D5.5 Socio-economic (integration of agriculture in rural society) impact assessment module

agricore

Deliverable Number	D5.5
Lead Beneficiary	UNIPR
Authors	UNIPR, IDE, AKD
Work package	WP5
Delivery Date	M42
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)

Version History

Version	Description	Organisation	Date
0.1	ToC definition	UNIPR	16-ene-2023
0.5	Content inclusion (first draft)	UNIPR	21-feb-2023
0.7	Revision and comments	AKD	01-mar-2023
0.9	Second draft	UNIPR	21-mar-2023
1.0	Final version (exportation and formatting)	IDE	31-may-2023

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). The AGRICORE suite stands out for being highly modular and customisable. Thanks to its open-source nature AGRICORE can be applied to a multitude of use cases and easily upgraded as future needs arise.

The modules in charge of assessing the impact of the simulated synthetic population in the frame of an agricultural policy are the impact assessment modules (IAMs), and one of them is presented in this deliverable: the socie-economic IAM. This module aims at measuring the impact of agricultural policies in rural society, such as the creation of employment and gross value added. To this end, a set of KPIs must be defined, which is described in this deliverable. First, the methodology on which the impact of agricultural policies is measured is presented in Section 2. To this end, the socio-economic indicators used in the literature are listed, and the cluster analysis methodology is explained, including several algorithms. In the third section, the AGRICORE approach is outlined, first, the method employed to select the indicators is explained, and the selected KPIs and how they will be calculated are described. In addition, the selection of the cluster analysis algorithm, k-means, is argued based on its simplicity, wide-spreading and easily interpretable results. Finally, the section includes a description of the input data and how to interpret the results of the IAM.

Abbreviations

Abbreviation	Full name
ABM	Agent-Based Model
ARDIT	Agricultural Research Data Index Tool
CAP	Common Agricultural Policy
DWH	Data Warehouse
FR	Functional Requirement
KPI	Key Performance Indicator
NFR	Non Functional Requirement
WP	Work Package
AoI	Attributes of interest
SP	Synthetic population
SPG	Synthetic population generation
GUI	Graphical user interface
IAM	Impact assessment module
TFP	Total factor productivity
AI	Artificial intelligence
SD	Sustainable development
IA	Impact assessment
LU	Livestock unit
LP	Linear programming
AH	Agricultural holding

List of Figures

List of Tables

Table of Contents

1 Introduction

The objective of Task 5.5 - Socio-economic (integration of agriculture in rural society) impact assessment module - is to develop an assessment tool performing a cluster analysis that will allow for evaluating the effects and impacts of agricultural policy measures incorporated into the ABM simulation from the economic and social perspective. Specifically, this module will enable assessing the relationship between policies and socio-economic indicators related to the integration of agricultural farms into the rural system.

1.1 Integration within the Agricore platform

The socio-economic impact assessment module is closely interconnected within the Agricore suite with the Agent-based simulation model. The simulations of the agent-based module performed on the synthetic population of agents take into account agents' behavioural components, economic and financial constraints, policy constraints and workforce availability constraints. The socio-economic impact assessment module uses agents' attributes derived from agent-based simulations as inputs for the implementation of the cluster analysis, as can be observed in Figure 1.

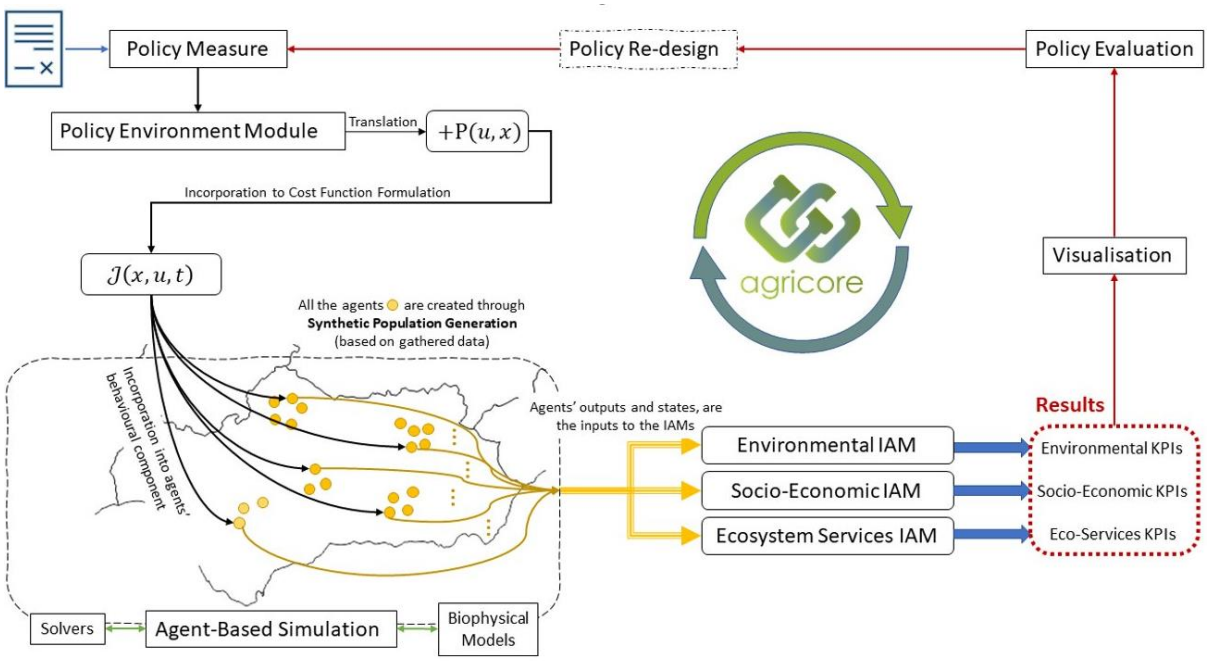

Figure 1. AGRICORE cycle of policy evaluation.

The main purpose of the socio-economic impact assessment module is to provide a reliable assessment of the social and economic impacts of EU agricultural policies. The output of the socio-economic IAM, together with that of the environmental IAM and the eco-system services IAM, will enable the economic, social and environmental impact assessment of a specific agricultural policy, and possibly provide a framework for policy re-design.

2 Methodology for socio-economic impact assessment

The methodology adopted is a double cluster analysis, where farms, subject to a specific agricultural policy are grouped through a clustering process. Clustering is done according to how each farm performs according to predefined socio-economic indicators. Clustering is double because:

- The first clustering is performed on observed data (or the initial synthetic population) before the ABM simulation run to accurately represent the starting situation.
- The second clustering is done on ABM simulation output data.

At the end of each cycle, the obtained clusters are analysed to determine each one intrinsic characteristic, such as farms holdings with particularly high profitability per hectare. Moreover, a comparison between the two clusters will allow for depicting how farms reacted to the adopted policy. Specific attention is given to how farms move from cluster to cluster.

Cluster analysis is a very common multivariate statistical analysis tool for the evaluation of agricultural holdings' socio-economic and agroecological performance or success strategies [1] and for comparison of cross-country agricultural performance [2]. It enables to compare average performance level of different farm clusters and to explore factors or variables (including policy interventions) related to performance . Minimizing diversity within a cluster allows for analysis of public policy changes on specific clusters and a comparison of differential effects of the changes across clusters. Data mining in agriculture is a relatively new research field, and the use of cluster analysis has almost just begun in this area. This statistical tool has regained importance in the evaluation of policy impacts in recent years.

2.1 Socio-economic indicators

Socio-economic indicators allow a quantitative measurement of predefined economic and/or social characteristics of a farm-holding, a region or a state. Socio-economic indicators can cover a wide range of topics, including income, employment, poverty, education, health, and quality of life. They can be used to evaluate the effectiveness of public policies, monitor progress towards sustainable development goals, and for analysis and future forecasting.

2.1.1 Economic indicators

Economic indicators, in the context of AGRICORE, are used to measure the economic performance of agricultural holdings and their integration into the rural system. Particular attention is given to economic sustainability, understood as the ability of a farm to survive and generate income in the long term and in an ever-changing context.

Commonly used indicators are:

- **Growth**: Revenues are the first performance indicator for measuring the state of a company; an increase in revenues can be a positive sign, but it must be accompanied by good profitability and proper monetary management.
- **Profitability**: The best-known indicator of profitability is the EBITDA margin. The EBITDA margin measures the gross profitability of sales, that is, the percentage of sales that remains after the monetary costs of current operations have been subtracted from revenues: consumption, labour costs, and services. This indicator is very useful both in intertemporal comparisons of whether or not management has improved over time and in comparisons between companies in the same sector.

- **Liquidity** in the sense of the company's ability to turn margins into cash through wise working capital management. Efficient companies are those with reduced working capital cycles since long collection and inventory times and short supplier payment times stand for cash absorption and, consequently, the creation of financial requirements.
- **Capital soundness**: The company's financial needs generated by current operations and investments are covered by internal and external sources of financing, the composition of which affects the health of the company, especially in periods of instability. The indicator that measures the company's level of soundness is the ratio of debt capital (D) to equity capital (E).
- **Solvency**: The last key element in monitoring the health of the company is the degree of solvency, that is, its ability to cover its financial debts through the cash flows generated by its operations.

[3] analysed the above categories from the perspective of economic sustainability of the agricultural enterprise. In that research, it emerged how Economic viability is measured primarily by profitability, productivity, liquidity and stability. Profitability is measured by comparing revenues and costs (either as a difference or ratio) or proxied by profit variables such as farm income. Stability is usually measured by the share and development of equity capital and liquidity, and liquidity is the ability to pay cash for immediate expenses or short-term obligations. Productivity, which measures the ability of factors of production to generate an output, is generally measured as a ratio of output to input, but also "*by measures that account for the possibility of input substitution or output substitution, such as total factor productivity (TFP) and technical efficien*cy" [3].

2.1.2 Social indicators

Social indicators are statistical time series that are "*used to monitor the social system, helping to identify changes and to guide intervention to alter the course of social change*" [4]. Due to the purposes of the Agricore project, the task of social indicators is to represent all aspects related to the quality of life of the rural population.

A systematic review of social indicators was developed by [5]. The goal of that paper was to inform the development of a set of social indicators to measure the level of participation of farmers with their agri-environmental schemes agreement and the social sustainability outcomes derived from their participation. In fact, it often happens that evaluation programs of agri-environmental schemes focus on the environmental impact and cost-benefit ratio of these schemes, while evaluation of the impact on social aspects on the rural population is limited.

2.2 Common socio-economic indicators for Impact Assessment

[6] performed a systematic review of the use of farm models for policy IA, based on 202 studies from the period 2007-2015. In their review, around half of the studies assessed impacts on indicators in two different SD dimensions, usually economic and environmental. Slightly less than a quarter included only one SD dimension, usually the economic one, and slightly more than a quarter included in all three SD dimensions. In the economic dimension, gross margin is the most used indicator, and in many studies, the calculation of gross margins was undertaken according to choices specific to the study or the author(s). The full list of economic indicators assessed in farm models, which were studied in D5.1, is presented in Table 1.

Table 1. Economic indicators assessed in farm models.

SD dimension	Indicators
Economic	Gross margin, gross income, net income, household income, net present value, income from different sources, the potential increase in earned income, value-added (all per farm, ha or labour unit)
	Crop prices, minimum subsidy level, subsidy, variable costs (e.g., seeds, water, pumping, fertiliser, pesticides, biomass, N surplus disposal, bought feed, livestock, maintenance and management, harvest, fixed, capital), compliance costs per ha, marginal abatement costs
	Crop yields, crop/milk/meat/energy/protein/carbohydrates/fat production, exploitable, livestock, woodstock, energy supply curves
	Investment (in land, farm buildings, tractors, tillage machinery, harvesting machinery), operational capital, farm income-investment elasticity, household worth, net worth growth, farm fixed investment, debt-to-asset ratio, long-term loans, the option value
	Allocative efficiency, economic efficiency, output-input efficiency, economic water efficiency, irrigation productivity
	Risk, risk efficiency, uncertainty, insurance, economic sustainability, CAP independency, business diversification
	Shadow price
	Consumption, wealth
	Costs of measures, cost-effectiveness of measures
	Return to a governmental body, regional consumption, equity, distribution of family farm income, distribution of farm subsidies, farm contribution to GDP
	Value for EU farmers, value for the seed sector
	Land price, land rent, tenure fee
	Agricultural trade, trade of roughage, total demand, net export
	Farm structure, farm size change
	Adaptability (wooded area/total, farm area with pasture only, subsidies/revenue, LU cattle/LU sheep, LU swine/LU total, cows per bull, ewes per ram, sows per boar)
	Stability (farm area in ownership, LU/ha, land fixed capital per ha, machinery fixed capital per ha, livestock fixed capital per ha, autochthonous cows/ewes per total, opportunity costs of owned resources)
	Economic viability (available income per worker compared with the national legal minimum wage, economic specialization rate)
	Independence (financial autonomy, reliance on direct subsidies from the CAP, and indirect economic impact of milk and sugar quota)
	Transferability (total assets minus land value by non-salaried worker units)
	Efficiency (operating expenses as a proportion of total production value)

Studies concerning social aspects usually refer to work use (hours worked, age, gender), but there have been cases where other indicators have been used, as shown in Table 2.

Table 2. Social indicators assessed in farm models.

SD dimension	Indicators
Social	Labour use (total/hired/family/men/women/harvest/seasonal/in mountain regions), labour productivity, labour intensity, labour allocation, off-farm employment, machinery use. Family consumption expenditure, caloric self-sufficiency

	Public expenditure, cost-effectiveness of measures, net social costs, global value for society, value for farmers and consumers in the rest of the world, welfare --> Ribaudo, F., 2011. Prontuario Di Agricoltura. Ulrico Hoepli Editor, Milano;
	Redistribution effects of payments, income distribution per farm types, income distribution per social groups
	(Average) farm size, farm size distribution, number of farms (total/single-holder/corporate), land ownership, abandoned land
	Nature area, landscape quality, cultural amenity, tourism, social valuation effects for environmental benefits, quality of life, odour, quality of the products and land (quality of foodstuffs produced, enhancement of buildings and landscape heritage, processing of non-organic waste, accessibility of space, social involvement)
	Organisation of space (short trade, services, multi-activities, contribution to employment, collective work, probable farm sustainability)
	Animal welfare, animal health
	Food safety, milk quality parameters (total bacterial count, somatic cell count, coliform count, freezing point, urea-N, fat content, protein content, and penalty points), seropositive pigs leaving the farm, carcass contamination after slaughter, PAHC of Salmonella
	Bankruptcy, sensitivity to technical and economic fluctuations, self-management (rented farm area, farm area with scrub only, farm area under crops, expenditure on animal feed, veterinary expenditure, intermediate consumption, reuse on-farm, resources used from environment/total resources needed by livestock)
	Ethics and human development (contribution to world food balance, training, labour intensity, quality of life, isolation, reception, hygiene, and safety)
	Population patterns, migration patterns
	Land rent, land demand
	Staying legal

2.3 Cluster analysis

Cluster analysis refers to a collection of techniques used to group n-units into k-groups, where $k << n$. The goal of classification is to study the relationships within a set of observations and determine if the data can be summarised into a small number of groups (clusters) with similar characteristics. If the data can be summarised by a small number of groups of observations, then the group labels may provide a very concise description of patterns of similarities in the data [7]. In general, cluster analysis is used when it is necessary to identify groups of units with similar behaviour. Cluster analysis aims to group objects (i.e., farm holdings) using numerical measures that reflect the properties of objects. The analysis is concerned with; i) deciding on the number of clusters, ii) identification of the membership of each group, and iii) profiling the characteristics of each group in terms of behaviour and characteristics. The criteria that are used to form the clusters are that objects within a group should be as '*similar'* as possible and objects belongings to different groups should be as '*dissimilar'* as possible. These criteria statistically imply that the variance within a group should be as small as possible, but the variance between groups should be as large as possible. These criteria are operationalized based on the measurement of closeness, likeness, or similarity between objects.

Cluster analysis should be considered a goal rather than a specific algorithm. There are numerous algorithms aimed at grouping, each with a different definition of the cluster concept and method of grouping. No single classification method is universally effective, as the effectiveness depends on the distribution and nature of the data in the dataset. Due to the diversity of classification methods, it is essential to compare them to determine under which conditions the most widely used algorithms produce the best results.

2.3.1 Classification methods

Cluster analysis is not a specific algorithm but rather a problem to be solved. Typically, a cluster is defined as a group with proximity between its members and dense regions in the data space. The main clustering algorithms can be categorised based on their cluster model. Types of clustering methodologies commonly used are:

- Centroid-based clustering
- Model-based
- Mixture-based clustering

Centroid-based clustering aims to maximise the difference between groups by assigning each unit to only one group. This method tries to find strong similarities within groups and strong dissimilarities between groups. In contrast, model-based and mixture model approaches to clustering posit a statistical model for the population from which the data is drawn, assuming it consists of several sub-populations, or clusters, with different or similar probability density functions. The advantage of these three models is that they can be formulated in a common way, allowing for easy switching between models by modifying the model parameters.

2.3.1.1 CENTROID-BASED CLUSTERING METHODS

In centroid-based clustering algorithms, groups are defined by a central vector which may not be an observation of the dataset. For a number of groups fixed to *k*, centroid-based clustering methods find the *k* groups centroids and assign the observations to the nearest cluster centre, such that the squared distances from the centroids are minimised.

2.3.1.2 K-MEANS

The most well-known centroid-based clustering technique is the k-means algorithm. The k-means clustering algorithm developed by MacQueen (1967) is one of the most widely used unsupervised machine learning algorithms for splitting a dataset into several clusters. It categorises items or objects in multiple clusters, such that items (objects) in the same clusters are related to each other with high intra-class similarity, while items from different clusters are distinct from each other with low inter-class similarity. It is popular due to its computational ease, speed, and memory efficiency. However, there are issues with the initial settings and stability of results. The algorithm is based on a series of iterations and begins by selecting *k* initial points to represent the centroids of the *k* clusters. Then, each point other than the *k* selected points are assigned to the closest cluster, which can then modify the cluster's centroid. However, only the points closest to the centroids are assigned to a cluster, so the centroids tend to stay relatively stable.

There are two preliminary assumptions in k-means:

- distances are measured with the Euclidean norm, formally:

$$d(x,y) = \|x-y\| = \left(\Sigma_{j=1}^{d}\,(x_j - y_j)^2\right)^{1/2},$$

where vectors $x, y \in R^d$;

- the number of groups, k, is fixed in advance.

We will generally be working with a set of d-dimensional statistical observations represented by the vectors ($x_1, x_2, \ldots, x_s$). It is important to note that each ($s = 1, \ldots, n$) is actually d-dimensional. To link this notation with the previous definition, we can replace x_1 with x and x_2 with y, for example. k-means aims to partition the *n* observations into *k* groups $G = \{G_1, G_2, \ldots, G_k\}$ with $k << n$.

The set inclusion of a d-dimensional vector x_i that belongs to the *j*-th cluster ($j = 1, 2, ..., k$) is denoted as $x_i \in G_j$. The partitioning is done by minimising the within-cluster sum of squares with respect to $\mu_1, \cdots, \mu_k$:

$$\arg\min_{\mu_1,\ldots,\mu_k} \sum_{j=1}^{k} \sum_{x_i \in G_j} \min_{j=1,\ldots,k} \|x_i - \mu_j\|^2$$

where μ_j is the mean of points in G_j;

The double sum $\sum_{j=1}^{k} \sum_{x_i \in G_j}$ is abbreviated by some authors with a single summation $\sum_{i=1}^{n}$. In what follows, the second formulation for the economy of space will be used.

The standard algorithm used for iterative allocations is the Voronoi iteration algorithm, which, given an initial set of k means $c(1)_1, \ldots, c(1)_k$, proceeds by alternating the following two steps where the superscript in parenthesis denotes the iteration step $t = 1, 2, \ldots$.

Below are the steps of the algorithm:

- Allocation step: assign each observation to the cluster with the closest mean

$$G_j^{(t)} = \left\{ x_i : \left\| x_i - c_j^{(t)} \right\| \leq \left\| x_i - c_l^{(t)} \right\| \forall 1 \leq l \leq k \right\}$$

where $i = 1, 2, \ldots, n, l \neq j$ and each x_i goes into G_j^t,

- Update step: define the new means to be the centroid of the observation in the cluster

$$c_j^{(t+1)} = \frac{1}{\left| G_j^{(t)} \right|} \sum_{x_i \in G_j^{(t)}} x_i$$

,

where $\left| G_j^{(t)} \right|$ denotes the number of observations which have been assigned to cluster G_j at iteration *t*.

The convergence of the algorithm shall be considered attained when the allocations do not change any more or when a prespecified rule stop is fulfilled. The quick convergence of the k-means algorithm, achieved using the Euclidean norm for measuring distances, is one of its most prized features. However, this ease of implementation and fast convergence can mask the complexity of the algorithm's behaviour, making it crucial to have a thorough understanding of it for correct usage and interpretation of results.

2.3.1.3 MODEL-BASED AND MIXTURE CLUSTERING

The k-means method is commonly used to identify groups of similar data points that are approximately spherical in shape and of equal size. To address the limitations of k-means in handling correlated variables, model-based and mixture clustering was introduced as efficient algorithms for determining the number of clusters and their optimal placement. Despite being effective for small to moderate-size datasets with correlated variables, these methods become computationally expensive for larger datasets. Therefore, other algorithms, such as hierarchical clustering, density-based clustering, and grid-based clustering, have been developed to address these limitations. The choice of clustering algorithm ultimately depends on the specific problem and dataset characteristics, and it is important to consider factors such as computational cost, scalability, and ease of implementation when making a decision. "*Model-based clustering methods have been found to be effective for determining the number of clusters, dealing with outliers, and selecting the best clustering method in datasets that are small to moderate in size*" [8]. However, the utilization of model-based clustering on large datasets can be challenging due to its computational demands, both in terms of time and memory requirements. As the size of the data increases, the calculation of maximum likelihood estimators becomes increasingly difficult,

leading to excessive computational costs that can make the direct application of model-based clustering prohibitive.

3 The AGRICORE approach

3.1 How to choose the most suitable indicators

A careful choice of indicators to be used to perform the cluster analysis is crucial for an effective impact assessment. [9] emphasised how the choice of the right indicators is crucial as it may influence the conclusions. A criterion for selecting the correct indicators was carried out by [3], where the researchers developed a review of indicators measuring sustainability from a threefold, economic, social and environmental perspective.

One of the most problematic aspects, also highlighted by [10], is the lack of general consensus among sustainability experts. There was often a lack of homogeneity in the selection of the most relevant indicators in the measurement of an economic, social or environmental aspect, and a widespread difficulty in ranking the indicators in order of significance.

[9] suggested three main principles for choosing the right indicator:

- **Relevance**: in the sense that the indicator must be appropriate to measure the context it is intended to describe.
- **Practicability**: related to the ability to obtain the necessary information, to be able to quantify it, to measure it, to interpret it and, eventually, to be able to transfer it.
- **End-user value**: related to the usefulness of the information provided by the indicator for the end user; this point is linked to stakeholders' expectations in terms of clarity, policy relevance and comprehensibility.

The presence of historical data regarding an indicator is certainly a relevant element since, as also reported by [11], it can be established the performance that the indicator has had over the years. In the same paper, the author points out how not only the theoretical aspect behind an indicator should be evaluated, but also how the expert community accepts such theoretical arguments should be taken into account.

Those described above are the "ideal" characteristics that an indicator should have. It will hardly be possible to obtain an indicator that perfectly meets the needs of the researcher that does not involve a high cost of data research. It may happen that the search for certain information can be too costly for the purpose of the research, and an indicator based on less specific data should be preferred.

As explained by [12], data should be available at an acceptable cost, and the cost related to the design and calculation of the indicator should also be tolerable. The author suggested considering the totality of costs, such as the implementation cost, the cost of using the indicator, and the cost of adapting it to changes in the context.

In order to have a complete impact assessment, many researchers suggest using a set of indicators rather than just one, whose representativeness of a phenomenon (economic or social in our case) may be limited. The guiding principle should be to select a set of indicators that together can represent the object of study in its complexity. [9] proposed three criteria for the selection of a set of indicators:

- **Parsimony**: in the sense that the indicators should be as few as possible and not redundant; this aspect is of paramount importance, especially in the case of cluster analysis as if too many indicators were chosen, it would then become too difficult, if not impossible, to identify the clusters.
- **Consistency**: i.e. they must incorporate all the measurements required for the impact assessment.

- **Sufficiency**: i.e. the set of indicators must cover all socio-economic aspects to be analysed.

More information concerning the choice of indicators for the creation of a cluster analysis allowing for an impact assessment suitable for the Agricore project is provided in the chapter dedicated to cluster analysis.

3.2 Indicators selected

The indicators to be used in the socio-economic impact assessment module within Agricore are determined from the agent attributes. This stems from the need to have economic and social data for each farm to develop a cluster analysis. During the simulation, the ABM modifies for each iteration only some of the attributes of the agents, so the indicators were also selected, considering which attributes are modified by the ABM. Following the principles listed in the previous paragraph and the need to refer to the attributes of the agent, the following indicators were selected.

Table 3. Selected socio-economic indicators and how they will be calculated.

	TYPE	INDICATOR	AGENT ATTRIBUTES	CAN BE CALCULATED	WHY HAS IT BEEN CHOSEN
AVAILABLE	**ECONOMIC INDICATORS**	PROFITABILITY	FARM NET INCOME or GROSS FARM INCOME	LP	A commonly used business profitability indicator indicates the company's ability to produce income.
		ECONOMIC DIMENSION (GROWTH)	UAA	SP+LP+LAND MARKET MODULE	UAA is used to estimate the economic size of the company and its growth over time.
		SUBSIDIES DEPENDENCY		SUBSIDIES (LP agent attributes) /total revenue	Used to assess the farm's dependence on subsidies, it assesses the impact of subsidies on total revenues. The revenues come from different sources, one of them the subsidies, which can vary over the years depending on the farm's activities. Therefore, the parameter is used on the agent's optimisation and varies during the 7-year cycle.
		LAND RENTED	SHARE OF TOTAL AH	LP	The share of rented land is important to understand farmer behaviour and AH's economic position.
	SOCIAL INDICATORS	RISK AVERSION	LIQUIDITY RATIO	LP	The liquidity ratio is used as an indicator to measure the risk aversion of the farm holder. The objective is to assess whether this propensity varies over time in accordance with the model.

		END OF BUSINESS		(n° of farms at t7) - (n° of farms at t0)	During the simulation, it may happen that companies that do not respect certain economic constraints close down. This indicator is used to measure social impact as farm shutdowns have an impact on rural society.
		INSOLVENCY RISK	SOLVENCY RATIO	LP	This indicator is included among the social impact indicators because the risk of insolvency can be used as a risk factor for bankruptcy resulting in stress on rural society.
		AGE	AH OWNER'S AGE	LP	Age, gender, education and experience of farm owners are important indicators for analyzing the characteristics of the agricultural entrepreneurial component and the AH's behaviour.
		GENDER	AH OWNER'S GENDER	LP	
		EDUCATION	AH OWNER'S EDUCATION	LP	
		EXPERIENCE	AH OWNER'S ACTIVE YEARS IN AGRICULTURE	LP	
TO BE IMPLEMENTED	**SOCIAL INDICATORS**	LABOR USE		It can be estimated from the average working hours required for agricultural activities in relation to the UAA. e.g. in Italy	Obtaining information on the number of hours worked would be an excellent tool for assessing the labour supply provided by agricultural enterprises, with the consequent impact on rural society.
		FARM EMPLOYMENT	SOURCE OF LABOR ON FARM	LP	Shift from on-farm to off-farm (and vice versa) employment is important to observe policy effects.
		POVERTY	AH OWNER'S INCOME LEVEL	regional poverty lines	Change in AH income after policy change is important to understand welfare effects.

As can be seen from Table 3, the social indicators Risk Aversion and Insolvency Risk refer to the agent attribute Liquidity Ratio and Solvency Ratio, respectively, which are economic indicators. The use of economic parameters to measure possible social impacts was necessary due to the lack of purely social parameters at the individual company level, as this model does not directly consider labour aspects such as hours worked, wages, accidents, etc. A future implementation of these parameters is desired.

Figure 2 shows the agent attributes that take part in financial optimisation.

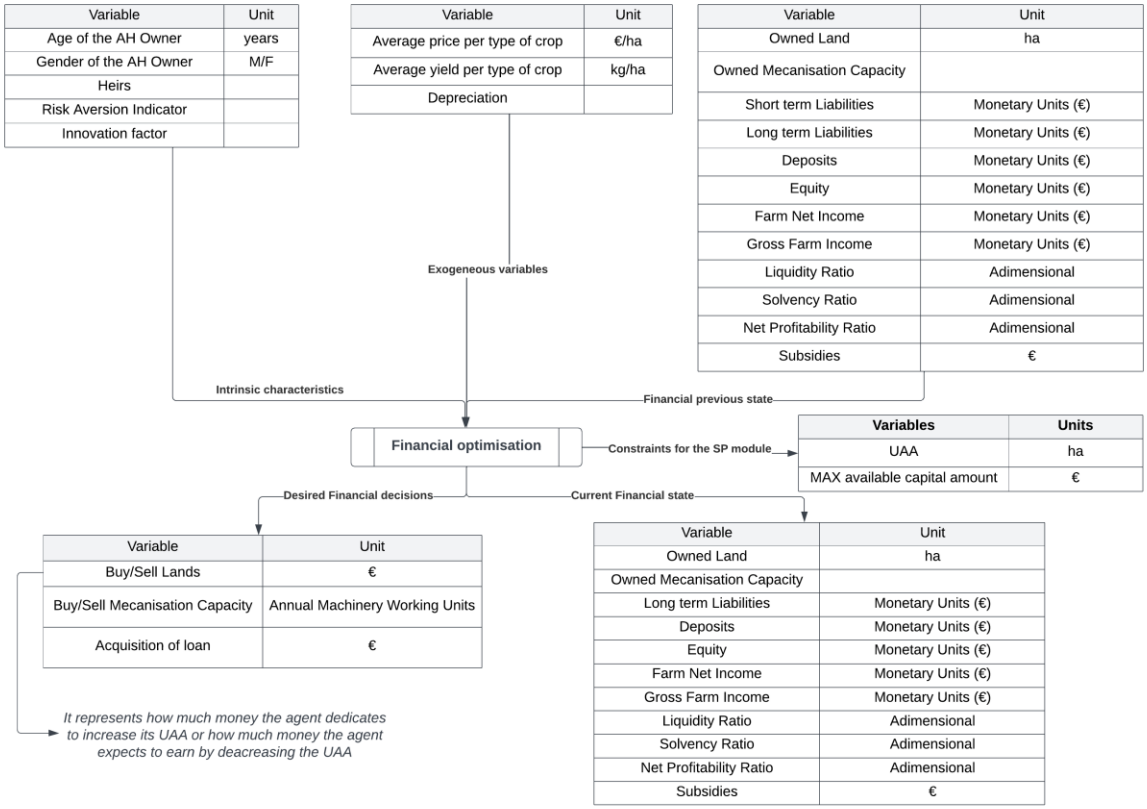

Figure 2. Agents attributes in financial optimisation.

3.3 Why K-means approach

There are several reasons for choosing to use the k-means methodology in their research or analysis.

Firstly, k-means is a simple and easy-to-implement algorithm that can handle large datasets efficiently [13]. The basic idea of the k-means algorithm is to randomly assign data points to clusters and then iteratively update the cluster centroids until convergence. This process is computationally efficient and can be easily parallelised, making it a suitable choice for large datasets. Additionally, k-means can handle data with different shapes and densities and can be extended to handle different types of distance metrics [14].

Secondly, the k-means algorithm provides interpretable results that are easy to visualise and understand. The algorithm produces *k* clusters, each with its own centroid, which can be plotted and analysed. This makes it easy to see how the data points are grouped and to identify any patterns or outliers. Moreover, the simplicity of the k-means algorithm means that it is easy to explain to non-technical stakeholders, making it an attractive choice for applications in business and other fields [14].

Thirdly, k-means is a versatile algorithm that can be applied to a wide range of problems. There are many variations and extensions of the k-means algorithm that can be used to address specific needs or requirements. For example, k-means can be extended to handle categorical data, missing values, and high-dimensional data [14].

Finally, k-means has been extensively studied and tested, and there are many resources available for learning and using the algorithm. The k-means algorithm was first introduced in the 1950s and has since become one of the most widely used clustering algorithms. As a result, there is a

vast body of literature on k-means, including many studies that compare its performance to other clustering algorithms [14]. Moreover, there are many software packages and libraries available for implementing the k-means algorithm, making it easy to use and integrate into existing workflows [15].

In conclusion, the k-means methodology is a simple, efficient, and versatile algorithm that provides interpretable results and can be applied to a wide range of problems. While there are other clustering algorithms available, k-means is a popular choice due to its ease of implementation, interpretability, and proven performance. Researchers and practitioners in various fields can benefit from using the k-means algorithm as a powerful tool for data analysis and machine learning.

3.4 Input data

Data formatting in k-means clustering analysis is a critical aspect of data analysis and the successful application of clustering techniques [16]. The input data must be organised in a consistent and coherent manner to be processed correctly by the clustering software. The input data format is dependent on the clustering software used, but generally, input data for k-means clustering must be numeric and in a tabular format. This means that the data should be organised so that each row represents an observation and each column represents a variable. Additionally, the input data must be normalised so that all variables have the same weight. This is necessary because k-means calculates the distance between observations using available variables, and if these variables have different scales, observations could be distorted, and the information could be misinterpreted. Data normalisation can be achieved using standard normalisation techniques, such as Min-Max normalisation or Z-score normalisation. The input data must be accurate and complete. Missing or inconsistent data can lead to incorrect results in the analysis. Therefore, before using the input data for k-means clustering analysis, data cleaning and quality checking must be performed. Finally, the input data format must be consistent with the type of problem being solved.

In summary, the input data format is a critical aspect of data analysis and the application of clustering techniques such as k-means clustering [16]. The input data must be organised in a consistent and coherent manner, normalised so that all variables have the same weight, accurate and complete, and consistent with the type of problem being solved.

The input to the clustering model will therefore be a dataset containing, for each row, a farm and each column, the attributes of the agent representing the variables chosen to be used to develop the impact assessment.

Since this is a double clustering, one at t_0 and one at t_7, a distinction must be made between the input data of the first clustering and those of the second:

- T_0 clustering: the input dataset comes from the initialisation process of the agents where all attributes have been assigned.
- T_7 clustering: the input dataset comes from the ABM model, which has modified all or some of the attributes as a result of its processing.

The attributes of the agents will be the same for each clustering process, the aim being to compare the results of the two processes in order to be able to develop an impact assessment.

Although the companies that will be clustered are the same at both t_0 and t_7 (only the values of the attributes change), there is, however, the possibility that the number of companies has decreased or increased as a result of the simulation performed by the ABM. This does not affect the robustness of the clustering process but should be considered when interpreting the results.

3.5 Guidelines to interpret the results

Interpreting the results of a before-after cluster analysis requires a critical evaluation and in-depth analysis of many different factors. Due to the high variability of the outputs, it is not possible to provide a methodology for interpreting the results that was universally valid. However, it is possible to follow a few guidelines in order to obtain as accurate an evaluation as possible.

- Evaluate the interpretability of the clusters in each analysis: consider whether the clusters make sense and are meaningful in the context of the sample of farms analyzed and the socio-economic context in which they are set [17].
- Compare cluster analysis results: Start by comparing the results of the two cluster analyses. Look for differences in the number and size of clusters as well as in the characteristics of the clusters themselves [18].
- Assess the quality of the clusters: once the results have been compared, assess the quality of the clusters. Look at the variation within the cluster and the variation between clusters for each analysis. A good cluster analysis will have low within-cluster variation and high between-cluster variation [19].
- Understanding the implementation of the policy: it is important to understand the nature of the policy and how it affects the data. This will help to understand the changes that have occurred in the data and how they might have affected the results of the cluster analysis [20].

4 Conclusions and future developments

This module represents an important tool for a socio-economic impact assessment within the AGRICORE project. By dividing farms into homogeneous groups, the cluster analysis allows for identifying the most vulnerable rural communities impacted by changes introduced through agricultural policies and assessing the effects of these policies on the redistribution of wealth in rural society. Although the results of the ABM are not yet available, it is believed that cluster analysis can be an effective methodology to improve the understanding of the effects of EU agricultural policies on rural society and the economy.

In summary, the outcome of this module can provide important insights for improving EU agricultural policies design and implementation, by fostering the sustainable development of rural communities and ensuring equitable access to resources and economic opportunities.

Future research in this field should continue to explore new assessment tools and refine existing methodologies to provide a sound basis for policy decisions and the well-being of rural communities. Future developments of this socio-economic impact assessment module should focus on the implementation of new indicators that can provide a more in-depth understanding of the agricultural landscape and the implementation of more refined cluster analysis methodologies.

5 References

1. C. Etumnu and A. W. Gray, "A Clustering Approach to Understanding Farmers' Success Strategies," Journal of Agricultural and Applied Economics, vol. 52, no. 3, pp. 335–351, 2020, [Online]. Available: https://EconPapers.repec.org/RePEc:cup:jagaec:v:52:y:2020:i:3:p:335-351_1
2. S. Kimura and C. Le Thi, Cross Country Analysis of Farm Economic Performance. 2013. doi: 10.1787/5k46ds9ljxkj-en.
3. L. Latruffe et al., "Measurement of sustainability in agriculture: a review of indicators," Studies in Agricultural Economics, vol. 118, pp. 123–130, Dec. 2016, doi: 10.7896/j.1624.
4. A. L. Ferriss, "The Uses of Social Indicators*," Social Forces, vol. 66, no. 3, pp. 601–617, Mar. 1988, doi: 10.1093/sf/66.3.601.
5. J. Mills et al., "Developing Farm-Level Social Indicators for Agri-Environment Schemes: A Focus on the Agents of Change," Sustainability, vol. 13, no. 14, 2021, doi: 10.3390/su13147820.
6. P. Reidsma, S. Janssen, J. Jansen, and M. K. van Ittersum, "On the development and use of farm models for policy impact assessment in the European Union – A review," Agricultural Systems, vol. 159, pp. 111–125, 2018, doi: https://doi.org/10.1016/j.agsy.2017.10.012.
7. Brian S. Everitt, Cluster Analysis. 2011. doi: 10.1002/9780470977811.
8. C. Fraley and A. E. Raftery, "Enhanced Model-Based Clustering, Density Estimation,and Discriminant Analysis Software: MCLUST," Journal of Classification, vol. 20, no. 2, pp. 263–286, Sep. 2003, doi: 10.1007/s00357-003-0015-3.
9. T. Lebacq, P. V. Baret, and D. Stilmant, "Sustainability indicators for livestock farming. A review," Agronomy for Sustainable Development, vol. 33, no. 2, pp. 311–327, Apr. 2013, doi: 10.1007/s13593-012-0121-x.
10. E. M. de Olde et al., "When experts disagree: the need to rethink indicator selection for assessing sustainability of agriculture," Environment, Development and Sustainability, vol. 19, no. 4, pp. 1327–1342, Aug. 2017, doi: 10.1007/s10668-016-9803-x.
11. J. Rice, "Environmental health indicators," Ocean & Coastal Management, vol. 46, no. 3, pp. 235–259, 2003, doi: https://doi.org/10.1016/S0964-5691(03)00006-1.
12. N. Pingault and B. Préault, "Indicateurs de développement durable : un outil de diagnostic et d'aide à la décision," 2007.
13. J. MacQueen, "Some methods for classification and analysis of multivariate observations," 1967.
14. T. Hastie, R. Tibshirani, and J. H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. in Springer series in statistics. Springer, 2009. [Online]. Available: https://books.google.es/books?id=eBSgoAEACAAJ
15. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
16. A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, 2010, doi: https://doi.org/10.1016/j.patrec.2009.09.011.
17. J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, no. 1, pp. 100–108, 1979, Accessed: May 30, 2023. [Online]. Available: http://www.jstor.org/stable/2346830
18. G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," Psychometrika, vol. 50, no. 2, pp. 159–179, Jun. 1985, doi: 10.1007/BF02294245.
19. D. Steinley, "Properties of the Hubert-Arable Adjusted Rand Index.," Psychological Methods, vol. 9, pp. 386–396, 2004, doi: 10.1037/1082-989X.9.3.386.

20. R. S. Bivand, E. Pebesma, and V. Gómez-Rubio, Applied spatial data analysis with R, Second edition. Springer, NY, 2013. [Online]. Available: https://asdar-book.org/

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

Deliverable Number	Deliverable Title	Lead beneficiary	Type	Dissemination Level	Due date
D5.1	State of the art review of agricultural policy assessment models, tools and indicators	UNIPR	Report	Public	M12
D5.4	Environmental and climate impact assessment module	IAPAS	Report	Public	M36