D2.3 Big Data Fusion Module

agricore

Deliverable Number	D2.3
Lead Beneficiary	AUTH
Authors	AUTH, IDE
Work package	WP2
Delivery Date	31/08/2022 (M36)
Dissemination Level	Public

www.agricore-project.eu

Document Information

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

Version History

Version	Description	Organisation	Date
0.1	ToC Proposal	AUTH	10/05/2022
0.2	ToC Approved	IDE	15/05/2022
0.3	Content inclusion (First Draft)	AUTH	19/07/2022
0.4	Revision and comments	IDE	07/08/2022
0.5	Implementation of corrections (Second Draft)	AUTH	20/08/2022
1.0	Exportation and formatting (Final version)	IDE	30/08/2022

Disclaimer

Executive Summary

AGRICORE is a research project funded by the European Commission under the RUR-04-2018 call, part of the H2020 programme, which proposes an innovative way to apply agent-based modelling to improve the capacity of policymakers to evaluate the impact of agricultural-related measurements under and outside the framework of the Common Agricultural Policy (CAP).
This deliverable presents the AGRICORE data fusion module, which allows the integration and blending of individual datasets (previously obtained by the data extraction module) to constitute enriched datasets that are used for the operation of the different AGRICORE modules.

The main data fusion operation required for the implementation of an AGRICORE use case is the one needed for producing the synthetic agents representing the agricultural and livestock holdings under study. Specifically, a mathematical artefact is needed to generate the values then assigned to the attributes that make up each of agent.

The mathematical tool chosen to perform this function is the Bayesian Network (BN). This deliverable introduces the Bayesian Network construction algorithm(s) that have been developed and or improved to be used within the AGRICORE project.

In order to test these algorithm(s), four synthetic samples of farms have been generated in three NUTS2 regions and one NUTS3 sub-region of Greece. This deliverable presents these example cases including the aggregations of specific variables, the structure of the resulting BN for each case, and the evaluation of the fit of the generated synthetic sample with respect to the real baseline sample.

Acronyms

Notation	Description
AOI	Attributes of Interest.
BN	Bayesian Network.
CAP	Common Agricultural Policy.
CPDAG	Complete Partially Directed Acyclic Graph.
DAG	Directed Acyclic Graph.
DEM	Data Extraction Module.
DFM	Data Fusion Module.
DWH	Data Warehouse.
ETL	Extraction-Transformation-Loading.
FEDHC	Forward Early Dropping Hill Climbing.
KDE	Kernel Density Estimate.
MMHC	Max-Min Hill Climbing.
MMPC	Max-Min Parents and Children.
PCA	Principal Component Analysis.
PCHC	PC Hill Climbing.
PDF	Probability Density Functions.
SP	Synthetic Population.
SPG	Synthetic Population Generator.
SS	Synthetic Sample.

List of Figures

List of Tables

Table of contents

1 Introduction

Data fusion is the process of combining multiple data sources to generate information that is more consistent, accurate, and useful than that the one provided by any of the individual data sources. It enhances the decision-making process by extracting value from data. This process improves data generation, storage, manipulation, and analysis.

The module in charge of this task in the AGRICORE project is presented as Data Fusion Module (DFM). This module aims to fuse individual datasets located and extracted by the Data Extraction Module (DEM) to obtain the mathematical artifacts (Bayesian Networks) that enable the Synthetic Population Generator (SPG) to produce the (pseudo) random values that are then assigned to each agent's attributes.

Then, data fusion refers here to the process of estimating the joint probability distribution(s) of some carefully selected Attributes of Interest (AoI). These are a collection of environmental, structural, plant and animal products, subsidies and grants, all listed later. The estimated joint distribution, represented in the form of a Bayesian Network, can be subsequently used for creating AGRICORE agents by assigning values to their skeleton of empty attributes. The objective is for these agents to mimic the statistical characteristics of the true population, represented by a sample of real farms, as close as possible. A synthetic sample (SS) is created when the number of synthetic agents generated is equal to the size of the real sample. When the number of synthetic agents generated is greater than the size of the real sample and equals the size of the real population, we call it a Synthetic Population (SP). The synthetic population must be targeted (contain only attributes of interest), microscopic (each entity is explicitly represented as an individual agent) and anonymised (it must be impossible to univocally identify a synthetic agent with any of the actual farms in the sample). The SP must match the aggregated statistical moments of the real population as close as possible, as this synthetic population will be the input for agent based models (ABMs) to simulate different policy scenarios and assess their potential impact.

The remaining of the deliverable is structured as follows: section 2 briefly summarises the exchange of information (input and output date) done between the DFM and the Data Warehouse (DWH). Section 3 introduces the concept of Bayesian Networks and explains the algorithms (MMHC, PCHC, FEDHC and MMPC) used for creating the Bayesian Networks and using them to generate values for the agents' attributes. These algorithms are applied in Section 4 to create synthetic samples of farms in four case studies for three Greek NUTS2 regions (Central Macedonia, Thessalia, Peloponnisos) and for one NUTS3 subregion (Thessaloniki). Finally, conclusions are presented in section 5.

2 Data Fusion module connection with the Data Warehouse (DWH)

As part of its functionality, DFM must communicate with the data repository (also known as DWH) to combine existing data to generate Bayesian Networks (BNs). Figure 1 depicts the inputs and outputs of the DFM at a high level. In accordance with the proposed methodology, the DWH will use distinct datasets in various formats previously loaded in the DWH. The resulting Bayesian Network, which is the output of the DFM, is also stored in the DWH.

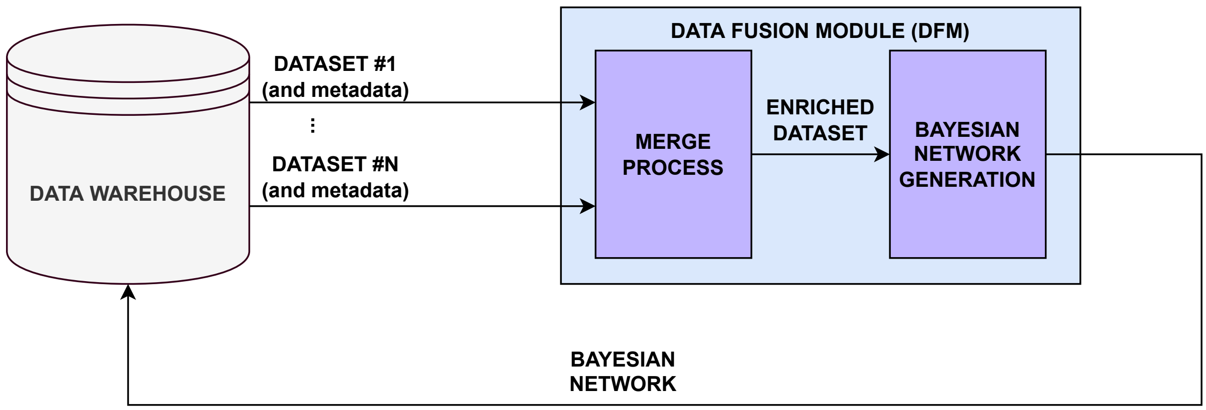

Figure 1: DFM connections with the DWH

Each necessary dataset for generating BNs should have been previously loaded into the DWH by the Data Extraction Module (DEM) through the execution of its respective ETL (Extraction-Transformation-Loading) script. The datasets typically required by AGRICORE are indexed in ARDIT along with an adequate ETL. Although DFM and DEM do not have a direct connection to one another, both are able to communicate indirectly through the DWH. From the point of view of the user, this communication takes place sequentially.

Once the datasets are stored in the DWH, they should be combined to produce an enriched dataset with the needed information for the next sub-module, which generates the Bayesian Network. The BN could be represented by one or several files containing the following information (Include reference to D6.1):

- The textual definition of the Bayesian Network.
- The order in which the values of correlated attributes should be generated, as well as a list of attributes that are completely independent and could be generated individually and in parallel.
- The Probability Distribution Functions (PDFs) necessary to generate those attributes, encoded as mathematical expressions or marginal tables.

To access the DWH, both to extract datasets and to ingest the resulting BN description files, the DFM must have the necessary permissions. The DWH itself provides native mechanisms for authentication and authorisation. The DFM is responsible for calling these services with the appropriate credentials.

3 Data fusion methods for Bayesian Network learning

Graphical models or probabilistic graphical models are probabilistic models that use a graph to visually express the conditional (in)dependencies between random attributes (V_i, $i = 1, \dots, D$). Nodes (or vertices) are used to represent the attributes V_i and edges between the nodes, for example $V_i - V_j$, indicate relationship between the attribute V_i and attribute V_j. Directed graphs are graphical models that contain arrows (arcs), instead of edges, indicating the direction of the relationship, for example $V_i \rightarrow V_j$. The parents of a node V_i are the nodes whose direction (arrows) points towards V_i. Consequently, the node V_i is termed child of those nodes. For instance, if $V_i \rightarrow V_j$, then V_i is the parent of V_j and V_j is the child of V_i. Directed acyclic graphs (DAG) are stricter in the sense that they impose no cycles on these directions, a crucial condition for the SPG task. For any path between V_i and V_j, $V_i \rightarrow V_k \rightarrow \dots \rightarrow V_j$, no path from V_j to V_i ($V_j \rightarrow \dots \rightarrow V_i$) exists.

A BN [1, 2] $B = \langle G, P \rangle$ consists of a DAG G over a collection of vertices (attributes) $\mathbf{V}$ and a joint probability distribution P. P is linked to G through the Markov condition, which states that each attribute is conditionally independent of its non-descendants given its parents. By using this condition, the joint distribution P can be factorised as the product of conditional distributions

$$P(V_1, \dots, V_D) = \prod_{i=1}^{D} P\left(V_i | Pa(V_i)\right), \tag{1}$$

where D is the total number of attributes and $\mathrm{Pa}(V_i)$ denotes the parent set of V_i in G. If G entails only conditional (in)dependencies in P and all conditional (in)dependencies in P are entailed by G, based on the Markov condition, then G, P and G are faithful to each other, and G is a perfect map of P [3].

A necessary assumption made by the BN learning algorithms is causal sufficiency; there are no latent (hidden, non observed) attributes among the observed attributes $\mathbf{V}$. The triplet (V_i, V_k, V_j) where $V_i \rightarrow V_k \leftarrow V_j$ is known as v-structure and V_k is termed collider (nodes V_1, V_3 and V_2 in Figure 2 is such an example). If there is no edge between V_i and V_j the node V_k is called unshielded collider. This translates to independence between X_i and V_j conditioning on V_k, if G and P are faithful to each other [2, 3]. Conversely, the triplet of nodes (V_i, V_k, V_j) such that $V_k \rightarrow V_i$ and $V_k \rightarrow V_j$ is called Λ-structure (nodes V_3, V_4 and V_5 in Figure 2 is such an example). The Λ-structure implies that V_i and V_j are conditionally independent given V_k.

Typically, multiple BNs encode the same set of conditional independences[1]. Such BNs are called Markov equivalent, and the set of all Markov equivalent BNs forms the Markov equivalence class. This class can be represented by a complete partially directed acyclic graph (CPDAG), which in addition to directed edges also contains undirected edges. Undirected edges may be oriented either way in some BNs in the Markov equivalence class (although not all combinations are possible), while directed and missing edges are shared among all equivalent networks.

[1] Two DAGs are called Markov equivalent if and only if they have the same skeletons and the same v-structures [4].

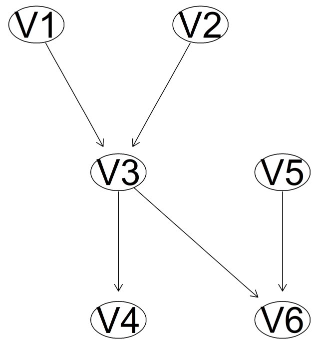

Figure 2: An example of a DAG. Nodes V_1 and V_2 are the parents of V_3, whose children are nodes V_4 and V_5. The common parents of a node are also called spouses. $V2$ is the spouse of $V1$ (and vice versa, $V1$ is the spouse of $V2$) and $V6$ is the spouse of $V3$.

As an example of (1) we will write the joint distribution of six attributes $V1 - V6$ based on the BN of Figure 2 as

$$P(V1, \ldots, V6) = P(V6|V3, V5) \times P(V4|V3) \times P(V3|V1, V2) \times P(V1) \times P(V2) \times P(V5),$$

where, in the BN terminology, the expression $P(V_i|V_k, V_j)$ indicates that the conditioning attributes V_k and V_j are the parents of V_i. In order to generate random values from a BN, the attributes must be topolocigally ordered first, as in Figure 2, where the BN is presented in tree-like structure. For instance, using the BN Figure 2 we can generate values first from the attributes with no parents, $V1, V2$ and $V5$. Those generated values of $V1$ and $V2$ are used, in combination, to generate values from $V3$. The values of $V6$ are generated using the generated values of $V3$ and $V5$, whereas the values of $V4$ are generated using the generated values of $V3$. The natural question of interest now is how to construct the BN of Figure 2 using observational data and hence factorise the joint distribution of the attributes.

3.1 The MMHC BN learning algorithm

BN learning algorithms are typically constraint-based, score-based or hybrid. Constraint-based learning algorithms, such as PC [5] and FCI [2] employ conditional independence (CI) tests to discover the structure of the network (skeleton), and then orient the edges by repetitively applying orientation rules. On the contrary, score-based methods [6, 7, 8], assign a score on the whole network and perform a search in the space of BNs to identify a high-scoring network. Hybrid algorithms, such as MMHC [9], PCHC [10] and FEDHC [11], combine both aforementioned methods; they first perform CI tests to discover the skeleton of the BN and then employ a scoring method to direct the edges in the space of BNs.

We particularly suggest the class of hybrid BN learning algorithms, and specifically MMHC [9], which first identifies the statistically significant associations between the attributes and then applies a scoring method to orient those relationships. FEDHC is a recently introduced algorithm that is designed to work mainly with large sample sizes and is available in the *R* package *bnlearn* [12]. At first, the MMPC attribute selection algorithm is applied to each attribute and secondly the Hill-Climbing (HC) scoring phase orients the directions of the statistically significant relationships. The algorithm is briefly discussed over the next four sub-sections followed by the importance of prior knowledge (e.g. theory), validation techniques and the advantage/disadvantages of BNs in general.

3.1.1 The MMPC attribute selection algorithm

In the classical forward selection algorithm all available predictor attributes are constantly examined and their statistical significance is assessed at each step. Assuming that out of $10,000$ predictor attributes only 10 are selected. This implies that almost $10,000 \times 10$ regression models must be fitted and the same amount of statistical tests must be executed. The computational cost is tremendous rendering this computationally expensive algorithm impractical and hence prohibitive. Secondly, this approach selects a relatively high number of non-significant attributes.

MMHC [13] is a hybrid method whose skeleton identification phase, also known as Max-Min Parents and Children (MMPC) algorithm, is presented in Algorithm 1. Given a target attribute (attribute of interest, V_i), a search for its statistically significantly associated attributes $\mathbf{V}_s$ is performed via statistical tests. The associations are stored and the attribute with the highest association (V_j) is chosen and an edge is added between V_i and V_j. In the second step, all CI tests between the target attribute and the other attribute, conditional upon all possible subsets of the previously selected attribute, are performed ($V_i \perp\!\!\!\perp V_m | V_j, \quad m \neq i, j$) and the non statistically significant attributes are neglected. The previously stored associations are updated, that is, for each attribute the minimum association between the old and the new variables is stored and the attribute with the highest association is selected[2]. In subsequent steps, while the set of the selected variables increases, the conditioning set does not, as its cardinality is at most equal to l. At the end, a backward selection in the using Max-Min heuristic is applied attempting to remove wrongly selected attributes.

The MMPC algorithm, acts as a speed-up modification of the traditional forward selection algorithm coupled with a variant of the backward selection algorithm [14] while retaining the false discovery rate (proportion of non significant attributes wrongly selected) at low levels [15]. At each step, non significant attributes are excluded from future searches and instead of conditioning on all selected attributes, thus reducing the computational cost. Secondly, the conditional independence (CI) test for the next attribute, conditions upon all possible subsets, up to a pre-specified cardinality l, of the already selected attributes. This property makes MMPC suitable for small sample sized datasets with numerous attributes, since a CI test involving many parameters has low power with small samples.

Algorithm 1 (H) *1: **Input**: Data set on a set of D variables $\mathbf{V}$.*
*2: **Repeat** for all variables $i = 1, \ldots, n$*
3: Let $l = 0$ and $\mathbf{S} = \emptyset$.
4: Select a variable V_i and keep all variables $V_j, j \neq i$ for which $V_i \perp\!\!\!\perp V_j$ holds true.
5: Chose the variable V_j with the highest association among these variables,
6: add and edge $V_i - V_j$ and add V_j to $\mathbf{S}$.

[2] This is the Max-Min heuristic.

Repeat
$l = l + 1$
If $\left(V_i \perp\!\!\!\perp V_j | \mathbf{S}_{(l)}\right)$ delete edge $V_i - V_j, j \neq i$, where $\mathbf{S}_{(l)}$ denotes all possible
subsets of the selected variables in $\mathbf{S}$, with cardinality less than or equal to l.
Chose the variable V_j with the highest minimum association among them,
add and edge $V_i - V_j$ and add V_j to $\mathbf{S}$.
Until *l has reached the pre-specified maximum value.*
***Return** G.*

3.1.2 Statistical tests of independence

The MMPC algorithm iteratively performs statistical tests to decide as for the significance of the relationships, so let us first briefly describe the concept of independence. Let X and Y be two random attributes, and $\mathbf{Z}$ be a (possibly empty) set of random attributes. X and Y are conditionally independent given $\mathbf{Z}$, if $P(X, Y|\mathbf{Z}) = P(X|\mathbf{Z}) \cdot P(Y|\mathbf{Z})$ holds for all values of X, Y and $\mathbf{Z}$. Equivalently, CI of X and Y given $\mathbf{Z}$ implies $P(X|Y, \mathbf{Z}) = P(X|\mathbf{Z})$ and $P(Y|X, \mathbf{Z}) = P(Y|\mathbf{Z})$. Such statements can be tested using CI tests.

A frequently employed independence test for two continuous attributes X and Y, conditional on a set of attributes $\mathbf{Z}$ is the partial correlation test [16] that assumes linear relationships among the attributes. The test statistic for deciding whether the partial Pearson correlation coefficient is zero is given by

$$T_p = \frac{1}{2}\left|\log\frac{1 + r_{X,Y|\mathbf{Z}}}{1 - r_{X,Y|\mathbf{Z}}}\right|\sqrt{n - |\mathbf{Z}| - 3}, \tag{2}$$

where n is the sample size, $|\mathbf{Z}|$ denotes the number of conditioning attributes and $r_{X,Y|\mathbf{z}}$ is the partial Pearson correlation[3] of X and Y conditioning on $\mathbf{Z}$. When $\mathbf{Z}$ is empty ($|\mathbf{Z}| = 0$), the partial correlation reduces to the usual Pearson correlation coefficient.

The p-value of the test is used to decide on the significance of the CI between X and Y. It is defined as $2\left(1 - F(Tp, df)\right)$, where $F(.)$ denotes cumulative distribution of the t distribution with degrees of freedom $df = n - |\mathbf{Z}| - \mathbf{3}$. The p-value lies within $(0, 1)$ with smaller values indicating higher strength of (un)conditional association between X and Y. If it is less than 0.05 the two attributes are claimed to be statistically significantly (conditionally) associated. In order to avoid numerical overflow problems, that could yield erroneous results, the logarithm of the p-value is computed instead, and subsequently the threshold of significance becomes $\log(0.05) = -2.995732$.

3.1.3 Skeleton identification phase of the MMHC algorithm

During the skeleton identification phase of MMHC, the MMPC algorithm is applied to each attribute (call it target attribute, V_i), performing the steps described below.

1. **Input**: Data set on a set of D attributes $\mathbf{V}$.

[3]The partial correlation is efficiently computed using the correlation matrix of X, Y and $\mathbf{Z}$ [16].

2. Let the adjacency matrix G be full of zeros.
3. Perform the MMPC algorithm for all attributes V_i, $i = 1, \ldots, D$, excluding the backward phase, and return $\mathbf{S}_i$, the set of attributes ($V_j,\ j \neq i$) related to V_i.
4. Set $G_{ij} = 1$ for all $j \in \mathbf{S}_i$.
5. **If** $G_{ij} \neq G_{ji}$ set $G_{ij} = G_{ji} = 0$.
6. **Output**: The (square) adjacency matrix G that contains 0s and 1s denoting the edges (statistically significant relationships) between pairs of attributes.

The final output is a matrix containing the edges (undirected relationships) discovered between each attribute, in an asymmetric way. The detected edges between any pair of attributes will remain only if they were identified by both attributes. If for example, V_j was found to associated with V_i ($G_{ji} = 1$), but V_i was not found to be associated with V_j ($G_{ij} = 0$), then no edge between V_i and V_j will be added, hence $G_{ij} = G_{ji} = 0$. The final output is the so called adjacency matrix G which contains 0s and 1s. If the element G_{ij} (and G_{ji}) equals zero this indicates that attributes V_i and V_j are not related, whereas if $G_{ij} = G_{ji} = 1$ indicates that attributes V_i and V_j are related.

3.1.4 Hill Climbing phase of the MMHC algortihm

During the second phase of MMHC a search for the optimal DAG is performed, where edges turn to arrows or are deleted towards maximisation of a score metric. This scoring phase performs a greedy HC search[4] in the space of BNs, commencing with an empty graph [9]. The edge deletion or direction reversal that leads to the largest increase in score, in the space of BNs[5], is applied and the search continues in a similar fashion recursively. The fundamental difference from standard greedy search is that the search is constrained to the orientation of the edges discovered by the skeleton identification phase[6].

The Bayesian Information Criterion (BIC) [17] is a frequent score used for continuous data, while other options include the multivariate normal log-likelihood, the Akaike Information Criterion (AIC) and the Bayesian Gaussian equivalent[7] [18] score. The Bayesian Dirichlet equivalent (BDE) [19], the BDe uniform score (BDeu) [7], the multinomial log-likelihood score [20] and the BIC score [17] are four options for scoring with discrete data. In this work we employed the BIC score $BIC(G, \Theta \mid \mathbf{V}) = \sum_{i=1}^{n} \log P(V_i \mid Pa(V_i), \Theta_{V_i}) - \frac{\log(n)}{2}\ |\Theta_{V_i}|$.

3.2 Prior knowledge required to build BNs

MMHC, as all BN learning algorithms, is agnostic of the true underlying relationships among the input data. It is customary though for practitioners and researchers to have prior knowledge of

[4]Tabu search is such an iterative local searching procedure adopted by [9] for this purpose.

[5]This implies that every time an edge removal, or arrow direction is implemented, a check for cycles is performed. If cycles are created, the operation is canceled regardless if it increases the score.

[6]For more information see [9].

[7]The term ”*equivalent*” refers to their attractive property of giving the same score to equivalent structures (Markov equivalent BNs) i.e., structures that are statistically indistinguishable [9].

the necessary directions (forbidden or not) of some of the relationships among the attributes. For instance, attributes such as manager's gender or age cannot be caused by any economic or demographic attributes. Economic theory (or theory from any other field) can further assist in improving the quality of the fitted BN by imposing or forbidding directions among some attributes. This prior information can be inserted into the scoring phase of MMHC leading to less errors and more realistic BNs.

Let us give an example of the importance and necessity of prior knowledge tailored in the needs of the current project. We know that the crop production cannot influence the cultivated area, or the irrigated area and the milk production cannot affect the livestock. The set of all forbidden directed relationships forms the prior knowledge that must be incorporated into the BN learning algorithm affecting only the HC phase. Non incorporation of this information would yield an unrealistic BN and as a result, an unrealistic joint distribution that fails to describe the true underlying joint distribution.

The statistical methods used to analyse the different variables included in the BN and to produce the prior knowledge needed to create the BN are presented in deliverable D2.2.

3.3 BN learning validation techniques

The strength of the significant relationships detected by the BN is defined as the decrease in the BIC score when a specific arrow (or arc or directed relationship) is deleted while fixing the structure of the BN stable. The higher the reduction in the score the higher the indications that this directed relationship is important or strong. This allows to order the relationships based on their strength.

Bootstrap can be implemented as a second measure (apart from the strengths) of the validity of the discovered (directed) relationships among the attributes. A set of of observations is sampled with replacement from the original sample (observed farms) and the BN was learned using MMHC. This process is repeated 1,000 times storing the discovered arcs of each repetition. The measure of interest is the proportion of times the observed directed relationships are discovered in the bootstrap samples. This acts as a metric of the confidence or the stability in the relationship of each discovered (directed) relationship in the original sample.

3.4 Generation of synthetic samples of farms

Generation of random values from BNs with continuous data leads to normally distributed values, which are far from reality as in our case where the distributions of most attributes are highly skewed to the right and most of them contain zero values. A more fine tuned method is required to simulate values whose distribution is close to the observed data distribution. To this end, we employed a complex generation scheme based on non-parametric regression relying on the BN structure learned using the attributes of the observed farms. The order of generation is sequential as mandated by the BN. That is, the values of each attribute are generated conditional upon its parent attribute(s).

For attributes with no parents, we computed the kernel density estimate (KDE) of the distribution of the non-zero values and generated non-zero values from this KDE, whereas zero values remained the same. For attributes with at least one parent, we utilised the the k-NN regression algorithm. The k-NN algorithm is a naive kernel regression that takes into account only the values of the k closest neighbours to a specific value.

Whenever values for an attribute are generated, we transform the data such that their mean is equal to the mean of the observed attribute values. However some post generation refinement was deemed necessary. Specifically for the crop production, when the synthetic cultivated land of a crop is zero, the corresponding (synthetic) irrigated land and crop production were set to zero. If the irrigated area of some crops being higher than the corresponding cultivated land, the irrigated area was set equal to the cultivated area. A similar refinement process took place for the animal products. For instance, the values of the animal products for the synthetic farms with no livestock were zeroed.

3.5 Evaluation of the generated synthetic samples

Researchers ordinarily assess the fit of the univariate distributions, that is, the distribution of each attribute. We employed a battery of both parametric and non-parametric testing procedure in order to evaluate the synthetically generated sample of farms. We applied a KDE hypothesis test of equality of two distributions (see Appendix) was applied to assess the equality of the distributions of each attribute, between the observed and the synthetic farms. We further applied a second non-parametric that is energy distance based [21]. The same energy distance test was applied to test the equality of the joint distributions, of the observed and of the synthetic farms. This inspects the equality of the distributions at the multi-attribute level, taking into account all attributes at once.

Secondly, the γ-OMP [22] and FBED [23] attribute selection algorithms were engaged in conjunction, to identify which attributes are responsible for separating between the two samples and how accurate their separation can be. Ideally, the two samples, the observed and the synthetic farms should be non-separable.

Thirdly we applied principal component analysis (PCA) in order to project the data into lower dimensions so as to visually inspect the two samples, the observed versus the synthetic farms.

4 Case studies

4.1 A synthetic sample for Central Macedonia (NUTS-2 level)

The greek FADN sample for Central Macedonia contains 1,017 farms, the largest sample available at NUTS-2 level. Due to sparsity (excessive amounts of zeros) in many attributes, aggregation of attributes, based on their proximity, resulting in 98 attributes, was deemed mandatory for the the BN learning and the SPG task subsequently[8]. Those 98 attributes, grouped according to the clusters presented previously, can be found in the Appendix.

- **Crop production**. Table A.1 shows the crop production of central Macedonia, where some crops have been aggregated due to sparsity (excessive amount of zeros), yielding 14 crops.
- **Animal products**. Table A.2 shows the condensed animal production, the weighted livestock, values of sold and slaughtered animals, values of animals left rearing-breeding and the total milk production.
- **Farm income, subsidies and grants**. Table A.3 contains information on the components that formulate the attribute termed "other farm income", the aggregation of the following characteristics: value of sold animals, value of sales of wool, eggs, honey and manure, other income from livestock (e.g. contract rearing), income from land (e.g. leasing), food processing (e.g. cow's milk), contractual work and income from other sources (e.g. tourism, production of renewable energy). Table B.9 shows the subsidies and grants grouped in 4 clusters, decoupled payments, crops and animals, exceptional support and rural development and subsidies on cost. Note that despite the subsidies on cost being listed in the FADN guide manual, this attribute was not applicable in the Greek use case.
- **Variable inputs cost**. Table A.4 contains 11 attributes (2 attributes were merged) representing the variable inputs cost.

As previously mentioned, BN learning algorithms are agnostic of the input data and require some prior knowledge to facilitate the production of more realistic results. A set of constraints must be imposed among these 98 attributes. These refer to rationally forbidden directions between the pairwise relationships (The attribute codings can be found in the Appendix).

1. Within crop production:
   - The production (CM-Xi.3) does not affect the cultivated area (CM-Xi.1) nor the irrigated area (CM-Xi.2), for all 14 products, i=1,...,14.
   - The irrigated area (CM-Xi.2) does not affect the cultivated area (CM-Xi.1), for all 14 products.
2. Within animal production:
   - The total milk production (CM-Z1.1) does not affect the weighted livestock (CM-Y1.1), the value of sold animals (CM Y1.3) and the value of slaughtered animals (CM-Y1.5).

[8] For instance, the 20 crops were merged into 14 crops.

- The value of animals for breeding (CM-Y1.7) does not affect the weighted livestock (CM-Y1.1), the value of sold animals (CM-Y1.3) and the value of slaughtered animals (CM-Y1.5).
- The value of slaughtered animals (CM-Y1.5) does not affect the weighted livestock (CM-Y1.1) and the value of sold animals (CM-Y1.3)
- The value of sold animals (CM-Y1.3) does not affect the weighted livestock (CM-Y1.1).

3. Other restrictions:
   - No attribute affects the soil, spatial and climatic data (Gi, i=1,...,12).
   - No attribute affects the manger's gender (L1.1), age (L1.2) and training (L1.3).
   - Xi.3, Yi.3 and M1 do not affect the farm labour attributes (L).

The MMHC BN learning algorithm discovered 121 statistically significantly associated relationships. These are presented in Table 1 along with their directions and their strength. For instance, the relationship between C5 and A2 is directed from C5 to A2 and hence in the BN terminology this is denoted by C5→A2. The same is true for all relationships. The results of bootstrap validation also appear in Table 1. The 83 out of the 121 (68.6%) identified directed relationships in the observed farms were observed more than 50% of the times in the bootstrap samples. This, rather low, number does not come by surprise as the data contain many attributes with high proportions of zero values. When sampling with replacement, the percentage of unique values in the bootstrap sample is on average equal to $1-((1-1/n))^n$, which in the current situation is equal to 63%. Hence the bootstrap sample of 1,017 farms contains around 63% unique farms. Attributes having more than 63% zeros may contribute only zeros to the bootstrap sample and hence no relationship can be discovered, even if there is one.

Table 1: The 121 statistically significant associations clustered according to the tables (see Appendix). The computed strengths were normalised with the strongest strength playing the role of the basis. The column "boot" refers to the proportion of times the observed directed relationships were discovered in the bootstrap samples.

from	to	strength	boot	from	to	strength	boot	from	to	strength	boot
C4	X12.1	0.0094	0.6100	L4.1	X13.3	0.0007	0.4840	X12.2	X12.3	0.1322	1.0000
C4	Y1.1	0.0069	0.6000	X1.1	X1.3	0.2231	1.0000	X13.1	X13.2	0.0435	1.0000
C5	S1	0.0255	0.2570	X2.1	X2.2	0.0044	0.6860	X13.1	X13.3	0.0636	1.0000
C5	Y1.1	0.0015	0.1030	X2.1	X2.3	0.1940	1.0000	X13.1	X14.3	0.0022	0.6010
C6	S1	0.1152	0.5270	X2.2	X2.3	0.0046	0.9570	X13.1	X2.3	0.0007	0.2020
C6	X11.1	0.0018	0.1890	X3.1	X3.2	0.6936	1.0000	X13.2	X1.2	0.0003	0.5370
C6	X2.1	0.0327	0.5300	X3.1	X3.3	0.3356	0.6570	X13.3	X9.3	0.0003	0.3630
C6	X8.1	0.0083	0.2630	X4.1	X4.2	0.0009	0.6720	X14.1	X14.2	0.1088	1.0000
G1	G4	0.0664	0.9990	X4.1	X4.3	0.2652	1.0000	X14.1	X14.3	0.0408	1.0000
G1	G7	0.0130	1.0000	X4.2	X4.3	0.0017	0.7600	X14.2	C4	0.0020	0.5330
G2	G3	0.0008	0.3530	X5.1	X5.2	0.7173	0.8600	X14.2	X14.3	0.0019	0.7870
G2	G7	0.0025	0.9480	X5.1	X5.3	0.3918	0.8550	Y1.3	V4	0.1117	0.3030
G2	X10.3	0.0019	0.6550	X5.3	V8	0.0238	0.1530	Y1.3	Y1.5	0.5004	1.0000
G3	G5	0.0026	0.8290	X6.1	C6	0.0074	0.3800	Y1.7	X11.1	0.0005	0.4650
G4	G3	0.0465	0.9460	X6.1	X6.2	0.0003	0.6750	Y1.7	Z1.1	0.0175	0.8420

continued....

from	to	strength	boot	from	to	strength	boot	from	to	strength	boot
G4	G5	0.0999	1.0000	X6.1	X6.3	0.2536	1.0000	Z1.1	X11.1	0.0015	0.3740
G4	X6.3	0.0002	0.3920	X6.2	X6.3	0.0323	0.9910	M1	V10	0.0160	0.7530
G7	C4	0.0059	0.9940	X6.3	V10	0.0006	0.2700	S1	A4	0.3061	0.9920
G7	X9.1	0.0014	0.3950	X7.1	S2	0.1559	0.8090	S1	S2	0.0738	0.8430
G7	Y1.7	0.0025	0.5130	X7.1	X7.2	0.3066	1.0000	V1	Y1.1	0.0050	0.3330
G8	G10	0.2725	1.0000	X7.1	X7.3	0.0205	1.0000	V3	C6	0.0222	0.4830
G8	G9	0.0196	0.9800	X7.2	V9	0.0070	0.5510	V3	V5	0.0110	0.2770
G9	C4	0.0122	0.4990	X7.2	X7.3	0.0053	0.9700	V3	V6	0.0150	0.0810
G9	G10	0.0730	0.9030	X8.1	X8.2	0.0289	0.8900	V3	V9	0.0167	0.4660
G10	G12	0.0284	0.7340	X8.1	X8.3	0.1898	1.0000	V5	C6	0.0094	0.2750
G11	G12	0.1543	0.6010	X8.2	X8.3	0.0238	1.0000	V5	V10	0.0133	0.6600
G11	X12.1	0.0003	0.2630	X9.1	X9.2	0.0435	1.0000	V6	Q2	0.1007	0.8510
Q1	V11	0.0960	0.5630	X9.2	V1	0.0003	0.2450	V6	Q3	0.0713	0.8170
Q1	V2	0.0349	0.7490	X9.2	V5	0.0041	0.4880	V6	V5	0.0172	0.7990
Q1	V6	0.0825	0.8400	X9.2	X9.3	0.0328	1.0000	V6	V7	0.0564	0.9680
Q3	V7	0.0075	0.3840	X9.3	S3	0.0038	0.5660	V7	V8	0.0350	0.8860
L1.1	L1.5	0.0120	0.9880	X10.1	X10.2	0.2315	1.0000	V8	V2	0.0105	0.2040
L1.2	L1.3	0.0027	0.9980	X10.1	X10.3	0.0180	0.9770	V8	V9	0.0192	0.1710
L1.3	L1.5	0.0000	0.4030	X10.2	X10.3	0.0002	0.6590	V9	C4	0.0035	0.6130
L1.5	L1.4	0.2072	0.9810	X11.1	X11.2	0.0386	0.9770	V9	S3	0.0043	0.3460
L2.1	L2.2	1.0000	0.9570	X11.1	X11.3	0.0081	1.0000	V9	X2.2	0.0017	0.3130
L3.1	L4.1	0.0117	0.4130	X11.2	V9	0.0004	0.3390	V10	L4.1	0.0015	0.3000
L3.1	L5	0.0137	0.6120	X11.2	X11.3	0.0285	1.0000	V10	V1	0.0070	0.5410
L3.1	X8.3	0.0005	0.3330	X11.2	X4.2	0.0026	0.7720	V11	V3	0.0867	0.6280
L3.1	X9.1	0.0034	0.3950	X11.2	X9.3	0.0001	0.3510				
L4.1	V1	0.1656	0.7500	X12.1	X12.2	0.2145	1.0000				

4.1.1 Evaluation of the synthetic sample generation in central Macedonia

Using the 98 attributes and the estimated BN structure we generated a sample of 1,017 synthetic farms whose characteristics match to a high a degree the characteristics of the observed farms. Application of the γ-OMP [22] and FBED [23] attribute selection algorithms indicated that the two samples (observed and synthetic farms) can be separated with accuracy 58.6%. These two algorithm were ordinarily identifying the attribute showing the years of education (*train*) was responsible for this level of separation. When this specific attributes was removed, γ-OMP could not separate the farms (accuracy = 50%).

The KDE hypothesis test of equality of two distributions applied to assess the equality of the distributions of each of the 95 attributes[9] between the observed and the synthetic farms showed that the majority of the associated p-values (74/95, 78%) were more than 0.05, indicating that the distributions of the synthetic farms are in close agreement with those of the observed farms. Figure 4 visualizes the distributions of the attributes measuring the crop production. These are the kernel density estimates of some attributes of the observed and of the synthetic farms. It can be observed that the densities of the attributes of observed and of the synthetic farms are in close agreement. The energy test is more sensitive and detected 63 out of 95 (66.3%) distributions of attributes as

[9] Three attributes had excessive amounts of zeros and the test was not applicable.

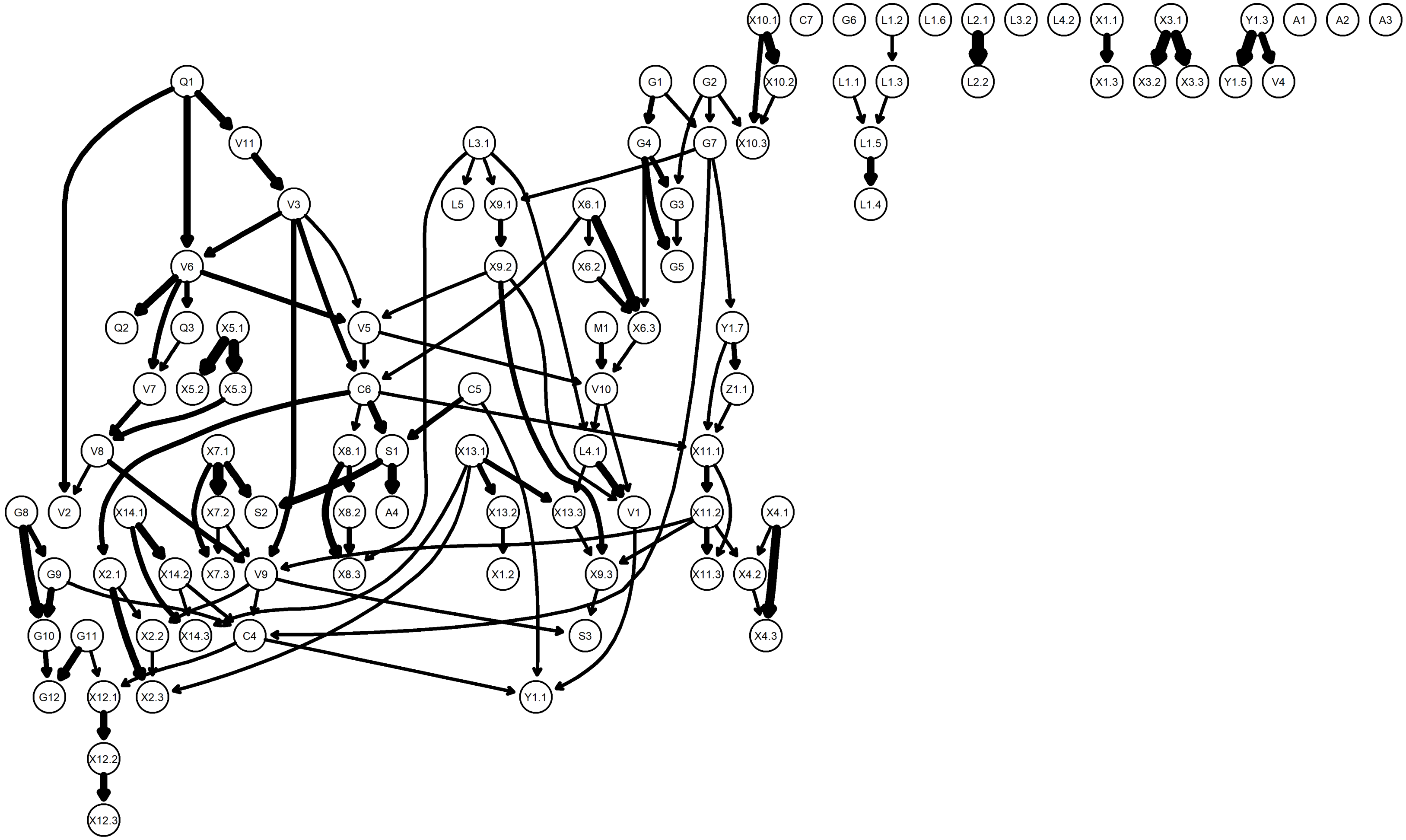

Figure 3: The BN structure of Central Macedonia. The strength of the directed relationships is denoted by the thickness of the arrows.

being statistically equal.

When applied to the joint distributions of the the observed and the synthetic farms, the energy test produced a p-value equal to 0.967 indicating a high similarity between the two joint distributions. However, the attributes are measured in different scales and different units of measurements. For this reason, the two groups (observed and synthetic farms) were standardised to have zero means and unity variances and the energy test was applied to the transformed data. The produced p-value was equal to 0.135, which corroborates the results of the sample generation process.
Figure 8 shows the data projected onto the first 5 principal components produced by PCA. It is evident that the synthetic farms cannot be distinguished from the observed farms.

4.2 A synthetic sample for Thessaloniki (Nuts-3 level) sub-region of central Macedonia

Thessaloniki region was included for (further) validation purposes, since it is a sub-region located within central Macedonia. In the greek FADN, Thessaloniki region contains the second largest collection of farms at the NUTS-3 level, equal to 325 farms. Chalkidiki region lies at the south of Thessaloniki and since the former contains only 20 farms we decided to include them in Thessaloniki region.

Specifically for the crop production, further aggregation was performed, again due to excessive number of zeros, merging two crops into one (see Table A.1 in the Appendix), thus leaving us with 13 attributes describing the crop production. Additionally, the same set of constraints imposed on the BN learning for the case of central Macedonia was also imposed among the 95 attributes of Thessaloniki.

Table 2 presents the strengths of the statistically significantly associated relationships discovered via the MMHC BN learning algorithm. Evidently, the BN identified 83 directed relationships in Thessaloniki, which are equal to 68.6% of the relationships identified in central Macedonia. This does not come by surprise for two reasons: a) the sample size of the farms in Thessaloniki is 1/3 of the number of farms in central Macedonia, b) the proportion of zero values is higher in some attributes and c) the merge of two crops in one leading to 13 crops. These three reasons combined, render many relationships undetectable resulting into only 44 common identified directed relationships between the two regions. The table also contains the percentage of times the detected relationships appeared in the bootstrap samples. The 72.3% of the detected relationships (60 out of 83) appeared more than 50% in the bootstrap samples.

Table 2: The 83 statistically significant associations clustered (see Appendix for the tables that group the attributes). We remind the reader that the Xis refer to TH. The computed strengths were normalised with the strongest strength playing the role of the basis. Numbers less than 1 show the strength of the relationship relevant to the strongest relationship, that between L2.1 and L2.2.

from	to	strength	boot	from	to	strength	boot
C4	Y1.1	0.0043	0.2500	X5.3	V8	0.2018	0.7700
C6	L3.1	0.0054	0.1400	X6.1	X6.3	0.2872	1.0000

G1	G4	0.0562	0.9400	X6.2	X3.3	0.0216	0.7000
G1	G5	0.0004	0.4300	X6.2	X6.3	0.0160	0.5050
G2	X10.3	0.0024	0.6000	X7.1	X7.3	0.0821	1.0000
G4	C4	0.0100	0.7900	X7.2	X7.3	0.0573	1.0000
G4	G5	0.0880	0.9950	X8.1	X8.3	0.1787	1.0000
G4	G7	0.0590	0.9700	X9.1	X12.1	0.0009	0.3700
G5	G7	0.0125	0.9350	X9.1	X9.2	0.0106	0.6900
G7	G6	0.0558	0.9650	X9.2	X9.3	0.0698	1.0000
G8	G10	0.6136	0.9950	X9.3	S3	0.0114	0.6000
G8	G11	0.0342	0.3900	X10.1	C4	0.0011	0.1950
G9	X7.3	0.0076	0.8950	X10.1	X10.2	0.2545	1.0000
G11	G12	0.1737	0.9850	X10.1	X10.3	0.0020	0.6200
G11	L1.6	0.0024	0.4300	X10.2	X10.3	0.0029	0.7000
G12	G3	0.0183	0.9450	X11.1	X11.3	0.0486	1.0000
G12	G9	0.0715	0.7150	X11.1	Y1.7	0.0032	0.1400
Q1	V6	0.1218	0.9250	X11.2	X11.3	0.0510	0.9850
Q1	V7	0.1267	0.4800	X12.1	X12.2	0.0618	0.9900
Q2	V5	0.0026	0.2350	X12.2	X12.3	0.0379	0.9700
Q3	V6	0.0113	0.1950	X12.3	Y1.1	0.0024	0.3150
L1.5	L1.4	0.2049	0.4900	X13.1	X13.2	0.1479	1.0000
L1.6	L1.4	0.0024	0.3800	X13.1	X13.3	0.0592	0.9850
L2.1	L2.2	1.0000	0.6900	X13.2	X13.3	0.0052	0.7700
L3.1	V10	0.0027	0.2700	Y1.1	V1	0.0107	0.3150
L4.1	V1	0.1891	0.9350	Y1.3	Y1.5	0.4643	1.0000
L4.1	V5	0.0065	0.1400	Y1.7	Z1.1	0.0162	0.5400
L4.1	Y1.1	0.0159	0.3650	Z1.1	L4.1	0.0134	0.2250
L5	L3.1	0.0309	0.3750	S1	A4	0.3864	1.0000
X1.1	X1.3	0.1901	1.0000	S2	L5	0.0024	0.1000
X2.1	X2.3	0.1568	1.0000	S2	X7.1	0.0411	0.2200
X2.2	X2.3	0.0091	0.8900	V3	C6	0.0663	0.3650
X2.2	X7.2	0.0053	0.1100	V3	V11	0.0892	0.5700
X3.1	X3.2	0.5958	0.9800	V5	V9	0.0035	0.2050
X3.1	X3.3	0.3353	0.7150	V6	Q2	0.0831	0.6150
X4.1	X2.3	0.0038	0.3200	V6	V5	0.0411	0.9650
X4.1	X4.3	0.2653	1.0000	V7	Q3	0.0640	0.6550
X4.2	X11.2	0.0077	0.3900	V7	S2	0.0857	0.2450
X4.2	X4.3	0.0011	0.5350	V9	C4	0.0043	0.7400
X4.3	X2.3	0.0077	0.4750	V9	V10	0.0069	0.4600
X5.1	X5.2	0.6599	0.8800	A4	V3	0.0917	0.4550
X5.1	X5.3	0.3612	0.7650				

4.2.1 Evaluation of the synthetic sample generation in Thessaloniki

Application of the γ-OMP [22] and FBED [23] attribute selection algorithms indicated that the two samples (observed and synthetic farms) can be separated with accuracy 75%. These two algorithms were ordinarily identifying the irrigated area of cotton and the values of animals for rearing or breeding as the two attributes responsible for this level of separation. When these two attributes were removed, γ-OMP could not separate the farms (accuracy = 50%). The mean of the irrigated area of cotton in the synthetic farms is less than the mean in the observed farms. Secondly, the values of animals for rearing or breeding contain 317 0s in the observed farms, but 325 0s in the synthetic farms. Especially for the second attribute, this excessive proportion of zeros has a great

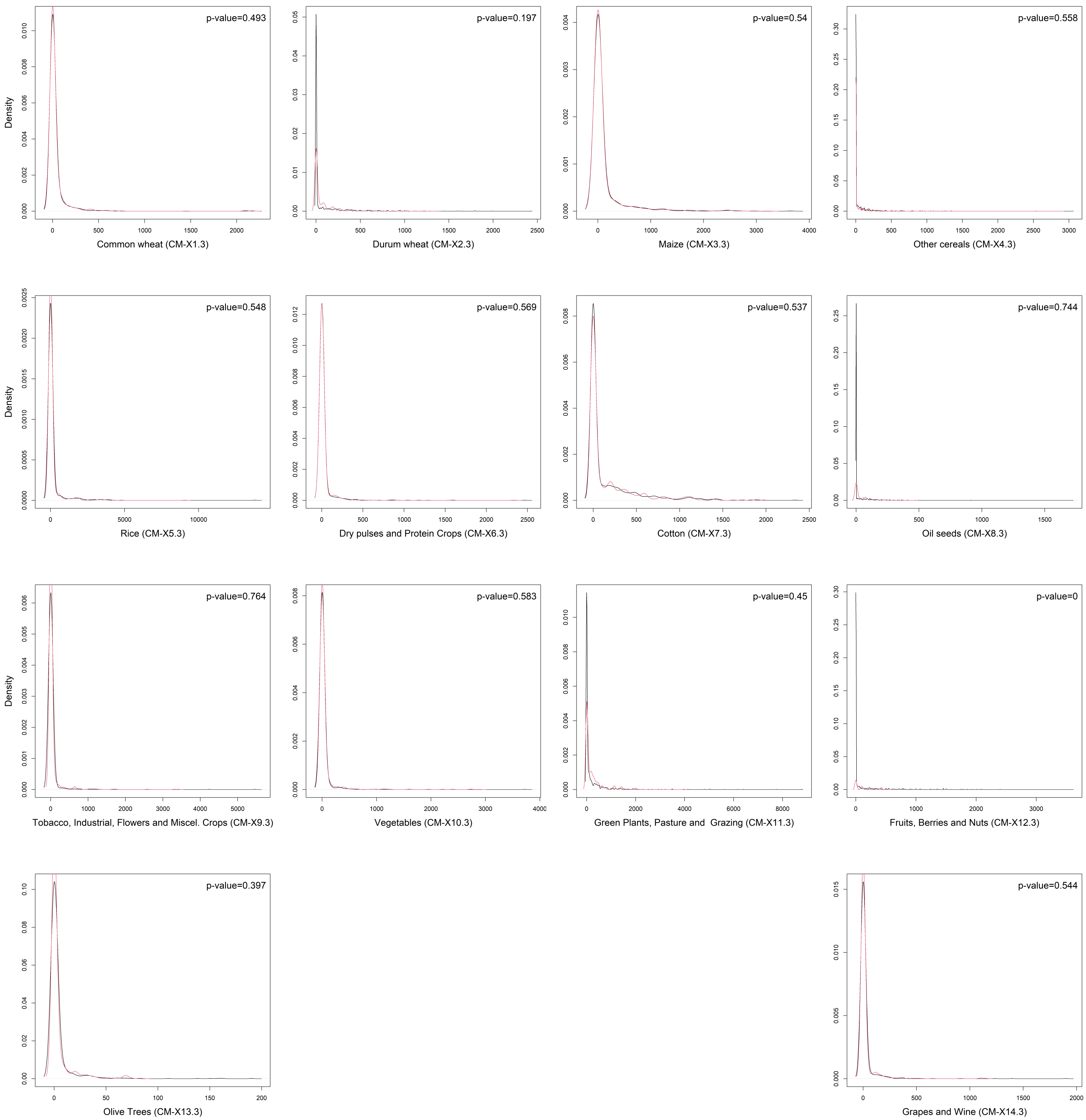

Figure 4: Distributions of the crop production (attributes CM-X1.3 - CM-X14.3). The black line refers to the observed farms, while the red line refers to the synthetic farms. The KDE test p-value appears on the top-right.

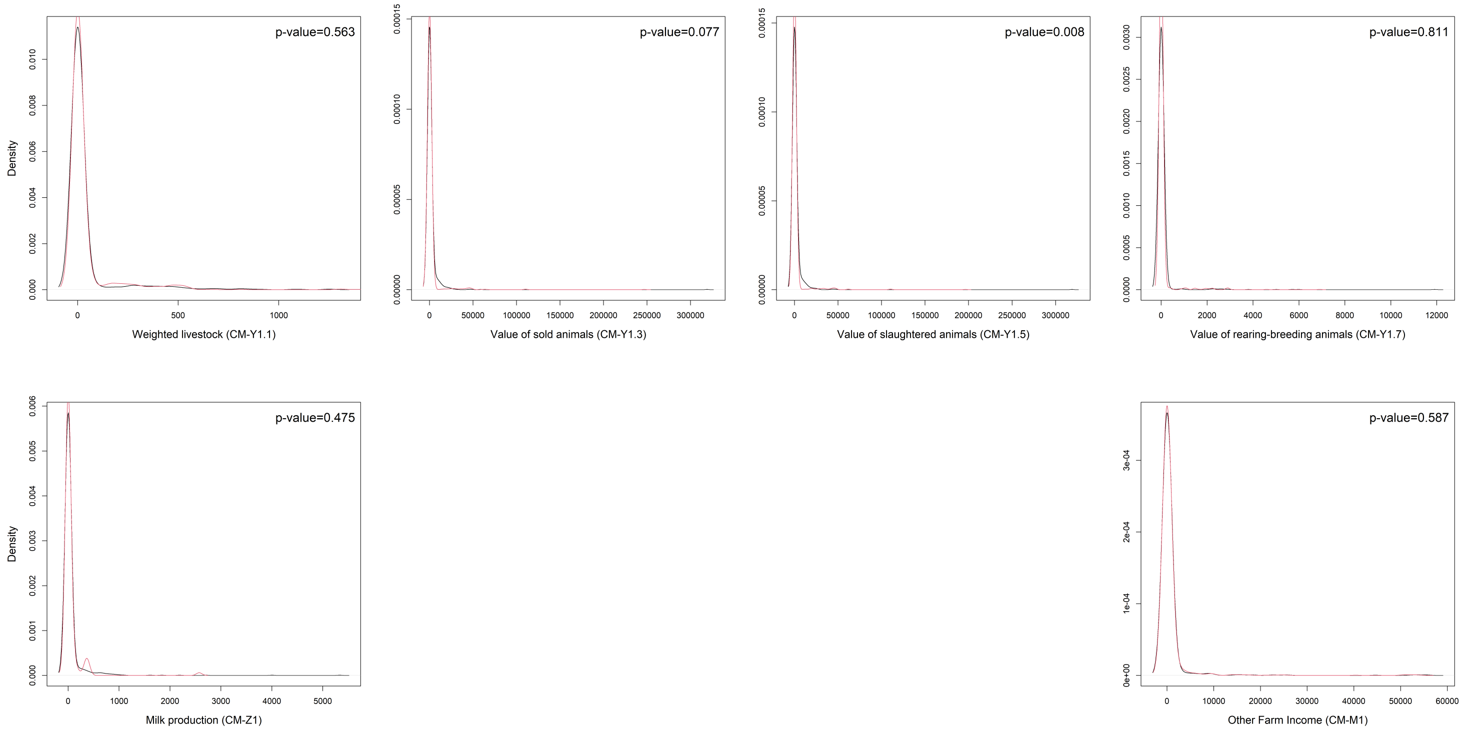

Figure 5: Distributions of the animal products (attributes CM-Y1.1, CM-Y1.3, CM-Y1.5, CM-Y1.7, CM-Z1) and of the other farm income (CM-M1). The black line refers to the observed farms, while the red line refers to the synthetic farms. The KDE test p-value appears on the top-right.

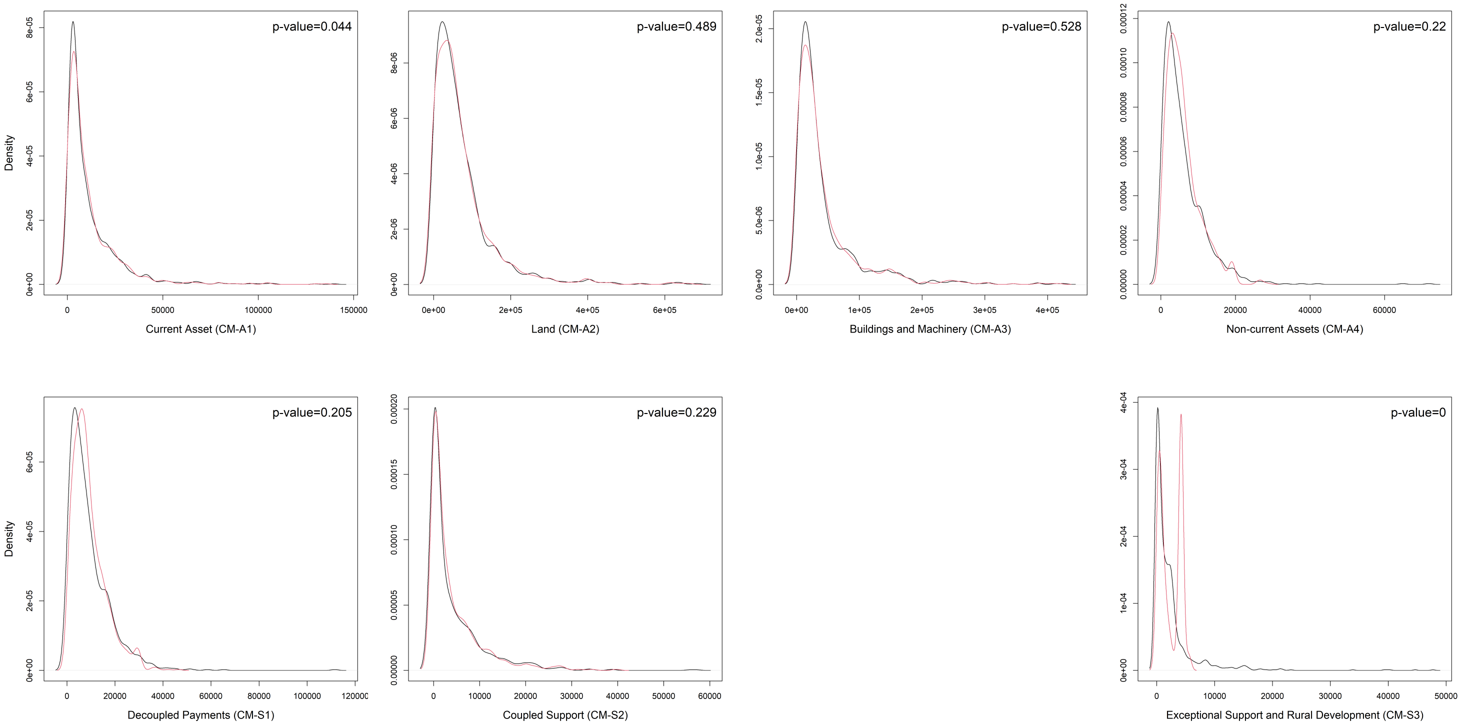

Figure 6: Distributions of the closing valuation of the farm assets (attributes CM-A1 - CM-A4) and of the subsidies and grants (CM-S1 - CM-S3). The black line refers to the observed farms, while the red line refers to the synthetic farms. The KDE test p-value appears on the top-right.

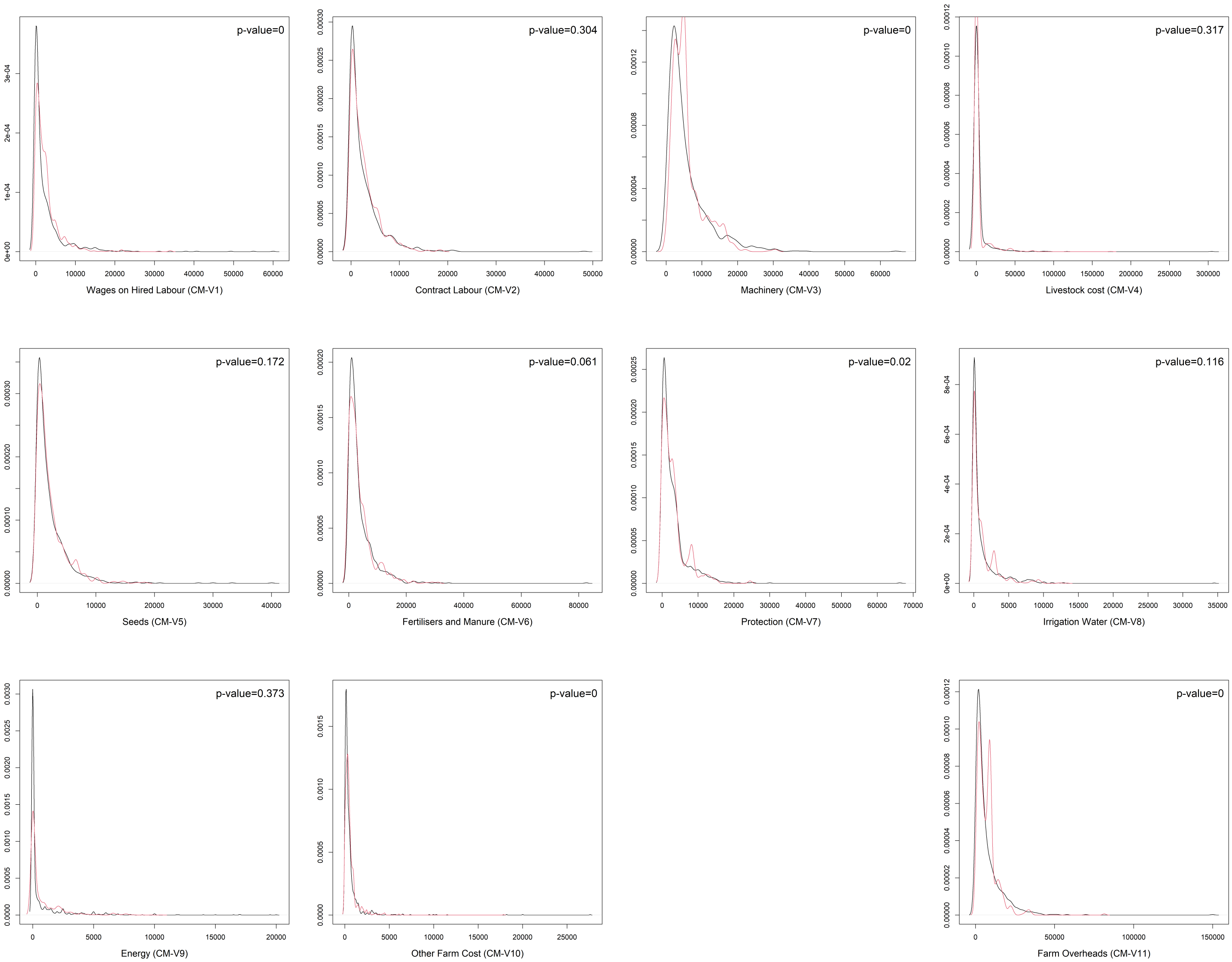

Figure 7: Distributions of the variable inputs cost (attributes CM-V1 - CM-V11). The black line refers to the observed farms, while the red line refers to the synthetic farms. The KDE test p-value appears on the top-right.

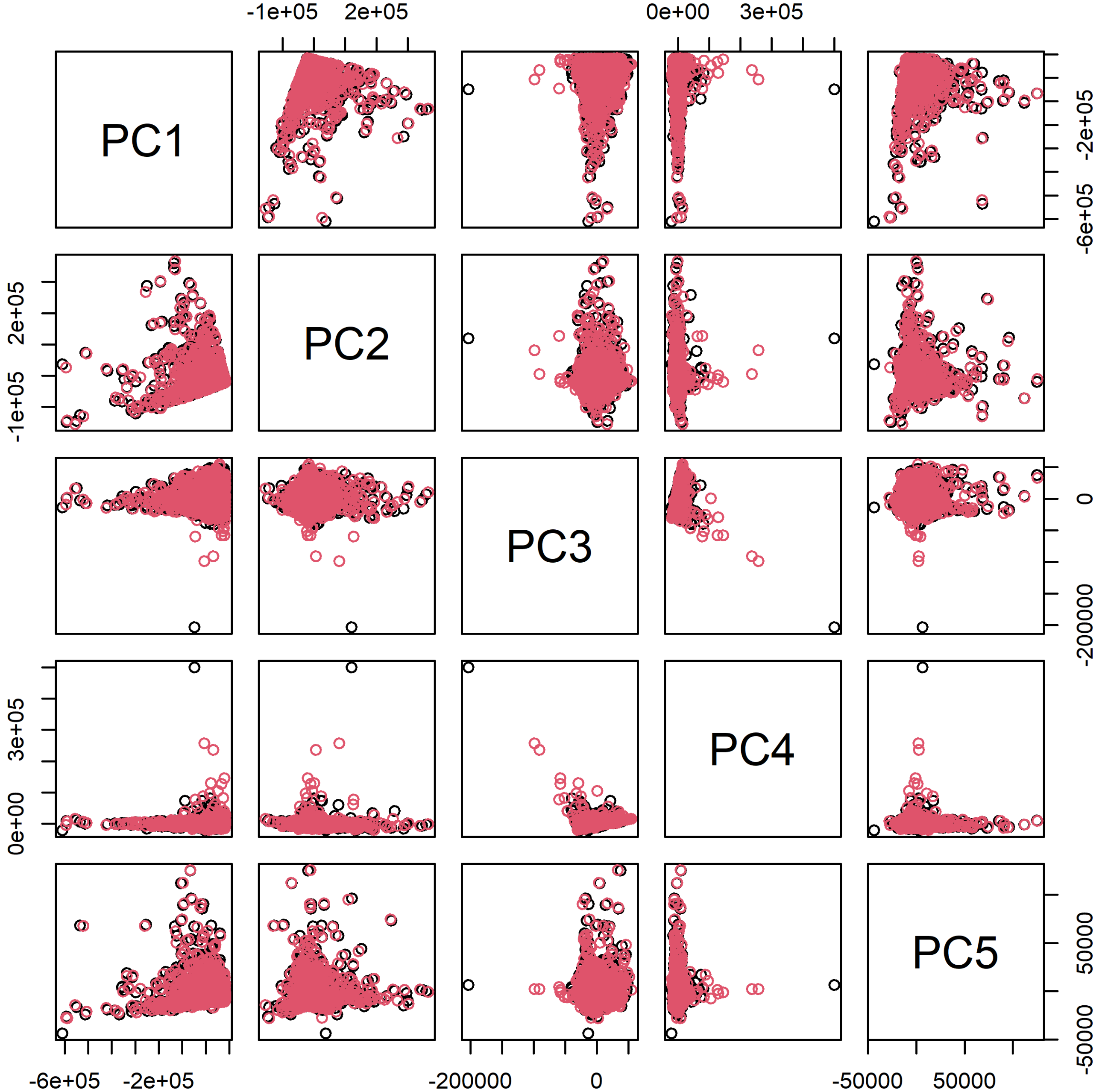

Figure 8: Central Macedonia: The data projected onto the first 5 principal components. The black circles refer to the observed farms whereas the red circles refer to the synthetic farms.

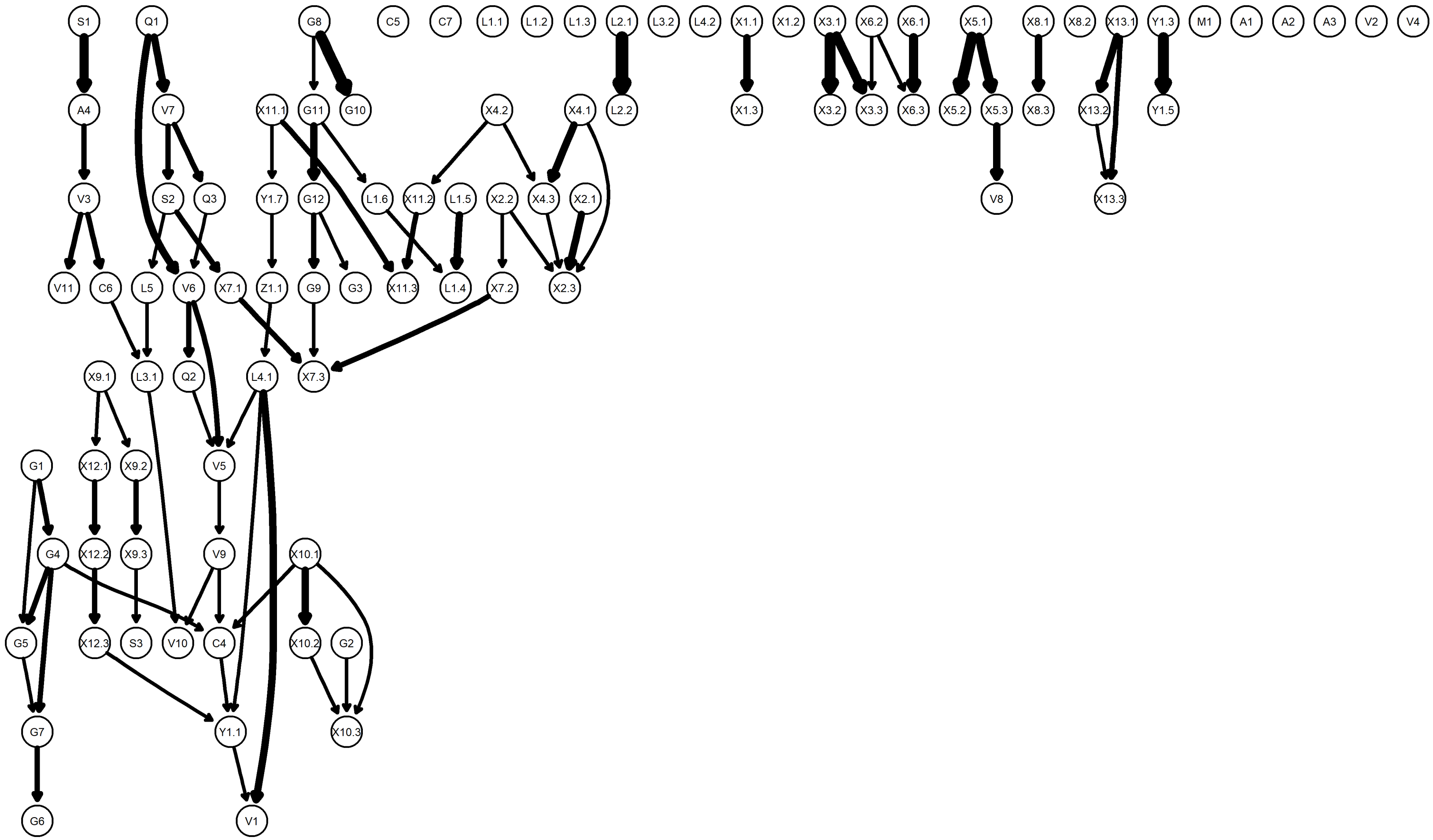

Figure 9: The BN structure of Thessaloniki. The strength of the directed relationships is denoted by the thickness of the arrows.

impact as already manifested due to the small sample size.

The KDE test assessing the distributions of the synthetic values of the attributes produced satisfactory results. The p-value for the distributions of 69 out of the 92 attributes (75%) was more than 0.05 indicating that these distributions, of the observed and and the synthetic farms, can be assumed (statistically) equal. The energy test was more conservative and produced 62 out of 92 p-values (67.4%) greater than 0.05.

Figure 10 presents the KDE of the observed and of the synthetic crop production in the 13 attributes. Evidently, the peaks of the estimated distributions of the cotton differ. This does not come by surprise as there are 255 observed farms with 0 production, whereas the synthetic farms contain 233 0s. The difference is small but due to the small sample size (only 325 farms) this difference is magnified.

When applied to the joint distributions of the the observed and the synthetic farms, the energy test produced a p-value equal to 0.861 indicating a high similarity between the two joint distributions. However, the attributes are measured in different scales and different units of measurements. For this reason, the two groups (observed and synthetic farms) were standardised to have zero means and unity variances and the energy test was applied to the transformed data and produced p-value equal to 0.070. The p-value seems small, yet it shows an acceptable agreement or fit.

Figure 11 shows the data projected onto the first 5 principal components produced by PCA. It is evident that the synthetic farms cannot be distinguished from the observed farms.

4.3 A synthetic sample for Thessalia (NUTS-2 level)

Thessalia is located at the center of the continental Greece and contributed to the Greek FADN sample with 509 farms. We used the same constraints as in central Macedonia for the BN learning process. Due to sparsity (excessive amounts of zeros) in many attributes, aggregation of attributes, based on their proximity, resulting in 86 attributes, was deemed mandatory for the the BN learning and the SPG task subsequently[10]. Those 86 attributes, grouped according to the clusters presented previously, can be found in the Appendix. Additionally, the same set of constraints imposed on the BN learning for the case of central Macedonia was also imposed among the 86 attributes of Thessalia.

- **Crop production**. Table A.1 shows the crop production of central Macedonia, where some crops have been aggregated due to sparsity (excessive amount of zeros), yielding 10 crops.
- **Animal products**. Table A.2 shows the condensed animal production, the weighted livestock, values of sold and slaughtered animals, values of animals left rearing-breeding and the total milk production.
- **Farm income, subsidies and grants**. Table A.3 contains information on the components that formulate the attribute termed ”other farm income”, the aggregation of the following characteristics: value of sold animals, value of sales of wool, eggs, honey and manure, other

[10] For instance, the 20 crops were merged into 10 crops.

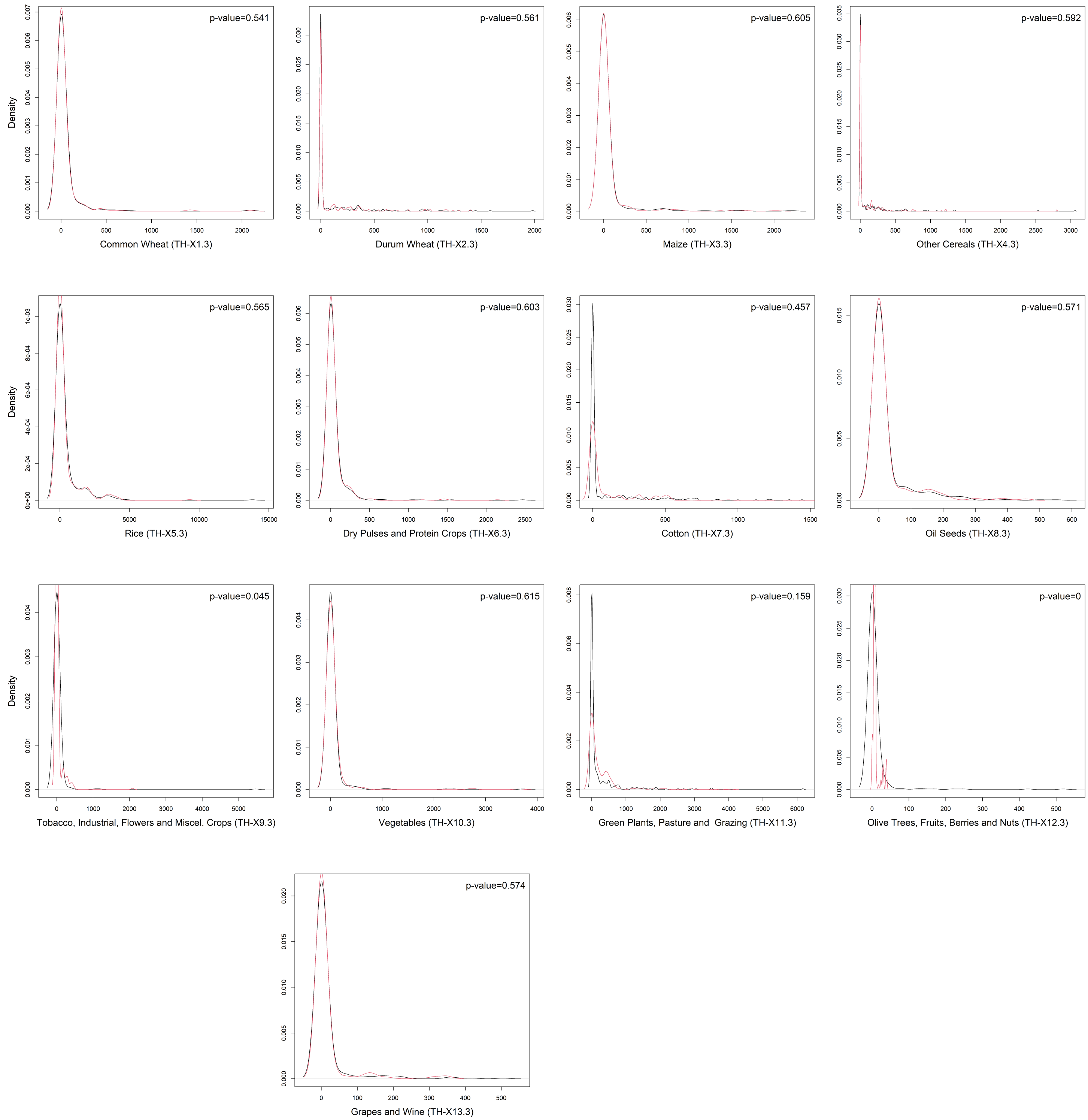

Figure 10: Distributions of the crop production (attributes TH-X1.3 - TH-X13.3). The black line refers to the observed farms, while the red line refers to the synthetic farms. The KDE test p-value appears on the top-right.

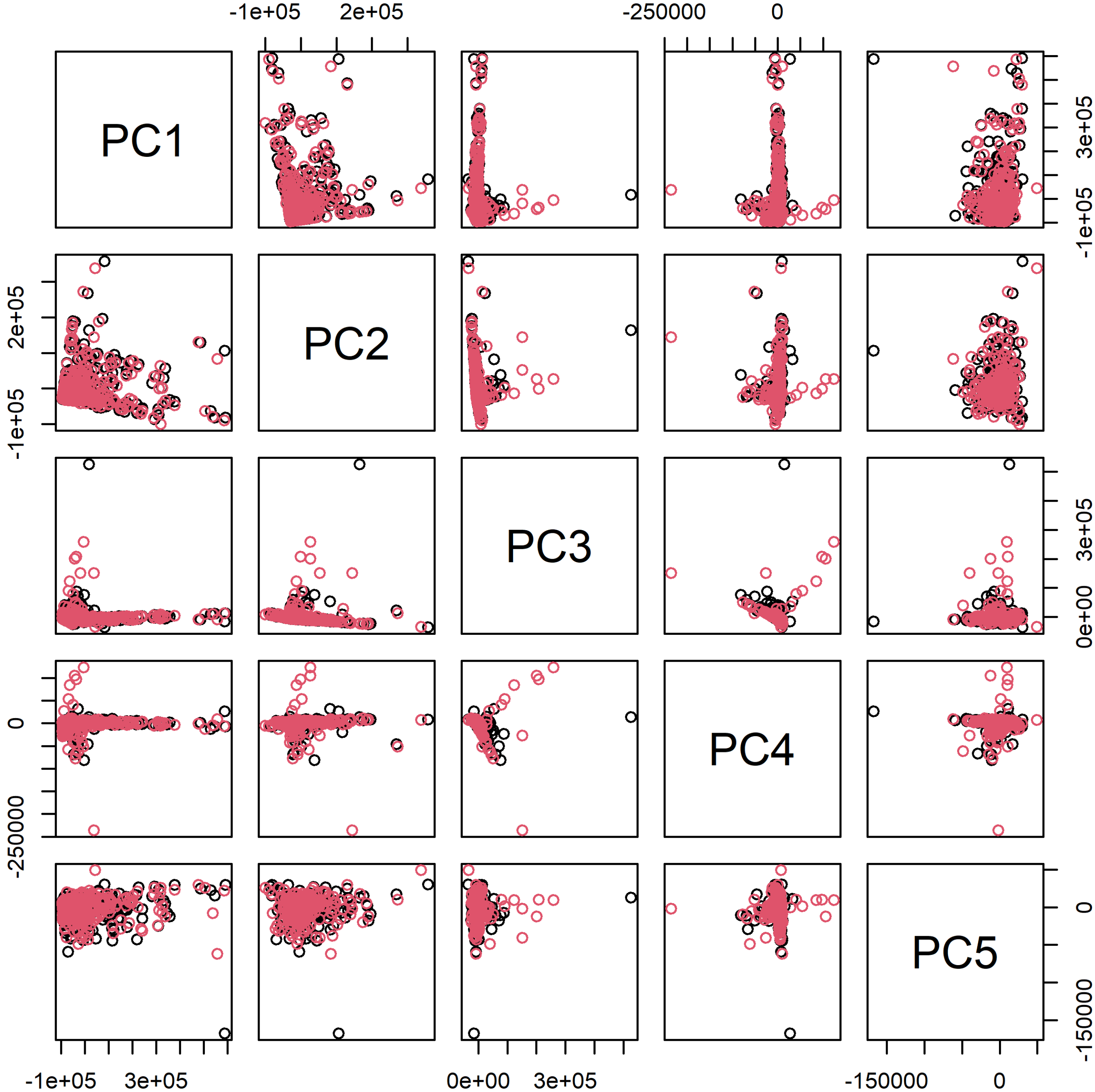

Figure 11: Thessaloniki: The data projected onto the first 5 principal components. The black circles refer to the observed farms whereas the red circles refer to the synthetic farms.

income from livestock (e.g. contract rearing), income from land (e.g. leasing), food processing (e.g. cow's milk), contractual work and income from other sources (e.g. tourism, production of renewable energy). Table B.9 shows the subsidies and grants grouped in 4 clusters, decoupled payments, crops and animals, exceptional support and rural development and subsidies on cost. Note that despite the subsidies on cost being listed in the FADN guide manual, this attribute was not applicable in the Greek use case.

- **Variable inputs cost**. Table A.4 contains 11 attributes (2 attributes were merged) representing the variable inputs cost.

Table 3 presents the strengths of the statistically significantly associated relationships discovered via the MMHC BN learning algorithm. Evidently, the BN identified 96 directed relationships in Thessalia presented in Table 3, along with their estimated strength, while Figure 12 shows the BN structure. The results of bootstrap validation also appear in Table 3. Out of the 96 identified directed relationships in the observed farms 61 (63.5%) were observed more than 50% of the times in the bootstrap samples.

Table 3: The 96 statistically significant associations clustered according to the tables (see Appendix). The computed strengths were normalised with the strongest strength playing the role of the basis. The column "boot" refers to the proportion of times the observed directed relationships were discovered in the bootstrap samples.

from	to	strength	boot	from	to	strength	boot	from	to	strength	boot
C5	Q1	0.0387	0.4700	L3.1	X1.1	0.0030	0.1600	X9.2	X9.3	0.0088	0.8600
C5	X2.1	0.0364	0.6300	L4.1	V1	0.2177	0.8000	X10.1	X10.2	0.1817	0.9850
C6	V11	0.0402	0.3350	L4.1	X6.3	0.0097	0.6550	X10.2	X10.3	0.2477	0.9950
C6	X7.1	0.0214	0.4750	X1.1	X1.3	0.2825	1.0000	X10.3	V7	0.0047	0.4800
C6	X7.3	0.0004	0.1100	X1.2	X10.3	0.0102	0.4900	Y1.1	L3.1	0.0193	0.8050
G1	G7	0.0131	0.9650	X1.2	X4.3	0.0141	0.6050	Y1.1	Y1.7	0.0631	0.3200
G1	G7	0.0131	0.9650	X1.3	Q2	0.0122	0.3400	Y1.3	V4	0.0684	0.6550
G1	V8	0.0034	0.5250	X2.1	X2.2	0.0018	0.7300	Y1.3	Y1.5	0.0628	0.5350
G3	C4	0.0031	0.4150	X2.1	X2.3	0.3126	1.0000	Y1.3	Z1	0.0082	0.7650
G3	G2	0.0060	0.3550	X2.2	X1.2	0.0015	0.5100	Y1.5	Z1	0.3913	0.3650
G4	G5	0.1626	0.9950	X3.1	X3.2	0.5271	0.9550	Y1.7	Z1	0.2116	0.8000
G5	G1	0.0433	0.4100	X3.1	X3.3	0.3566	0.9500	M1	V8	0.0021	0.2500
G5	G3	0.0999	1.0000	X4.1	X4.2	0.0253	0.9050	S1	A4	0.3914	1.0000
G7	C4	0.0050	0.9850	X4.1	X4.3	0.0018	0.7500	S1	C6	0.0654	0.1250
G7	S3	0.0073	0.8600	X4.2	V10	0.0045	0.3450	S1	S2	0.0318	0.3700
G7	X9.3	0.0007	0.5200	X4.2	X4.3	0.0028	0.8650	S2	S3	0.0329	0.3400
G8	G9	0.3444	0.8650	X5.1	X5.2	1.0000	0.9950	S2	V9	0.0275	0.6250
G9	G12	0.3791	1.0000	X6.1	X4.1	0.0003	0.1450	S3	V3	0.0206	0.3850
G10	G11	0.7616	0.9850	X6.1	X6.2	0.1488	1.0000	S3	Y1.1	0.0039	0.3200
G10	G8	0.0032	0.2450	X6.1	X6.3	0.0010	0.4500	V1	V7	0.0294	0.6000
G10	G9	0.3025	0.3600	X6.2	Q3	0.0249	0.4950	V5	S1	0.0689	0.0950
G11	G8	0.0300	0.8500	X6.2	X6.3	0.0272	0.9950	V5	S2	0.0157	0.4000
Q1	Q2	0.0036	0.7350	X6.3	V1	0.0040	0.2500	V5	V11	0.0307	0.4700
Q1	V5	0.0837	0.5700	X6.3	V5	0.0097	0.2500	V5	V2	0.0477	0.9750
Q1	V6	0.1149	0.8850	X7.1	X7.3	0.0294	1.0000	V5	X4.1	0.0086	0.5150
Q1	X5.3	0.0606	0.3300	X7.2	Q2	0.0041	0.4500	V6	Q2	0.0048	0.6750
Q3	Q2	0.0039	0.2400	X7.2	X7.3	0.0629	1.0000	V6	Q3	0.0364	0.6150

continued....

from	to	strength	boot	from	to	strength	boot	from	to	strength	boot
L1.1	L1.5	0.0109	0.7800	X8.1	X8.2	0.2505	1.0000	V6	V7	0.0217	0.6300
L1.1	M1	0.0008	0.5100	X8.2	X8.3	0.1101	0.9950	V7	V9	0.0075	0.2000
L1.2	L1.5	0.0083	0.5150	X8.3	V7	0.0065	0.5750	V9	C4	0.0058	0.7100
L1.5	L1.4	0.2302	0.8350	X9.1	X9.2	0.1209	1.0000	V10	M1	0.0533	0.5850
L3.1	L5	0.0882	0.7600	X9.1	X9.3	0.0117	0.9950	V10	V8	0.0323	0.3450

4.3.1 Evaluation of the synthetic sample generation in Thessalia

Using the estimated BN structure we generated a sample of 509 synthetic farms whose characteristics match to a high a degree the characteristics of the observed farms. Application of the γ-OMP [22] and FBED [23] attribute selection algorithms indicated that the two samples (observed and synthetic farms) can be separated with accuracy 71%. These two algorithm were ordinarily identifying the irrigated are for green plants, pasture and grazing, the irrigation system, the annual unpaid labour time worked and the houseehold size as the four attributes responsible for this level of separation. When these attributes were removed, γ-OMP could not separate the farms (accuracy = 51%). The mean of the irrigated area of that crop in the synthetic farms is less than the mean in the observed farms, while for the irrigation system the synthetic farms contained a higher number of farms without irrigation system than the the actual number observed. Secondly, the annual unpaid labour time had smaller values in the synthetic sample and we generated households with smaller sizes than the ones observed.

KDE test assessing the distributions of the synthetic values of the attributes produced satisfactory results. The p-value for the distributions of 49 out of the 84 attributes[11] (58.3%) was more than 0.05 indicating that these distributions, of the observed and and the synthetic farms, can be assumed (statistically) equal. The energy test was more conservative and produced 43 out of the 84 p-values (51.2%) greater than 0.05. When applied to the standardised data, the energy test of equality of the joint distributions between the observed and the synthetic farms produced a p-value equal to 0.479 providing evidence of a very good fit.
Finally, Figure 17 shows the data projected onto the first 5 principal components produced by PCA. It is evident that the synthetic farms cannot be distinguished from the observed farms.

4.4 A synthetic sample for Peloponnisos (NUTS-2 level)

Peloponnisos is located at the south of the continental Greece and contributed to the Greek FADN sample with 697 farms. Due to sparsity (excessive amounts of zeros) in many attributes, aggregation of attributes, based on their proximity, was deemed mandatory for the the BN learning and the SPG task subsequently[12] resulting in 85 attributes. Those 85 attributes, grouped according to the clusters presented previously, can be found in the Appendix. Additionally, the same set of constraints imposed on the BN learning for the case of the previous regions was also imposed among the 85 attributes of Peloponnisos.

[11] Two attributes had excessive amounts of zeros and the KDE test was not applicable

[12] For instance, the 20 crops were merged into 8 crops.

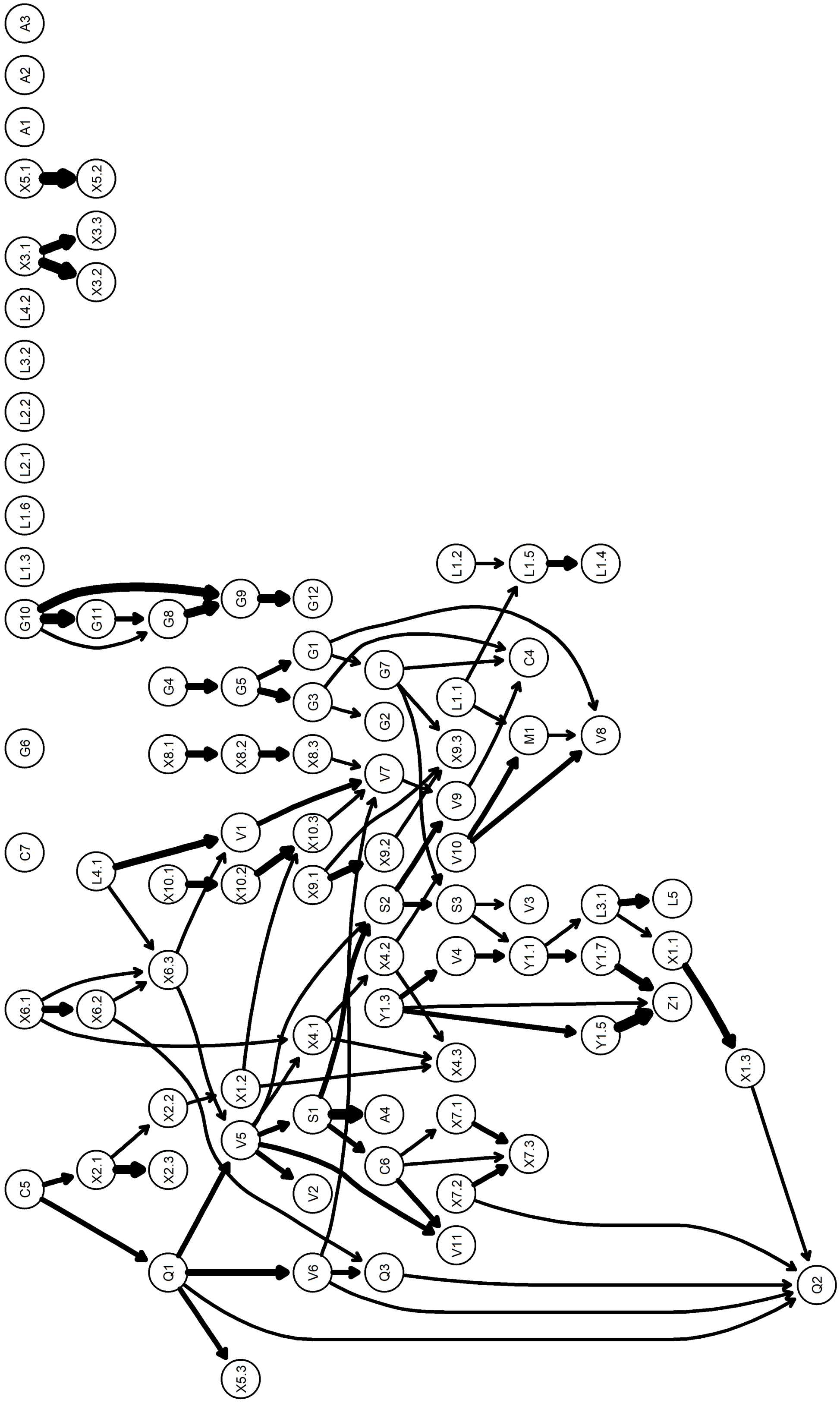

Figure 12: The BN structure of Thessalia. The strength of the directed relationships is denoted by the thickness of the arrows.

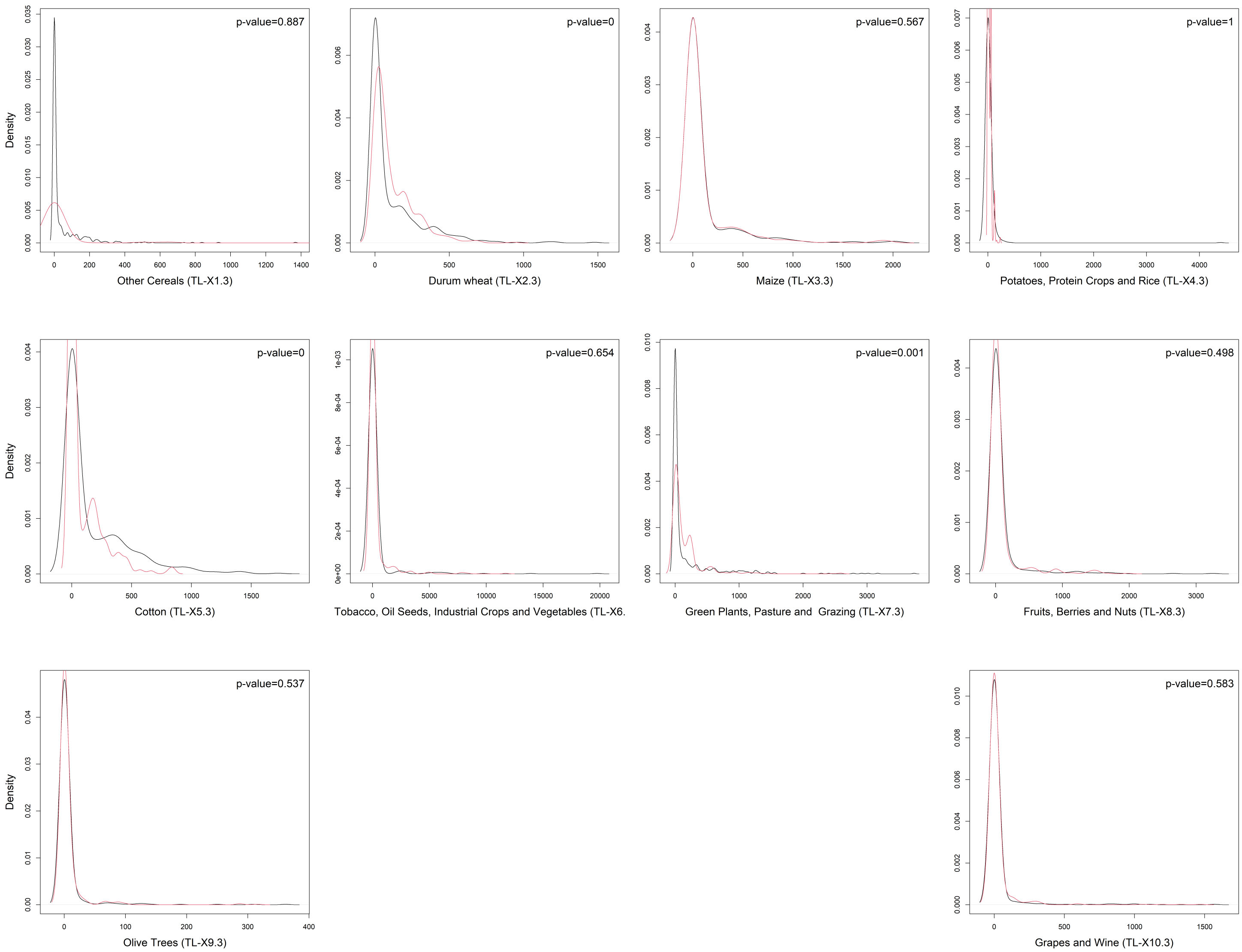

Figure 13: Distributions of the crop production (attributes TL-X1.3 - TL-X10.3). The black line refers to the observed farms, while the red line refers to the synthetic farms. The KDE test p-value appears on the top-right.

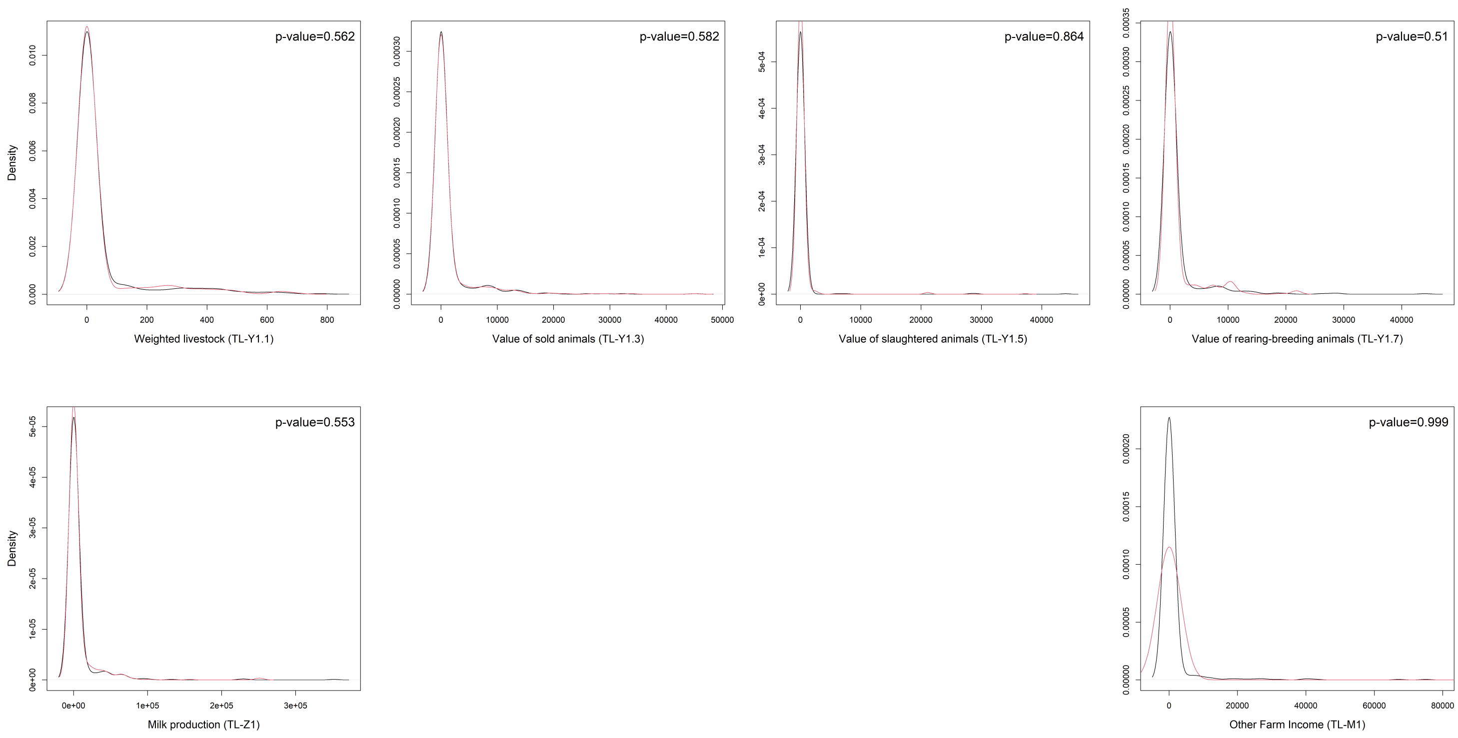

Figure 14: Distributions of the animal products (attributes TL-Y1.1, TL-Y1.3, TL-Y1.5, TL-Y1.7, TL-Z1) and of the other farm income (TL-M1). The black line refers to the observed farms, while the red line refers to the synthetic farms. The KDE test p-value appears on the top-right.

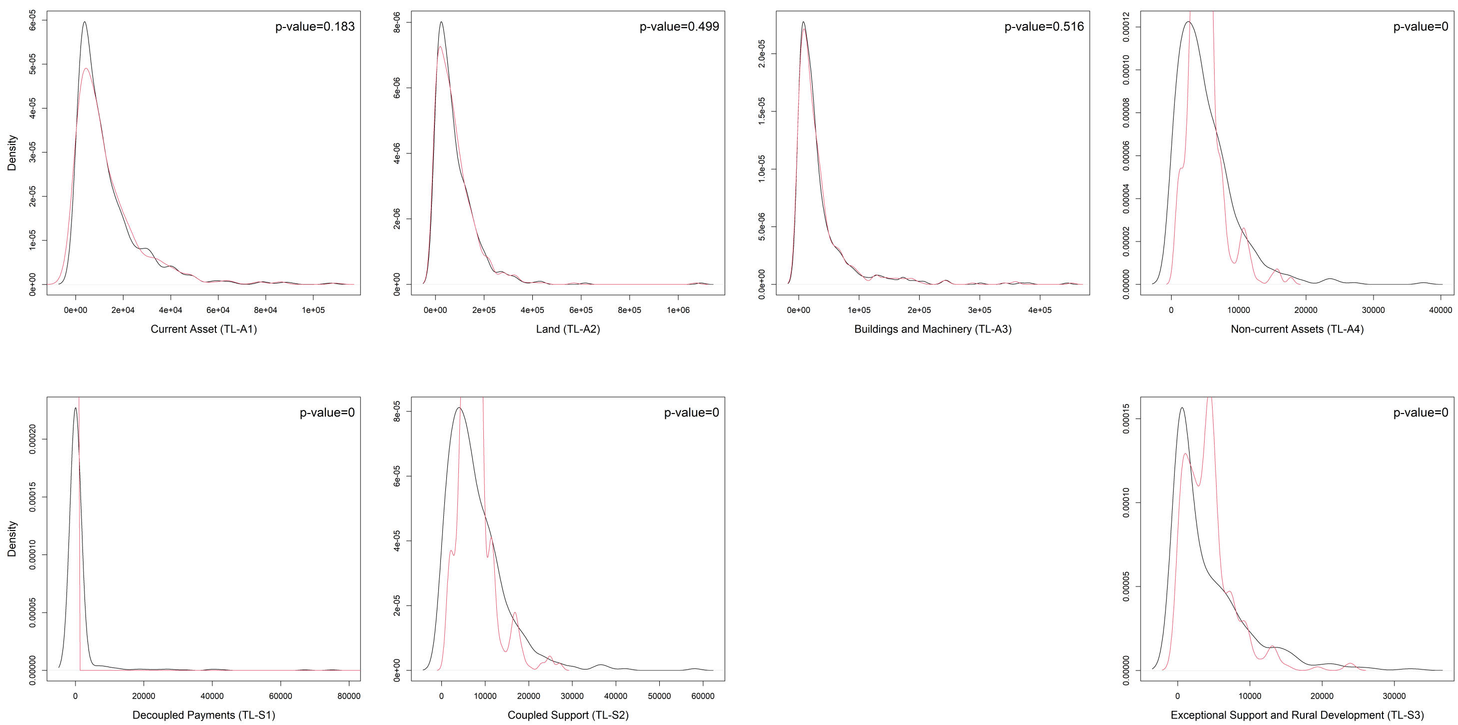

Figure 15: Distributions of the closing valuation of the farm assets (attributes TL-A1 - TL-A4) and of the subsidies and grants (TL-S1 - TL-S3). The black line refers to the observed farms, while the red line refers to the synthetic farms. The KDE test p-value appears on the top-right.

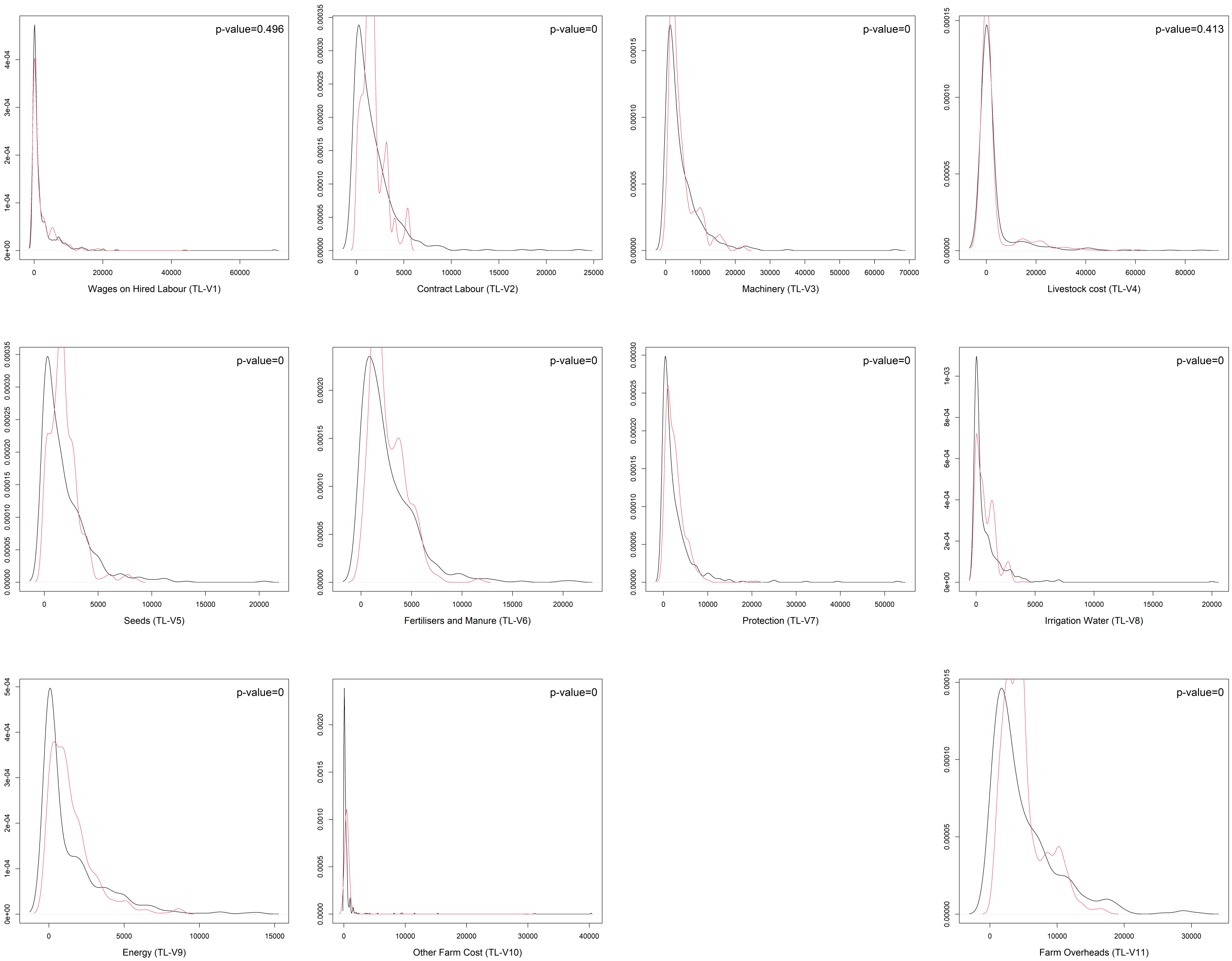

Figure 16: Distributions of the variable inputs cost (attributes TL-V1 - TL-V11). The black line refers to the observed farms, while the red line refers to the synthetic farms. The KDE test p-value appears on the top-right.

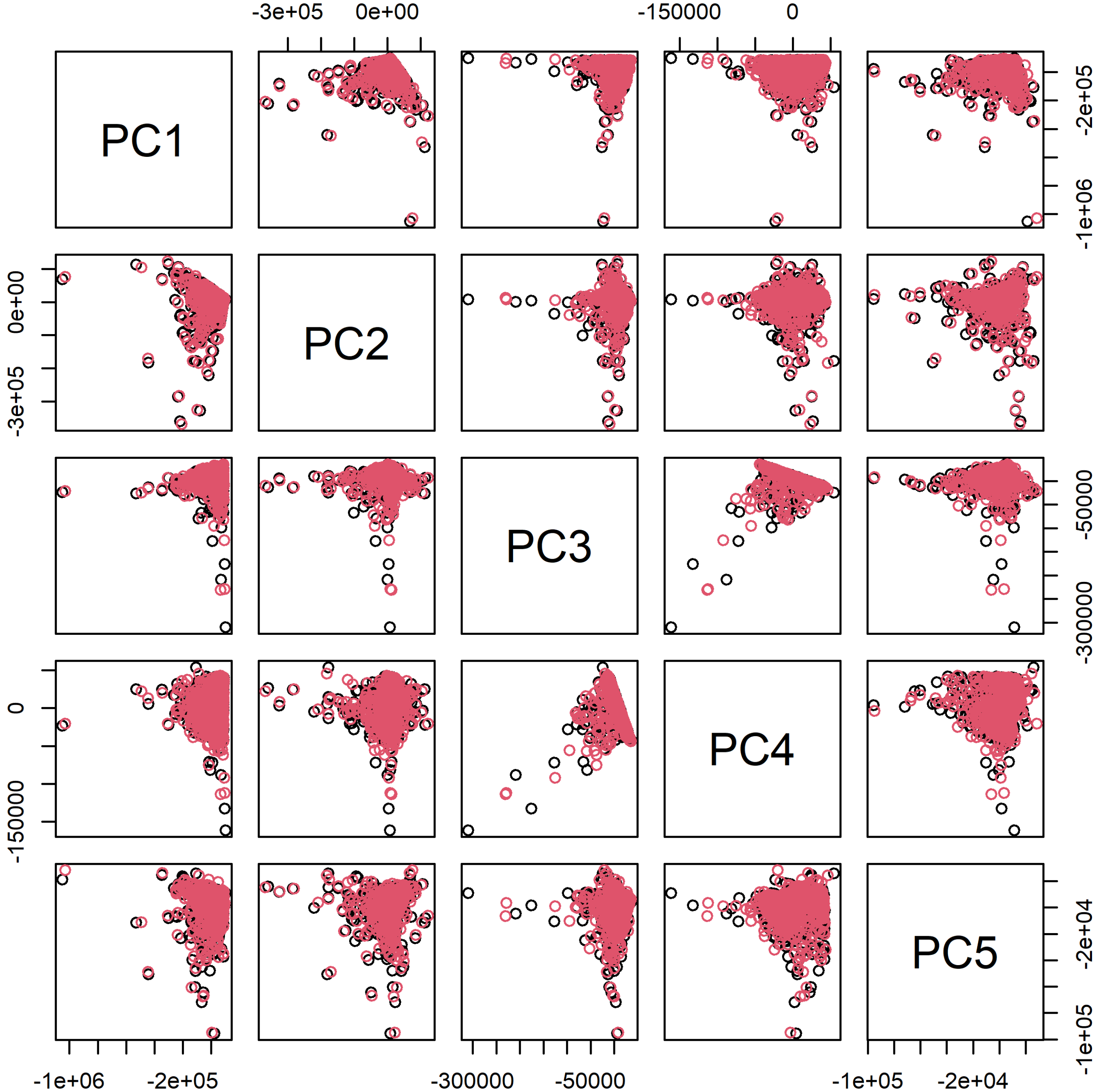

Figure 17: Thessalia: The data projected onto the first 5 principal components. The black circles refer to the observed farms whereas the red circles refer to the synthetic farms.

- **Crop production**. Table A.1 shows the crop production of central Macedonia, where some crops have been aggregated due to sparsity (excessive amount of zeros), yielding 10 crops.
- **Animal products**. Table A.2 shows the condensed animal production, the weighted livestock, values of sold and slaughtered animals, values of animals left rearing-breeding and the total milk production.
- **Farm income, subsidies and grants**. Table A.3 contains information on the components that formulate the attribute termed ”other farm income”, the aggregation of the following characteristics: value of sold animals, value of sales of wool, eggs, honey and manure, other income from livestock (e.g. contract rearing), income from land (e.g. leasing), food processing (e.g. cow's milk), contractual work and income from other sources (e.g. tourism, production of renewable energy). Table B.9 shows the subsidies and grants grouped in 4 clusters, decoupled payments, crops and animals, exceptional support and rural development and subsidies on cost. Note that despite the subsidies on cost being listed in the FADN guide manual, this attribute was not applicable in the Greek use case.
- **Variable inputs cost**. Table A.4 contains 11 attributes (2 attributes were merged) representing the variable inputs cost.

Table 4 presents the strengths of the statistically significantly associated relationships discovered via the MMHC BN learning algorithm. Evidently, the BN identified 119 directed relationships in Peloponnisos presented in Table 4, along with their estimated strength, while Figure 18 shows the BN structure. The results of bootstrap validation also appear in Table 4. Out of the 119 identified directed relationships in the observed farms, 75 (63%) were observed more than 50% of the times in the bootstrap samples.

Table 4: The 119 statistically significant associations clustered according to the tables (see Appendix). The computed strengths were normalised with the strongest strength playing the role of the basis. The column ”boot” refers to the proportion of times the observed directed relationships were discovered in the bootstrap samples.

from	to	strength	boot	from	to	strength	boot	from	to	strength	boot
C4	X7.1	0.0025	0.2900	L2.1	V3	0.0009	0.5350	Y1.5	X4.1	0.0056	0.4700
C5	S3	0.0168	0.2850	L3.2	L3.1	0.0022	0.1850	Y1.5	Y1.7	0.1512	0.9900
C6	V11	0.0220	0.8200	L3.2	M1	0.0253	0.9300	Y2.1	V4	0.0339	0.9650
C6	X4.1	0.0147	0.5250	L3.2	V10	0.0141	0.5400	Y2.1	Y2.3	0.2765	0.9650
G1	G3	0.0037	0.8850	L3.2	Y2.1	0.0382	0.2200	Y2.1	Y2.5	0.0217	0.5300
G1	G4	0.0549	0.9100	L4.1	V1	0.4480	0.7800	Y2.1	Y2.7	0.0245	0.6700
G1	S1	0.0041	0.3200	L4.2	M1	0.0012	0.3300	Y2.1	Z1	0.0582	0.9300
G1	X6.3	0.0015	0.4550	L4.2	X4.1	0.0144	0.5350	Y2.3	Y2.5	0.0344	1.0000
G2	G3	0.0084	0.6150	L5	L3.1	0.0191	0.8900	Y2.3	Y2.7	0.0583	1.0000
G2	Y1.5	0.0003	0.3800	X1.1	X1.3	0.0447	1.0000	Y2.5	C6	0.0014	0.3950
G3	L4.2	0.0057	0.6100	X1.2	X1.3	0.1029	1.0000	Y2.5	V10	0.0383	0.3250
G4	G5	0.1655	1.0000	X2.1	X2.2	0.3736	1.0000	Y2.5	Z1	0.0071	0.5750
G4	G6	0.0500	0.7400	X2.2	Q2	0.0032	0.3300	Y2.7	C5	0.0143	0.4350
G4	V10	0.0009	0.2500	X2.2	X2.3	0.0531	0.9600	Z1	L3.1	0.0146	0.3700
G4	X5.1	0.0014	0.4750	X2.2	X4.2	0.0053	0.2300	Z1	V4	0.0173	0.7200

continued....

from	to	strength	boot	from	to	strength	boot	from	to	strength	boot
G5	G3	0.0605	0.9900	X2.3	V5	0.0074	0.7800	S1	A4	0.6501	0.9250
G7	V9	0.0038	0.3000	X2.3	V7	0.0121	0.6300	S2	S3	0.0437	0.6000
G7	X7.1	0.0027	0.7050	X3.1	X3.2	0.6123	1.0000	A4	C5	0.0553	0.4550
G8	G9	1.0000	1.0000	X3.1	X3.3	0.0616	1.0000	A4	S2	0.0481	0.4500
G8	X8.1	0.0008	0.6750	X3.2	X3.3	0.0369	1.0000	A4	V3	0.0188	0.3750
G9	G10	0.6974	1.0000	X4.2	X4.3	0.1652	1.0000	V1	V11	0.0162	0.4550
G9	X8.1	0.0007	0.5500	X4.3	Y1.5	0.0015	0.3750	V2	C4	0.0013	0.1900
G10	G11	0.1915	1.0000	X5.1	X5.2	0.2089	1.0000	V2	L5	0.0019	0.3300
G11	G12	0.3326	1.0000	X5.1	X5.3	0.0018	0.7600	V2	V10	0.0087	0.3100
G12	G2	0.0007	0.3600	X5.2	X5.3	0.0335	0.9950	V3	L3.1	0.0143	0.5450
G12	V5	0.0153	0.6450	X5.3	V7	0.0030	0.5000	V4	C6	0.0027	0.2950
Q1	V6	0.0722	0.4450	X6.1	X6.2	0.9965	1.0000	V5	X1.1	0.0047	0.3450
Q2	Q1	0.0830	0.2550	X6.1	X6.3	0.0002	0.5750	V6	L4.1	0.0369	0.2550
Q3	Q2	0.1040	0.5000	X6.2	X6.3	0.0004	0.5900	V6	V11	0.0094	0.4350
Q3	V3	0.0022	0.3800	X6.3	V5	0.0028	0.0450	V6	V7	0.0739	0.7350
Q3	V6	0.0828	0.4750	X7.1	X7.2	0.0530	1.0000	V6	V9	0.1085	0.5050
L1.1	L1.5	0.0130	0.5500	X7.1	X7.3	0.0173	1.0000	V7	V3	0.0217	0.6450
L1.2	L1.3	0.0057	0.9600	X7.2	X7.3	0.0281	1.0000	V8	C4	0.0046	0.5250
L1.2	L1.5	0.0072	0.9450	X8.1	X8.3	0.1373	1.0000	V9	L4.1	0.0238	0.1850
L1.2	V2	0.0017	0.6100	X8.2	X8.3	0.0419	0.9800	V9	V11	0.0055	0.3350
L1.3	C4	0.0076	0.7150	X8.3	V7	0.0246	0.5350	V9	V3	0.0103	0.3800
L1.3	L1.6	0.0134	1.0000	Y1.1	Y1.3	0.0122	0.4800	V9	X7.2	0.0041	0.2850
L1.3	M1	0.0012	0.3800	Y1.3	X4.1	0.0021	0.1100	V10	M1	0.2139	0.3900
L1.3	X7.1	0.0010	0.6350	Y1.3	Y1.5	0.2341	1.0000	V11	V5	0.0262	0.5650
L1.5	L1.4	0.3613	0.9150	Y1.3	Y1.7	0.2420	0.9950				

4.4.1 Evaluation of the synthetic sample generation in Peloponnisos

Using the estimated BN structure and those 85 attributes we generated a sample of 697 synthetic farms whose characteristics match to a relatively low degree the characteristics of the observed farms. Application of the γ-OMP [22] and FBED [23] attribute selection algorithms indicated that the two samples (observed and synthetic farms) can be separated with accuracy 78.3%. These two algorithm were ordinarily identifying seven attributes responsible for this level of separation. When these attributes were removed, γ-OMP could not separate the farms adequately (accuracy = 55%).

The KDE test assessing the distributions of the synthetic values of the attributes produced satisfactory results. The p-value for the distributions of 44 out of the 83 attributes[13] (53.4%) was more than 0.05 indicating that these distributions, of the observed and and the synthetic farms, can be assumed (statistically) equal. The energy test was more conservative and produced 30 out of the 83 p-values (36.1%) greater than 0.05. The energy test applied to the, independently, standardised observed and synthetic farms produced a low p-value equal to 0.002.
Finally, Figure 23 shows the data projected onto the first 5 principal components produced by PCA. It is evident that the synthetic farms cannot be distinguished from the observed farms.

[13] Two attributes had excessive amounts of zeros and the KDE test was not applicable

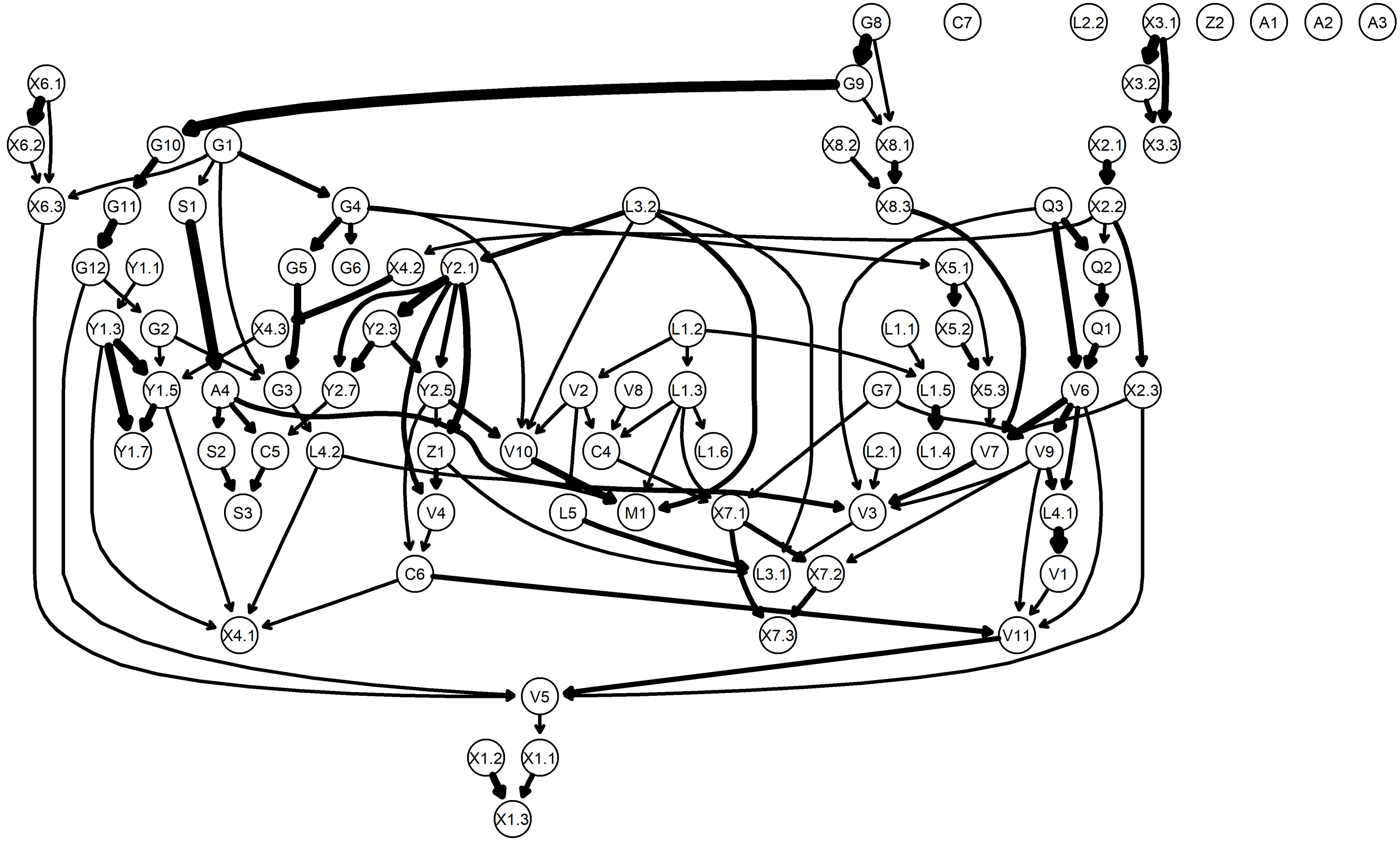

Figure 18: The BN structure of Peloponnisos. The strength of the directed relationships is denoted by the thickness of the arrows.

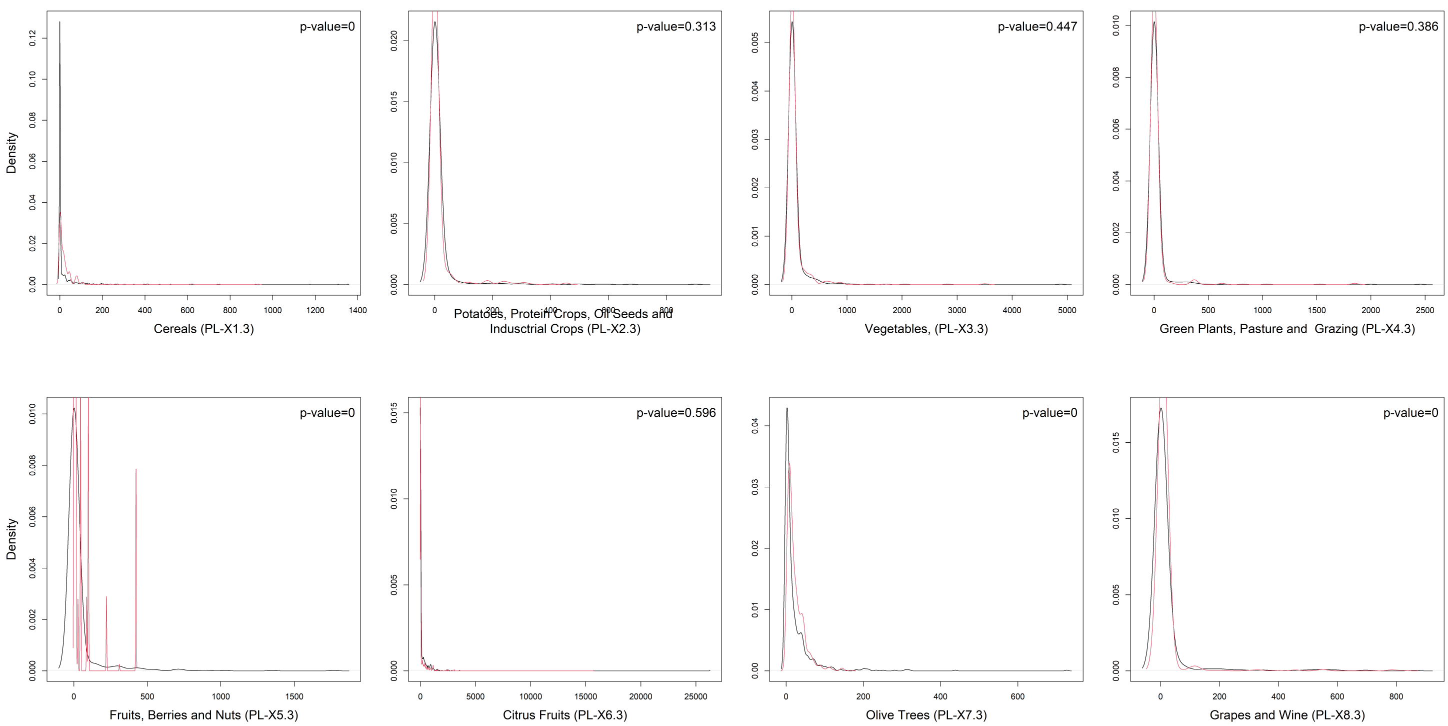

Figure 19: Distributions of the crop production (attributes PL-X1.3 - PL-X8.3). The black line refers to the observed farms, while the red line refers to the synthetic farms. The KDE test p-value appears on the top-right.

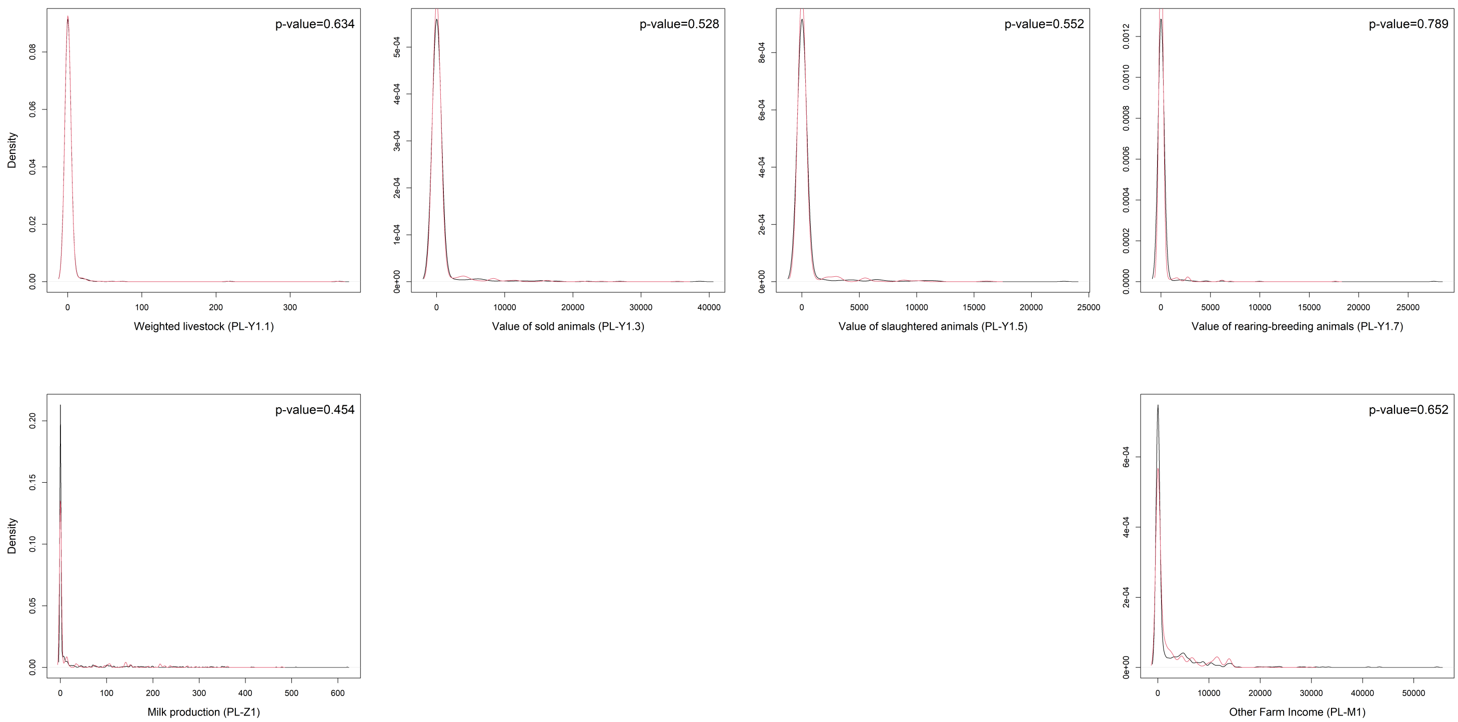

Figure 20: Distributions of the animal products (attributes PL-Y1.1, PL-Y1.3, PL-Y1.5, PL-Y1.7, PL-Z1) and of the other farm income (PL-M1). The black line refers to the observed farms, while the red line refers to the synthetic farms. The KDE test p-value appears on the top-right.

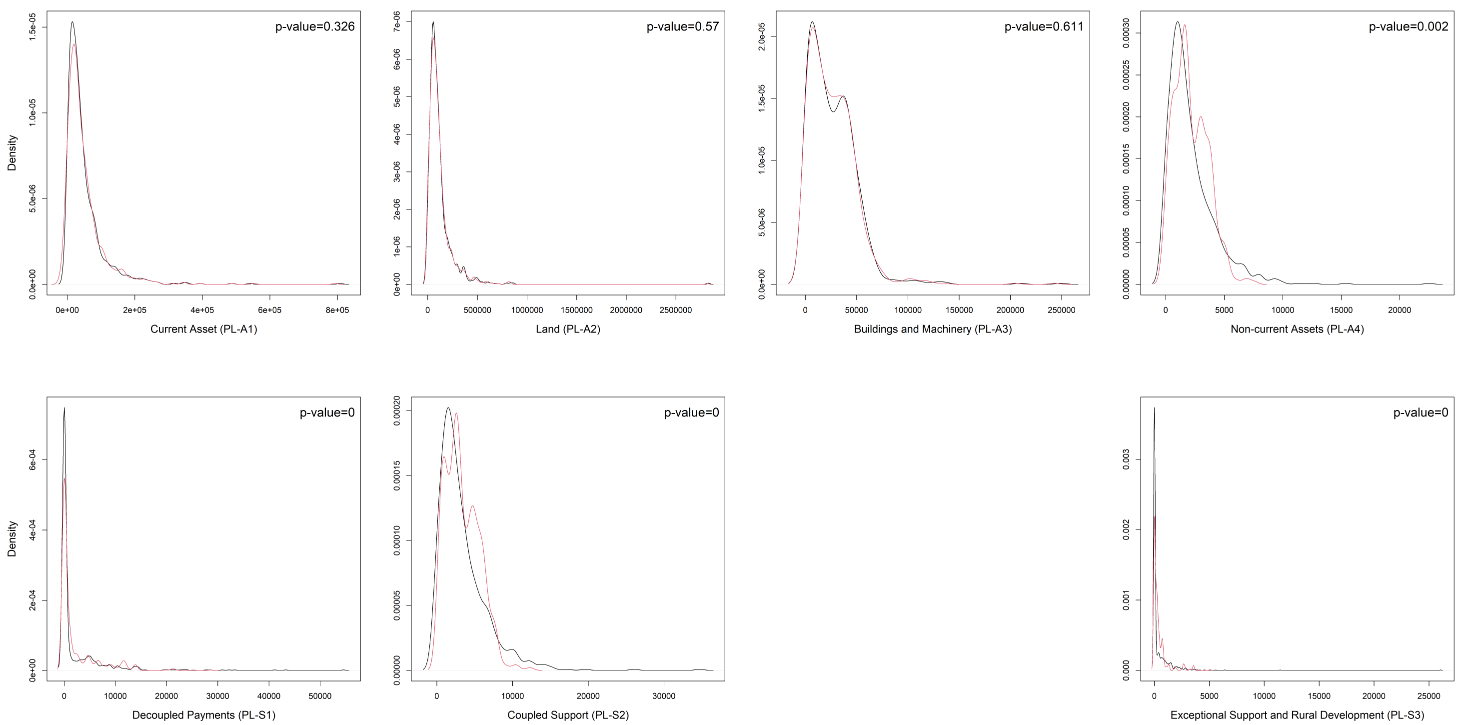

Figure 21: Distributions of the closing valuation of the farm assets (attributes PL-A1 - PL-A4) and of the subsidies and grants (PL-S1 - PL-S3). The black line refers to the observed farms, while the red line refers to the synthetic farms. The KDE test p-value appears on the top-right.

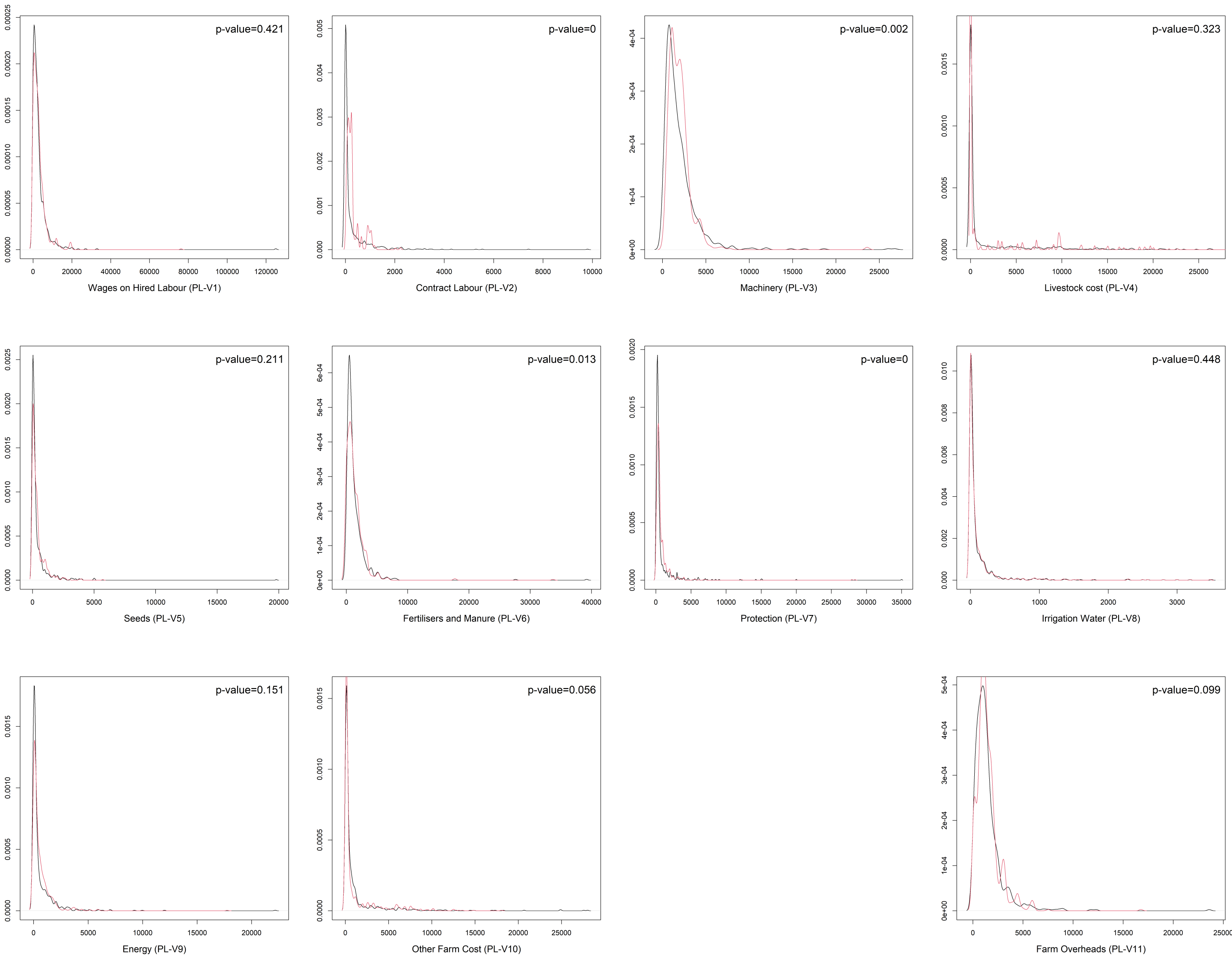

Figure 22: Distributions of the variable inputs cost (attributes PL-V1 - PL-V11). The black line refers to the observed farms, while the red line refers to the synthetic farms. The KDE test p-value appears on the top-right.

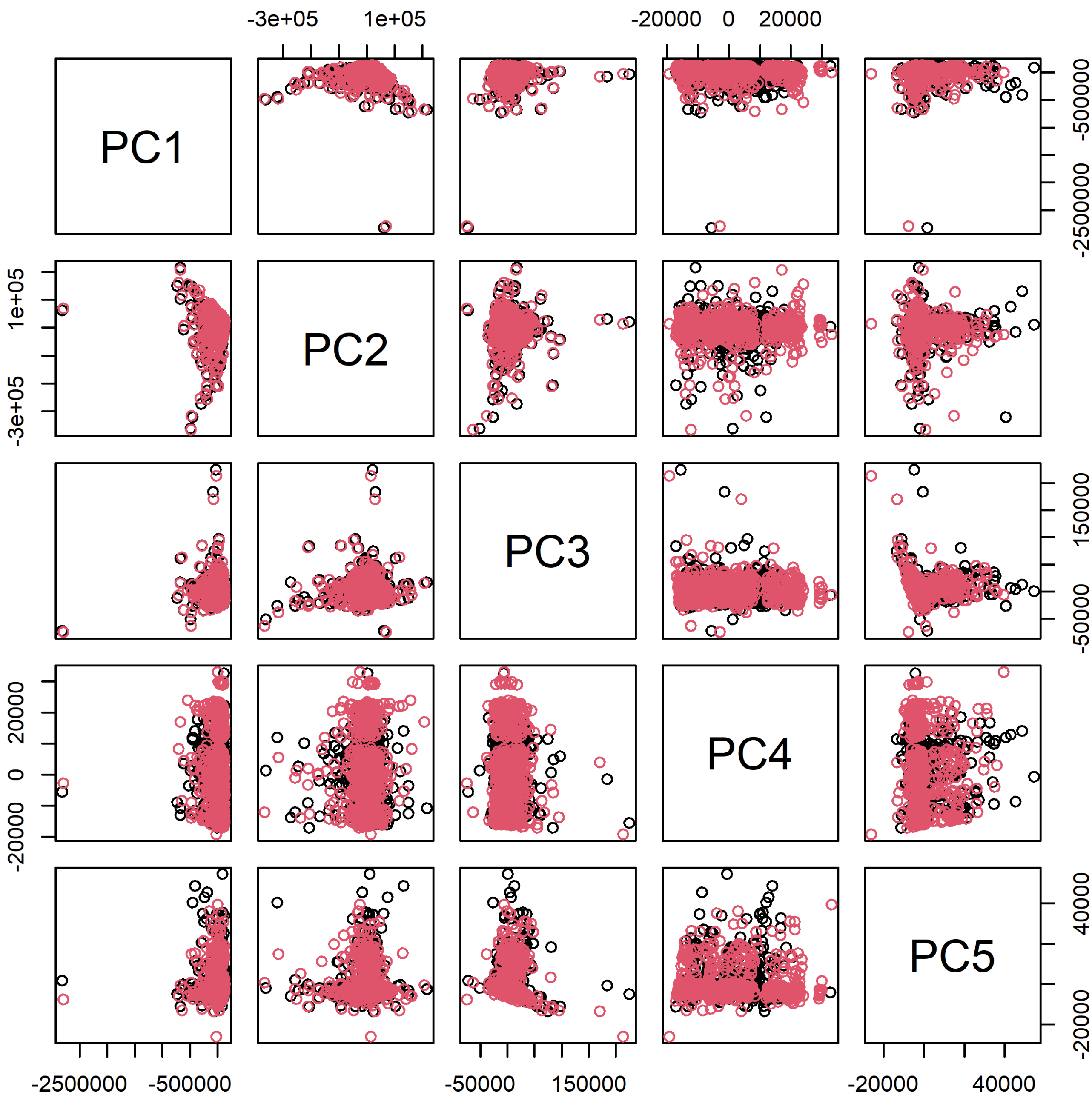

Figure 23: Peloponnisos: The data projected onto the first 5 principal components. The black circles refer to the observed farms whereas the red circles refer to the synthetic farms.

5 Conclusions

This deliverable presented the Data Fusion Module (DFM) of the AGRICORE suite. The DFM is responsible for accessing the Data Warehouse to access the individual datasets previously extracted and transformed by the DEM, as well as their metadata (statistical characterisation and forbidden relationships) also obtained through the DEM. Once the necessary data is loaded, the DFM executes a series of procedures to generate enriched datasets by integrating the individual datasets. These enriched datasets are used for various processes in AGRICORE. The most important of these is the construction of anonymised agents that form the synthetic population that is subsequently simulated by the ABM engine. This construction requires generating, for each agent, pseudo-random values and assigning them to each of its attributes. Given that the variables associated with certain attributes show correlation with the variables of other attributes, the assignment of a value for one attribute conditions the range of values assignable to other attributes. Therefore, it is necessary to have a mathematical object to determine the order in which attributes are generated, as well as the joint probability densities of these attributes.

The mathematical artefact chosen for AGRICORE is the Bayesian Network. This deliverable describes the Max-Min Hill Climbing (MMHC) algorithm and the variants incorporated into it to adapt it to the particular needs of the project.

In order to test the performance of the BNs built using the MMHC, 4 example cases have been implemented at regional (NUTS2) and sub-regional (NUTS3) level belonging to the Greek use case of the AGRICORE Project. Specifically, based on the regionalised subsamples of the Greek FADN, equivalent synthetic subsamples were constructed and their goodness-of-fit was analysed.

The results show that the fit is very accurate in all four cases, making the proposed procedure very promising for application in the Synthetic Population Generator (SPG). The next steps are the integration and packaging of the BN construction scripts for its execution from the SPG, and the testing of the generation of complete synthetic populations for the 3 use cases contemplated in the project.

References

[1] Judea Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible reasoning*. Morgan Kaufmann Publishers, Los Altos, 1988.

[2] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

[3] Richard E Neapolitan. *Learning Bayesian networks*. Pearson Prentice Hall Upper Saddle River, NJ, 2003.

[4] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, pages 220–227, 1991.

[5] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

[6] Gregory F Cooper and Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

[7] David Heckerman, Dan Geiger, and David M Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[8] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

[9] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

[10] Michail Tsagris. A new scalable Bayesian network learning algorithm with applications to economics. *Computational Economics*, 57(1):341–367, 2021.

[11] Michail Tsagris. The FEDHC Bayesian network learning algorithm. *arXiv preprint arXiv:2012.00113*, 2021.

[12] Marco Scutari and Radhakrishnan Nagarajan. Identifying significant edges in graphical models of molecular networks. *Artificial intelligence in medicine*, 57(3):207–217, 2013.

[13] Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. Time and sample efficient discovery of Markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–678. ACM, 2003.

[14] Norman R Draper and Harry Smith. *Applied Regression Analysis*. John Wiley & Sons, 1998.

[15] Ioannis Tsamardinos and Laura E Brown. Bounding the False Discovery Rate in Local Bayesian Network Learning. In *AAAI*, pages 1100–1105, 2008.

[16] Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.

[17] Wai Lam and Fahiem Bacchus. Learning Bayesian belief networks: An approach based on the MDL principle. *Computational intelligence*, 10(3):269–293, 1994.

[18] Dan Geiger and David Heckerman. Learning Gaussian networks. In *Proceedings of the 10th international conference on Uncertainty in Artificial Intelligence*, pages 235–243. Morgan Kaufmann Publishers Inc., 1994.

[19] Wray Buntine. Theory refinement on bayesian networks. In *Uncertainty Proceedings 1991*, pages 52–60. Elsevier, 1991.

[20] Remco Ronaldus Bouckaert. *Bayesian belief networks: from construction to inference*. PhD thesis, 1995.

[21] Gábor J Székely, Maria L Rizzo, et al. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.

[22] Michail Tsagris, Zacharias Papadovasilakis, Kleanthi Lakiotaki, and Ioannis Tsamardinos. The γ-OMP algorithm for feature selection with application to gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(2):1214–1224, 2022.

[23] Giorgos Borboudakis and Ioannis Tsamardinos. Forward-backward selection with early dropping. *The Journal of Machine Learning Research*, 20:276–314, 2019.

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

Deliverable Number	**Deliverable Title**	**Lead beneficiary**	**Type**	**Dissemination Level**	**Due date**
D6.1	AGRICORE architecture	IDE	Report	Public	M23
D2.2	Data Extraction Module	AUTH	Report	Public	M36

Appendix A Specific aggregations of variables for each case study

A.1 Aggregation of attributes-linked variables for Central Macedonia (NUTS-2 level)

For the variables within Farm Labor, Subsidies and Farm Assets categories we use the national aggregation scheme.

Table A.1: Aggregation of Crop Production for Central Macedonia case example

Code	Crop	National Coding
CM-X1	Common Wheat	X1
CM-X2	Durum Wheat	X2
CM-X3	Maize	X3
CM-X4	Other Cereals	X4
CM-X5	Rice	X5
CM-X6	Dry pulses and Protein Crops	X6
CM-X7	Cotton	X9
CM-X8	Oil Seeds	X10
CM-X9	Tobacco, Other Industrial, Flowers and Miscellaneous Crops	X7-X8 & X11
CM-X10	Vegetables	X12-X13
CM-X11	Green Plants, Pasture and Grazing	X14
CM-X12	Fruits, Berries and Nuts	X15
CM-X13	Olive Trees	X17
CM-X14	Grapes and Wine	X18-X20

Crop data include Xi.1: cultivated area; Xi.2: irrigated area; Xi.3: crop production; Xi.4: quantity sold; Xi.5: value of sales, where i=1,...,14. Citrus fruit production is negligible in the region and therefore it was excluded.

Table A.2: Aggregation of Animal Products variables for Central Macedonia case example

Code	Product	National Coding
CM-Y1	All types of meat	Y2-Y4
CM-Z1	All types of milk	Z1-Z3

Meat production: Yi.1: Weighted average of livestock; Yi.3: Value of sold animals; Yi.5: Value of slaughtered animals; Yi.7 Value of animals for breeding. **Milk production:** Z1: Total production.

Table A.3: Aggregation of Other Farm Income variables for Central Macedonia case example

Code	Product	National Coding
CM-M1	Other Farm Income	Y1, Y5, Z4-Z7, M1-M5

Data include Yi.3: Value of sold animals; Zi.3: Value of sales.

Table A.4: Aggregation of Variable Inputs Cost variables for Central Macedonia case example

Code	Attribute	National Coding
CM-V1	Wages on Hired Labour	V1
CM-V2	Contract Labour	V2
CM-V3	Machinery	V3
CM-V4	Livestock Cost	V4-V5
CM-V5	Seeds	V6
CM-V6	Fertilisers and Manure	V7
CM-V7	Protection	V8
CM-V8	Irrigation Water	V9
CM-V9	Energy	V10
CM-V10	Other Farm Cost	V11
CM-V11	Farm Overheads	V12

A.2 Aggregation of attributes-linked variables for Thessaloniki (NUTS-3 level)

Table A.1: Aggregation of Crop Production variables in the Thessaloniki case example

Code	Crop	National Coding
TH-X1	Common Wheat	X1
TH-X2	Durum Wheat	X2
TH-X3	Maize	X3
TH-X4	Other Cereals	X4
TH-X5	Rice	X5
TH-X6	Dry pulses and Protein Crops	X6
TH-X7	Cotton	X9
TH-X8	Oil Seeds	X10
TH-X9	Tobacco, Other Industrial, Flowers and Miscellaneous Crops	X7-X8 & X11
TH-X10	Vegetables	X12-X13
TH-X11	Green Plants, Pasture and Grazing	X14
TH-X12	Olive Trees, Fruits, Berries and Nuts	X15 & X17
TH-X13	Grapes and Wine[1]	X18-X20

Crop data include Xi.1: cultivated area; Xi.2: irrigated area; Xi.3: crop production; Xi.4: quantity sold; Xi.5: value of sales, where i=1,...,13. Citrus fruit production is negligible in the region and therefore it was excluded.

[1] Only information on Xi.1, Xi.2 and Xi.5.

A.3 Aggregation of attributes-linked variables for Thessalia (NUTS-2 level)

Table A.1: Aggregation of Crop Production variables for the Thessalia case example

Code	Crop	National Coding
TL-X1	Other Cereals	X1 & X4
TL-X2	Durum Wheat	X2
TL-X3	Maize	X3
TL-X4	Potatoes, Protein Crops and Rice	X5-X7
TL-X5	Cotton	X9
TL-X6	Tobacco, Oil Seeds, Industrial Crops and Vegetables	X8 & X10-X13
TL-X7	Green Plants, Pasture and Grazing	X14
TL-X8	Fruits, Berries and Nuts	X15-X16
TL-X9	Olive Trees	X17
TL-X10	Grapes and Wine[1]	X18-X20

Crop data include Xi.1: cultivated area; Xi.2: irrigated area; Xi.3: crop production; Xi.4: quantity sold; Xi.5: value of sales, where i=1,...,10. Citrus fruit production is negligible in the region and therefore it was excluded.

[1] Only information on Xi.1, Xi.2 and Xi.5.

Table A.2: Aggregation of Animal Products variables for Thessalia case example

Code	Product	National Coding
TL-Y1	All types of meat	Y2-Y4
TL-Z1	All types of milk	Z1-Z3

Meat production: Yi.1: Weighted average of livestock; Yi.3: Value of sold animals; Yi.5: Value of slaughtered animals; Yi.7 Value of animals for breeding. **Milk production:** Z1: Total production.

Table A.3: Aggregation for Other Farm Income variables for Thessalia case example

Code	Product	National Coding
TL-M1	Other Farm Income	Y1, Y5, Z4-Z7, M1-M5

Data include Yi.3: Value of sold animals; Zi.3: Value of sales.

Table A.4: Aggregation of Variable Inputs Cost variables for Thessalia case example

Code	Attribute	National Coding
TL-V1	Wages on Hired Labour	V1
TL-V2	Contract Labour	V2
TL-V3	Machinery	V3
TL-V4	Livestock Cost	V4-V5
TL-V5	Seeds	V6
TL-V6	Fertilisers and Manure	V7
TL-V7	Protection	V8
TL-V8	Irrigation Water	V9
TL-V9	Energy	V10
TL-V10	Other Farm Cost	V11
TL-V11	Farm Overheads	V12

A.4 Aggregation of attributes-linked variables for Peloponnisos (NUTS-2 level)

Table A.1: Aggregation of Crop Production variables for the Peloponnisos case example

Code	Crop	National Coding
PL-X1	Cereals	X1 - X4
PL-X2	Potatoes, Protein Crops, Tobacco, Oil Seeds and Industrial Crops	X6-X11
PL-X3	Vegetables	X3
PL-X4	Green Plants, Pasture and Grazing	X14
PL-X5	Fruits, Berries and Nuts	X15
PL-X6	Citrus Fruits	X16
PL-X7	Olive Trees	X17
PL-X8	Grapes and Wine[1]	X18-X20

Crop data include Xi.1: cultivated area; Xi.2: irrigated area; Xi.3: crop production; Xi.4: quantity sold; Xi.5: value of sales, where i=1,...,8. Citrus fruit production is negligible in the region and therefore it was excluded.
[1] Only information on Xi.1, Xi.2 and Xi.5.

Table A.2: Aggregation of Animal Products variables for the Peloponnisos case example

Code	Product	National Coding
PL-Y1	All types of meat	Y2-Y4
PL-Z1	All types of milk	Z1-Z3

Meat production: Yi.1: Weighted average of livestock; Yi.3: Value of sold animals; Yi.5: Value of slaughtered animals; Yi.7 Value of animals for breeding. **Milk production:** Z1: Total production.

Table A.3: Aggregation of Other Farm Income variables for Peloponnisos case example

Code	Product	National Coding
PL-M1	Other Farm Income	Y1, Y5, Z4-Z7, M1-M5

Data include Yi.3: Value of sold animals; Zi.3: Value of sales.

Table A.4: Aggregation of Variable Inputs Cost variables for Peloponnisos case example

Code	Attribute	National Coding
PL-V1	Wages on Hired Labour	V1
PL-V2	Contract Labour	V2
PL-V3	Machinery	V3
PL-V4	Livestock Cost	V4-V5
PL-V5	Seeds	V6
PL-V6	Fertilisers and Manure	V7
PL-V7	Protection	V8
PL-V8	Irrigation Water	V9
PL-V9	Energy	V10
PL-V10	Other Farm Cost	V11
PL-V11	Farm Overheads	V12

Appendix B Greek FADN variables common for all case examples

Table B.1: Structural Characteristics

Attribute	FADN Coding and Definitions		Code
Latitude	20	Latitude in degrees	C1
Longitude	30	Longitude in degrees	C2
Size Class	90	Economic size class	C3
Irrigation system	210	Main irrigation system used on the farm	C4
Owned UAA	10	farm is the owner, lifelong tenant or leaseholder	C5
Rented UAA	20	Land not owned by the holder for which a fixed rent is paid	C6
Sharecropped UAA	30	Land farmed jointly by the grantor	C7

Table B.2: Soil, Spatial and Climatic Data

Attribute	Information	Code
Human Influence Index (Direct human influence on ecosystems)	Values 0 - 51.6. Zero value represents no human influence and 64 represents maximum human influence possible, using all 8 measurements of human presence: Population Density/km^2, Score of Railroads, Score of Major Roads, Score of Navigable, Rivers, Score of Coastlines, Score of Nighttime Stable Lights Values, Urban Polygons, Land Cover Categories.	G1
Soil pH (CaCl2)	Values 0 - 7.5	G2
Topsoil organic carbon content	(SOC) content (%) in the surface horizon of soils Values 0-10.1	G3
Altitude	in meters. Values 0 - 1723.	G4
Slope	Values 0% - 70.2%. 100% is horizontal line.	G5
Coast distance	in meters. Values 0 - 135758.	G6
Erosion	% of land downgraded. Values 0 - 50.8	G7
Average Annual Temperature	in ^{o}C. Values 13.3 - 21.3	G8
Maximum Annual Temperature	in ^{o}C. Values 33.9 - 39.4	G9
Minimum Annual Temperature	in ^{o}C. Values -8.7 - 6.8	G10
Humidity	in %. Values 55.4 - 73.6	G11
Total Rainfall	in mm. Values 86.8 - 926.3	G12

Table B.3: Soil and Water Contamination

Attribute	FADN Coding and Definitions		Code
Nitrogen	3031	Quintals of N used in mineral fertilisers	Q1
Phosphorous Pentoxide	3032	Quintals of P2O5 used in mineral fertilisers	Q2
Potassium Oxide	3033	Quintals of K2O used in mineral fertilisers	Q3

Table B.4: Farm Labour

Attribute	FADN Coding and Definitions		Code
Manager	G	Gender	L1.1
Characteristics[1]	B	Age	L1.2
	T	Training	L1.3
	Y1	Hours worked annually	L1.4
	W1	Number of Annual Work Units (AWU)	L1.5
	W2	Share of work for OGA directly related to the holding	L1.6
Holder Characteristics	Y1	Hours worked annually	L2.1
	W1	Number of Annual Work Units (AWU)	L2.2
Unpaid Labour	40	Y1 Annual time worked	
	50	Y1 Annual time worked	L3.1
	60	Y1 Annual time worked	
	40	Y2 % of annual time worked	
	50	Y2 % of annual time worked	L3.2
	60	Y2 % of annual time worked	
Paid Labour	50	Y1 Annual time worked	
	60	Y1 Annual time worked	L4.1
	70	Y1 Annual time worked	
	50	Y2 % of annual time worked	
	60	Y2 % of annual time worked	L4.2
	70	Y2 % of annual time worked[1]	
Household Size	40	Spouse of holder	L5
	50	Other unpaid	

[1] When manager is paid labourer, we report W2 not Y2.

Table B.5: Crop Production

Variable	FADN	Coding and Definitions	Code
Common Wheat	10110	Common wheat and spelt	X1.1-X1.5
Durum Wheat	10120	Durum wheat	X2.1-X2.5
Maize	10160	Grain maize	X3.1-X3.5
Other Cereals	10130	Rye	
	10140	Barley	X4.1-X4.5
	10150	Oats	
	10190	Other cereals for grain production	
Rice	10170	Rice	X5.1-X5.5
Dry pulses and	10210	Peas, field beans and sweet lupines	
Protein Crops	10220	Lentils, chickpeas and vetches	X6.1-X6.5
	10290	Other protein crops	
Potatoes and	10300	Potatoes	
Root Crops	10310	Potatoes for starch	X7.1-X7.5
	10390	Other potatoes	
	10400	Sugar beet	
	10500	Other fodder roots and brassicats	
Tobacco	10601	Tobacco	X8.1-X8.5
Cotton	10603	Cotton	X9.1-X9.5
Oil Seeds	10605	Sunflower	
	10604	Rape and turnip rape	
	10606	Soya	X10.1-X10.5
	10607	Linseed[1]	
	10608	Other oil seed crops	
Other Industrial and	10609	Flax[1]	
Miscellaneous Crops	10610	Hemp	
	10611	Other fiber plants[1]	
	10602	Hops[1]	
	10612	Aromatic, medical & cullinary	
	10690	Other industrial crops	
	10613	Sugar cane[1]	X11.1-X11.5
	10810	Open field flower and ornamental plants	
	10820	Greenhouse flower and ornamental plants	
	40500	Nurseries	
	40600	Other permanent crops	
	60000	Mushrooms[1]	
	40610	Christmas trees[1]	
	40700	Permanent crops under glass[1]	
Vegetables (open field)	10711	Fresh vegetables, melons and strawberries	X12.1-X12.5
	10712	Market gardening	
Vegetables (greenhouses)	10720	Fresh vegetables, melons and strawberries	X13.1-X13.5
Green Plants, Pasture	10910	Temporary grass	X14.1-X14.5
and Grazing	10921	Green maize	
	10922	Leguminous plants	
	10923	Other green plants	
	11000	Seed and seeding	
	11100	Other arable land crops	
	11210	Fallow land without subsidies	X14.1-X14.2, X14.5
	30100	Pasture and meadow	X14.1-X14.3, X14.5
	30200	Rough grazing	X14.1-X14.3, X14.5
Fruits, berries and nuts	40111	Apples	

continued....

X15.1-X15.5

Variable	FADN Coding and Definitions		Code
	40112	Pears	
	40113	Peaches and nectarines	
	40114	Other fruit of temperate zones	
	40115	Subtropical or tropical fruits	
	40120	Berry species	
	40130	Nuts	
Citrus Fruits	40210	Oranges	X16.1-X16.5
	40220	Tangerines, mandarins & clementines	
	40230	Lemons	
	40290	Other citrus fruit	
Olive Trees	40310	Table olives	X17.1-X17.5
	40320	Olives for oil production	
	40330	Olive-oil	
	40340	Olive by-products	X17.3-X17.5
Grapes for wine	40451	Grapes for wine PDO	X18.1-X18.5
	40452	Grapes for wine PGI	
	40460	Grapes for other wine	
	40470	Miscellaneous products of vines	X18.3-X18.5
	40480	Vine by-products	X18.3-X18.5
Table grapes and raisins	40430	Table grapes	X19.1-X19.5
	40440	Raisins	
Wines	40411	Wine PDO	X20.1-X20.5
	40412	Wine PGI	
	40420	Other wines	

Xi.1: cultivated area; Xi.2: irrigated area; Xi.3: crop production; Xi.4: quantity sold; Xi.5: value of sales

[1] Non applicable in the Greek FADN dataset.

Table B.6: Livestock Production

Variable	FADN Coding and Definitions		Code
Equidae	100	Equidae	Y1.1-Y1.3, Y1.6-Y1.7
Bovine	210	Bovine animals <1 yr old male-female	Y2.1-Y2.7
	220	Bovine animals 1-2 yr old male	
	230	Bovine animals 1-2 yr old female	
	240	Male bovine animals >2 yr old	
	269	Other cows	
	261	Dairy cows	
	252	Heifers for fattening	Y2.1-Y2.5
	262	Buffalo cows	Y2.1-Y2.5
	251	Breeding heifers	Y2.1-Y2.3, Y2.6-Y2.7
Sheep and Goats	311	Ewes, breeding females	Y3.1-Y3.7
	319	Other sheep	
	321	Goats, breeding females	
	329	Other goats	
Pigs, Poultry etc	410	Piglets having weight <20 Kgs	Y4.1-Y4.7
	420	Breeding sows having weight >50 Kgs	
	491	Pigs for fattening	
	499	Other pigs	
	510	Poultry-boilers	Y4.1-Y4.5
	520	Laying hens	
	530	Other poultry	
	610	Rabbits, breeding females	Y4.1
	699	Other rabbits	Y4.1-Y4.5
Bees	700	Bees	Y5.1-Y5.3

Yi.1: No of animals; Yi.2: No of animals sold; Yi.3: Value of sold animals; Yi4: No of animals for slaughtering; Yi.5: Value of slaughtered animals; Yi.6: No of animals for rearing-breeding; Yi.7: Value of animals for rearing-breeding.

Table B.7: Animal Products

Variable	FADN Coding and Definitions		Code
Cow milk	261	Cows' milk	Z1.1-Z1.3
	262	Buffalo's cows' milk	
Sheep milk	311	Sheep milk	Z2.1-Z2.3
Goat milk	321	Goat's milk	Z3.1-Z3.3
Wool	330	Wool	Z4.1-Z4.3
Eggs	531	Eggs for consumption	Z5.1-Z5.3
	532	Eggs for hatching	
Honey	700	Honey and products of bee-keeping	Z6.1-Z6.3
Manure	800	Manure	Z7.3

Zi.1: total production; Zi.2: production sold; Zi.3: value of sales.

Table B.8: Values of Sales of Other Farm Income Sources

Variable	FADN	Coding and Definitions	Code
Income from Land	11210	Fallow land without subsidies	
	11300	Leased land	
	90100	Receipts from renting out land	
	90200	Compensation by crop insurance	
	90300	By-products other than olive and vine	M1
	90310	Straw	
	90320	Sugar beet tops[1]	
	90330	Other by-products	
	90900	Other	
Income from Livestock	1100	Contract rearing[1]	
	1120	Cattle under contract[1]	
	1130	Sheep and goats under contract[1]	
	1140	Pigs under contract[1]	M2
	1150	Poultry under contract[1]	
	1190	Other animals under contract[1]	
	1200	Other animal services	
Food Processing	261	Processing of cow's milk	
	262	Processing of buffalo's milk[1]	
	311	Processing of sheep's milk	
	321	Processing of goat's milk	M3
	900	Processing of meat or other animal products[1]	
	1010	Processing of crop	
	1020	Forestry and wood processing	
Contractual work	2010	Contract work for others	M4
Other Income Sources	2020	Tourism, accommodation, catering etc.	
	2030	Production of renewable energy	M5
	9000	Other gainful activities related to farm	

[1] Non applicable in the Greek FADN dataset.

Table B.9: Subsidies and Grants

Variable	FADN	Coding and Definitions	Code
Decoupled Payments	1150	Basic payment scheme	S1
	1200	Single area payment scheme[1]	
	1300	Redistributive payment[1]	
	1400	Practices beneficial for environment	
	1500	Payment for areas with natural constraints[1]	
	1600	Payment for young farms	
	1700	Small farms scheme	
Coupled Support on:			S2
Crops	23111	Cereals	
	23112	Oilseeds[1]	
	23113	Protein crops	
	2312	Potatoes[1]	
	23121	Starch potatoes[1]	
	2313	Sugar beet	
Industrial Crops	23141	Flax[1]	
	23142	Hemp[1]	
	23143	Hops[1]	
	23144	Sugar cane[1]	
	23145	Chicory[1]	
	23149	Other industrial crops	
	2315	Vegetables	
	2316	Fallow land[1]	
	2317	Rice	
	2318	Grain legumes	
	2319	Arable crops not defined[1]	
	2320	Permanent grassland[1]	
	2321	Dried fodder	
	2322	Crop specific payment for cotton	
	2323	National program for cotton[1]	
	2324	Seed production	
Permanent Crops	23311	Berries[1]	
	23312	Nuts	
	2332	Pome and stone fruit	
	2333	Citrus plantations	
	2334	Olive plantations[1]	
	2335	Vineyards	
	2339	Other permanent crops[1]	
Animals	2341	Dairy[1]	
	2342	Beef and veal	
	2343	Cattle (type not specified)[1]	
	2344	Sheep and goat	
	2345	Pigs and poultry[1]	
	2346	Silkworms	
	2349	Other animals[1]	
	2410	Short rotation coppices[1]	
	2490	Other coupled payments	
Exceptional Support and Rural Development	2810	Disaster payments[1]	
	2890	Other grants and subsidies	
	2900	Other direct payments	

continued....

Variable	FADN Coding and Definitions		Code
	3100	Agriculture	
	3300	Agri-environment-climate and animal welfare	
	3350	Organic farming	
	3400	Natura 2000 and WFD[1]	
	3500	Areas facing natural constraints	
	3610	Viability of forests	
	3620	Forest conservation[1]	
	3750	Restoration of agricultural products[1]	
	3900	Other	
Subsidies on Cost[1]	4100	Wages and social security	S4
	4200	Motor fuels	
	4310	Livestock	
	4320	Feed and grazing livestock	
	4330	Other livestock costs	
	4410	Seeds	
	4420	Fertilisers	
	4430	Crop protection	

[1] Non applicable in the Greek FADN dataset.

Table B.10: Closing Valuation of Farm Assets

Variable	FADN Coding and Definitions		Code
Current Asset	1010	Cash and equivalents	A1
	1020	Receivables	
	1030	Other current assets	
	1040	Inventories	
	2010	Biological assets - plants	
Land	3010	Agricultural land	A2
	3020	Land improvements	
	5010	Forest land	
Buildings and Machinery	3030	Farm buildings	A3
	4010	Machinery and equipment	
Non-current Assets	7010	Intangible assets, tradable	A4
	7020	Intangible assets, non-tradable	
	8010	Other non-current assets[1]	

[1] Non applicable in the Greek FADN dataset.

Table B.11: Variable Inputs Cost

Variable	FADN Coding and Definitions		Code
Wages on Hired Labour	1010	Wages and social security costs for paid labour	V1
Contract Labor	1020	Contract work and machinery hire	V2
Machinery	1030	Current upkeep of machinery and equipment	V3
	1040	Motor fuels and lubricants	
	1050	Car expenses	
Feedstuff	2010	Feedstuffs for grazing stock	V4
	2020	Purchased coarse fodder for grazing stock	
	2030	Purchased feedstuffs for pigs	
	2040	Purchased feedstuffs for small animals	
	2050	Purchased feedstuffs for for grazing stock	
	2060	Farm-produced feedstuffs for pigs	
	2070	Farm-produced feedstuffs for small animals	
Livestock Cost	2080	Veterinary expenses	V5
	2090	Other specific livestock costs	
Seeds	3010	Seeds and seedlings purchased	V6
	3020	Seeds and seedlings produced and used on the farm	
Fertilizers and Manure	3030	Fertilisers and soil improvers	V7
	3034	Purchased manure	
Protection	3040	Crop protection products	V8
Irrigation Water	5040	Irrigation water cost	V9
Energy	5020	Electricity	V10
	5030	Heating fuels	
Other Farm Cost	3090	Other specific crop costs	V11
	4010	Costs for forestry and wood processing	
	4020	Costs for crop processing	
	4030	Costs for cow's milk processing	
	4040	Costs for buffalo's milk processing[1]	
	4050	Costs for sheep's milk processing	
	4060	Costs for goat's milk processing	
	4070	Costs for animal meat processing	
	4090	Costs for other gainful activities	
Farm Overheads	5010	Current upkeep of land and buildings	V12
	5051	Agricultural insurance	
	5055	Other farm insurance	
	5061	Taxes and other dues	
	5062	Taxes on land and buildings	
	5070	Rent paid, total	
	5080	Interest and financial charges paid	
	5090	Other farming overheads	

[1] Non applicable in the Greek FADN dataset.