to be utilised as well as data layers that are available. This section of the first draft of the D2.2 – UC Analysis report will be updated as WP3 and WP4 make progress in this area.

More information can be found in deliverable *D3.2 – Machine learning strategy specific for each use case*, which focuses on the description of the ML algorithms that are being used or planned to use for each use case.

Further information will be provided in the next report updates expected in months 30.

4 Analysis and description of similarities on UC analysis approaches

This task aims to analyse and describe similarities on UC analysis approaches across UCs.

In the first reporting period and based on the initial description and initial progress, the similarities of UC1 (Urban adaptation to climate change) and UC4 (Spatial and temporal assessment of neighbourhood building stock) had already resulted in pairing UC1 and UC4 on EOxHub services. On the other hand, UC2 (Agriculture and Biodiversity Nexus), UC3 (Biodiversity occurrence cubes – Drosophila landscape genomics) and UC5 (Validation of Phytosociological Methods through Occurrence Cubes) are grouped on Rasdaman services. However, in the second reporting period, there is a focus on the productivity of the use cases and the allocation of use cases to designated infrastructure stacks is lifted. All use cases can choose the corresponding FAIRiCUBE Hub service at own will and according to the concrete execution tasks. That can imply for example a mix of FAIRiCUBE Hub services together, data request from rasdaman database from a Jupyter Notebook hosted on EOX Hub.

These commonalities highlight shared approaches, methodologies, and objectives across the use cases within the FAIRiCUBE project, contributing to the project's overarching goals of integrating and analysing diverse datasets for improved understanding and decision-making.

A difficulty shared between all use cases is the need to fill gaps in their data. For use cases needing to integrate data at multiple scales, e.g., local and regional or European levels (UC1, UC2), gap filling methods will need to be investigated to successfully integrate the different scales. On the other hand, as described in section 2.1.1.13 and in deliverable *D3.2 Machine learning strategy specific for each use case*, UC3 will investigate different approaches to deal with missing allele frequency data in the genomic datasets. To test the gap filling methodology, gaps were created in a subset of non-missing data to be able to compare estimated with existing values.

Based on the dimensionality of data sources to be ingested, use cases are facing similar difficulties when transforming source data to grid data. On the one hand, UC2, UC3 and UC5 need to transform point data (species data and genetic variance, respectively) to grid data. On the other hand, UC1, UC4 and UC5 will need to transform vector data (urban data, building stock data, and vegetation community distributions respectively) to grid data.

Regarding similarities on use cases focus, synergies have emerged between UC1 and UC3. Whereas UC1 focuses directly on climate change impacts and adaptation, UC3 will investigate how climate and anthropogenic changes can affect variation in genomics.

On the other hand, sharing the focus on biodiversity, UC2, UC3 and UC5 are using the concept of occurrence cubes as an approach to explore the relationship between environmental data (climate, agriculture, etc.) with species abundance (UC2), genomic data (UC3) or species occurrence (UC5).

4.1 Summary of each use case

In summary, UC1 analyses "cities" as spatial entities on the European level, considering administrative, morphological, and functional units. It involves data collection of climate-related and land use data, creating a data cube. Machine learning is applied to calculate indicators for cities, and the results are visualised. The process includes gap filling for socio-economic data, and the new data cube is assessed for quality and compliance. Local data with higher resolution are integrated, and simulations of decision-making processes are developed.

UC2 focuses on biodiversity, environmental, and agricultural data. Separate processing flows generate data cubes, merged using causal machine learning. The approach includes data acquisition from local to global scales, building biodiversity indicators, and estimating the impact of farmers' interventions on biodiversity. The integration of datasets and advanced techniques provides insights into the causal mechanisms linking agriculture and biodiversity.

At the same time, UC3, leveraging genome-wide sequencing data of Drosophila melanogaster, explores spatial, temporal, and seasonal patterns. An existing analysis pipeline is adapted, and an API is developed for data compatibility. Statistical methods, machine learning, and imputation address missing allele frequency data. The study investigates factors influencing genetic differentiation and identifies drivers of selection using advanced population genetics and machine learning approaches.

UC4 aims to create 3D models for energy and environmental performance in four European cities. It uses vectorised and rasterised data from public platforms and local bodies. Three data types (ground-based, remote sensing, and tabular) undergo a cascade of modelling to estimate building stocks. The process involves evaluating machine learning suitability and converting vector to grid data, ensuring comprehensive 3D models for energy retrofitting.

And UC5 aims to study plant community distribution patterns using habitat-related taxa, vegetation units, and known distributions. Diagnostic taxa and vegetation units are collected, and data cubes are produced based on (i) occurrences of taxa and environmental data, and (ii) distributions of vegetation communities. The investigation explores distribution patterns, and machine learning and AI are applied to establish relationships and predict plant community presence in specific environmental conditions. The use case integrates habitat-based taxa, vegetation units, and known plant community distributions for a comprehensive analysis.

4.2 Similarities between use cases

The similarities between the five use cases are key to understand and exploit the synergies between them.

Data Cube Generations

UC1, **UC2**, **UC4**, **UC5** involve the generation of Data Cubes by combining various datasets. **UC1** creates a data cube integrating climate, land use/land cover, and socio-economic indicators. **UC2** generates individual data cubes for biodiversity-related, environmental, and agricultural data, which are later merged. **UC4** produces 3D models and stores building stock information as a data cube. **UC5** creates Data Cubes based on taxa occurrences and habitat-related diagnostic data.

Multi-Scale Data Integration

UC1, **UC2**, **UC3** and **UC5** integrate data at multiple scales. **UC1** integrates data from European-level sources and local data with higher resolution. **UC2** considers data at local, national, and global scales. **UC3** utilises genomic data from a global scale, while **UC5** integrates habitat-based taxa data on a European scale.

Machine Learning (ML) and Advanced Techniques

All apply machine learning and advanced techniques. **UC1** and **UC4** utilise ML for calculating indicators, **UC2** employs causal machine learning for biodiversity impact assessment, **UC3** uses statistical methods and ML approaches for genetic variation analysis and **UC5** applies ML and AI for investigating distribution patterns and predicting plant community presence.

Evaluation and Assessment

UC1 and **UC4** include evaluation and assessment steps. **UC1** assesses the new data cube's applicability, usefulness, and compliance with European Green Deal data space requirements. **UC4** evaluates the suitability of using machine learning for specific processes.

Spatial and Temporal Analysis

UC1, **UC2**, **UC3** and **UC5** involve spatial and temporal analysis: **UC1** focuses on cities as spatial entities and will at a later stage integrate time series information. **UC2** considers the impact of agriculture on biodiversity over time. **UC3** explores genetic variation in populations on a global scale. **UC5** studies distribution patterns of plant communities within habitats.

Use of Local Publicly Available Data

UC1, UC2 and UC4 rely on publicly available data. **UC2** incorporates local and national datasets, and **UC4** utilises vectorised and rasterised data from public platforms.

Integration of Environmental and Climate Data

UC1, **UC2**, **UC3** and **UC5** integrate environmental and climate data. **UC1** integrates climate data with land use/land cover for cities. **UC2** incorporates environmental factors and climate data. **UC3** intersects genetic variation data with gridded climate and agriculture data. **UC5** applies ML and AI to establish relationships between plant.

4.3 Differences between use cases

Before summarising the synergies of use cases, the difference between them is summarised. These differences highlight the diverse nature of the use cases, ranging from urban analysis and biodiversity impact assessment to genomic studies, 3D modeling, and plant community distribution analysis. Each use case addresses distinct challenges and contributes to the overarching goals of the FAIRiCUBE project in unique ways.

Scope and Focus

UC1 focuses on the analysis of cities, considering administrative, morphological, and functional units at the European level, with an emphasis on climate, land use, and socio-economic factors. **UC2** concentrates on the impact of agriculture on biodiversity, incorporating biodiversity-related, environmental, and agricultural data, utilising causal machine learning for meaningful explanations.**UC3** deals with genomic data and its intersection with climate and agriculture data to understand genetic variation in Drosophila populations on a global scale.

UC4 revolves around creating 3D models for energy and environmental performance in selected European cities, emphasising architectural diversity and climate zones. **UC5** studies the distribution patterns of plant communities within selected habitats, combining habitat-related diagnostic taxa, vegetation units, and known plant community distributions.

Data Types and Sources

UC1 integrates data from Copernicus services, including land-based data (CLMS), climate data (C3S), and socio-economic data (Urban Audit), while **UC2** incorporates local, national, and global datasets related to species occurrences, environmental factors, and agricultural practices. **UC4** uses vectorized and rasterised data from public platforms, including OpenStreetMap and INSPIRE geoportal, and relies on publicly available data for 3D model creation.

UC3 leverages genome-wide sequencing data from DrosEU, covering over 271,000 populations worldwide, while **UC5** selects habitats from the EUNIS classification and obtains diagnostic taxa from GBIF, Virtual Herbaria, and vegetation units in Mucina et al. (1993).

Methods and Techniques

UC1 employs ML for calculating indicators and information products for cities, integrating thematic data into a data cube. While **UC2** utilises causal ML for insights into the impact of agriculture on biodiversity, creating biodiversity indicators and indices. **UC3** employs statistical methods, API development, and ML to analyse genomic and environmental data. **UC4** applies ML techniques for gap filling, modeling, and evaluating the 3D models' suitability. **UC5** combines biotic and abiotic data to generate Data Cubes, using ML and AI for investigating distribution patterns and predicting plant community presence.

Geographic Scale

UC1 operates at the European level, considering various interpretations of cities while **UC2** encompasses regional and national scale for agriculture and biodiversity in The Netherlands (only, though any developed methodology can potentially be applied in other countries with similar circumstances and data availability).

UC3 spans a global scale, focusing on Drosophila populations across continents. **UC4** selects specific European cities (Oslo, Barcelona, Luxembourg, Vienna) for 3D modeling. **UC5** considers selected habitats from the EUNIS classification on a European scale.

Primary Objectives

UC1 aims to analyse cities as spatial entities and evaluate the compliance with European Green Deal data space requirements, while **UC4** aims to create 3D models for energy and environmental performance in specific European cities. **UC2** focuses on understanding causal mechanisms in agriculture-biodiversity interactions.

UC3 explores genetic variation in Drosophila populations and identifies interactions between genomic and environmental factors, while **UC5** investigates distribution patterns of plant communities within selected habitats, applying ML and AI for predictions.

4.4 Synergies between use cases

At the current stage of the project, there are real synergies between use cases, but also potential synergies which are currently being discussed. Both synergies between the five use cases are summarised. They demonstrate opportunities for collaboration, data sharing, and the application of advanced analytics across different use cases, fostering a more comprehensive understanding of the complex interactions within the FAIRiCUBE project.

Integration and sharing

UC1 focuses on city-related data, including socio-economic indicators and land use/land cover. This data could be valuable for **UC4** in assessing building stocks and environmental performance within cities. The local data collected in **UC4** can also be integrated into **UC1** for higher-resolution analyses. In addition, UC3 can benefit from the availability of land use related and climatic data from UC1 to analyse Drosophila populations in cities.

Machine Learning and Analytics:

The machine learning approaches used in **UC1** for calculating indicators and information products for cities can inspire similar approaches in **UC4** for estimating energy and environmental performance, or vice versa. Moreover, techniques developed for filling data gaps in **UC1** may be applicable to address missing data in other use cases, such as **UC4**.

Spatial Analysis and Modeling

UC1 deals with cities as spatial entities, and the data cube generated can be utilised for spatial analysis and modeling, which might be relevant for understanding the spatial patterns explored in **UC5**. Techniques for creating 3D models in **UC4** can potentially be extended or adapted for spatial analysis in other use cases, especially those involving geographical data like **UC3** and **UC1**. The **UC5** will make use of the Worm Picker tool developed by the **UC3** to obtain rasterised data for specific coordinates.

Environmental Data Sharing

Environmental data, especially climate-related and land use/land cover data collected in **UC1**, could be shared with other use cases, such as **UC2** and **UC5**, which also involve environmental considerations.

Advanced Analytics and Machine Learning in Genetics

Techniques developed in **UC3** for analysing genetic data, including machine learning-supported imputation methods, could inspire similar approaches in **UC2** for understanding the impact of agriculture on biodiversity and **UC5**.

Global Data Sources

The global-scale genome-wide sequencing data used in **UC3** from the DrosEU consortium can potentially contribute to global-scale analyses in other use cases, enhancing the scope and applicability of findings.

Common Data Infrastructure

The data cube generation, processing, and storage methodologies developed in **UC1** and **UC4** can serve as a basis for creating a common data infrastructure that could be shared among various use cases.

Cross-Disciplinary Collaboration

The integration of genetic, environmental, and geographical data in **UC3** exemplifies a cross-disciplinary approach. Insights gained from this collaboration could be applied to enhance the understanding of interactions in other use cases.

Predictive Modeling and Explanatory Power

The emphasis on causal machine learning in **UC2** to provide meaningful explanations for the impact of agriculture on biodiversity aligns with the goal of enhancing interpretability. Similar approaches may be considered in other use cases to improve the explanatory power of predictive models.

Application of FAIR Principles

The API development in **UC3** for compatibility between genomic data formats and Rasdaman interface data structures follows FAIR principles, and similar approaches could be explored in other use cases to facilitate interoperability.

This section of the report will be updated in the subsequent reports as WP3 and WP4 make progress in the data analysis. More information will be provided in the next report expected in month 30.