D2.2. Big data extraction module

agricore

Deliverable Number	D2.2
Lead Beneficiary	AUTH
Authors	AUTH, IDE
Work package	WP2
Delivery Date	31/08/2022 (M36)
Dissemination Level	Public

www.agricore-project.eu

The Agricore project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No. 816078

Document Information

Project title	Agent-based support tool for the development of agriculture policies
Project acronym	AGRICORE
Project call	H2020-RUR-04-2018-2019
Grant number	816078
Project duration	1.09.2019-31.8.2023 (48 months)
Deliverable authors	Michail Tsagris (AUTH), Vangelis Tzouvelekas (AUTH), IDENER Team
Deliverable reviewers	IDENER Team

Version History

Version	Description	Organisation	Date
0.1	ToC Proposal	AUTH	10 May 2022
0.2	ToC Approved	IDE	15 May 2022
0.3	Content inclusion (First Draft)	AUTH	19 Jul 2022
0.4	Revision and comments	IDE	07 Aug 2022
0.5	Implementation of corrections (Second Draft)	AUTH	20 Aug 2022
1.0	Exportation and formatting (Final version)	AUTH	31 Aug 2022

Disclaimer

All the contributors to this deliverable declare that they:

- Are aware that plagiarism and/or literal utilisation (copy) of materials and texts from other Projects, works and deliverables must be avoided and may be subject to disciplinary actions against the related partners and/or the Project consortium by the EU.
- Confirm that all their individual contributions to this deliverable are genuine and their own work or the work of their teams working in the Project, except where is explicitly indicated otherwise.
- Have followed the required conventions in referencing the thoughts, ideas and texts made outside the Project.

Executive Summary

AGRICORE is a research project funded by the European Commission under the RUR-04-2018 call, part of the H2020 programme, which proposes an innovative way to apply agent-based modelling to improve the capacity of policymakers to evaluate the impact of agricultural-related measurements under and outside the framework of the Common Agricultural Policy (CAP).

This deliverable presents the AGRICORE data extraction module (DEM), which allows the location, extraction and storage of all types of data necessary for the use of the different tools and modules of the AGRICORE platform. It also presents the main types of information that are necessary for the initialisation, calibration and use of the different AGRICORE modules, as well as the main data sources from which this information can be extracted.

Among the functionalities of the data extraction module is also the statistical analysis of the variables contained in the imported datasets, as well as the detection of correlation relationships between them. This deliverable introduces the processing sequence performed on each of the variables individually, as well as on all of them to detect interdependencies.

The results of these operations generate statistical metadata and forbidden directions between variables. Both outputs are arguments that are also passed as input to the data fusion module, presented in deliverable D2.3.

Abbreviations

Abbreviation	Full name
ABM	Agent-based Model.
AH	Agricultural Holding.
DEM	Data Extraction Module
DFM	Data Fusion Module
EDA	Exploratory Data Analysis
EFS	Economic-Financial Statement.
ES	Economic Size Class
ETL	Extract-Transform-Load
FADN	Farm Accountancy Data Network.
HDFS	Hadoop Distributed File System
MS	Member State
NUTS	Nomenclature of Territorial Units for Statistics
PDF	Probability Density Function
SO	Standard Output
SP	Synthetic Population.
SPEI	Standardised Precipitation and Evapotranspiration Index
SPG	Synthetic Population Generator
SS	Synthetic Sample.
TF	Type of Farming

List of Figures

List of Tables

Table of Contents

1 Introduction

Data extraction is the process of retrieving information from various sources. It is commonly referred to as the first step in the ETL/ELT process in Big Data.

- The extract stage in the ETL refers to copying raw data from different data sources to a staging area. Some examples of these data sources could be SQL/NoSQL databases, ERP systems, and web pages.
- Transformation in the ETL take places in the stagging area, where the raw data undergoes to data processing techniques. The data is transformed and consolidated for its intended analytical use case. During this stage, different methods can be applied. Some examples of these methods are filtering, removing outliers, validating data, conducting audits, encrypting data, and formatting data.
- The final step of ETL involves loading processed data into a target system. This target system for the AGRICORE project is the Data Warehouse, where all the data is stored for future use. Typically, this processing entails loading all the data, but it may also include incremental data ingestion if the source data is continuously growing.

The order of operations is the most distinguishing characteristic between ETL and ELT. ELT copies or exports source data, but instead of loading it into a staging area for transformation, it loads the raw data directly into the target system for transformation as required. ELT typically refers to data lakes, where it is essential to retain unprocessed data for future applications. ETL is typically employed in data warehouses.

In AGRICORE, the module responsible for performing this operations is the Data Extraction Module (DEM). The DEM also obtains the probability distributions for each variable based on the aggregation of data corresponding to the relevant variables extracted from the DWH. In addition, it can obtain the joint probability distributions (marginal or conditional distributions) of those groups of variables for which a statistical correlation is detected. Automated parametric and non-parametric fitting methods are considered to adjust these distribution functions.

2 Data required for using the AGRICORE tool

The generation of the synthetic population and its simulation require a set of information inputs. All these inputs must be initialised, even despite some of them are not the main focus of the use case analysis. For this reason, it is necessary to gather a wide variety of data sources to cover all inputs. In this process, information gaps could be detected if there are no data sources for some required input data. A guideline to fill such information gaps is presented in deliverables D1.7 and D1.8. The main AGRICORE data inputs can be divided into three groups, corresponding to the subsections below.

2.1 Attributes of interest - ABM

Attributes of Interest (AOIs) are properties that define the Agricultural Holding (AH) agent and the context where it takes its actions. The AOIs (see the Table 1): are grouped according to the sub-element of the AH they are associated to, such as AH structure, AH owner, AH manager, AH parcel(s), AH crop enterprises, AH livestock enterprises, AH output products, AH economic-financial statement (EFS), and AH ecosystem context. From a control theory perspective, these AH's sub-elements are translated into SW objects whose attributes can be classified as:

- Parameters: these attributes determine the essential aspects of the agent or its sub-elements, and they are either fixed or change only in the long term.
- States: they are attributes represented by variables that evolve in time as a result of the actions taken by the agent during the simulation. These actions might affect not only the persons working on the holding but also all aspects related to the simulation frame, which are subject to change, like the own structure of the AH, its EFS, etc.
- Agro-management decisions: these attributes are the agent's actions, i.e. the decision variables determined by solving, at each simulation step, the optimisation problem(s) that repre sent the rationale of the AH manager.
- Disturbances: these attributes represent external actions and events that cannot be modified or controlled by the AH, but which directly affect its operation and results. The variables representing these attributes are driven by external modules (pilicy module, climate module, market module, etc.) or by the actions of other neighbour agents.
- Outputs: these attributes are the results of the operation of the AH. At the agronomic level, they represent the total production of the different farm enterprises. At the financial level, they represent the balance sheet and the profit & losses account for the current accounting year. Outputs are the result of an agent's actions subject to its initial states (at the beginning of the current simulation interval) and to external disturbances. While they are the outcomes of the simulation, they might be initialised to represent the last known situation of the AH prior to the initial year of the simulation.

Table 1: Attributes of interest of the objects forming each agent (Source: own elaboration).

Objects	Agricultural Holding Structure	Agricultural Holding Owner(s)	Agricultural Holding Manager	Parcel (optional)	Crop enterprises	Livestock enterprises	Output Products	Economic Financial Statement	Ecosystem
Parameters	Number of owners, probability of generational renewal, geographic location (coordinates, NUTS3, further granularity).	Gender, Grade of innovativeness, Risk aversion level.	Gender, Grade of innovativeness, Risk aversion level, education level.	Geographic location (centre), shape (coordinates of the polygon vertices), area, allowed uses, etc. Soil quality parameters	Type, regional cultivation standards, average regional yield.	Type, regional breeding standards.	Type	(re)investment propensity, size synergies, rate of interest, tax rates, WD_{min}	Soil properties: number of layers, layer thickness, max. bulk density, clay, sand, silt, organic carbon Soil types. Aquifers' quality
States	Economic size (FADN), Type of Farming (FADN), land structure (total area, parcels), livestock units, machinery capacity, regular workforce.	Age, probability of generational renewal	-	Current use (crop/livestock/mixed enterprise) Soil quality status	Status of permanent crops (age and health).	Livestock herd status (species and ages).	Stock levels	Assets, liabilities, equity. Solvency, Liquidity and Profitability indicators.	Soil properties: vol. water content, bulk density, nitrate levels, erosion level. Current classification of ecosystems that make the Agricultural Holdings located in them potential recipients of AES. (e.g. nitrates-polluted areas)
	Structural (LP): land ownership management (buy/sell), available			Chemical management: total amount of manure,	Production Technology (if more than one	Production Technology (if more than one	Production utilization (sales, farm	Investment, loans, withdrawals	

Agromanagement decisions	capital, livestock management (buy/sell), Workforce management, Machinery capacity management, quotas (milk, manure, etc.) **Agroeconomic (SP):** allocation of resources to enterprises, land management (rent/lease), contracted machinery,	-	-	ammonia amount, nitrate amount	has been considered in the model)	has been considered in the model)	use, farm consumption, changes in stock)		-
Disturbances	Land prices, production factors prices, output product prices, public policies.	-	-	Influence of weather conditions on soil status, external pollution.	Plagues, meteorological conditions	Unexpected deaths, meteorological conditions	Outputs from external agents (and imports)	Taxes, accountancy regulations.	Deviations of abiotic Factors (temperature, rainfall, etc), plagues, patogens, others.
Outputs	Socio-economic impact (labour, rent), environmental impact (land use, emissions, water intake, pollution)	-	-	-	Actual Yield	Actual Yield	Product Revenue	Cash flow, profit/loss, balance sheet.	Ecosystem services impact

2.2 Aggregated data of the real population to be synthesised

These data are used to check the correct fit of the generated synthetic population with the real population, i.e. whether the statistical distributions of the attributes of the synthetic population and the real population match up to a certain minimum threshold of goodness of fit (GoF).

The typically required aggregated information on the real population includes: (i) the total number of holdings, (ii) the area devoted to each enterprise, and (iii) the total output derived from each activity. These three sets of variables are usually required on an aggregated basis by NUTS2 or NUTS3 regions. In order for the simulation to be statistically representative and the results of the Impact Assessment to be usable, the aggregated population information must correspond to the agricultural and/or livestock activities that are the subject of the policy assessed, but also to the most representative crops and livestock of the geographical region(s) included in the analysis.

This type of information is usually obtained from official institutions, such as the regional/national ministry of agriculture, the national statistical institute and other public bodies. In addition, in order to be able to assess the synthetic population more accurately, it is desirable that the data are from official censuses, rather than estimates made from statistical samples.

2.3 Initialisation of external modules

External modules of the AGRICORE tool provide the agents with the simulation context where they take their decisions and interact between them. These modules are:

- Agricultural policy module: it introduces the legal context into the simulation by translating the agricultural policy(ies) to be analysed. This module determines the requirements to be a beneficiary of a measure, the subsidy amount and the incompatibilities with other measures, among others. Therefore, the description of these policies of interest is a data input that might be prepared while setting up an AGRICORE use case.
- Land market module: this module simulates the market through which agents exchange land. Its initialisation may require historical series of data on average purchase/sale/rental prices in the past, and/or predictions of future price developments. Data on buying pressure from non-agricultural agents may also be needed.
- Product market module: it simulates the market context that determines the prices for agricultural inputs and outputs. Its initialisation may require historical series of data on average selling prices in the past, and/or predictions of future price developments.
- Biophysical module: this module computes how crops and livestock evolve over time as a function of climatic conditions and farm management actions associated with the technology chosen by the farmer. If a pre-existing external biophysical model is used, all data required for its initialisation will be necessary. If no external model is used, historical series of climate-input factors-output data are needed from which to build an extrapolation mechanism.
- Meteorological module: this module simulates regional meteorological patterns that can be used either as inputs for the BioPhysical module or as inputs for the Impact Assessment Modules (IAMs). For its initialisation, both historical meteorological observations and prediction models relevant for the time period and geographical scope covered by the use case will therefore be necessary.

The aforementioned Impact Assessment Module (IAM) includes an environmental & climate IAM, a socio-economic IAM and an ecosystem services IAM. These sub-modules evaluate the impact of the simulation results through some KPIs, which are part of the output of the AGRICORE tool.

The IA sub-modules are made of external models that must be calibrated and initialised using data obtained from adequate data sources. A partial list of currently required data inputs for these modules is presented in deliverables D5.4, D5.5, and D5.6. Moreover, due to the modular approach and open-source nature of the AGRICORE tool, the models within the IAMs can be modified or substituted by newer or more accurate models that may be developed in the future, provided that the format of the communication interface between the IAMs and the ABS engine is maintained.

3 Data sources commonly used by AGRICORE

3.1 The European Farm Accountancy Data Network (FADN)

3.1.1 Description

The Farm Accountancy Data Network (FADN) is a European system for accountancy data collection from agricultural holdings which was established in 1965 (Council Regulation EEC/79/65)[1]. Farm-level data is essential for monitoring and evaluating the achievements of the CAP and for better targeting of CAP support. FADN is the only source of microeconomic data based on harmonised accounting principles for 27 EU Member States (MS). It is based on respective national surveys in these MS and only covers EU agricultural holdings which, due to their size, can be considered commercial. A commercial farm is defined as a farm which is large enough to provide a main activity for the farmer and a level of income sufficient to support his or her family. In practical terms, in order to be classified as commercial, a farm must exceed a minimum economic size. However, because of the different farm structures across the European Union, a different threshold is set for each Member State. Consequently, the set of farms which constitute the FADN field of observation in a given country is represented by those agricultural holdings surveyed by the FSS, with an economic size exceeding the threshold set for that country [1].

Farms participating in FADN are classified according to Community Typology for Agricultural Holdings. Since 2015, the legal provisions on Community Typology for Agricultural Holdings have been integrated together with the FADN legislation. Commission Regulation (EC) No 1242/2008 has been repealed by Commission Delegated Regulation (EU) No 1198/2014 of 1 August 2014 supplementing Council Regulation (EC) No 1217/2009 of 30 November 2009 setting up a network for the collection of accountancy data on the incomes and business operation of agricultural holdings in the European Community [2].

Classification of agricultural holdings is carried out according to three criteria:

- FADN Region: FADN regions are geographical regions roughly coinciding with the European regional (States/Autonomous Communities) delimitation (equivalent in size to NUTS1 or NUTS2 territorial units, depending on the size of each country).
- Economic Size (ES): The Economic Size of a holding is expressed in a sum of all Standard Outputs (SO) for all agricultural activities existing in that farm. Standard Output is the average monetary value of the agricultural output of an agricultural product (crop or livestock) over the reference period of 5 years, per 1 ha or 1 head of livestock per year, in average production conditions in particular geographical units (regions).
- Type of Farming (TF): The type of farming of agricultural holding is based on the share of SO for each group of agricultural activities in the total SO of the farm.

Within the Farm to Fork strategy, the Commission announced its intention to convert the FADN into a Farm Sustainability Data Network (FSDN). The FSDN initiative will expand the scope of the current FADN network to also collect farm-level data on environmental and social farming practices, as well as provide farmers with reporting on their farm performance. The data collection methodology will be in line with the existing one for the current FADN. The three sister projects of the Agrimodels Cluster (of which AGRICORE is a member) have provided their feedback in the public consultation on the Commission's FSDN Roadmap.

[1] http://fadn.pl/en/organisation/european-fadn/european-fadn-organization/

3.1.2 Sampling Methodology

There is a wide variety of farming within FADN's statistical universe. In terms of economic size, some farms are very large, while others may be very small. Farms can specialize in crop production, raise livestock exclusively, or practice mixed farming, which involves raising both crops and livestock. Based solely on these two criteria, ES and TF, the field of observation of farms in the European Union is highly heterogeneous. Liaison Agencies stratify [2] the field of observation before choosing the sample of farms, to make sure it accurately reflects this heterogeneity. If this weren't done, some specific types of farms would not be sufficiently (or at all) represented by the sample.

Numerous classification cells make up the theoretical grid used for stratification (140 FADN regions x 62 types of farming x 14 economic size classes = 121 520 cells) (Figure 1). In some Member States, specific cells (type of farming x economic size classes) do not exist or are very unusual.

Member States ensure that all key categories of holdings, or all categorization cells containing holdings in the field of the survey, are represented in the FADN farm samples by implementing selection plans. However, in reality, the targeted sample might not be obtained, leaving certain survey cells unrepresented in the sample.

The Commission services, with assistance from the appropriate Member State Liaison Agency, are able to determine which farm types cells in the sample may be empty based on their understanding of the field of observation and selection methods. Then, while calculating the weights, cells with similar features are clustered (aggregated) and regarded as a single cell.

The level of representativeness achieved by the FADN stratification methodology is of fundamental importance for the AGRICORE Project. This is because the level of fit of the synthetic populations (which are generated from FADN samples) to the real populations will be proportionally accurate (or inaccurate) to how accurate (or inaccurate) the statistical representativeness of the source FADN sample is.

[2] Stratification is a statistical technique which is used to increase sampling efficiency (i.e. to minimise the number of farms required to represent the variety of farms in the field of observation).

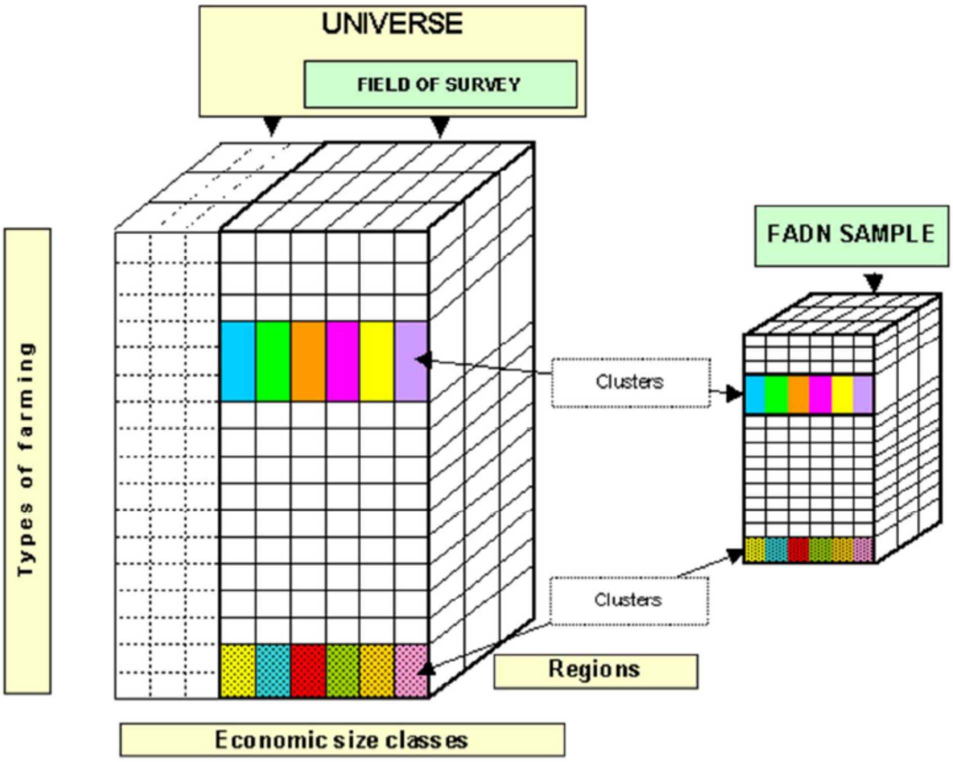


Figure 1 FADN Stratification Methodology. Taken from [1].

3.1.3 Types of FADN data

There are two fundamental types of datasets associated with the FADN:

- **The Farm Return Tables:** The farm return is generated for each farm belonging to the sample of agricultural holdings approved by the European Commission. The data to be collected is classified by tables and broken down into groups, categories and columns. The convention used to refer to a specific data field is: <table letter>_<group>_<category>_<column>. Tables are represented by one letter, groups by one or more letters, categories by numeric codes and columns by one or more letters. The tables of the Farm Return are the following. The complete description of Tables structure and variables is published by the EC on a yearly basis [3].
    - **Table A - General information on the holding:** Identification and classification of the farm.
    - **Table B - Type of occupation:** Breakdown of the farm area: owned, rented or sharecropped.
    - **Table C - Labour:** All labour, paid and unpaid, which has contributed to work on the farm during the accounting year.
    - **Table D - Assets:** Value of all non-capital inputs used in the production of non-capital products during the accounting year.
    - **Table E - Quotas and other rights:** Quotas and other rights included those acquired free if they can be traded separately from linked land.

  - **Table F - Debts:** Outstanding amounts i.e. loans contracted minus the repayments already made.
  - **Table G - Value added tax (VAT):** The VAT system applying and in certain cases VAT payments and receipts.
  - **Table H - Inputs:** Costs in cash and in kind, quantities of selected inputs.
  - **Table I - Crops:** The area, quantity and value of all crops, animal products and other activities.
  - **Table J - Livestock production:** Opening and closing valuations (in number and value) and the average number of livestock, the value of transactions together with the value of any farmhouse consumption of livestock, purchases and sales.
  - **Table K - Animal products and services:** per animal category.
  - **Table L - Other gainful activities directly related to the farm:** The definition of OGA is the same as used in the Farm Structure Surveys and in the Community typology for agricultural holdings.
  - **Table M - Subsidies:** Defined as specific payments made directly to the farm business from public funds, excluding those for investment in land, plant, machinery and equipment. Detailed data concerning CAP arable crops area payments and direct payments for beef.

  These tables are filled in individually for each farm in the sample at the end of each accounting year and processed by the respective national Liason Agencies to obtain harmonised formats in all participating countries. The Farm Return Tables, also known as FADN microdata, are private and are accessible only upon request and authorisation by the file producer (FADN National Office or Unit C.3 - Farm Economics of DG AGRI).

- **The FADN Standard Results:** Standard results are sets of nationally/regionally aggregated variables calculated on the basis of data derived from Farm Returns submitted by Liaison Agencies of the European Union Member States. At present all financial results are expressed in euro. This enables grouping and analyzing results of particular Member States at Community level and direct comparison of results between Member States. In order to avoid identification of particular holding, which participate in the FADN, Commission does not publish averaged results data from the set comprising fewer than 15 farms. The Standard Results are publicly available on the European Commission website and updated regularly after each data amendment done by Member States.

3.1.4 Use of FADN data in AGRICORE

In order to create a synthetic population with a given geographical scope, it is essential to have the FADN microdata for that geographical area. Table 2 shows the FADN variables of the Farm Return Tables that are useful for assigning values to the attributes of the AGRICORE agents. The additional filtering of such microdata makes it possible to narrow down the economic dimension classes and/or the types of farming that are to be included in the synthetic population generated.

Once filtered according to the 3 stratification categories, the resulting variables are processed to calculate their probability density functions (PDFs), individually or jointly with other variables. Finally, the variables are compiled in a single file, which together with the calculated PDFs form the inputs of the DFM.

Table 2: FADN variables relevant for the initialisation of agents' attributes

AGENT ATTRIBUTE (Type)	USEFUL FADN VARIABLES
Agricultural Holding Structure	
Number of owners (Parameter)	*C_UR_10_P, C_UR_20_P*
Economic Size Class (State)	*A_TY_90_ES*
Type of Farming (State)	*A_TY_90_TF*
Mechanisation capacity (State)	*D_OV_4010_V, D_AD_4010_V, D_DY_4010_V, D_IP_4010_V, D_S_4010_V, D_SA_4010_V, D_CV_4010_V*
Regular workforce (State)	*C_UR_40_P, C_UR_40_Y1, C_UR_50_P, C_UR_50_Y1, C_PR_50_P, C_PR_50_Y1*
Workforce management (Agro-management decision)	*C_UC_60_Y1, C_PC_60_Y1, H_LM_1010_V*
Contracted machinery (Agro-management decision)	*H_LM_1020_V, H_LM_1030_V, H_LM_1040_V, H_LM_1050_V*
Agricultural Holding Owner	
Age (State)	*C_UR_10_B, C_UR20_B*
Gender (Parameter)	*C_UR_10_G, C_UR20_G*
Agricultural Holding Manager	
Age (State)	*C_UR_10_B, C_UR_30_B, C_PR_70_B*
Gender (Parameter)	*C_UR_10_G, C_UR_30_G, C_PR_70_G*
Parcel	
Coordinates (Parameter)	*A_LO_40_N*
Area (Parameter)	*B_UO_10_A, B_UT_20_A, B_US_30_A*
Chemical management: total amount of manure (Agro-management decision)	*H_SC_3034_V*
Chemical management:	*H_SC_3032_Q, H_SC_3033_Q, H_SC_3030_V*

ammonia amount (Agro-management decision)	
Chemical management: nitrate amount (Agro-management decision)	H_SC_3031_Q
Land price (Disturbance)	H_FO_5071_V
Crop	
Current activity levels (State)	I_FC_40320_TC_MD_V, I_FC_40130_TC_MD_V, I_A_40320_TC_MD_TA, I_OV_40320_TC_MD_V, I_CV_40320_TC_MD_V, I_A_40310_TC_MD_TA, I_OV_40310_TC_MD_V, I_CV_40310_TC_MD_V, I_A_40330_TC_MD_TA, I_OV_40330_TC_MD_V, I_CV_40330_TC_MD_V, I_A_30200_TC_MD_TA, I_OV_30200_TC_MD_V, I_CV_30200_TC_MD_V, I_A_10120_TC_MD_TA, I_OV_10120_TC_MD_V, I_CV_10120_TC_MD_V, I_A_10605_TC_MD_TA, I_OV_10605_TC_MD_V, I_CV_10605_TC_MD_V, I_A_11220_TC_MD_TA, I_A_10110_TC_MD_TA, I_OV_10110_TC_MD_V, I_CV_10110_TC_MD_V, I_A_50200_TC_MD_TA, I_A_40130_TC_MD_TA, I_OV_40130_TC_MD_V, I_CV_40130_TC_MD_V, I_A_10220_TC_MD_TA, I_OV_10220_TC_MD_V, I_CV_10220_TC_MD_V, I_A_10603_TC_MD_TA, I_OV_10603_TC_MD_V, I_CV_10603_TC_MD_V, I_A_50210_TC_MD_TA, I_OV_50210_TC_MD_V, I_CV_50210_TC_MD_V, I_A_10190_TC_MD_TA, I_OV_10190_TC_MD_V, I_CV_10190_TC_MD_V, I_A_10140_TC_MD_TA, I_OV_10140_TC_MD_V, I_CV_10140_TC_MD_V, I_A_1015_TC_MD_TA, I_OV_10150_TC_MD_V, I_CV_10150_TC_MD_V, I_A_30100_TC_MD_TA, I_OV_30100_TC_MD_V, I_CV_30100_TC_MD_V, I_A_10210_TC_MD_TA, I_OV_10210_TC_MD_V, I_CV_10210_TC_MD_V, I_A_40210_TC_MD_TA, I_OV_40210_TC_MD_V, I_CV_40210_TC_MD_V, I_A_11210_TC_MD_TA, I_OV_11210_TC_MD_V, I_CV_11210_TC_MD_V, I_A_50900_TC_MD_TA, I_A_30300_TC_MD_TA, I_A_40460_TC_MD_TA, I_OV_40460_TC_MD_V, I_CV_40460_TC_MD_V, I_A_10290_TC_MD_TA, I_OV_10290_TC_MD_V, I_CV_10290_TC_MD_V, I_A_10300_TC_MD_TA, I_OV_10300_TC_MD_V, I_CV_10300_TC_MD_V, I_A_40451_TC_MD_TA, I_OV_40451_TC_MD_V, I_CV_40451_TC_MD_V, I_A_10604_TC_MD_TA, I_OV_10604_TC_MD_V, I_CV_10604_TC_MD_V, I_A_10737_TC_MD_TA, I_OV_10737_TC_MD_V, I_CV_10737_TC_MD_V, I_A_40452_TC_MD_TA, I_OV_40452_TC_MD_V, I_CV_40452_TC_MD_V, I_A_10711_TC_MD_TA, I_OV_10711_TC_MD_V, I_CV_10711_TC_MD_V, I_A_10712_TC_MD_TA, I_OV_10712_TC_MD_V, I_CV_10712_TC_MD_V, I_A_40411_TC_MD_TA, I_OV_40411_TC_MD_V, I_CV_40411_TC_MD_V, I_A_10922_TC_MD_TA, I_OV_10922_TC_MD_V, I_CV_10922_TC_MD_V, I_A_10160_TC_MD_TA, I_OV_10160_TC_MD_V, I_CV_10160_TC_MD_V, I_A_40114_TC_MD_TA, I_OV_40114_TC_MD_V, I_CV_40114_TC_MD_V, I_A_40113_TC_MD_TA, I_OV_40113_TC_MD_V, I_CV_40113_TC_MD_V, I_A_20000_TC_MD_TA
Irrigated area (per activity) (Agro-management decisions)	I_A_40320_TC_MD_IR, I_A_40310_TC_MD_IR, I_A_40330_TC_MD_IR, I_A_30200_TC_MD_IR, I_A_10120_TC_MD_IR, I_A_10605_TC_MD_IR, I_A_10110_TC_MD_IR, I_A_40130_TC_MD_IR, I_A_10220_TC_MD_IR, I_A_10603_TC_MD_IR, I_A_50210_TC_MD_IR, I_A_10190_TC_MD_IR, I_A_10140_TC_MD_IR, I_A_10150_TC_MD_IR, I_A_30100_TC_MD_IR, I_A_10210_TC_MD_IR, I_A_40210_TC_MD_IR, I_A_11210_TC_MD_IR, I_A_40460_TC_MD_IR, I_A_10290_TC_MD_IR, I_A_10300_TC_MD_IR, I_A_40451_TC_MD_IR, I_A_10604_TC_MD_IR, I_A_10737_TC_MD_IR, I_A_40452_TC_MD_IR, I_A_10711_TC_MD_IR, I_A_10712_TC_MD_IR, I_A_40411_TC_MD_IR, I_A_10922_TC_MD_IR, I_A_10160_TC_MD_IR, I_A_40114_TC_MD_IR, I_A_40113_TC_MD_IR,
Irrigation system (Agro-management decisions)	A_OT_210

Livestock	
Current activity levels Livestock units (State)	*J_CV_100_N, J_CV_210_N, J_CV_220_N, J_CV_240_N, J_CV_252_N, J_CV_269_N, J_CV_311_N, J_CV_319_N*
Purchased livestock (Agro-management decisions)	*J_PU_100_N, J_PU_100_V, J_PU_210_N, J_PU_210_V, J_PU_220_N, J_PU_220_V, J_PU_240_N, J_PU_240_V, J_PU_252_N, J_PU_252_V, J_PU_269_N, J_PU_269_V, J_PU_311_N, J_PU_311_V, J_PU_319_N, J_PU_319_V*
Sold livestock (Agro-management decisions)	*J_SA_100_N, J_SA_100_V, J_SA_210_N, J_SA_210_V, J_SA_220_N, J_SA_220_V, J_SA_240_N, J_SA_240_V, J_SA_252_N, J_SA_252_V, J_SA_269_N, J_SA_269_V, J_SA_311_N, J_SA_311_V, J_SA_319_N, J_SA_319_V*
Breed livestock (Agro-management decisions)	*J_SR_100_N, J_SR_100_V, J_SR_210_N, J_SR_210_V, J_SR_220_N, J_SR_220_V, J_SR_240_N, J_SR_240_V, J_SR_269_N, J_SR_269_V, J_SR_311_N, J_SR_311_V, J_SR_319_N, J_SR_319_V*
Slaughtered livestock (Agro-management decisions)	*J_SS_100_N, J_SS_100_V, J_SS_210_N, J_SS_210_V, J_SS_220_N, J_SS_220_V, J_SS_240_N, J_SS_240_V, J_SS_252_N, J_SS_252_V, J_SS_269_N, J_SS_269_V, J_SS_311_N, J_SS_311_V, J_SS_319_N, J_SS_319_V*
Products	
Stocks (State)	*K_CV_330_Q, K_CV_330_V, J_CV_100_V, J_CV_210_V, J_CV_220_V, J_CV_240_V, J_CV_252_V, J_CV_269_V, J_CV_311_V, J_CV_319_V*
Production level (Agro-management decisions)	*I_PR_40320_TC_MD_Q, I_PR_40310_TC_MD_Q, I_PR_40330_TC_MD_Q, I_PR_30200_TC_MD_Q, I_PR_10120_TC_MD_Q, I_PR_10605_TC_MD_Q, I_PR_10110_TC_MD_Q, I_PR_40130_TC_MD_Q, I_PR_10220_TC_MD_Q, I_PR_10603_TC_MD_Q, I_PR_10190_TC_MD_Q, I_PR_10150_TC_MD_Q, I_PR_30100_TC_MD_Q, I_PR_10210_TC_MD_Q, I_PR_40210_TC_MD_Q, I_PR_11210_TC_MD_Q, I_PR_40460_TC_MD_Q, I_PR_10290_TC_MD_Q, I_PR_10300_TC_MD_Q, I_PR_40451_TC_MD_Q, I_PR_10604_TC_MD_Q, I_PR_10737_TC_MD_Q, I_PR_40452_TC_MD_Q, I_PR_10711_TC_MD_Q, I_PR_10712_TC_MD_Q, I_PR_40411_TC_MD_Q, I_PR_10922_TC_MD_Q, I_PR_10160_TC_MD_Q, I_PR_40114_TC_MD_Q, I_PR_40113_TC_MD_Q, I_PR_10140_TC_MD_Q, K_PR_330_Q,*
Selling prices (Disturbance)	*I_SA_40113_TC_MD_Q, I_SA_40113_TC_MD_V, I_SA_40114_TC_MD_Q, I_SA_40114_TC_MD_V, I_SA_10160_TC_MD_Q, I_SA_10160_TC_MD_V, I_SA_10922_TC_MD_Q, I_SA_10922_TC_MD_V, I_SA_40411_TC_MD_Q, I_SA_40411_TC_MD_V, I_SA_10712_TC_MD_Q, I_SA_10712_TC_MD_V, I_SA_10711_TC_MD_Q, I_SA_10711_TC_MD_V, I_SA_40452_TC_MD_Q, I_SA_40452_TC_MD_V, I_SA_10737_TC_MD_Q, I_SA_10737_TC_MD_V, I_SA_10604_TC_MD_Q, I_SA_10604_TC_MD_V, I_SA_40451_TC_MD_Q, I_SA_40451_TC_MD_V, I_SA_10300_TC_MD_Q, I_SA_10300_TC_MD_V, I_SA_10290_TC_MD_Q, I_SA_10290_TC_MD_V, I_SA_40460_TC_MD_Q, I_SA_40460_TC_MD_V, I_SA_11210_TC_MD_Q, I_SA_11210_TC_MD_V, I_SA_40210_TC_MD_Q, I_SA_40210_TC_MD_V, I_SA_10210_TC_MD_Q, I_SA_10210_TC_MD_V, I_SA_30100_TC_MD_Q, I_SA_30100_TC_MD_V, I_SA_10150_TC_MD_Q, I_SA_10150_TC_MD_V, I_SA_10140_TC_MD_Q, I_SA_10140_TC_MD_V, I_SA_10190_TC_MD_Q, I_SA_10190_TC_MD_V, I_SA_50210_TC_MD_V, I_SA_10603_TC_MD_Q, I_SA_10603_TC_MD_V, I_SA_10220_TC_MD_Q, I_SA_10220_TC_MD_V, I_SA_40130_TC_MD_Q, I_SA_40130_TC_MD_V, I_SA_10110_TC_MD_Q, I_SA_10110_TC_MD_V, I_SA_10605_TC_MD_Q, I_SA_10605_TC_MD_V, I_SA_10120_TC_MD_Q, I_SA_10120_TC_MD_V, I_SA_30200_TC_MD_Q, I_SA_30200_TC_MD_V, I_SA_40330_TC_MD_Q, I_SA_40330_TC_MD_V, I_SA_40310_TC_MD_Q, I_SA_40310_TC_MD_V, I_SA_40320_TC_MD_Q, I_SA_40320_TC_MD_V, K_SA_330_Q, K_SA_330_V*

Economic Financial Statement	
Assets (State)	*D_OV_1010_V, D_CV_1010_V, D_OV_1020_V, D_CV_1020_V, D_OV_1030_V, D_CV_1030_V, D_OV_1040_V, D_S_1040_V, D_SA_1040_V, D_CV_1040_V, D_OV_2010_V, D_S_2010_V, D_SA_2010_V, D_CV_2010_V, D_OV_3010_V, D_S_3010_V, D_SA_3010_V, D_CV_3010_V, D_OV_3020_V, D_AD_3020_V, D_DY_3020_V, D_S_3020_V, D_SA_3020_V, D_CV_3020_V, D_OV_3030_V, D_AD_3030_V, D_DY_3030_V, D_S_3030_V, D_SA_3030_V, D_CV_3030_V, D_OV_5010_V, D_S_5010_V, D_SA_5010_V, D_CV_5010_V, D_OV_7010_V, D_S_7010_V, D_SA_7010_V, D_CV_7010_V, D_OV_7020_V, D_AD_7020_V, D_DY_7020_V, D_S_7020_V, D_SA_7020_V, D_CV_7020_V, D_OV_8010_V, D_AD_8010_V, D_DY_8010_V, D_S_8010_V, D_SA_8010_V, D_CV_8010_V*
Liabilities (State)	*F_OV_1010_S, F_OV_1010_L, F_CV_1010_S, F_CV_1010_L, F_OV_1020_S, F_OV_1020_L, F_CV_1020_S, F_CV_1020_L, F_OV_1030_S, F_OV_1030_L, F_CV_1030_S, F_CV_1030_L, F_OV_2010_S, F_CV_2010_S, F_OV_3000_S, F_OV_3000_L, F_CV_3000_S, F_CV_3000_L, A_CL_130_C*
Investments (Agro-management decisions)	*D_IP_1040_V, D_IP_2010_V, D_IP_3010_V, D_IP_3020_V, D_IP_3030_V, D_IP_4010_V, D_IP_5010_V, D_IP_7010_V, D_IP_7020_V, D_IP_8010_V*
Subsidies (Disturbance)	*M_S_1150_FI_BU_N, M_S_1150_FI_BU_V, M_S_1400_FI_BU_V, M_S_2334_FI_BU_N, M_S_2334_FI_BU_V, M_S_3300_FI_BU_N, M_S_3300_FI_BU_V, M_S_3500_FI_BU_N, M_S_3500_FI_BU_V, M_S_3500_FI_BU_N, M_S_3500_FI_BU_V, M_S_9000_FI_BU_N, M_S_9000_FI_BU_V, M_AI_10000_FI_BU_T, M_AI_10100_FI_BU_N, M_AI_10100_FI_BU_T, M_AI_10200_FI_BU_N, M_AI_10200_FI_BU_T, M_AI_10210_FI_BU_N, M_AI_10210_FI_BU_T, M_AI_10220_FI_BU_N, M_AI_10220_FI_BU_T, M_AI_10300_FI_BU_N, M_AI_10300_FI_BU_T*
Taxes (Disturbance)	*E_TX_60_T, G_VA_1010_C, G_VA_1010_NI, G_VA_1010_I, G_VA_1020_C, G_VA_1020_NI, G_VA_1020_I*
Input costs (Disturbance)	*H_SL_2010_V, H_SL_2020_V, H_SL_2080_V, H_SL_2090_V, H_SC_3010_V, H_SC_3020_V, H_SC_3040_V, H_SC_3090_V, H_FO_5010_V, H_FO_5020_V, H_FO_5030_V, H_FO_5040_V, H_FO_5051_V, H_FO_5055_V, H_FO_5061_V, H_FO_5062_V, H_FO_5070_V, H_FO_5080_V, H_FO_5090_V*

Once the synthetic population has been constructed, the standard results of the real FADN sample used for its generation are compared with the standard results of the synthetic population, the difference between the two SOs being an additional indicator of the goodness of fit of the SP generation process.

3.2 National Accountancy Data Networks

The European FADN is the compilation of the national versions of the 27 member states that are part of the network. The legislation associated with the FADN sets out a list of variables and indicators that all countries must collect, as well as the methodology for doing so.

However, each country can add additional variables, or collect some of them using methodologies that increase the level of detail of the data. This is for example the case of the RICA[3], the Italian FADN, which surveys the use and costs of each productive factor by activity, instead of doing it globally for all the activities of the holding like in the EU FADN. This greatly facilitates not only the generation of the synthetic population but above all the calibration of the optimisation models that govern the actions of each agent.

Therefore, when preparing a use case in AGRICORE it is important to check whether the national version of the FADN has a higher level of detail than the European one, in order to decide which of the two to request access to.

3.3 Additional data sources

Below is a (non-exhaustive) list of other data sources whose exploration may be required for the initialisation and operation of the AGRICORE tool and its modules:

- The European Farm Structure Survey (FSS): is the methodological basis for censuses of agricultural operations in the European Union. The FSS is carried out every 10 years as a European census of agriculture, and as a sample in some of the intermediate years.
- EUROSTAT: Most Euro indicators are produced by Eurostat. They are complemented by selected monetary and financial indicators produced by the European Central Bank and by data from economic tendencies surveys produced by the Directorate-General for Economic and Financial Affairs of the EC. The main topics covered by Eurostat which might be useful in AGRICORE are: International Trade, Labour market (e.g., unemployment, labour costs, job vacancies), Financial indicators (specifically interest rates) and Prices (e.g. inflation rates, land prices).
- FAOSTATS: The statistical portal of the Food and Agriculture Organisation (FAO) offers datasets on Production, Trade, Food Balance, Prices, Inputs, Population, Investment, Agri-Environmental Indicators, Climate Change and Forestry. The FAO database does not provide predefined tables, but it offers the possibility for the user to build his own with the desired data.
- OECD: The OECD database still has a subdivision by themes: tables under the 'Agriculture and Fisheries' theme concern Agricultural Outlook, Agricultural Policy Indicators and Environmental Indicators for Agriculture; tables under the 'Environment' theme concern Air and Climate, Water, Material Resources, Forest, Biodiversity, Land Resources, Innovation in environment-related technologies, Environmental policy, Agri-Environmental indicators and Green Growth.

[3] https://rica.crea.gov.it/

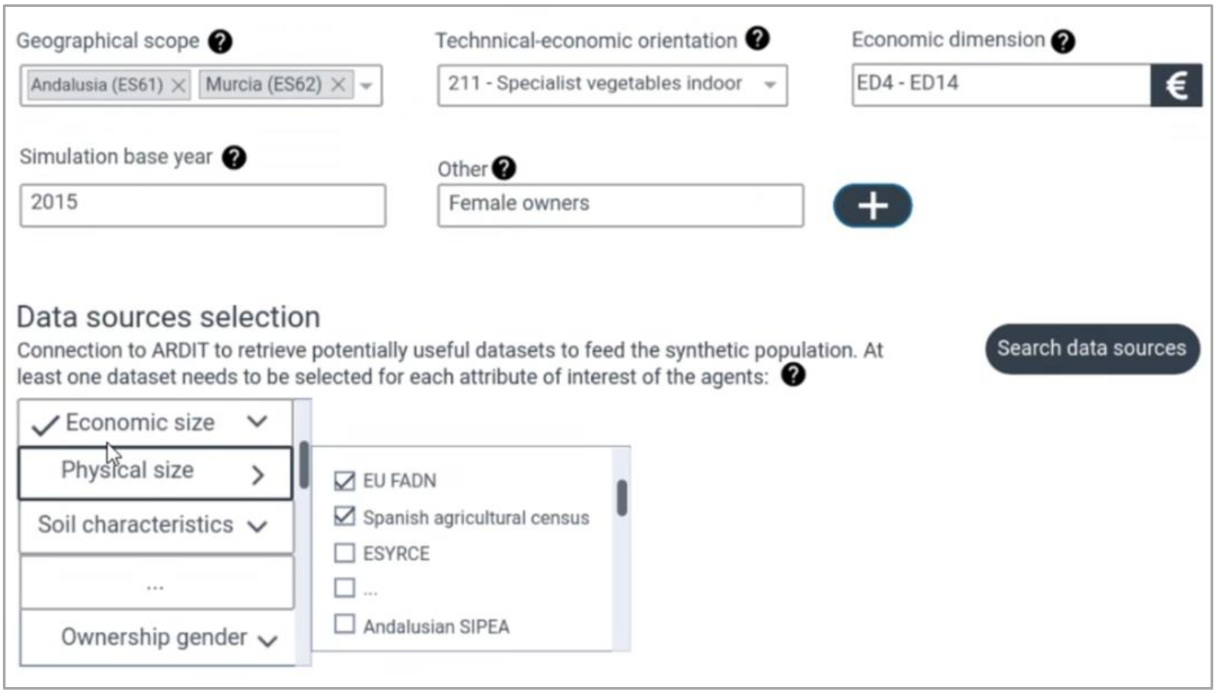
- Data sources from MS Ministries of Agriculture: Basically, the agricultural census corresponding to the base year of the synthetic population, with the aim of being able to compare the adjustment of the synthetic population with the real census population. Other datasets such as the national surveys of crop surfaces and yields, surveys of fertiliser and machinery use, data from the agricultural labour force survey, etc. are also of interest.
- Georeferenced data sources on Land Use and Land Cover such as LUCAS or CORINE.
- Georeferenced data sources on Climate and Meteorological Conditions such as AgMERRA or the SPEI, which will be used later as an example in section 5.
- Any dataset resulting from participatory research activities developed to fill information gaps related to specific attributes. This case is usually limited to attributes related to the behaviour of farm managers, such as risk aversion, community cohesion, or the degree of innovativeness, for which no standardised indicators or indeed harmonised surveys at the European level exist yet.

Some of the aforementioned data sources were characterised and indexed in ARDIT within tasks T1.3-T1.6, as explained in the respective deliverables D1.3-D1.6.

4 Data extraction module connections

Setting up an AGRICORE use case requires a series of data and information elements, which have to be previously located, stored, and possibly processed to produce derived data. The most obvious example of such derived data is the synthetic population of agents to be simulated, but as explained in the previous section, there are others (regionalised climate data, data for the initialisation of bio-physical models, etc.). DEM aims to provide the functionality required to transfer these required data from external data sources to the DWH, generally using ARDIT as a search engine for the discovery of useful datasets.

The construction of the use cases in AGRICORE will be primarily guided by the graphical user interface (GUI). As an example for the case of synthetic population generation, Figure 2 shows the mockup of the GUI screen that assists in the search of datasets of interest for the generation of each agent attribute. The user can narrow the scopes of the population to be synthesised: temporal, geographical, size of farms, and type(s) of farming. Once the filters have been set, a series of parallel searches are launched in ARDIT (through its API) which should return the available datasets matching the selected scopes.



Geographical scope: Andalusia (ES61) × Murcia (ES62) ×

Technnical-economic orientation: 211 - Specialist vegetables indoor

Economic dimension: ED4 - ED14

Simulation base year: 2015

Other: Female owners

Data sources selection

Connection to ARDIT to retrieve potentially useful datasets to feed the synthetic population. At least one dataset needs to be selected for each attribute of interest of the agents:

- ✓ Economic size
- Physical size
- Soil characteristics
- ...
- Ownership gender

☑ EU FADN
☑ Spanish agricultural census
☐ ESYRCE
☐ ...
☐ Andalusian SIPEA

Figure 2: Mock-up of the AGRICORE GUI aiding the search of data sources for simulation set-up.

It is important to remark that the DWH is the central data repository for the AGRICORE project. This means that, except ARDIT, which has its own database, all the information produced by any module is stored in the DWH. The rationale behind this decision is saving computational cost by reusing previously produced data and avoiding the need of having dedicated databases for individual modules, centralizing all the data in the DWH instead.

Even though the DEM only communicates directly with ARDIT and the DWH, the information that the DEM generates is subsequently used by the DFM. Thus, Data Fusion Module (DEM) and DFM communicate indirectly through the DWH.

4.1 Connection with ARDIT

As explained in deliverable D1.9, ARDIT is an index of agricultural data sources that let the general public access the characterisation of different datasets. The users can use ARDIT to locate data sources containing specific variables with certain geographical scope and spatial/statistical resolution. Also, the users can propose the inclusion of new data sources by means of a characterisation procedure. Once the new characterisation is accepted by the ARDIT maintainers, the data source is recorded in the index. ARDIT does not contain any raw data, but only the metadata characterising such data for each dataset. It also contains, for each dataset, an ETL script that allows the extraction and storage of the raw data, transformed adequately to meet AGRICORE needs. The ETL can also be directly provided by the user. The Data Extraction Module (DEM) connects to ARDIT to allow the use of its indexed datasets by the rest of the AGRICORE tools. If these datasets have not been loaded into the DWH before, the DEM downloads and executes the respective ETLs linked in ARDIT to store this information in the DWH.

4.2 Connection with Data Warehouse (DWH)

As explained in deliverable D2.1 within the AGRICORE IT architecture, DWH is the central data repository and exchange point for non-volatile data. DWH contains a large quantity of distributively stored data. Each piece of data produced within an AGRICORE utilisation is stored in the DWH. Specifically, the DWH plays a crucial role in the process of constructing synthetic populations as it is the repository for the raw data of all datasets required to populate the agent attributes, as well as for the intermediate data derived from the DEM (definitions of the simple and compound probability distributions, etc) and the DFM (the definition of the Bayesian Network, etc.). Once the synthetic population of agents has been created, its representation in file format is also stored in the DWH (see deliverable D6.1).

DEM is responsible for launching the execution of the ETLs stored in ARDIT, as ARDIT lacks the capabilities to execute ETLs itself. To achieve this, the DEM is implemented on top of the DWH and utilizes part of its native capabilities. DWH is not only a location for storing data; it also contains all the data processing engines (e.g., Apache Spark) necessary to execute ETLs. These engines allow users to execute distributed queries and thus provide the required scalability to ingest data independent of its size.

Thanks to these engines, DEM could extract the data from the original sources, perform the required pre-processing operations, and load these data into the storage layer of the DWH, technically known as Hadoop Distributed File System (HDFS). Following this methodology, once a new dataset is added to the DWH, its metadata (which could be a database or even a simple text file) are managed and updated by the DEM.

To access the DWH to ingest the extracted datasets, the DEM must have the necessary permissions. The DWH itself provides native mechanisms for authentication and authorisation. The DEM is responsible for calling these services with the appropriate credentials.

5 Data extraction module data pipeline

This section presents the data workflow performed by AGRICORE's Data Extraction Module. The process is very similar for all datasets (only the libraries needed to handle the basic data types of each dataset change). However, a case dataset required by AGRICORE's biophysical module will be used as an example: the calculation of SPEI12 by NUTS3.

The Standard Precipitation Evapotranspiration Index (SPEI) [4] is a drought monitoring index that measures the normalised standard deviation of the difference between Precipitation and Potential Evapotranspiration ($D_i = P_i - PET_i$) for a time period *i*, with respect to the average of that difference over all previous isotemporal periods. This means that SPEI can be measured for different time ranges, measured in months (e.g. SPEI12 July 2020 measures the deviation of $D_{July2020}$ in the 12 months prior to (and including) July 2020 from the average of $D_{JulyYYYY}$ in the 12 months prior to July of all years $YYYY$ for which there are data. Typical SPEI versions are SPEI3, SPEI6, SPEI12, SPEI24 and SPEI48.

SPEI is relevant for AGRICORE because it has proven to be a good indicator for predicting variations in crop yields due to weather conditions [5] [6] [7] [8] [9]. Therefore, when AGRICORE's Biophysical Module does not have a connection to a detailed biophysical model, it can use the simulated SPEI12 value (among others) to generate simulated yield values. The problem is that to calculate these yield predictions, AGRICORE needs the SPEI at the level of NUTS3 regions (matching the geographical resolution of the FADN microdata). However, the SPEI is given as a global gridded dataset with rectangular cells of 0.5ºx0.5º in latitude/longitude, at timescales between 1 and 48.

It is, therefore, necessary to download the SPEI and calculate, for each NUTS3 region, the average of the SPEI12 values of all cells whose centre is contained within the contour of that NUTS3. To do this, as shown in Figure 3, the following steps are necessary:

1. The user defines, from the DEM GUI, the geographic region and the time interval of interest for the simulation. The DEM performs a Data Search using the ARDIT API, which returns a list of appropriate datasets identifiers. These would include SPEI and NUTS3 contours.
2. Once selected, ARDIT responds by sending the ETLs associated with both datasets.
3. The DEM runs the ETL corresponding to the SPEI, which is downloaded from the data servers of the Higher National Research Council (CSIC for its acronym in Spanish), and the ETL corresponding to the NUTS3 contours, which is downloaded from the EUROSTAT servers.
4. After individual transformation and filtering (temporal for SPEI and geographical for the NUTS3 contours), both datasets are stored in the DWH.
5. The DFM extracts both datasets and performs the necessary operations to compute the SPEI12 by NUTS3.
6. The new enriched dataset resulting from data fusion is also separately stored in the DWH.
7. In case the user wants to make such an enriched dataset public, an appropriate ETL can be added to the characterisation of the SPIE12_NUTS3 dataset included in ARDIT.

The result is that the DEM provides the DFM (indirectly through the DWH) with the necessary inputs (dashed thick line in Figure 3) to calculate, through data fusion operations, the enriched dataset needed by the AGRICORE biophysical module. A similar process occurs for the remaining required data required to set up AGRICORE use cases.

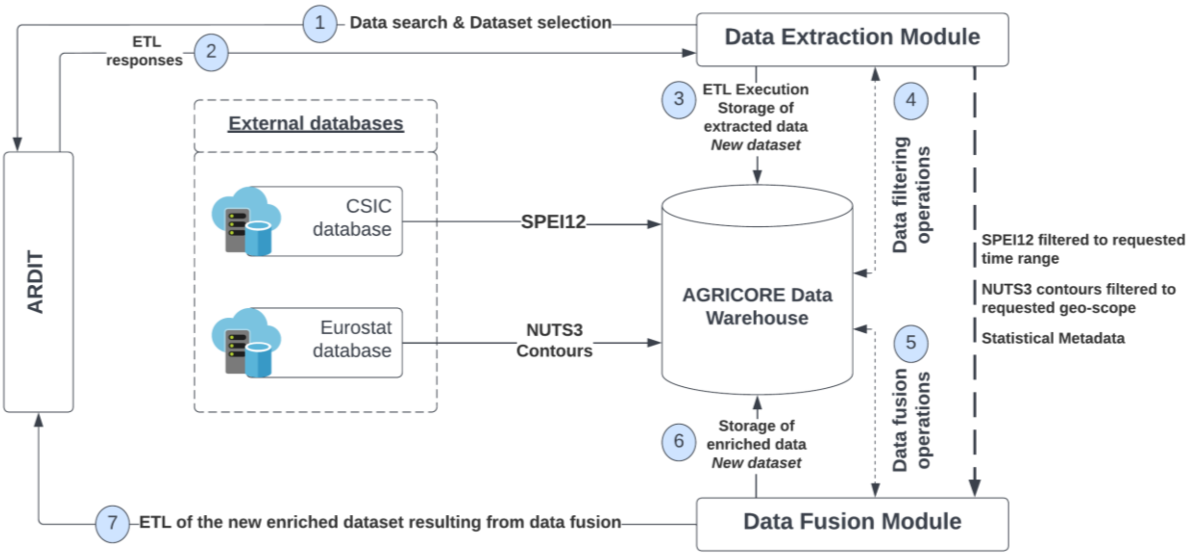


Figure 3 SPEI12 ETL Workflow

Sticking exclusively to point 2, which is the relevant one for this deliverable, below is the code execution flow necessary to extract, transform and load one of the two datasets in the DWH. SPEI12 is taken as an example, but the sequence would be similar for NUTS3 contours, varying only the libraries that need to be imported in order to work with data in GeoJSON format instead of the netCDF4 format of SPEI.

5.1 Import Dependencies

The following code snippet imports the Python libraries necessary to execute requests to external servers, to handle netCDF4 files, as well as those necessary to perform mathematical and statistical operations on the variables present in these files.

```
import os
import pyspark
from pyspark.sql import SparkSession
from pandas_profiling import ProfileReport
from pyspark import SparkFiles
import requests
from netCDF4 import Dataset
import pandas as pd
```

Code Block 1: Import Dependencies

5.2 Set Spark Configuration Properties

The following code snippet initialises and configures the data processing engine (Apache Spark) of the DWH, and establishes the connection ports between it and the DWH's own file management system (Hadoop Distributed File System)[4].

```
conf = pyspark.SparkConf()
# Establish Spark Master
conf.setMaster("spark://spark-master:7077")
# Set hardware for the execution of the ETLs
conf.set("spark.executor.memory", "4g")
conf.set("spark.executor.cores", "1")
conf.set("spark.cores.max", "2")
# Create Application -> ETL
spark = SparkSession.builder.appName('AGRICORE-ETL').config(conf=conf).
→getOrCreate()
sc = spark.sparkContext
# Connecto to the DWH
sc._jsc.hadoopConfiguration().set("dfs.nameservices", "agricorebatchcluster")
sc._jsc.hadoopConfiguration().set("dfs.client.failover.proxy.provider.
→agricorebatchcluster", "org.apache.hadoop.hdfs.server.namenode.ha.
→ConfiguredFailoverProxyProvider")
sc._jsc.hadoopConfiguration().set("dfs.ha.namenodes.agricorebatchcluster",␣
→"namenode-1,namenode-2")
sc._jsc.hadoopConfiguration().set("dfs.namenode.rpc-address.
→agricorebatchcluster.namenode-1", "namenode-1:8020")
sc._jsc.hadoopConfiguration().set("dfs.namenode.rpc-address.
→agricorebatchcluster.namenode-2", "namenode-2:8020")
```

Code Block 2: Set Spark Configuration Properties

5.3 Extraction

Once the data processing tools are configured, the extraction section of the ETL script is executed, which simply makes a request operation to the URL of the CSIC server where the original dataset is hosted.

```
URL = 'https://indecis.csic.es/nc/spei12_month.nc'
data = requests.get(URL).content
data = Dataset('Dataset', memory=data)
data
```

Code Block 3: Data extraction: define and execute ETL

Returning the basic description of the variables contained in the file, their data types and their dimensions.

[4] https://spark.apache.org/docs/latest/configuration.html

```
<class 'netCDF4._netCDF4.Dataset'>
root group (NETCDF4 data model, file format HDF5):
    dimensions(sizes): longitude(464), latitude(201), time(840)
    variables(dimensions): float64 longitude(longitude), float64
latitude(latitude), float64 time(time), float32 spei12(time, latitude,
longitude)
    groups:
```

Code Block 4: Output: extracted dataset

5.4 Transformation

The following snippets extract the dictionary of keys for each of the variables in the file, through which the values of these variables can be accessed and converted into lists (which is an iterable data type) as a preliminary step to their transformation and individualised filtering.

```
data.set_auto_mask(False)
data.set_auto_scale(False)
print(data.variables.keys())
```

Code Block 5: Data transformation: print dictionary of keys

```
dict_keys(['longitude', 'latitude', 'time', 'spei12'])
```

Code Block 6: Output: dictionary of keys

```
spei12 = data.variables["spei12"]
time = data.variables["time"]
latitude = data.variables["latitude"] # North-South
longitude = data.variables["longitude"] # West-East
```

Code Block 7: Data transformation: assignation of individual variables

```
# Convert to list
time_list = time[:].tolist()
latitude_list = latitude[:].tolist()
longitude_list = longitude[:].tolist()
spei12_list = spei12[:].tolist()
```

Code Block 8: Data transformation: conversion to iterable data types

```
spei12_list_flat = []
for sublist_1 in spei12_list:
    for sublist_2 in sublist_1:
        for item in sublist_2:
            spei12_list_flat.append(item)
```

Code Block 9: Data transformation: variables rearrangement through list iteration

5.4.1 Transform time variable into year/month format

The time period is between 1950 to 2020 -> 70 years * 12 months, resulting in 840 instancies. Time variables takes negative values for y/m before 1970/01 and positive values for dates after. Therefore the value 0 corresponds to position 240 in the list. The variable is transformed to be indexed by month and year from January 1950 to January 2020.

```
data.variables["time"]
```

Code Block 10: Request for description of the variable "time"

```
<class 'netCDF4._netCDF4.Variable'>
float64 time(time)
    units: days since 1970-01-01 00:00
    calendar: standard
unlimited dimensions: time
current shape = (840,)
filling on, default _FillValue of 9.969209968386869e+36 used
```

Code Block 11: Output: "time" characteristics

```
# Generate a list with all the dates from 1950-01 to 2020-01 with month
frequency. The result will subtitute time_list
ini_date = "1950-01-01"
end_date = "2020-01-01"

time_list_formated = pd.date_range(start = ini_date, end = end_date, freq="M",
normalize=True)
```

Code Block 12: Data Transformation: "time" format change

5.4.2 SPEI12 data indexed by time - latitude - longitude

SPEI12 of all cells are indexed by Lat/Lon, for each time value (month).

```
spei12_data_indexes_tll = pd.MultiIndex.from_product([time_list_formated,
latitude_list, longitude_list], names = ["Time", "Latitude", "Longitude"])
spei12_data_tll = pd.DataFrame(spei12_list_flat, columns = ["spei12"], index =
spei12_data_indexes_tll)
```

Code Block 13: Data transformation: Python Pandas dataframe declaration

```
# SPEI12 data indexed by latitude - longitude and time in columns --> It takes
more time to generate the dataframe, aprox 10-12 min.

spei12_data_indexes = pd.MultiIndex.from_product([latitude_list,
longitude_list], names = ["Latitude", "Longitude"])
spei12_data_indexed = pd.DataFrame(columns = time_list, index =
spei12_data_indexes)

for t, i in zip(time_list, range(len(time_list))):
    spei12_t_list = spei12[:][i].tolist()
# Save as a list spei values in t for all the coordinates
    spei12_t_list_flat = [item for sublist in spei12_t_list for item in
sublist]   # Flat list of lists: [[] []] --> []
    # Save data in the dataframe
    spei12_data_indexed[t] = spei12_t_list_flat
```

Code Block 14: Data Transformation: rearrangement of georeferenced information

5.4.3 Filter NaN values

Data values equal to -100000000 are equivalent to NaN, therefore those values are eliminated from the transformed dataset.

```
spei_data = spei12_data_tll[spei12_data_tll["spei12"] != -100000000.0]
spei_data
```

Code Block 15 Data transformation: deletion of NaNs

Time	Latitude	Longitude	spei12
1950-12-31	35.125	24.625	-0.938001
		24.875	-1.168773
		25.125	-1.287677
		25.375	-1.138038
		25.625	-1.024847
...	...	...	...
2019-12-31	70.875	27.875	0.508895
		28.125	0.512309
		28.375	0.526519
		28.875	0.536696
	71.125	25.625	0.759034

10395660 rows × 1 columns

5.4.4 Filter data according to temporal scope of the simulation

According to the user selection used as an example (Figure 2) the data is restricted to the period 2015-2019. Note that the data are dated the last day of each month.

```
spei_data_2015_2019 = spei_data.loc[spei_data.index.get_level_values(0) >
'2015-01-01']
spei_data_2015_2019
```

Code Block 16 Data transformation: narrowing down the data to the time period selected for the simulation.

Time	Latitude	Longitude	spei12
2015-01-31	35.125	24.625	0.414102
		24.875	0.668824
		25.125	0.703793
		25.375	0.308237
		25.625	-0.155889
...	...	...	...
2019-12-31	70.875	27.875	0.508895
		28.125	0.512309
		28.375	0.526519
		28.875	0.536696
	71.125	25.625	0.759034

150480 rows × 1 columns

5.4.5 Transform Pandas Dataframe to Spark Dataframe

The Python Pandas dataframe is transformed into a Spark dataframe, which is the data processing engine that runs on top of the Data Warehouse (DWH).

```
# Create DataFrame by changing schema
sparkDF = spark.createDataFrame(spei_data_2015_2019)
sparkDF.printSchema()
sparkDF.show()
```

Code Block 17: Data transformation: Spark dataframe declaration

```
root
 |-- spei12: double (nullable = true)

[Stage 0:>  (0 + 1) / 1]
+---------------------+
|               spei12|
+---------------------+
|   0.4141019880771637|
|   0.6688240170478821|
|   0.7037929892539978|
|   0.3082369863986969|
|-0.15588900446891785|
|  -0.5789200067520142|
|  -0.7917490005493164|
|   1.0996030569076538|
|   0.8911499977111816|
|   0.6742630004882812|
|  0.37740400433540344|
|   0.5595840215682983|
|0.028463000431656837|
|   0.7147989869117737|
|   0.8020840287208557|
|   0.6456530094146729|
|   1.2385159730911255|
|   0.7900689840316772|
|   0.7884250283241272|
|   0.7935389876365662|
+---------------------+
only showing top 20 rows
```

Code Block 18: Output: Spark dataframe

5.5 Exploratory Data Analysis

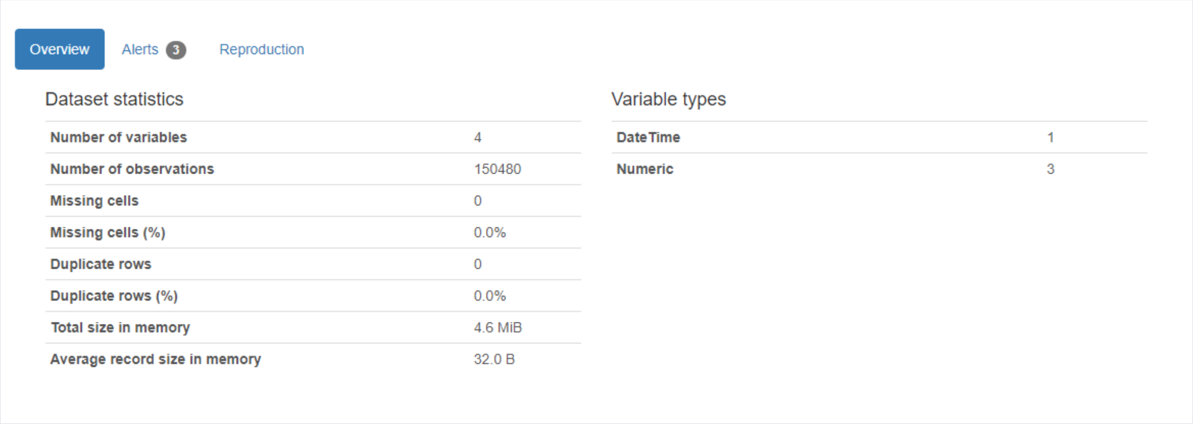
The Pandas library offers among its functionalities the possibility of generating an exploratory data analysis (EDA) report in HTML format, which can be displayed by a web browser. This report includes an overview of the dataset structure (Figure 4), the individual statistics of the dataset variables (Figure 5), a selectable applet to view interactions between pairs of variables (Figure 6, see how a pixeled version of Europe is displayed, as Longitude and Latitude values of SPEI cells are logically concentrated in European+British Islands+Iceland mainland areas), and a report on detected correlations indexes between variables (Figure 7).

```
eda = ProfileReport(spei_data_2015_2019)
display(eda)
```

Code Block 19 Data exploration: generation of profile report

```
Summarize dataset: 100%|*******| 26/26 [00:04<00:00, 5.26it/s, Completed]
Generate report structure: 100%|*******| 1/1 [00:03<00:00, 3.56s/it]
Render HTML: 100%|*******| 1/1 [00:00<00:00, 2.21it/s]
<IPython.core.display.HTML object>
```

Code Block 20 Output: profile report generation



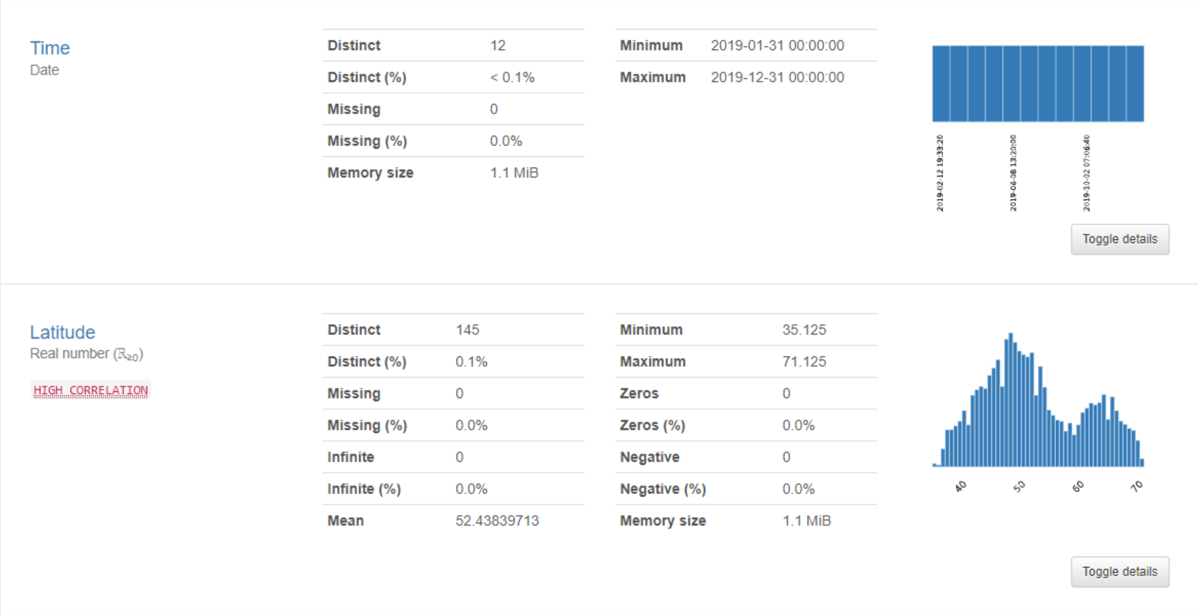
Dataset statistics

Number of variables	4
Number of observations	150480
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	4.6 MiB
Average record size in memory	32.0 B

Variable types

DateTime	1
Numeric	3

Figure 4 EDA: Overview of Dataset Statistics



Time
Date

Distinct	12
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	1.1 MiB

Minimum	2019-01-31 00:00:00
Maximum	2019-12-31 00:00:00

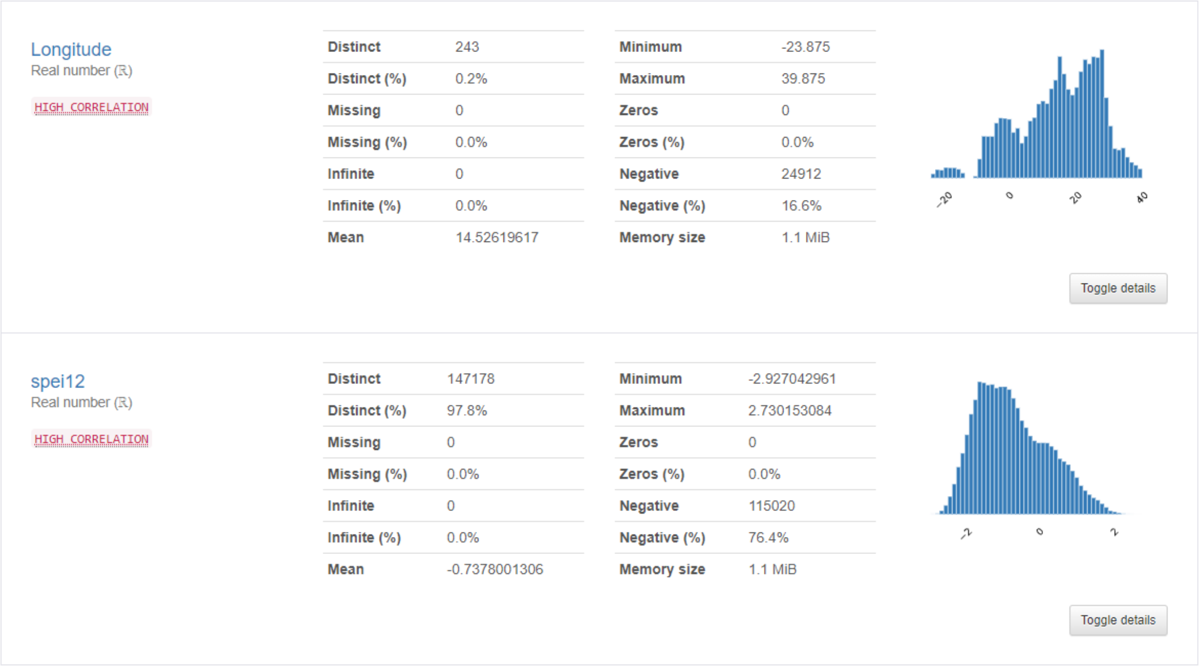
Latitude
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION

Distinct	145
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	52.43839713

Minimum	35.125
Maximum	71.125
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	1.1 MiB

(a)



Longitude
Real number (ℝ)

HIGH CORRELATION

Distinct	243	Minimum	-23.875
Distinct (%)	0.2%	Maximum	39.875
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	24912
Infinite (%)	0.0%	Negative (%)	16.6%
Mean	14.52619617	Memory size	1.1 MiB

spei12
Real number (ℝ)

HIGH CORRELATION

Distinct	147178	Minimum	-2.927042961
Distinct (%)	97.8%	Maximum	2.730153084
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	115020
Infinite (%)	0.0%	Negative (%)	76.4%
Mean	-0.7378001306	Memory size	1.1 MiB

(b)

Figure 5 (a-b): FADN variables relevant for the initialisation of agents' attributes

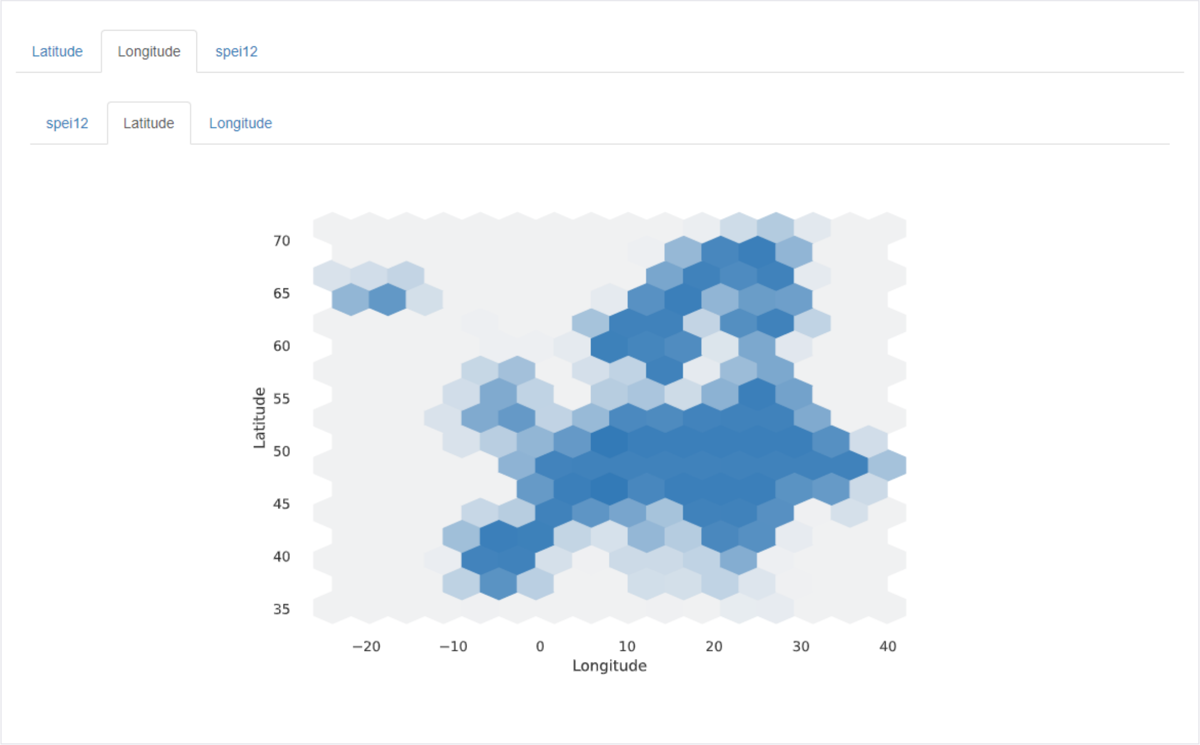


Figure 6 EDA: Interactions

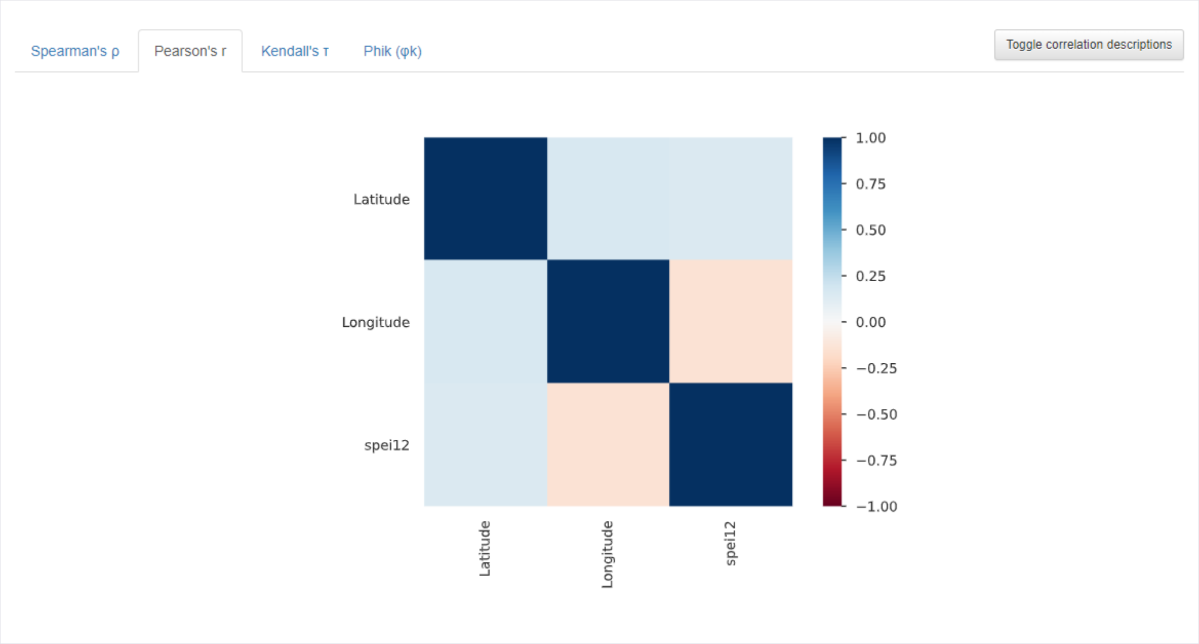


Figure 7 EDA: Correlation Levels

This report can be saved for later viewing, or to be passed as additional information input to the DFM.

```
eda.to_file(output_file="report.html")
```

Code Block 21: Data exploration: saving profile report

```
Export report to file: 100%|******| 1/1 [00:00<00:00, 406.54it/s]
```

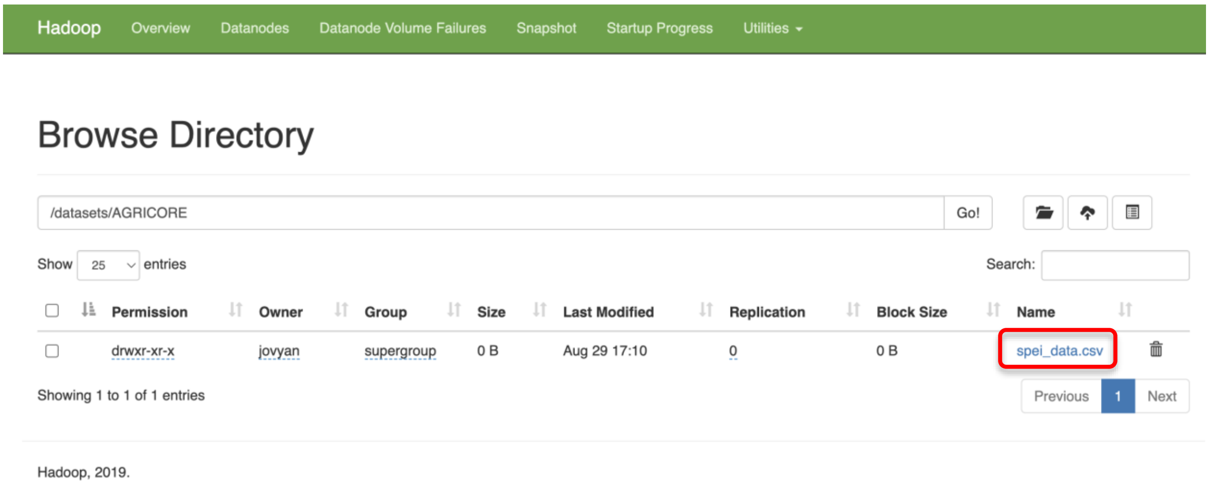
Code Block 22: Output: profile report saved

5.6 Load Data in the DWH

Finally, the Spark data frame with the extracted and transformed dataset is loaded into the distributed file system (Hadoop) that builds on the storage layer of the DWH. Figure 8 shows a screen of the Hadoop file browser showing the ingested dataset already stored in the DWH.

```
# Write data in CSV format into HDFS
sparkDF.write.mode("overwrite").csv("hdfs://agricorebatchcluster/datasets/
→AGRICORE/spei_data.csv", header=True)
```

Code Block 23: Data loading



Browse Directory

/datasets/AGRICORE

Show 25 entries

Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	drwxr-xr-x	jovyan	supergroup	0 B	Aug 29 17:10	0	0 B	spei_data.csv

Showing 1 to 1 of 1 entries

Hadoop, 2019.

Figure 8: Hadoop file browser showing the loaded dataset

6 Conclusions

This deliverable presents the functionalities of AGRICORE's data extraction module, including the obtention of data from local or cloud locations, basic transformation and exploration of the obtained datasets, and loading them into the Data Warehouse. The connections of the DEM with ARDIT and with the DWH itself, which are explained in section 4, are fundamental for the correct functioning of the dataset search and collection process.

The generic flow of datasets from their localisation through ARDIT to their eventual merging process to produce enriched derived datasets has also been presented. For this purpose, a climatological indicator (SPEI12) whose spatial resolution must be changed ({Lat,Lon} to NUTS3) to adapt to the format required by the Biophysical Module of AGRICORE has been used as an example.

This deliverable also presents the types of information that are necessary to initialise and configure the modules that make up the AGRICORE tool. It also briefly describes the main data sources from which the required data are normally extracted.

Besides being a necessary element for the operation of the rest of the modules, the DEM has a fundamental importance in the process of generating synthetic populations, as it is (together with ARDIT) the tool that allows obtaining the raw data from which the DFM will generate the mathematical object (Bayesian Network) that allows assigning numerical values to the attributes of each agent.

The next steps in this task T2.2 are the integration of the DEM code in the DWH deployment, and the functional tests on the synthetic population generation process, once the SPG module is completed (future deliverable D2.4).

7 References

[1]^ European Commission. Directorate General for Agriculture. C.3 Unit (Farm Economics), "FADN - A Metodology From A To Z," Brussels.

[Online]. Available: https://ec.europa.eu/agriculture/rica/pdf/site_en.pdf

[2]^ Polish Institute of Agricultural and Food Economics - National Research Institute, European FADN organization.

[Online]. Available: http://fadn.pl/en/organisation/european-fadn/european-fadn-organization/

[3]^ European Commission. Directorate General for Agriculture. C.3 Unit (Farm Economics), "FADN - Farm Return Data Definitions For Accounting Year 2017," Brussels, Apr. 2018.

[Online]. Available: https://fadn.pl/wp-content/uploads/2012/12/RICC-1680-v1.3-accounting-year-2014.pdf

[4]^ S. Beguería and S. Vicente-Serrano, SPEI Database, The Standardised Precipitation-Evapotranspiration Index.

[Online]. Available: http://hdl.handle.net/10261/268088

[5]^ J. A. Mathieu and F. Aires, "Assessment of the agro-climatic indices to improve crop yield forecasting," Agricultural and Forest Meteorology, vol. 253–254, pp. 15–30, May 2018, doi: 10.1016/j.agrformet.2018.01.031.

[6]^ R. Lecerf, A. Ceglar, R. López-Lozano, M. Van Der Velde, and B. Baruth, "Assessing the information in crop model and meteorological indicators to forecast crop yield over Europe," Agricultural Systems, vol. 168, pp. 191–202, Jan. 2019, doi: 10.1016/j.agsy.2018.03.002.

[7]^ T. Ben-Ari, J. Adrian, T. Klein, P. Calanca, M. Van der Velde, and D. Makowski, "Identifying indicators for extreme wheat and maize yield losses," Agricultural and Forest Meteorology, vol. 220, pp. 130–140, Apr. 2016, doi: 10.1016/j.agrformet.2016.01.009.

[8]^ F. F. Jabbi, Y. Li, T. Zhang, W. Bin, W. Hassan, and Y. Songcai, "Impacts of Temperature Trends and SPEI on Yields of Major Cereal Crops in the Gambia," Sustainability, vol. 13, no. 22, p. 12480, Nov. 2021, doi: 10.3390/su132212480.

[9]^ S. Mohammed et al., "Assessing the impacts of agricultural drought (SPI/SPEI) on maize and wheat yields across Hungary," Scientific Reports, vol. 12, no. 1, May 2022, doi: 10.1038/s41598-022-12799-w.

For preparing this report, the following deliverables have been taken into consideration:

Deliverable Number	Deliverable Title	Lead beneficiary	Type	Dissemination Level	Due date
D1.7	Identification and filling of information gaps through participatory research actions	AXIA	Report	Public	M35
D1.8	Use case participatory research actions	CAAND	Report	Public	M18
D1.9	Agricultural Research Data Index Tool (ARDIT)	AAT	Tool + Report	Public	M31
D2.3	Data Fusion Module	AUTH	Report	Public	M36
D6.1	AGRICORE architecture	IDE	Report	Public	M23