D2.4 Synthetic population generation module

agricore

Deliverable Number	D2.4
Lead Beneficiary	IDE
Authors	IDE, AUTH
Work package	WP2
Delivery Date	
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	IDE	27/10/2022
0.2	Content inclusion	IDE	13/01/2023
0.3	Complementary content inclusion (first draft)	AUTH	27/01/2023
0.4	Revision and comments	IDE	16/02/2023
0.5	Reviewed version	IDE	28/02/2023
1.0	Exportation and formatting (final version)	IDE	26/05/2023

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).

One of the essential inputs to execute this agent-based modelling approach is the synthetic population of agents. This is a set of autonomous decision-makers entities defined by some attributes of interest that mimics the distribution and features of the real farmers' population of interest. This deliverable presents the AGRICORE synthetic population generation (SPG) module, finalising the developments in WP4. To this end, a Bayesian network learning algorithm has been designed to model the visible and hidden relationships between the attributes of the agents based on the output of the data fusion module (D2.3), FADN data and other data sources.

In this deliverable, the complete procedure, including the employed techniques, is explained. This must guarantee a resulting representative synthetic population whose agents cannot be identified with any of the sample farms or farmers in FADN or the real population. To this end, algorithms and procedures have been defined to be able to scale up from synthetic samples that can be completely evaluated with eh FADN data to synthetic populations that must be realistic and whose validation is limited by the available census data. For this purpose, a representation weights calculation approach has been proposed, obtaining more accurate results than applying the FADN weights.

Abbreviations

Abbreviation	Full name
ABM	Agent-based model
AoI	Attributes of interest
BIC	Bayesian Information Criterion
BN	Bayesian network
CI	Conditional independence
CO	Combinatorial optimisation
DAG	Directed acyclic graph
DEM	Data extraction module
DFM	Data fusion module
DWH	Data Warehouse
FADN	Farm Accountancy Data Network
HC	Hill Climbing
IPF	Iterative Proportional Fitting
IPU	Iterative Proportional Updating
KDE	Kernel density estimation
MCMC	Markov Chain Monte Carlo
MMPC	Max-Min Parents and Children
PCA	Principal component analysis
SP	Synthetic population
SPG	Synthetic population generation
SR	Synthetic reconstruction
SS	Synthetic sample

List of Figures

List of Tables

Table of Contents

1 Introduction

The synthetic population generation (SPG) module is the last step before the use case simulation. This module finalises the work done in WP4 since the beginning of the project, which encompasses the data storage in the Data Warehouse (DWH), processing of those data and their transformation into valuable information in the form of a synthetic population (SP). The whole proposed workflow is carried out by four modules, as can be observed in Figure 1: AGRICORE DWH (D2), Big Data Extraction Module (D3), Big Data Fusion Module (D4) and Synthetic Population Generator (D5). The first three modules were already developed and presented in deliverables 2.1, 2.2 and 2.3, respectively. This deliverable addresses the development of the final module, which outputs a that mimics the distribution and features of the real farmers' population of interest.

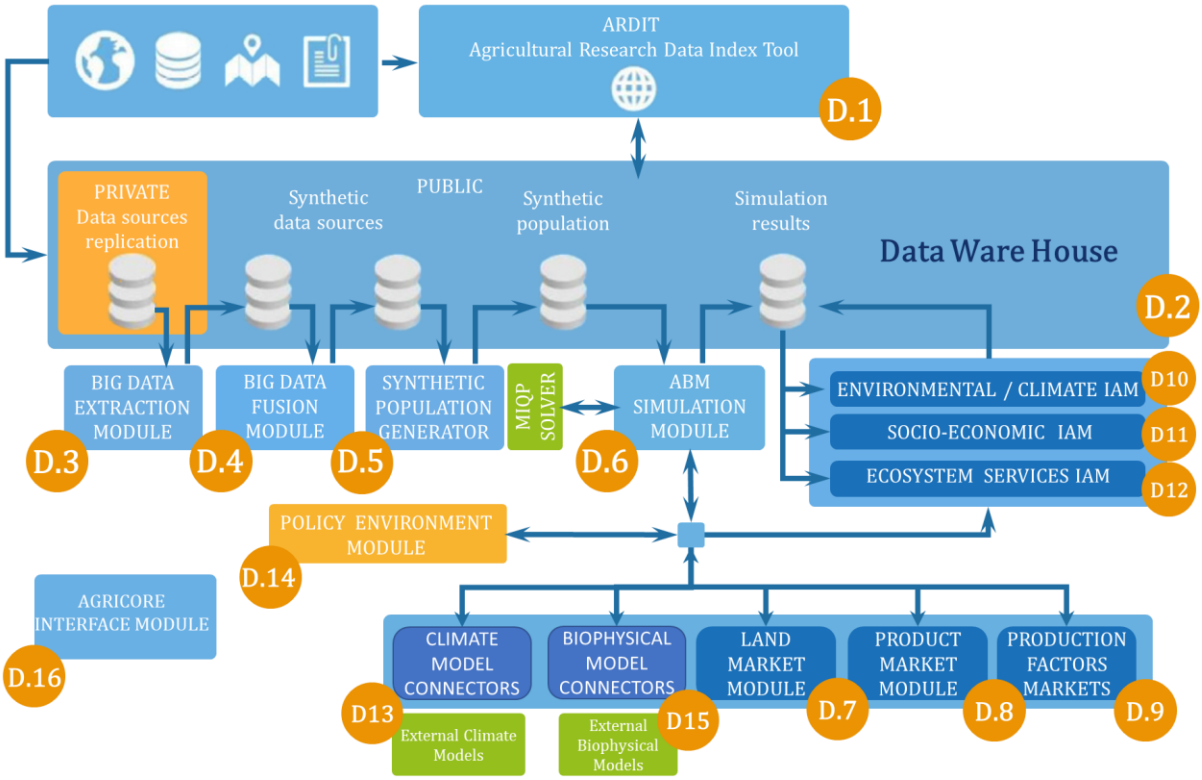

Figure 1. Modular architecture for AGRICORE.

The generation of the SP is towards a Bayesian Network (BN), mainly based on the available FADN data of the population of interest. As was explained in D2.3, Data Fusion Module (DFM) is in charge of extracting the joint probability distributions of some carefully selected Attributes of Interest (AoI). These define each synthetic farmer determining his/her behaviour during the simulation. Then, this information is processed with a set of techniques to determine the relationships between attributes. This results in a BN, such as Figure 2. This is a probabilistic graphical model that represents the conditional (in)dependencies between attributes (nodes or vertices) through arcs that indicate the direction of the relationship. For the SPG task, the model takes the form of a directed acyclic graph (DAG), where parents and child nodes can be observed, and the thickness of the arcs represents the strength of the relationships.

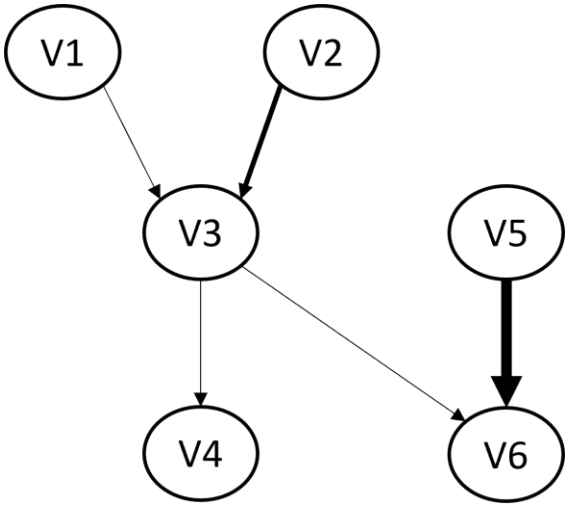

Figure 2. Example of DAG.

In D2.3, the extraction of the relationships between attributes was explained, which was illustrated with the generation of the synthetic sample (SS) of three Greek NUTS-2 regions. D2.4 goes one step further and describes the procedure to obtain a fully realistic SP. This consists of increasing the size of the SS until it coincides with the size of the real population, avoiding incurring biases and maintaining representativeness. For this purpose, the obtained SP must fulfil the following characteristics: targeted (contain just AoI without unnecessary information), microscopic (each agent represents an individual farmer) and anonymised (it must be impossible to univocally identify a synthetic agent with any of the actual farms or farmers in the sample). This SP is the input of the agent-based model (ABM) that simulates the evolution of the SP under specific conditions of the use case. In it, the agents will behave as individual and autonomous decision-making entities that asses their context and act according to their situation, expectations and objectives.

The remaining part of the deliverable is structured as follows: section 2 briefly summarises the exchange of information between the SPG module and DFM (input), and ABM module (output). Section 3 reviews the main concepts and methods for the development of BNs presented in D2.3 but elaborates on a few aspects. Section 4 explains the main concepts of SPG and introduces different state-of-the-art approaches with BNs and how they are evaluated. The advantages and limitations of those approaches are addressed in Section 5. The next section illustrates the results of applying the selected methodologies in the generation of SP, as well as the most relevant issues detected. Finally, conclusions are presented in section 7.

2 Module connections

The connection of the GPS module is in line with the other modules developed in WP4. That is, there is no direct connection between the modules, but they communicate through the DWH. For this purpose, the SPG module must have the necessary permissions to upload and download files from the DWH. This internal operation gives the generation of SPs greater modularity, although the user perceives it as if the DEM, DFM and SPG modules are cascaded from the datasets stored in the DWH to the generation of the SP. In this way, the connections to the SPG module are reflected in Figure 3.

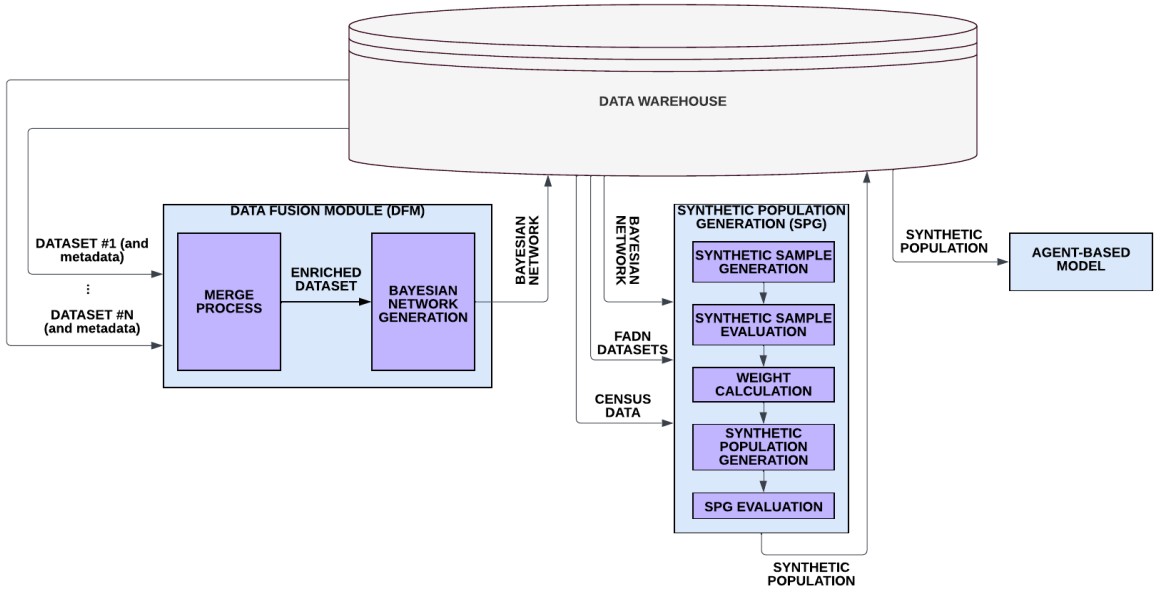

Figure 3. Connections with the DFM and ABM modules.

On the one hand, the input to the SPG module is the file(s) defining the BN, which was obtained from the processing carried out by the DFM. These file(s) are stored in the DWH and contain the following information:

- The textual definition of the BN.
- The list of attributes that are entirely independent and could be generated independently and in parallel, as well as the sequence in which the values of correlated attributes should be generated.
- The probability distribution functions (PDFs), expressed as mathematical expressions or marginal tables, required to provide such attributes.

On the other hand, the output of the SPG module will also be stored in the DWH and will be accessed by the ABM module to simulate the SPs. This output could be composed exclusively of the values of the attributes of interest of each agent in the SP; however, other intermediate outputs are stored to ensure replicability and increase the transparency of the process.

- Representation weight of each agent in terms of cultivated area and livestock in the actual population. These weights allow scaling from the SS to the SP, generating as many agents as the real population.
- Changelog file of the attributes. After generating the first version of the SP, its representativeness is checked with the total data of the real population of interest. If it is

necessary to adjust the population, the value of some attributes in certain agents is modified. These changes would be saved in a plain text file.

- Synthetic population. A CSV or similar file where the numerical values of the AoI for all agents in the SP are stored.

The last file is really the data the ABM module will access to run the simulation. The procedure of the whole simulation is described in WP3 deliverables in more detail. Basically, once each simulation year starts, each agent plans the allocation of crops and livestock based on its financial status and context (agricultural policies, available land, soil conditions, risk aversion, willingness to innovate, etc.), which are determined by its attributes and expectations regarding external factors, such as climate. This process includes a set of short-term agro-management decisions, which are constrained by long-term decisions on the economic level. Afterwards, the simulation starts, and the status of each agent changes according to its planning and external conditions. With this new scenario, which can match or not with the initial expectations, each farmer re-evaluates its context and plans for the next simulation year.

3 Introduction to Synthetic Population Generation

Synthetic population generation is a field with significant growth in recent years, especially due to the use of ABM models. Although BNs are the cornerstone of the SPG approach employed in AGRICORE, there are other widely-known methods that have been applied to similar cases. This section includes two distinctive sub-sections that present the necessary knowledge about SPG techniques to decide the most suitable one for AGRICORE. Firstly, the following sub-section describes useful key concepts of BNs addressed in D2.3. Moreover, some other important concepts and methods used in the development of BNs for SPG are explained in detail, such as value assignation and evaluation. On the other hand, the second sub-section reviews generic SPG techniques useful for the development of the AGRICORE SPG module in the literature, illustrating their advantages and disadvantages with the well-known application of household generation.

3.1 Prior knowledge of BNs

3.1.1 Bayesian networks overview

A BN [1] [2] consists of a directed acyclic graph, G, over a collection of vertices (attributes), V, and a joint probability distribution, P. This latter is related to G by the Markov condition, which states that each attribute is conditionally independent of its non-descendants given its parents. For the development of a BN learning algorithm, an essential assumption is causal sufficiency. This means that there are no latent (hidden, non-observed) attributes within the observed attributes V and all relevant ones are included.

Regarding the graphical model, some basic concepts must be highlighted in Figure 2. Firstly, independent vertices must be highlighted as the only vertices that do not conditionally depend on others, such as V1, V2 and V5. On the other hand, we must differentiate between parents and children vertices. For example, V3 is a children vertice of the parents vertices V1 and V2 because it directly depended on them. In the same line, a children vertice can also be a parent vertice. For instance, V3 is the parents vertice for V4 and V6. Finally, two graphical structures can be observed in the triplets of a BN. The first one is the v-structure, where the children vertice is called collider (for example, the triplet composed by V1, V2 and V3 (collider)). The second structure is the Λ-structure, where two vertices conditionally depend on one vertice (for example, the triplet of V3, V4 and V6).

For the SPG with BNs, BN learning algorithms that automatically construct (in)dependency relationships from observational data are necessary. Among these algorithms, it can be distinguished between constraint-based, score-based or hybrid algorithms. For this application, hybrid BN learning algorithms have been considered the most suitable ones, selecting the MMHC algorithm [3]. This algorithm consists of two phases where, first, the statistically significant links between the variables are detected, and then, a scoring approach is used to orient those relationships.

The first phase is carried out by a method called Max-Min Parents and Children (MMPC) algorithm. This iterative method first analyses the statically significant associations of an attribute of interest with other attributes via statistical tests, stores them and selects the strongest association. The second step is to perform conditional independence (CI) tests between the attribute of interest and the unselected ones, removing the stored relationships of the statically insignificant ones. Moreover, the associations are updated with the minimum value between the resulting ones of the first step and the CI tests. Finally, the loop starts again by selecting the highest association. After all this process, the end result is a matrix that asymmetrically contains the edges (undirected relationships) that were found between each attribute. Only if they were identified for both attributes, the observed edges between them

would be maintained. Further details of the procedure, including the statistical tests of independence, are included in D2.3.

The second phase of MMHC is the scoring approach with Hill Climbing (HC), where edges are converted to arrows or removed in order to maximise a score metric. This latter is a parameter of the HC algorithm to be selected among a wide range of metrics. In this case, for working with continuous data, the Bayesian Information Criterion (BIC) was selected. Regarding the HC procedure, starting with an empty graph with the exclusive constraint of the edges detected in the previous phase, a greedy HC search is conducted in the space of BNs [3]. Pursuing the maximum score increase, the edges are deleted and re-oriented, always avoiding creating cycles in the BN space. This whole process is recursively performed until determining the direction of all edges and maximising the scoring metric. The final result is a complete DAG that reflects the existing relationships between the attributes based on the input data.

However, it is a common practice to introduce external knowledge to enhance the resulting DAG from executing the MMHC algorithm. This knowledge takes the form of pre-defined directions of the association between some attributes. In this case, these relationships are automatically extracted from the input data as forbidden directions by the DEM (deliverable D2.2). This prior information is included in the scoring phase, enhancing the fit of the BN to the reality to be modelled.

Finally, two techniques are employed to evaluate the resulting DAG. On the one hand, the importance of the detected relationships is measured through the reduction in the BIC score when a specific arrow is deleted while maintaining the rest of the BN structure. The higher the reduction in the score, the higher the indications that this directed relationship is important or strong. This allows ordering the relationships based on their strength. On the other hand, the bootstrap technique is used to measure the confidence or stability of the discovered relationships. To this end, the holdings are sampled with replacements from the observed holdings and a BN is learned with the aforementioned methods. This process is repeated 1,000 times, measuring the proportion of times that the same relationships are discovered.

3.1.2 Assignation values to attributes

This sub-section explains in more detail the methods used for the value assignation for attributes in SPG. Since there are different types of values and distributions, most of them highly skewed to the right and with zero values, random value assignation will lead to an unrealistic SP. For this reason, the solution was to design a sophisticated generation method based on non-parametric regression that relied on the BN structure learnt using the characteristics of the observed farms. As required by the BN, the order of generation is sequential. In other words, each attribute's values depend on the value of its parent attribute(s). For the value assignation, different techniques have been considered depending on whether the attribute has parents or not. In the case of not having parents, the values are assigned with kernel density estimation (KDE). On the contrary, the values of attributes are estimated with the k-NN algorithm. Moreover, a special case is the attributes related to climate and soil conditions, whose values are assigned with 1-NN or k-NN algorithms depending on the resolution of the available datasets. The following sub-sections explain the selected techniques.

3.1.2.1 1-NN algorithm

In the first approach, the environmental attributes, such as soil pH, temperature, humidity, rainfall, and temperature measurements, were matched with the available FADN data using the 1-NN. Each farm was assigned the set of measurements corresponding to the location closest to the farm, with their proximity being computed via the Euclidean metric. Nonetheless, based on the same metric, their values could be assigned through a k-NN algorithm if more data were available, which provides more accurate values. Thus, using one algorithm or another will depend on the data resolution.

The spherical coordinates (latitude and longitude) are transformed into Euclidean coordinates. The following transformation is applied using the locations of the sample farms U^{farm} and of the locations of the environmental attributes U^{env}.

$$U = (u_1, u_2, u_3) = (\cos(lat), \sin(lat)\cos(long), \sin(lat)\sin(long))$$

Based on both distances, the Euclidean distance between the location of each farm and the location of the environmental attributes is calculated.

$$D(U^{farm}, U^{env}) = \sqrt{\left(u_1^{farm} - u_1^{env}\right)^2 + \left(u_2^{farm} - u_2^{env}\right)^2 + \left(u_3^{farm} - u_3^{env}\right)^2}$$

Finally, the location of the environmental attributes with the smallest Euclidean distance from each farm is selected, and their environmental conditions are assigned to that farm.

3.1.2.2 Kernel density estimate

Kernel density estimation (KDE) is a well-known technique for estimating a probability density function based on a finite sample population [4]. In the AGRICORE project, this technique is used for the value assignation of independent attributes, that is, those with no parents. To this end, the KDE of the distribution of the non-zero values is computed, and the non-zero values to be assigned are generated based on it, with zero values remaining the same.

Suppose there is an attribute, X taking values $x_1, x_2, \ldots, x_n$, where n denotes the sample size. Its KDE is given by

$$\hat{f}(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} e^{-\frac{(x-x_i)^2}{2h^2}},$$

where $h = 0.9\min\left(\hat{\sigma}, IQR/1,34\right) n^{-1/5}$, where $\hat{\sigma}$ denotes the sample standard deviation and IQR is the interquartile range ($Q_3 - Q_1$, where Q_1 and Q_3 denote the 25th and the 75th quartile of the values, respectively). In simple words, one should compute the above expression for every observation $x_i \ (i = 1, \ldots, n)$. Random number generation from the attribute X that has no parents occurs using the following formula $_i = \bar{x} + (x_i - \bar{X} + \hat{h}z_i)/\left(1 + \hat{h}^2/\hat{\sigma}^2\right)^{1/2}$, where z_i are random values generated from the standard normal distribution, $\hat{h}$ is the estimated bandwidth and $\bar{x}$ and $\hat{\sigma}^2$ denote the sample mean and variance, respectively, of the observed attribute values x_i and $\hat{x}_i$ denotes the values of the synthetic attribute.

3.1.2.3 k-NN regression and classification

For attributes with at least one parent, the k-NN algorithm is used to generate values from. Using the observed farms, if the attribute is continuous, its estimated value is given by $\tilde{x} = \sum_{i \in \mathcal{C}_k} x_i$, where $\mathcal{C}_k$ denotes the k closest neighbours of, whose proximity is computed via the Euclidean distance of the parent attributes. Using the observed attributes, the value of k is chosen towards minimising the sum of squares of the errors of the fitted values, $\sum_{j=1}^{n}\left(\tilde{X}_{ij} - X_{ij}\right)^2$. We generate values for this attribute based on the selected value of k and the synthetically generated values of the parent attribute(s).

3.1.3 Evaluation of the synthetic sample generation

Regardless of whether a synthetic sample or a synthetic population is generated, a thorough evaluation process is required to ensure that the population of agents represents the real population of interest. To this end, two testing procedures, a parametric one and a non-parametric one, will be used to assess the fit of univariate distributions to the true distributions. Although these methods were introduced and employed in the SS evaluations presented in deliverable 2.3, they are further described in the following sub-sections according to their application order. After their application, which attribute(s) causes the inconsistency and what degree of accuracy could be reached, the γ-OMP [5] and FBED [6] attribute selection algorithms are jointly employed. Furthermore, a principal component analysis (PCA) is applied to reduce the dimension of the sample and set visual comparisons.

3.1.3.1 KDE hypothesis test of equality of two distributions

In order to test the null hypothesis $H_0 : f_1 = f_2$, where f_1 and f_2 denote the probability density function of an attribute of the observed and of the synthetic farms, this non-parametric test that makes no assumptions about the functional form of the f_s is applied. The measure of discrepancy between the two f_2 is the integrated squared error $\int \left(\hat{f}_1(x) - \hat{f}_2(x)\right)^2 dx$, where $\hat{f}_1$ and $\hat{f}_2$ indicate the kernel density estimate of f_1 and f_2, respectively. The asymptotic null distribution of this test statistic is the standard normal N(0, 1), i.e. the normal distribution with mean zero and variance one, and hence the asymptotic p-value is computed.

The test statistic of the energy test of equality of distributions has the following formula

$$e(S_1, S_2) = \frac{n_1 n_2}{n_1 + n_2} \left(2M_{12} - M_{11} - M_{22}\right),$$

where $M_{ij} = \frac{1}{n_i n_j} \sum_{p=1}^{n_i} \sum_{q=1}^{n_j} \|V_{ip} - V_{jq}\|$, for $i, j = 1, 2$ and $\|\cdot\|$ denotes the Euclidean norm.

Specifically for the irrigation system, the size of the economic class, the manager's gender and training, the X^2 test is applied because these attributes take discrete values

$$X^2 = \sum_{k=1}^{K} \frac{\left(A_k - \hat{A}_k\right)^2}{\hat{A}_k},$$

where A_k and $\hat{A}_k$ denote the frequency of the k-th possible value of the attribute of the observed and of the synthetic farm, respectively and K is the number of possible values of the attribute. If $X^2 > \chi^2_{0.95,K-1}$ the equality of the distributions of the observed and the synthetic farms is rejected. The $\chi^2_{0.95,K-1}$ denote the 95% quantile of the χ^2 distribution with $K - 1$ degrees of freedom, and the p-value is computed using this χ^2 distribution.

For either testing procedure, if the p-value is less than 0.05, the H_0 is rejected, and hence the distributions cannot be assumed statistically equal.

3.1.3.2 Energy distance test of equality of two distributions

The second test of equality is the non-parametric energy distance-based test presented in [7], which was also applied to test the equality of the joint distributions of the sample and synthetic farms. This test enables studying the equality of the distributions at the multi-attribute levels, that is, considering all attributes. Moreover, the process is based on Euclidean distance between sample elements and performs particularly well in high dimensions. Indeed, the computational

complexity of the algorithm does not depend on the dimension and number of samples. Finally, a key feature of the presented testing procedure is that it is multivariate and distribution-free, which is not very common among classical approaches. Thus, distributional assumptions are not required, increasing its robustness.

This approach to testing the k-sample hypothesis calculates the $\mathcal{E}$ statistic (defined below) based on Euclidean distances between sample elements. This statistic is more general than tests based on ranks of neighbours in the sense that no assumptions about the continuity of the underlying distributions are necessary. Indeed, this testing approach has shown better performance than the nearest neighbour test in higher dimensions.

Suppose there are two samples, $\mathcal{A}$ and $\mathcal{B}$, composed by the elements $X_1, X_2, \ldots, X_{n1}$ and $Y_1, Y_2, \ldots, Y_{n2}$, respectively, where n_1 and n_2 are the sample sizes, and X_i and Y_j are vectors in R^d, with d corresponding to the number of attributes. The two-sample test statistic, $\mathcal{E}_{n_1,n_2} = \epsilon(\mathcal{A}, \mathcal{B})$, is defined as follows:

$$\mathcal{E}_{n_1,n_2} = \frac{n_1 n_2}{n_1+n_2}\left(\frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{m=1}^{n_2}\|X_i - Y_m\| - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\|X_i - X_j\| - \frac{1}{n_2^2}\sum_{t=1}^{n_2}\sum_{m=1}^{n_2}\|Y_t - Y_m\|\right)$$

Let $\alpha \in (0,1)$ be fixed, and let c_α be the constant satisfying $\lim_{n\to\infty} P(\mathcal{E}_n > c_\alpha) = \alpha$, where P is the limiting probability for two independent random samples size n_1 and n_2. The size α test of $H_0 : f_X = f_Y$ based on $\mathcal{E}$ rejects the null hypothesis if $\mathcal{E}_n > c_\alpha$.

3.2 Synthetic population generation

The task of SPG has attracted research interest in the 21st century, and many suggestions have been proposed. The present section conducts a literature review on SPG techniques, spanning from 2001 to 2019, emphasising that most papers present generation techniques that focus on social characteristics, such as household generation. These techniques can be divided into two broad categories, synthetic reconstruction and combinatorial optimisation approaches [8]. Moreover, these techniques are illustrated and compared between them with their performance in the problem of synthetic household generation [9].

Synthetic reconstruction (SR) is perhaps the safest option, as it generates new observations. The rationale is to sequentially create observations using the information on some attributes and then exploit those generated observations to fill the gaps and generate values for the other attributes. This process is continued until the population values are all filled. The cornerstone of this process is some true population constraints that must be satisfied in the SP as well. Those constraints are usually taken from the available census data but are not the sole constraints. They are, however, the only feasible ones, given the accessible information. A standard algorithm for this technique is the Iterative Proportional Fitting (IPF) which requires two-way contingency tables, or in more technical terms, the joint (bivariate) distribution of two attributes. In the case of more than two attributes, IPF considers pairs of attributes conditioning on the values of the other attributes.

Combinatorial optimisation (CO), on the other hand, uses the publicly available data and samples from them (with replacement) until the value of a stress (fit) criterion is minimised. Similarly to SR, a list of constraints must also be satisfied (related to the stress criterion), but their difference lies in their generated output. SR simulates new values, whereas CO reproduces different combinations of the observed values. SR proceeds hierarchically, simulating attributes in a specific order, whereas CO uses all attributes in an iterative process. CO starts from a randomly chosen set of observations and replaces an observation with a new one if the fit is improved until the fit can not be further improved. As the name reveals, this is an optimisation strategy which, however, has no guarantees of achieving the optimal solution (the optimal fit). The SR approach

is evidently faster and perhaps produces an SP that is more realistic than CO, but the latter can yield an SP that better fits some known constraints of the true population.

The drawback of both approaches though is that the SP is calibrated against some known tabular information (a number of cross-tabulated attributes) which in practice may not be representative of the characteristics of the true population, e.g. full relationships among the attributes. In this project, we are interested in estimating the joint distribution of the attributes, and hence the SR approach is to be followed. In the next subsection, the literature on the SR approaches is reviewed.

3.2.1 Synthetic reconstruction approaches

The SR approach is more popular than the CO, as depicted by the vast literature. [10] generalised the IPF algorithm to the Iterative Proportional Updating (IPU) algorithm in order to better capture the overall joint distribution of the attributes of the true population. They claim to have solved the zero-cell problem, that is, the case of a zero frequency of a household with some specific characteristics. They also claim to have solved the zero-marginal problem. For example, in some regions, there are no low-income households. The drawback of the minimisation process of the IPU is that there is a positive probability that no solution exists. According to [10], IPU will fail in extreme cases, such as when all persons of certain types completely fall into a single household type. Their solution is to consolidate and aggregate such cells. Further, IPU may reach a solution that lies outside the feasible region. In those cases, IPU will iterate until a corner/border solution is found.

[11] proposed an iterative approach to generate statistically realistic populations of households matching a few attributes only, type and size of household and age of participants. Their idea can be extended to more attributes, but it suffers from the usual disadvantage of the synthetic reconstruction approaches; that is, the order to generate the attributes is unknown. The advantage is that they consider no sample data, only the tabular information, which on the other hand, does not take into account the relationship among the attributes. [12] utilised the IPF algorithm in order to generate households controlling for a few characteristics only. [13] was the closest work to our case study as they created an SP of poultry farms in the USA. Given the vast availability of attributes, they identified the most important factors that influence the siting of poultry houses. However, not much technical information is given, but the authors do mention many pre-processing steps and the use of a geographical information system (coordinates) as a highly important asset. [14] compared the sample-free method of [11] to the IPU algorithm [10] in generating individuals and households in France with very few characteristics only and found that the differences between the two approaches are very small.

[15] described a sample-free SPG procedure, again in the context of household generation. [16] examined the problem from the game theoretic scope. We consider this approach unfeasible for many attributes, which in turn contain more than just two values. Finally, [17] reviewed some SPG approaches but focused on papers that were published in a specific journal between 20014 and 2018.

On a different route, [18] proposed a hierarchical generation using Markov Chain Monte Carlo (MCMC) simulation and, in particular, the Gibbs sampler. Gibbs sampler draws values from an attribute conditioning on all other attributes. Respecting the known hierarchy of the attributes, [18] carefully generated the values of the attributes of the households. For instance, the age and gender of a spouse are generated after the household owner is generated, and their kids are generated after. The generation of an attribute conditional of previous attributes takes place using a regression model. A multinomial regression is fitted, and the estimated probabilities are fed into the multinomial distribution that generates values for the attribute of interest. Starting from an attribute, all attribute values are sequentially generated, and the process starts over again. According to the MCMC theory, hundreds of thousands, or perhaps millions of values,

must be generated, and then only a fraction of them (10%, for instance) is used. The final SP does not satisfy some known marginal constraints, and a post-process of the generated data must be applied.

4 Bayesian networks for SPG

This section analyses the use of BNs in the literature for the generation of SPs. Although the MMHC BN learning algorithm was selected as the most suitable one for SPG in AGRICORE, the analysis intends to consider further factors, such as the scalability from sample populations to SPs. In addition, it encompasses a more theoretical analysis (presented in this section) and a practical one, which consists of the application of MMHC for the use case SPG (described in the following section). Finally, it should be highlighted that BNs have been successfully adopted for the purpose of SPG [19] but with discrete data, and their application with the FADN continuous data is a challenge, requiring appropriate treatment, as stated earlier.

Along the lines of [20], there lies a rather more promising approach which relies upon Bayesian networks (BNs) [21] [19]. BNs have been used by [22] to analyse data extracted from the British general household survey. More importantly, [23], [24], and [25] used BNs to generate synthetic privacy data, population synthesis and social media profiles data, respectively. Following [21] and [19], the use of BNs is proposed because they take into account the conditional distribution of some attributes formulating a suitable way to generate an SP. Further, BNs have been successfully coupled with ABM models [26] [27].

The rationale is to first create a network of the attributes that can be represented via a graph where all attributes appear with nodes (vertices) and can either be connected with an arrow indicating the direction of their relationship or not connected at all. This yields two advantages over the previous SPG approaches: a) it provides information on the joint distribution of all attributes, and b) it shows which attributes depend upon which in an ordered fashion. For example, one will be able to identify the attributes that affect a given attribute and use this information to generate values from that given attribute. The population is then hierarchically generated as in the SR approaches (e.g. [20]), but the estimated conditional distributions will be more accurate than in the MCMC-based approach of [20].

4.1 Advantages and limitations of BNs and SPG framework

After the study of different SPG techniques, their suitability for the AGRICORE project has been checked based on the project's needs and available data. It can be concluded that BNs could fill the gap to obtain a realistic landscape of the population of interest to work with the ABM model and obtain feasible results. Despite this, the suitability analysis has reflected a series of advantages and limitations of SPG techniques for their application in the project. They are listed in this section.

4.1.1 Advantages

- Two interrelated advantages of BN learning are the detection of statistically significant associations among the attributes and the topological order of the attributes (a tree-like structure, see Figure 2 for an example). This feature is very helpful in order to generate values (hierarchically) from the attributes.
- BN learning algorithms are data-driven and agnostic of any theory; hence, the estimated parent-child relationship between the attributes might be wrong. In this case, the addition of prior knowledge is necessary to avoid these erroneous directions in the relationships of the attributes and facilitate the construction of a realistic hierarchical structure.
- BNs have proved useful for the SPG task. The task requires the specification of the joint distribution of the data, and BNs accomplish this [21] [19]. Based on the Markov condition, the joint distribution can be written down explicitly, allowing for SPG in sequential order. The

generation process commences by generating values for attributes that have no parents. These values are used to generate values from their children attributes, and the process continues until values for all attributes have been generated.

- BNs can be adopted by ABMs [26] [27].

4.1.2 Limitations

- BNs cannot handle attributes measured at different levels when the data are heterogeneous in the sense that they are measured at different levels of aggregation. For example, data for some attributes might exist at the household level, whereas for others, they exist at zip code levels, city level, county level and so on. In this case, one should be able to reduce/disaggregate the data to a lower level. If most attributes are available at the city level, but some are not, for example, the funding is available at a prefecture-level, it could be lowered down to the city level by leveraging or eliciting the information from other agricultural attributes.
- In all regression models, if important attributes are missing from the equation, bias is added. The same is true for BNs that assume that all relevant attributes have been included.
- The MMHC, similarly to its competitors, discovers only the linear or monotonic relationships among the attributes.

5 SPG of farms

The generation of the synthetic population relies heavily on the corresponding synthetic sample, so the generation of a realistic SS according to the actual sample data is essential. Thus, the SPG task will be executed similarly to the SSG, whose whole process is illustrated in Figure 4. This process starts with the creation of the BN based on the FADN data of the population of interest. Using the MMHC algorithm described in Section 3.1.1., the DAG is obtained, and the strength of the detected relationships is validated, resulting in the BN capable of generating synthetic agents with the characteristics of the real population of interest. The next step is the assignment of the attribute's values which is done attribute by attribute and with an algorithm that depends on whether the attribute has parents or not. For attributes with parents, the k-NN algorithm is used, whereas, in the case of not having parents, the KDE algorithm is employed. In the flow diagram of this second step in Figure 4, a final validation step can be observed. This consists of checking if the average value of the synthetic attribute is similar to the mean value calculated from the observed farmers in FADN. If any bias is detected, the values of that attribute for all agents are assigned again. Finally, once the entire value assignment process finishes, the resulting SS is validated. To this end, three strict tests are carried out: i) γ-OMP and FBED tests, ii) KDE hypothesis test for comparing individual attributes distributions, and iii) Energy distance test for comparing joint distributions. Thanks to this validation, a realistic SS is obtained.

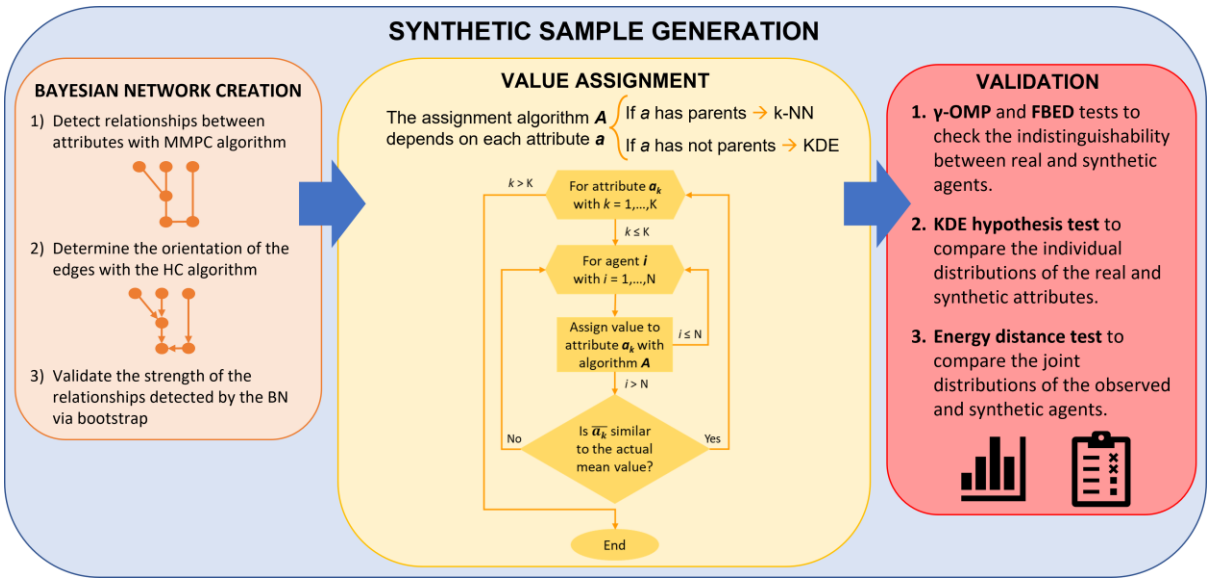

Figure 4. Synthetic sample generation process.

On the basis of the generated SS and the census data is possible to scale up to an SP. For that purpose, an intermediate step is needed, the calculation of representation weights. These are calculated with the exponential empirical likelihood, which is explained in the section below, and allow for estimating how many agents of each type are in the real population according to the census data. The types of agents are defined for each NUTS2 region and based on its crop(s) and livestock. Thus, for a type of agent i whose representation weight is w_i, $i = 1, 2, \ldots, K$ agents, where K is the total number of farmers in the real population, must be generated. To do this, several synthetic samples are generated until the number of agents of each type estimated by their weights is reached. Lastly, the SP is validated by comparing some parameters with the totals included in the census, mainly cultivated area, production and the number of heads of livestock. This whole procedure is depicted in Figure 5.

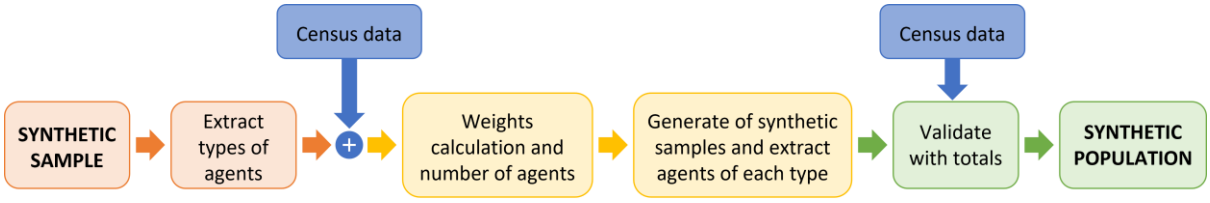

Figure 5. Synthetic population generation process.

5.1 Weights calculation

The FADN data contain information on the representation weight of each farm; however, when used, the results were not accurate. The totals in some attributed exceeded, significantly, the known totals of some attributed. For this reason, the representation weights have been estimated relying on the total cultivated area of the crops and not on any other attribute of each NUTS-2 region. The weights were estimated using the exponential empirical likelihood, a.k.a. exponential tilting [28] or the empirical likelihood [29] (explained below). The reasoning behind these two non-parametric likelihoods is to put some positive weights (p_i), which sum to one, on the observations such that the weighted sample mean ($\bar{x}$) is equal to some pre-specified population mean (μ). The mean, in this case, consists of the mean total cultivated area in each of the crop products of each NUTS-2 region. Initially, the mean production was also included, but none of the two non-parametric likelihoods converged, and that is why only the land has been retained.

In the exponential empirical likelihood, the choice of p_i will minimise the following objective function

$$\sum_{i=1}^{n} p_i \log(np_i),$$

subject to the constraint defined in [28]

$$\sum_{i=1}^{n} p_i x_i = \mu,$$

where n denotes the sample size. Thus, with the introduction of the Lagrangian parameters on the constraints and after some algebra, the form of the constraint becomes

$$\frac{\sum_{i=1}^{n} e^{\lambda x_i}(x_i-\mu)}{\sum_{j=1}^{n} e^{\lambda x_j}} = 0 \Rightarrow \frac{\sum_{i=1}^{n} x_i e^{\lambda x_i}}{\sum_{j=1}^{n} e^{\lambda x_j}} - \mu = 0.$$

A numerical search over λ is necessary to find the probabilities that will minimise the objective function (7) or solve Eq. (9). The drawback with this approach is that there might be no solution for Eq. (9), which comes from the fact that μ does not lie within the convex hull constructed from the data. The empirical likelihood [29] is an alternative method for this case, and the equation to be solved is

$$\sum_{i=1}^{n} \frac{1}{n} \frac{x_i-\mu}{1+\lambda(x_i-\mu)} = 0.$$

Empirical likelihood does not solve the convex hull problem, but it rather addresses it in a more efficient way. The disadvantage is that the estimated weights are less accurate compared to those estimated by the exponential empirical likelihood.

In the case of agricultural holdings with livestock, the calculation of the weights is different. For that, the variables of livestock production must be considered. These variables are gathered in the following table, where $i = 1, 2, \ldots, K$ with K the number of livestock species considered

in the use case. For instance, in the Greek use case, the analysed livestock species are 1: equidae; 2: bovine; 3: sheep and goats; 4: pigs and poultry; and 5: bees.

Table 1. Variables of livestock production.

Variable	Description
Yi.1	Number of animals
Yi.2	Number of animals sold
Yi.3	Value of sold animals
Yi.4	Number of animals for slaughtering
Yi.5	Value of slaughtered animals
Yi.6	Number of animals for rearing-breeding
Yi.7	Value of animals for rearing-breeding

Based on those variables, a set of ratios is calculated: $R_1 = \frac{Yi.3}{Yi.2}$, $R_2 = \frac{Yi.5}{Yi.4}$ and $R_3 = \frac{Yi.7}{Yi.6}$. These are the basis for calculating other ratios:

$$a_1 = \frac{R1}{R1+R2+R3} Yi.1, \; a_2 = \frac{R2}{R1+R2+R3} Yi.1 \text{ and } a_3 = \frac{R3}{R1+R2+R3} Yi.1$$

The weighted livestock is the sum of a_1, a_2 and a_3.

5.2 Results of generating a synthetic population for UC#

The generation of real use case synthetic population allows for verifying the suitability of the techniques and procedures described in this deliverable. The main output, in this sense, is the calculation of weights. Before their calculation, the FADN representation weights of each farm were used to pass from the SS to the SP. The application of those weights led to significant biases, exceeding by far the known totals of the real population of interest. On the contrary, the representation weights calculated with the equations described before achieve a closer adjustment of the SP in terms of the total values of the attributes. Proof of this can be seen in Table 1, where the ratios of estimated to true totals of cultivated area and production of 14 crops using the FADN representation weights and the estimated ones with exponential empirical likelihood are presented. The closer these ratios are to 1, the better represented the real population is and the better the SP is. Thus, comparing one by one, it can be observed that the majority of the ratios resulting from applying the estimated weights are closer to 1 than those calculated with the FADN weights. The ratios of Table 1 were calculated for the NUTS2 region of Central Macedonia for 2018.

Table 2. Ratios of estimated totals to true totals using the FADN representation weights and the estimated weights.

Attribute	Ratio with FADN weights	Ratio with estimated weights	Attribute	Ratio with FADN weights	Ratio with estimated weights
X1.1	2.000	1.027	X8.1	1.448	1.059
X1.3	2.302	1.009	X8.3	1.922	0.983
X2.1	2.139	1.011	X9.1	16.978	1.076
X2.3	1.670	0.977	X10.1	0.985	0.697
X3.1	3.633	3.726	X10.3	0.954	1.058
X3.3	1.537	1.183	X11.1	1.446	0.968
X4.1	1.084	0.951	X11.3	2.788	0.877

X4.3	2.011	1.064	X12.1	1.587	1.204
X5.1	2.397	1.043	X12.3	2.282	1.151
X5.3	1.996	0.962	X13.1	11.791	0.871
X6.1	1.972	1.150	X13.3	3.704	2.582
X6.3	0.902	0.999	X14.1	2.232	1.730
X7.1	8.488	5.873	X14.3	1.212	1.011
X7.3	3.073	1.024			

Nonetheless, some deficiencies in the initial procedure were detected. Firstly, the evaluation cannot be performed with all totals because the totals of some attributes are not available in census data. In addition, this would be computationally inefficient because the evaluation and adjustment of some attributes do not have a significant impact on the output of the model. Secondly, the initial idea of evaluating each attribute during the process of hierarchically generating the SP is not feasible, and it is applicable only upon completion of the SPG. The reason is that we generate sample FADN data for each NUTS-2 region in order to match the characteristics of the synthetic farms to those of the observed farms. To generate the population, we simply generate multiple samples until the desired number of farms of each type is reached, that is, $i = 1, 2, \ldots, K$ farms, where K is the total number of farmers in the population and p_i is the calculated representation weight for each type of farm i. Finally, the values of some attributes are compared with the known totals of these attributes.

6 Conclusions

This deliverable presented the complete Synthetic Population Generation process for simulating the AGRICORE use cases and closing the work in WP2. This process is the completion of the SPG task that started with the development of the Data Warehouse to store all the necessary datasets (D2.1), processing and transformation of those data into useful information by the Data Extraction Module (D2.2) and the generation of Bayesian networks from those data by the Data Fusion Module (D2.3). On the basis of the algorithms presented in D2.3 to generate synthetic samples of anonymised agents, the processes to scale up from those synthetic samples to synthetic populations have been proposed. This includes the detailed definition of the algorithm already presented in D2.3 and the development of others for that scale-up. Regarding the latter, the main contribution has been the representation weights estimation through empirical likelihood.

The results show that the mathematical artefact to estimate the representation weights improves the FADN ones, resulting in a more realistic synthetic population. This makes the proposed procedure very promising for application in the AGRICORE use cases. The next step is the integration of this procedure as a service in the AGRICORE suite to test the generation of complete synthetic populations for the 3 use cases contemplated in the project.

7 References

1. J. Pearl, Probabilistic reasoning in intelligent systems: Networks of plausible reasoning. Morgan Kaufmann Publishers, Los Altos, 1988.
2. P. Spirtes, C. N. Glymour, and R. Scheines, Causation, Prediction, and Search. MIT press, 2000.
3. I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," Machine Learning, vol. 65, no. 1, pp. 31–78, 2006.
4. S. Węglarczyk, "Kernel density estimation and its application," in ITM Web of Conferences, EDP Sciences, 2018, p. 00037.
5. M. Tsagris, Z. Papadovasilakis, K. Lakiotaki, and I. Tsamardinos, "The γ-OMP algorithm for feature selection with application to gene expression data," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 2, pp. 1214–1224, 2022.
6. G. Borboudakis and I. Tsamardinos, "Forward-backward selection with early dropping," The Journal of Machine Learning Research, vol. 20, pp. 276–314, 2019.
7. G. J. Székely, M. L. Rizzo, and others, "Testing for equal distributions in high dimension," InterStat, vol. 5, no. 16.1, pp. 1249–1272, 2004.
8. Z. Huang and P. Williamson, "A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata," Department of Geography, University of Liverpool, 2001.
9. T. Arentze, H. Timmermans, and F. Hofman, "Creating synthetic household populations: Problems and approach," Transportation Research Record, vol. 2014, no. 1, pp. 85–91, 2007.
10. X. Ye, K. Konduri, R. M. Pendyala, B. Sana, and P. Waddell, "A methodology to match distributions of both household and person attributes in the generation of synthetic populations," in 88th Annual Meeting of the transportation research Board, Washington, DC, 2009.
11. F. Gargiulo, S. Ternes, S. Huet, and G. Deffuant, "An iterative approach for generating statistically realistic populations of households," PloS one, vol. 5, no. 1, p. e8828, 2010.
12. J. Auld and A. Mohammadian, "Efficient methodology for generating synthetic populations with multiple control levels," Transportation Research Record, vol. 2175, no. 1, pp. 138–147, 2010.
13. M. C. Bruhn et al., "Synthesized population databases: a geospatial database of US poultry farms," Methods report (RTI Press), p. 1, 2012.
14. M. Lenormand and G. Deffuant, "Generating a synthetic population of individuals in households: Sample-free vs sample-based methods," arXiv preprint arXiv:1208.6403, 2012.
15. J. Barthelemy and P. L. Toint, "Synthetic population generation without a sample," Transportation Science, vol. 47, no. 2, pp. 266–279, 2013.
16. C. Proietti and A. Franco, "Social norms and the dominance of low-doers," Journal of Artificial Societies and Social Simulation, vol. 21, no. 1, 2018.
17. K. Chapuis and P. Taillandier, "A brief review of synthetic population generation practices in agent-based social simulation," in submitted to SSC2019, Social Simulation Conference, 2019.
18. D. Casati, K. Müller, P. J. Fourie, A. Erath, and K. W. Axhausen, "Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking," Transportation Research Record, vol. 2493, no. 1, pp. 107–116, 2015.
19. L. Sun and A. Erath, "A Bayesian network approach for population synthesis," Transportation Research Part C: Emerging Technologies, vol. 61, pp. 49–62, 2015.
20. D. Casati, K. Müller, P. J. Fourie, A. Erath, and K. W. Axhausen, "Synthetic opulation generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking," Transportation Research Record, vol. 2493, no. 1, pp. 107–116, 2015.

21. J. Young, P. Graham, and R. Penny, "Using Bayesian networks to create synthetic data," Journal of Official Statistics, vol. 25, no. 4, p. 549, 2009.

22. P. Sebastiani and M. Ramoni, "On the use of Bayesian networks to analyze survey data," Research in Official Statistics, vol. 4, no. 1, pp. 53–64, 2001.

23. J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: Private data release via Bayesian networks," ACM Transactions on Database Systems (TODS), vol. 42, no. 4, pp. 1–41, 2017.

24. A. Ilahi and K. W. Axhausen, "Integrating Bayesian network and generalized raking for population synthesis in Greater Jakarta," Regional Studies, Regional Science, vol. 6, no. 1, pp. 623–636, 2019.

25. I. Deeva, P. D. Andriushchenko, A. V. Kalyuzhnaya, and A. V. Boukhanovsky, "Bayesian Networks-based personal data synthesis," in Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good, 2020, pp. 6–11.

26. V. Kocabas and S. Dragicevic, "Agent-based model validation using Bayesian networks and vector spatial data," Environment and Planning B: Planning and Design, vol. 36, no. 5, pp. 787–801, 2009.

27. V. Kocabas and S. Dragicevic, "Bayesian networks and agent-based modeling approach for urban land-use and population density change: a BNAS model," Journal of Geographical Systems, vol. 15, no. 4, pp. 403–426, 2013.

28. B. Efron, "Nonparametric standard errors and confidence intervals," Canadian Journal of Statistics, vol. 9, no. 2, pp. 139–158, 1981.

29. A. B. Owen, Empirical likelihood. Boca Raton: Chapman & Hall/CRC, 2001.

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

Deliverable Number	Deliverable Title	Lead beneficiary	Type	Dissemination Level	Due date
D2.2	Big data extraction module	AUTH	Other	Public	M36
D2.3	Big data fusion module	AUTH	Other	Public	M36
D3.1	Non-linear dynamic model of the farm agents	IDE	Report	Public	M31
D3.2	AI-based farmer's behavioural foundation	IDE	Report	Public	M31
D3.3	Model interaction capabilities for the ABM	IDE	Report	Public	M31