RESEARCH PROGRAM ON Integrated Systems for the Humid Tropics

CGIAR

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Typology construction, a way of dealing with farm diversity

General guidelines for Humidtropics

Stéphanie Alvarez
Wim Paas
Katrien Descheemaeker
Pablo Tittonell
Jeroen Groot

Wageningen, The Netherlands
2014

Humidtropics, a CGIAR Research Program led by IITA, seeks to transform the lives of the rural poor in tropical America, Asia and Africa. Research organizations involved in core partnership with Humidtropics are AVRDC, Bioversity International, CIAT, CIP, FARA, icipe, ICRAF, ILRI, IITA, IWMI and WUR. humidtropics.cgiar.org

Published by Humidtropics
http://humidtropics.cgiar.org

Acknowledgements

We acknowledge Humidtropics, a CGIAR Research Program that aims to develop new opportunities for improved livelihoods in a sustainable environment, for partially funding this work.

Humidtropics is a CGIAR Research Program led by IITA. It seeks to transform the lives of the rural poor in tropical America, Asia and Africa, and uses integrated systems research and unique partnership platforms for impact on poverty and ecosystems integrity. Research organizations involved in core partnership with Humidtropics are AVRDC, Bioversity International, CIAT, CIP, FARA, *icipe,* ICRAF, IITA, ILRI, IWMI, and WUR. humidtropics.cgiar.org

CGIAR is a global agricultural research partnership for a food secure future. Its science is carried out by the 15 research centers who are members of the CGIAR Consortium in collaboration with hundreds of partner organizations. www.cgiar.org

The mission of Wageningen UR (University & Research centre) is 'To explore the potential of nature to improve the quality of life'. Within Wageningen UR, nine specialised research institutes of the DLO Foundation have joined forces with Wageningen University to help answer the most important questions in the domain of healthy food and living environment. With approximately 30 locations, 6,000 members of staff and 9,000 students, Wageningen UR is one of the leading organisations in its domain worldwide. The integral approach to problems and the cooperation between the various disciplines are at the heart of the unique Wageningen Approach. www.wageningenur.nl/en.htm

Reference

Alvarez, S., Paas, W., Descheemaeker, K., Tittonell, P., Groot, J.C.J., 2014. *Constructing typologies, a way to deal with farm diversity: general guidelines for the Humidtropics.* Report for the CGIAR Research Program on Integrated Systems for the Humid Tropics. Plant Sciences Group, Wageningen University, the Netherlands.

Table of contents

Summary

To capture variability of farming systems, typologies are an often-used tool. Typologies in research for development projects are used to effectively derive best-fit farm adjustments, improved policies and innovations in order to meet Humidtropics' goals.

The objective of this document is to provide guidelines for the construction of typologies. The general typology approach proposed here is to combine expert knowledge obtained through a participative approach with multivariate statistics.

After briefly sharing some background information on typology making, the different steps of the typology construction are described. The steps are:

1. Precisely state the objective of the typology;
2. Formulate a hypothesis on farming systems diversity;
3. Select the variables characterizing the farming systems;
4. Design a sampling method for data collection;
5. Cluster the farming systems using multivariate statistics;
6. Compare the typology result with the hypothesis and validate the typology with local experts.

In the appendices, detailed examples are given of the statistical methods that can be applied for typology construction.

1 Introduction

Farming systems in the humid tropics exist across a wide variety of cultures and landscapes. The biophysical, institutional, social and economic drivers differ between contexts, resulting in different responses of farmers and communities between and within areas. Moreover, farms are in different development stages, and farmers have different skills and ambitions. Over time, these differences in drivers and in farm features lead to temporal and spatial variability between and within farming systems.

The existing farming systems variability is challenging to fully comprehend, leading to partial representation of reality. Various tools and methods (e.g. wealth rankings, farm typologies, distributions) have been developed to understand and deal with farming systems diversity. When using these tools and methods, a trade-off is made between the quality of representing reality and the level of detail required. An often-used methodology to deal with variability and diversity is typology construction, i.e. the grouping of farms/households.

In this document we provide guidelines for constructing typologies. Our objectives are to present a step-wise approach to structure this process, to provide practical advices and suggestions on useful techniques and to raise awareness of opportunities and pitfalls that arise during typology construction.

1.1 Purposes

Typologies respond to research questions that require **taking into account the agricultural heterogeneity within a region** (e.g. Alary et al., 2002; Righi et al., 2011; cf. Figure 1).
The four main reasons to develop a typology are:

1. *Targeting*: the distinction between farming systems is aimed at identifying appropriate interventions per farming system type;
2. *Scaling-out*: typologies contribute to understanding how appropriate interventions can be disseminated at a large scale;
3. *Selection*: typologies support the selection of representative farms or the formulation of (average) prototype farms for detailed analyses.
4. *Scaling-up*: typologies support the extrapolation of ex-ante impact assessments to larger spatial or organizational scales (Ewert et al., 2011).

1.2 Methods

Farm typologies can be constructed using various methods:

- *Step by step comparison of farm functioning* (Capillon, 1993; Landais, 1998): for a delimited area, this classification method is based on extensive data about farm functioning (family, objectives, history, productions, management, techno-economic results, biophysical constraints, etc.), which can be obtained from surveys of a stratified sample of farms. The grouping into types is made using a "step by step" comparison of neighbouring farms (for more details on the method, see Landais, 1998).

- *Expert knowledge*: the typology construction is based on aggregating farms in clusters defined by local experts, key informants, or farmers (Giller et al., 2011; Landais, 1998; Pacini et al., 2013). This approach leads to the establishment of a common reference base (Landais, 1998). Generally, the typology approach based on expert knowledge requires little time and costs (Landais, 1998).

- *Participatory rankings*: the ranking of households, mostly according to wealth (wealth ranking), by experts and/or farmers themselves in a participatory process. Observable assets are important when ranking is based on wealth (Kebede, 2007).

- *Multivariate analysis including ordination and clustering methods:* this method can be seen as the quantitative equivalent of the 'expert knowledge approach'. Statistical methods (e.g. Principal

components analysis, Multiple correspondence analysis, Multiple factorial analysis, Multidimensional scaling) are used to classify objects (here farms). In the ideal case no hierarchy or preconceptions are projected on the objects (Alary et al., 2002, Giller et al. 2011). This kind of methods are also called 'dimension reduction' or 'data-reduction' techniques (Pacini et al., 2013) because they have the advantage of capturing the complexity of farming systems through taking into account, at the same time, numerous farm dimensions and then highlighting a few dimensions that are more explanatory of farm diversity (Alary et al., 2002).

In projects we have to meet at least two important standards: (i) the standards of science in which accuracy, objectivity and reproducibility are important, and (ii) the standards of project outcomes, which are dependent on different needs, perceptions, interests, etc. of stakeholders.

Multivariate statistics methods are often preferred over expert knowledge based approaches because of the reproducibility inherent to their statistical foundations (Pacini et al., 2013), contributing mostly to standard (i). However, to also contribute to standard (ii) and enhance the success of projects, typologies have to be relevant to stakeholders. Therefore, the different typology methods could be used in a complementary way, here using multivariate statistics in addition to participatory approaches (Alary et al., 2002; Pacini et al., 2013; Righi et al., 2011).

Use multivariate statistics and expert knowledge in a complementary way

2 Typology guidelines

The structure of the typology construction framework is presented in Figure 1. It comprises six steps to go from a heterogeneous population of farms to the grouping into coherent farm types. The six steps are:

1. Precisely state the objective of the typology;
2. Formulate a hypothesis on farming systems diversity;
3. Select the variables characterizing the farming systems;
4. Design a sampling method for data collection;
5. Cluster the farming systems using multivariate statistics;
6. Compare the typology result with the hypothesis and validate the typology with local experts.

2.1 Typology objectives

A **farm typology is dependent on the research question** (e.g. Köbrich et al., 2003). Typologies can be constructed for a specific research objective for a specific area (e.g. "to improve forage supply in the highlands of Madagascar") or for a global objective for a broad zone (e.g. "to improve food security in the humid tropics"). In both cases, keeping in mind the objective is important when a typology is constructed and in particular during the selection of variables (Figure 1).

Farms are moving targets (Giller et al., 2011), while **typologies (based on one-time measurements) give a snapshot of farm situations at a certain period of time** (Kostrowicki, 1977). Because of farm dynamics, typologies could quickly become obsolete and hence it is preferable to regularly update typologies (Landais, 1998; Valbuena et al., 2014). Therefore, typologies should be continuously evaluated and updated .

Continuously evaluate and update your typology

Another point of attention is that we might face data scarcity and time constraints. In that case a "simple classification" based mainly on resource endowment (a so-called structural typology), might after all be the best option (Giller et al., 2011). This fast approach provides a starting point to further explore existing constraints and drivers.

2.2 Hypothesis on typology structure

As a starting point of the typology development, it is advised **to establish a hypothesis on the farm diversity of the studied area** (Tittonell, 2014a). The hypothesis can be structured using expert knowledge, participatory methods and/or previous studies in the area or field observation. The hypothesis should be related to the purpose of the typology construction and, preferably, based on agricultural knowledge and theories (Whatmore et al., 1987). The hypothesis may concern the number of farming system types, their main characteristics and their proportion in the studies area.

Make a hypothesis on farm diversity

The effectiveness of the typology development could be improved by the participation of local stakeholders in the hypothesis construction process (Righi et al., 2011). We propose to use participatory approaches, in which local stakeholders (local researchers, actors, farmers) are included in order to formulate the hypothesis together, resulting in an **ex-ante description of different farm types.** The hypothesis on the farm types should reflect the criteria selected by local stakeholders to describe the local farming systems (Alary et al., 2002; Pacini et al., 2013). These criteria

should be part of the questionnaire used for the farm surveys in the following step. An added advantage of including local stakeholders is that communication and involvement can be increased.

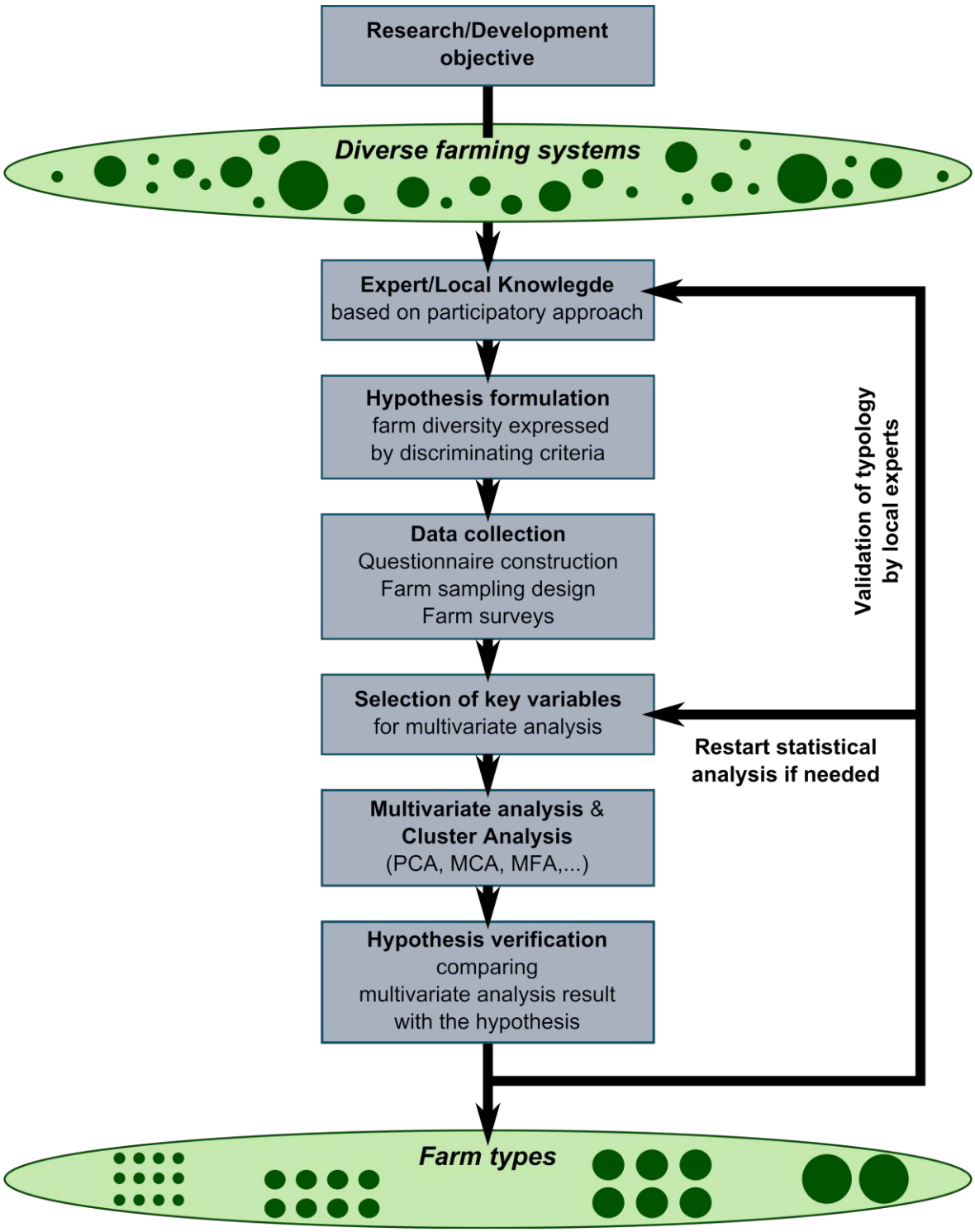

Figure 1: General framework of the typology process (PCA: Principal components analysis; MCA: Multiple correspondence analysis; MFA: Multiple Factorial Analysis)

2.3 Key variable selection for the statistical methods

Typologies could be grouped into two main classes: (i) **structural typologies** based mainly on variables that describe resources and asset levels, and, (ii) **functional typologies** based on variables that describe livelihood strategies and household dynamics (Tittonell, 2014a). The purpose of the typology should drive the typology development process, and hence **the variable**[1] **selection** (Pacini et al., 2013).

The collection of data from farms is an essential step in the typology construction. It is advised to design a survey questionnaire to capture the whole farming system (Giller et al., 2011; Tittonell et al., 2010). The variables used in the farm surveys could be grouped into specific categories, for example:

- Variables of structural characteristics and variables of farm functioning (Sanago et al., 2010);
- Farm resources availability and Management (Pacini et al., 2013)
- Biophysical resources, Socio-economic aspects and Equipment (Righi et al., 2011);
- Cultivation, Chemical inputs, Harvest, density and fruit quality, Economic resources, Farming system nature, Physical context, Personal ambition, Social, Performance Agronomical, Performance Economic and Performance Environmental (Blazy et al., 2009).

Here, in order to ensure a systematic approach, we advise to consider variables related to the main components of the farming system (i.e. household/family, cropping system, livestock system) and their interactions with the outside/environment (e.g. environmental context, economic context, socio-cultural context). As crop–livestock farming systems are the backbone of smallholder agriculture in developing countries (Thornton and Herrero, 2001), here we present an example of a variable set commonly used in the context of crop-livestock farming systems (Table 1). Naturally, the selection of key variables for the multivariate analysis should be adapted to the purpose of the typology, the area and the farming system context, following the first step of exchanges with local stakeholders and hypothesis formulation. Blazy et al. (2009), for instance, studied crop management innovations per farm type and therefore focused more on variables that indicate the technical nature of the crop management system, and less on the variables that are indicative for the social factors.

Before the **selection of the key variables for the multivariate analysis**, it is important to evaluate the required quality of the data (amount, accuracy). Household survey data in developing countries often is erroneous (Howe and McKay, 2007; examples are provided in Appendix A), undermining the reliability of the statistics. Therefore it is good to check the data and to choose certain variables wisely. A step of data control for the Principal Component Analysis is described in Appendix C.

[1] Variable as an attribute of farms in the farm population that we want to study

Table 1: Example of variables describing crop-livestock farming systems. The column 'common set' proposes a set of variables that could be included in typology making across sites in Humidtropics

Attr.	Category	Variable	Unit	Common set
R	Household	Family size	Capita	✓
R	Household	Household head age	year	
R	Household	Family labour on farm activities	capita or man-day/year [a]	✓
R	Household	Labour hired	capita or man-day/year [a]	✓
O	Household	Months food self-sufficiency	months	
R	Household	Total gross margin of the household	local currency	
R	Household-Environment	Off-farm activities	classes [b]	✓
R	Household-Environment	Total gross margin of the household (income - expenditure)	local currency	
R	Household-Environment	Off-farm income (% of the household income)	%	✓
O	Household-Environment	Food purchase (% of the household expenditure)	%	
R	Cropping system	Area owned by the household	ha	
R	Cropping system	Area farmed by the household [c]	ha	✓
R/O	Cropping system	Area with food crops	ha or % cropped area	
R/O	Cropping system	Area with fodder crops	ha or % cropped area	✓
R/O	Cropping system	Area with cash crops	ha or % cropped area	✓
O	Cropping system-Environment	Crop production sales (% of the income)	%	
O	Cropping system-Environment	Purchase of mineral or organic fertilizers (% of the expenditure)	%	
O	Cropping system-Environment	Purchase of pesticide (% of the expenditure)	%	
R	Livestock system	Total number of livestock	TLU	✓
R/O	Livestock system	Number of local cattle	no.	
R/O	Livestock system	Number of improved-bred cattle	no.	
R/O	Livestock system	Number of small ruminants	no.	
R/O	Livestock system	Number of small animal (pig and/or poultry)	no. or TLU	
O	Livestock system	Milk production	l/year	
O	Livestock system-Environment	Total animal products sales (% of the income)	%	
O	Livestock system-Environment	Manure sales (% of the income)	%	
O	Livestock system-Environment	Concentrate/Fodder purchases (% of the expenditure)	%	
O	Household-Environment	Production objective/strategies (withstanding exterior constraints)	classes [d]	✓

Attr.: Attribute; O: Orientations; R: Resources

a: man-day/year could allow to consider different kinds of labour (e.g. full time person, children, woman)

b: classes to be defined according to the set of the survey results (e.g. Yes/No, Agricultural/Non-agricultural, Agricultural/Urban)

c: farmed area could include cropped, grazing or plantation areas

d: classes to be defined according to the set of the survey results (e.g. Increase/Maintain/Decrease production, Increase/Improve/Diversify/Change production)

External variables (e.g. distance to the road, availability of market access) can be used **to explain the diversity** amongst farms/households, and it drivers. Some examples in literature combine on-farm variables and variables on the external environment in the actual typology making (e.g. Ansoms and McKay, 2010; Tittonell et al., 2010). The availability of external variables supplies an excellent opportunity to test theories on the drivers of diversity. However, it could be preferable not to use those external variables for the identification of the farm types (i.e. as key variables for the multivariate analysis) in order to distinguish more clearly the variables describing farming systems diversity and the ones explaining this diversity.

Distinguish between the variables that describe farm diversity and the indicators that explain diversity

It is advised to use "a small number" of key variables (Kostrowicki, 1977) and to make sure that the number of surveyed farms is at least five times larger than the number of key variables used for the multivariate analysis (Hair et al., 2010). Hence the number of key variables required and selected for the multivariate statistics could differ from the number of variables asked during the survey. Table 1 provides some variables that might be selected as key variables for the multivariate analysis. Furthermore, other integrative variables (e.g. ratio Labour/Land, ratio TLU/Fodder area, ratio Food crops area/Cropped area) could be calculated from the variables asked during the survey.

Use the ratio 5 observations (farms) for 1 key variable

The number of key variables used in multivariate statistics for a typology purpose is highly variable; from 5 to 46 variables, with an average of about 15 variables[2]. An overview of 21 studies[2] showed an average of 9 observations per farm variable. It is important to note that the selection of key variables introduces a degree of subjectivity in the typology making process.

It is important to keep in mind that, as a result of the multivariate analysis, not all the key variables fed into the multivariate analysis will necessarily be retained as discriminating variables. The **discriminating variables are the variables resulting from the multivariate analysis as describing best the farm variability** (cf. example on Appendix C). It may be noted that intermediate typologies per variable class can be constructed in order to gain more insight in the diversity per variable class (e.g. Maton et al., 2005; Moreno-Pérez et al., 2011).

2.4 Sampling

The farm sampling should cover the farm diversity of the studied area (Pacini et al., 2013). Thus the sampling should be elaborated based on the initial hypothesis, and notably on the expected farm types proportions. If the sampling is completely randomized, a large sampling size is necessary. To reduce the sampling size, methods to capture diversity along a gradient can be used (e.g. Y-shaped method described by Tittonell et al., 2010), or methods based on stratification or along transects (e.g. "transect following an intensification gradient" used by Pacini et al., 2013).

It is not recommended to ask to "all farmers" to come to a meeting place to make them fill the survey: the farm sample could be biased by the ability and/or motivation of farmers to come to the meeting place. Moreover, farm visits allow some additional checks, for instance on field area cultivated, crops grown, tools owned and livestock kept.

[2] Based on 21 typologies studies

Usually, for statistical reasons it advised to sample at least 50 farms (Hair et al., 2010). In practice the sample of farms for typology studies ranges from 18 farms to 2746 farms, with a median of 138 farms surveyed[3]. Besides, additional information on sampling methodology is given by Kumar (2014).
It is important to keep in mind that the size of the sample and the sampling method can impact on the proportion of farms belonging to each resulting farm type. For instance, when a sample contains 100 farms, a farm type that actually combines about 10% of the farming systems of the area may be represented by only one or two surveyed farms due to the sampling process (Hair et al., 2010). Moreover, during the multivariate process, these two farms may be considered as outliers or they may be combined in other farm types.

2.5 Multivariate statistics

Multivariate and cluster analysis are used **to identify explanatory variables (discriminating variables) and to group farms in homogeneous types**. Multivariate statistics allow reducing the number of variables and preserving the maximum of the total variability of the sample. According the nature of the selected key variables (quantitative and/or categorical) different multivariate statistics should be used:

- *Principal Components Analysis* (PCA) for quantitative (continuous or discrete) variables (e.g. Bidogeza et al., 2009; Sanogo et al., 2010; Tittonell et al., 2010) (cf. example of PCA in Appendix C);
- *Multiple Correspondence Analysis* (MCA) for categorical variables (e.g. Blazy et al., 2009);
- *Multiple Factorial Analysis* (MFA) for categorical variables organized in multi-table and multi-block data sets (Alary et al., 2002);
- *Hill and Smith Analysis* for mixed quantitative and qualitative variables (e.g. Rueff et al., 2012);
- *Multidimensional scaling* to build a classification configuration in a specific dimension (e.g. Pacini et al., 2013; Righi et al., 2011).

As mentioned previously, the first step of the multivariate analysis concerns the selection of key variables. It is necessary to check that a certain category of variables (i.e. a group of variables describing a same aspect of the system; cf. examples of categories in session 2.3) is not over-represented (i.e. the number of variables in this category is much larger than for the other categories); otherwise that could give more weight to this category of variables and so bias the analysis (Blazy et al., 2009; Kostrowicki, 1977).

Another precaution, specifically for PCA, is to standardize all the selected (quantitative) variables, using for example percentages, "to avoid the influence of different levels of variation due to the unit of measurement" (Pacini et al., 2013). This precaution improves the comparison of variables with different units, e.g. the cultivated area ranging from 0.5 to 2.0 ha, and the agricultural income ranging from 150 $/year to 5 000 $/year. The MCA and MFA methods are sensitive to low numbers of observations or unbalanced classes; hence here it could be necessary to combine some classes.

Furthermore, it is required to test the independence of variables, with for instance Pearson's Chi-squared Test. In fact, if two variables were strongly correlated it would give two times more weight in the multivariate analysis to the information given by these variables.

Moreover, MCA could be more difficult to interpret than the PCA or MFA analysis; the interpretability of MCA is higher when the number of selected variables is limited (preferably ≤ 20 variables; Hervé, 2011).

Multivariate analyses are quite sensitive to outliers (potential error or 'exceptional' observations), so it is advised to remove them from the analysis (Hair et al., 2010). If 'exceptional farms' are removed from the multivariate-analysis, as suggested (Appendix C), it may be useful to highlight them in a final report presenting the farming systems diversity.

[3] Median observed on a sample of 22 typology studies

The number of axes (i.e. principal component or factors) for the PCA, MCA and MFA can be determined according to a criterion that is fixed before the analysis (e.g. the number of axes that explain a minimum of x % of the variability – "usually 60% or higher" (Hair et al., 2010)) - or using the Kaiser's criterion for the PCA (i.e. all axis with an eigenvalue higher than 1 are chosen; Hervé, 2011).

Select the axes with a eigenvalue greater than 1.0 and/or axes that explain more than 60% of the variability

The Cluster Analysis (CA) aims to group farms into classes/types that are as "homogeneous" as possible. There are two main methods of CA commonly used:

- *Non-hierarchical clustering*, i.e. a separation of observations/farms space into disjoint groups/types where the number of groups (k) is fixed;
- *Hierarchical clustering*, i.e. a stepwise aggregation of observations/farms space into disjoint groups/types. First each farm is a group all by itself, and then at each step, the two most similar groups are merged until only one group with all farms remains. The visual result of these steps (algorithm) is a dendrogram or classification tree (Figure 2). The height of the dendrogram branches represents the average distance (dissimilarity) between the observations within the groups and between groups. Therefore, the dendrogram provides a visual representation of the variability of data and a useful tool for justifying the choice of a partition, i.e. the number of clusters. The choice of number of clusters is a trade-off between reducing dissimilarity and increasing the number of clusters. The partition of the dendrogram could be done based on: (i) the overall appearance of the dendrogram, (ii) the number of clusters and (iii) their interpretability, and (iv) the examination of the heights delta. Therefore, starting from the top of the dendrogram (the highest level of height or "root nodes"), the dendrogram structure suggests a division into *n* clusters when the decrease of the level of dissimilarity passing from a *(n-1)* clusters to *n* clusters (i.e. $\Delta Height_{(n-1) \text{ to } n \text{ clusters}}$) is much greater than passing from *n* clusters to *(n+1)* clusters (Husson et al., 2011). Finally, it is important to note, that despite the use of criteria to support the partition of the dendrogram, such as ΔHeight, subjectivity remains in the choice of the partition.
  The *Agglomerative Hierarchical Clustering* algorithm is often used in the typology construction process (e.g. Alary et al., 2002; Blazy et al., 2009; Pacini et al., 2013; Sanogo et al., 2010).

The two clustering methods can be used together to combine the strengths of the two approaches (Michielsens et al., 2002; Iraizoz et al., 2007). In the combination, hierarchical clustering is used to estimate the number of clusters, while the non-hierarchical clustering is used to calculate the cluster centres. The number of farm types typically ranges from 3 to 7, with a median of 5 farm types[4].

[4] Data observed on a sample of 20 typology studies

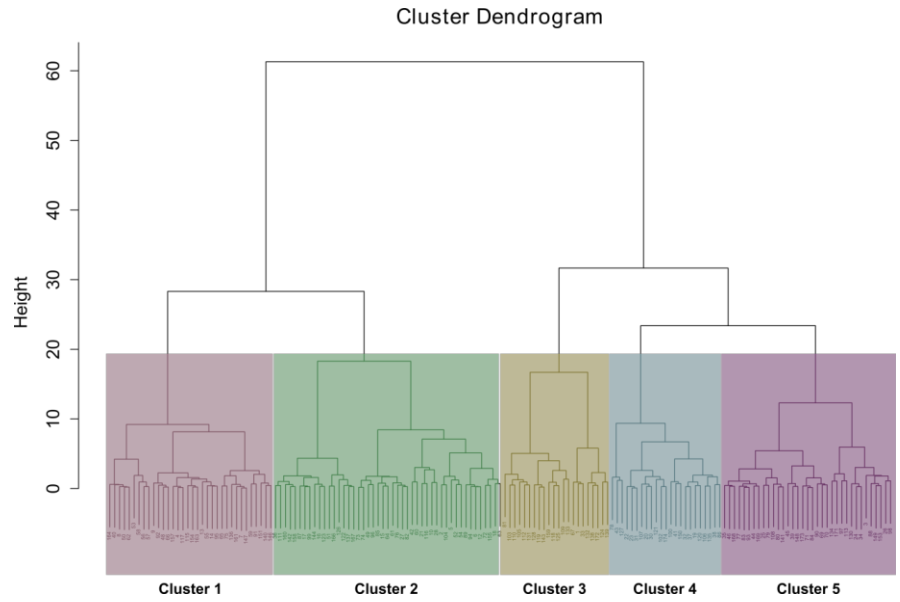

Figure 2: Example of a dendrogram from Agglomerative Hierarchical Clustering on 161 farms from Tanzania (Appendix C) showing five clusters (farm types).

The results from the multivariate analysis and the clustering analysis are used together **to interpret the meaning of each cluster** (Figure 3). A concrete example of PCA and CA is detailed in Appendix C.

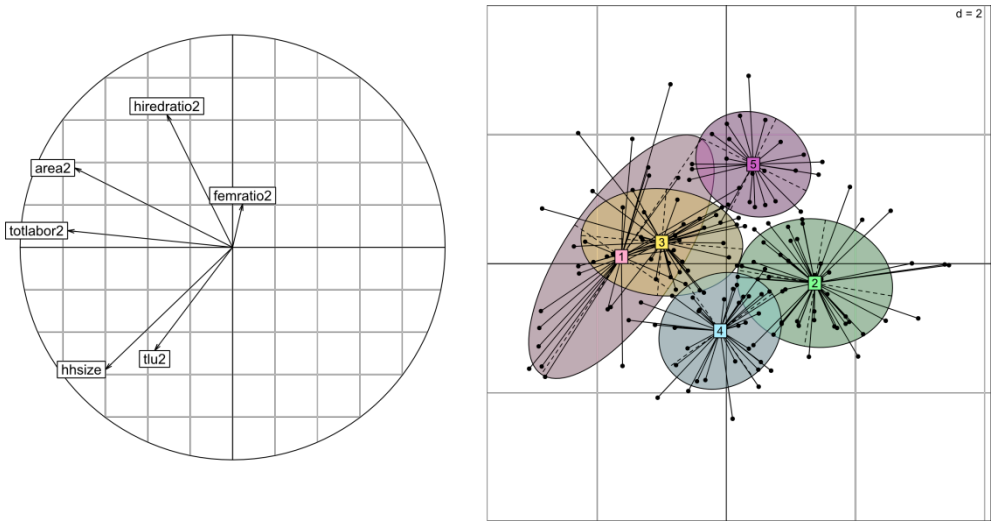

Figure 3: Example of results from the Principal Component Analysis and the Hierarchical Clustering on 161 farms from Tanzania (more details are provided in Appendix C).

2.6 Hypothesis verification

Farm types should be selected on the basis of their explanatory value, i.e. they have to be conceptually meaningful (Moreno-Pérez, 2011). The farm types resulting from the multivariate and cluster analysis should be compared with the initial hypothesis (cf. 2.2). It is necessary to discuss and to try to understand differences between the hypothesis and the results of the multivariate analysis. In case of unexpected results, the multivariate and cluster analysis may need to be repeated and/or the discussion and feedback sessions with local stakeholders may need to be re-initiated. Finally, a validation of the typology results by local experts and/or farmers is desired (Figure 1).

Furthermore, a comparison between the resulting farm types and findings from other research present an opportunity for gaining a better understanding of the agricultural sector.

3 Quick classification, dynamics and spatial distribution

3.1 Derivation of decision trees for quick farm classification

The typology results can be visualized using boxplots. Boxplots can support the farm type interpretation but also the identification of variables with distinctive power (Figure 4).

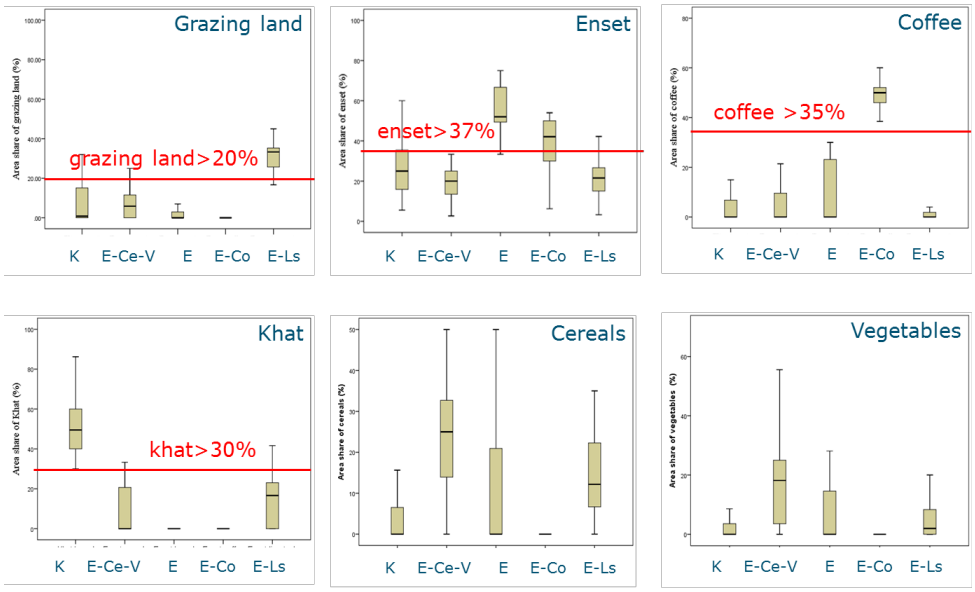

Figure 4: Boxplots of distinctive indicators for different farm types based on land-use (%) for certain crops. By following the selection order it is possible to distinguish farm types based on only a few indicators. Source: van de Ven and Melisse (2014). Letters on the X-axis represent crops (K for khat, E for enset, Ce for cereals, V for vegetables, Co for coffee) and livestock (Ls)

These distinctive indicators could be used to develop a classification tree providing thus a tool for a "quick classification" of additional farms based on a reduced amount of variables (here 4, area share of grazing land, of enset, of coffee and of khat). For instance, in Figure 4, more than 35% land used for coffee is distinctive for farm type E-Co, in a next step, more than 30% land used for khat is distinctive for farm type K, etc. (Figure 5).

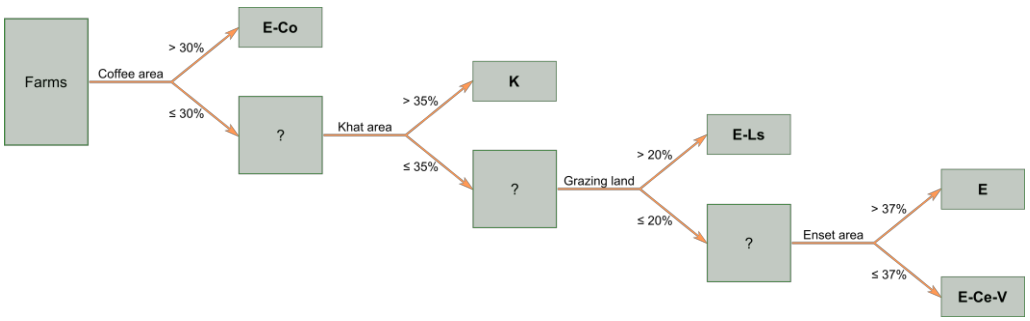

Figure 5: Example a "quick classification" using a classification tree. The classification tree was developed based on boxplots of the typology results and on the distinctive indicators identified (K=khat, E=enset, Ce=cereals, V=vegetables, Co= coffee and Ls=livestock)

3.2 Farm/household dynamics

As mentioned before, households/farms are recognized as moving targets (Giller et al., 2011). Hence, a farm/household survey and the resulting typology are snapshots in time (Kostrowicki, 1977; Laurent et al., 1999). Therefore typologies need to be employed with caution. It is good to know the temporal context in which the survey has been conducted in order to reflect on the representativeness of certain variables.

In literature there are some examples that take into account dynamics:

- The movements of individual farms can be projected into trajectories of farm types, which have been hypothesized to be predictable, i.e. following general trends (Laurent et al., 1999; Tittonell, 2014b). However, it should be kept in mind that individual farms might opt/be forced to follow different pathways (Valbuena et al., 2014).
- When the data of multiple years is available, the method of Falconnier et al. (submitted) is interesting. Falconnier et al. (submitted) conducted a multivariate analysis for the current situation. After that, they analysed the land use for crops, and constructed a decision tree for easy farm classification, like Melisse and van de Ven (2014; Figures 5 and 6). Based on the decision tree and a few variables from farm data from previous years, farms could be classified and compared with the current classification.
- Including knowledge of stakeholders/experts provide an opportunity to capture some of the farm dynamics, as they can evaluate farm types in the light of long term trends. Their expertise can help to trace the farms history/evolution leading to actual farm structure and so to the resulting farm types.
- For prioritizing within action sites and/or field sites and for comparisons between areas household types that are below the poverty line can be identified (Davis et al., 1997; Howe and McKay, 2007; Tittonell et al., 2010). In addition, identifying types that are situated in a poverty trap, reveals some of the (im)possibilities of households to escape poverty (Howe and McKay, 2007).
- The flows of resources (e.g. food, labour, money, knowledge) and interdependency between household types in a community are another indicator for the potential for dynamics. Laurent et al. (1999) present a flow diagram of resources between household types, which supports the development of an understanding of the situation.

3.3 Geo Information Systems

Farming is a spatial activity. Several articles plead for spatially linked farm typologies, i.e. connecting the farm types to their position in the landscape, notably for land-use planning (Landais, 1998; Carmona et al., 2010; Madry et al., 2013). However, conducting a spatial analysis requires a data-rich environment, which is often not the case in developing countries (Carmona et al., 2010). The use of the farms GPS-coordinates to map the farm types allows to assess the relationships between the types and the landscape elements, for instance roads. Figure 6 presents a map where farm types are represented by different coloured dots, helping to visualize the farm diversity, i.e. the spatial organisation of the different farm types.

A: Farm types description

Farm Types	*Main characteristics*
1	Medium arable land area, largest household and animal herd + ample off-farm activities
2	Largest arable land area, high legume ratio, large animal herd + market oriented
3	Large arable land area, highest legume- and small ruminant ratio + most on-farm labour
4	Medium land area, small animal herd, high legume ratio, most hired labour + market oriented
5	Smallest arable land area, highest maize ratio, lowest legume ratio + lowest off-farm income
6	Small arable land area, smallest animal herd + most livestock sales

B: Spatial organization of the farm types

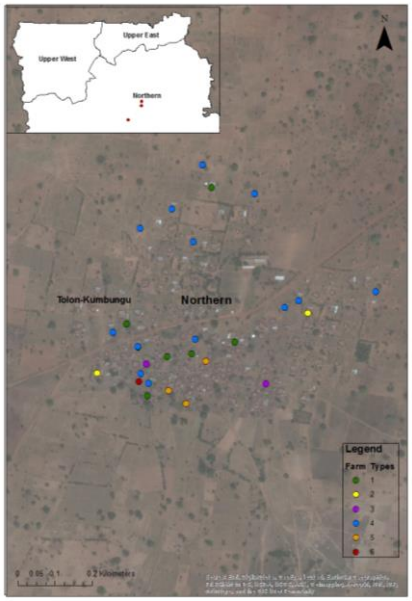

Figure 6: Spatial distribution of farm types in a village in Ghana's Northern Region. Source: Kuivanen (2014).

4 Including development outcomes in typologies

4.1 Food Security and Nutrition

As improved nutrition is an important program outcome of research for development projects, it is key to include indicators on food security nutrition in typology making, especially when the research question is specifically addressing food security and nutritional status of households. Undernourishment and hidden hunger are two forms of malnutrition.

Some studies include undernourishment by indicating the months of food security (e.g. Table 1). This could be further specified by asking in which month the household eats food from the farm (e.g. Tittonell et al., 2010), and from other sources (e.g. Ebanyat et al., 2010). Also the months in which households struggle to get food could be indicated. It is useful to have indicators that specify the 'other sources', which might be food-aid or food bought with salary from off-farm labour..

Hidden hunger refers to the deficiencies in diets, when calorie and/or protein intake are sufficient. Important deficiencies in diets in development countries include iron, iodide, vitamin A, zinc and folate deficiencies (Kennedy et al., 2003; Muthayya et al., 2013; both articles present global maps on nutrient deficiencies). Deficiencies are partly due to monotonous diets. Hence, the diversity of the diet is an important indicator. To assess the diversity of the diet, the Dietary Diversity Score (i.e., the number of certain food groups consumed by an individual or the household (Kennedy et al., 2011)) or the Functional Diversity (Remans et al., 2011) could be used. The adequacy of the diet can be evaluated by the balance of intakes versus requirements; the 24H-recall method is recommended to quantify the human intakes. Nutritional content of most tropical crops can be found via the websites of the FAO (1968).

4.2 Gender

We were not able to find many examples in literature where indicators on gender were explicitly included in typology study. De Lima Vidal (2013) developed a typology with only female household heads in an area in Brazil. Molua (2011) presents a study where the only determinant for the typology was the gender of the head of the household. Djurfeldt et al. (2008) found that gender issues in relation to farm productivity are context specific, and do not always depend on harshness of the environment, or affluence in a region. Hence, although simple, the distinction of Molua (2011) might be a first quick quantitative assessment of the relative importance of gender issues in a region.

In the current ImpactLite household survey of ILRI, attention has been given to gender aspects in the households. The gathered indicators give opportunities to evaluate the influence of gender in farm types. It would be interesting to see whether gender specific indicators within a household are associated with different farm structures and productivity levels. Especially, in regard to targeting technical interventions that improve working conditions and labour efficiency of women, it could be desirable to distinguish groups of households based on gender aspects. Possible variables could be: the female and children ratio in the household; the total male and female labour per farm; cattle, land and other assets owned by male and female.

References

Ansoms, A., McKay, A., 2010. A quantitative analysis of poverty and livelihood profiles: The case of rural Rwanda. Food Policy 35, 584-598.

Alary V., Messad S., Taché C., Tillard E., 2002. Approach to the diversity of dairy farm systems in Reunion. Revue Elev Méd Vét Pays Trop 55, 285-297.

Bidogeza, J. C., Berentsen, P. B. M., Graaff, J., Oude Lansink, A. G. J. M., 2009. A typology of farm households for the Umutara Province in Rwanda. 1(3): 321-335.

Blazy, J.-M., Ozier-Lafontaine, H., Doré, T., Thomas, A., Wery, J., 2009. A methodological framework that accounts for farm diversity in the prototyping of crop management systems. Application to banana-based systems in Guadeloupe. Agricultural Systems 101, 30-41.

Capillon, A., 1993. Typologie des exploitations agricoles, contribution à l'étude régionale des problèmes techniques. Doctoral thesis, INA P-G, Paris.

Carmona, A., Nahuelhual, L., Echeverría, C., Báez, A. 2010. Linking farming systems to landscape change: An emperical and spatially explicit study in southern Chile. Agriculture, Ecosystems and Environment 139: 40-50.

Davis, J., Mack, N., Kirke, A., 1997. New perspectives on farm household incomes. Journal of Rural Studies 13, 57-64.

De Lima Vidal, D., 2013. Work division in family farm production units: Feminine responsibilities typology in a semi-arid region of Brazil. Journal of Arid Environments 97, 242-252.

Djurfeldt, G., Larsson, R., Holmquist, B., Jirström, M. and Andersson, A., 2008. African farm dynamics and the sub- continental food crisis – the case of maize. Food Economics, Acta Agriculturae Scandanavica C 5, 75–91.

Ebanyat, P., De Ridder, N., De Jager, A., Delve, R.J., Bekunda, M.A., Giller, K.E., 2010. Drivers of land use change and household determinants of sustainability in smallholder farming systems of Eastern Uganda. Population & Environment 31, 474-506.

Ewert, F., Van Ittersum, M.K., Heckelei, T., Therond, O., Bezlepkina, I. and Andersen, E., (2011) Scale changes and model linking methods for integrated assessment of agri-environmental systems. "Agriculture, Ecosystems and Environment" 142, 6–17.

Falconnier, G., Descheemaeker, K., van Mourik, T.A., Sanogo, O.M., Giller, K.E., *summited.* Two decades of change in southern Mali: understanding farm trajectories and development pathways. Submitted to Agricultural Systems.

FAO, 1968. Food Composition for use in Africa. http://www.fao.org/docrep/003/X6877E/X6877E00.htm

Giller, K.E., Tittonell, P., Rufino, M.C., van Wijk, M.T., Zingore, S., Mapfumo, P., Adjei-Nsiah, S., Herrero, M., Chikowo, R., Corbeels, M., Rowe, E.C., Baijukya, F., Mwijage, A., Smith, J., Yeboah, E., van der Burg, W.J., Sanogo, O.M., Misiko, M., de Ridder, N., Karanja, S., Kaizzi, C., K'ungu, J., Mwale, M., Nwaga, D., Pacini, C., Vanlauwe, B., 2011. Communicating complexity: Integrated assessment of trade-offs concerning soil fertility management within African farming systems to support innovation and development. Agricultural Systems 104, 191-203.

Hair, J. F., Black, W.C, Babin, B.J, Anderson,R.E., 2010. Multivariate Data Analysis: A Global Perspective, Seventh Edition, Pearson.

Hervé M., 2011. Aide-mémoire de statistique appliquée à la biologie – Construire son étude et analyser les résultats à l'aide du logiciel R, 2ème version.

Howe, G., McKay, A., 2007. Combining quantitative and qualitative methods in assessing chronic poverty: The case of Rwanda. World Development 35, 197-211.

Husson F., Lê S., Pagès J., 2011. Exploratory Multivariate Analysis by Example Using R. Chapman & Hall/CRC.

Iraizoz, B., Gorton, M., Davidova, S., 2007. Segmenting farms for analysing agricultural trajectories: A case study of the Navarra region in Spain. Agricultural Systems 93(1–3): 143-169.

Kebede, B., 2007. Community wealth-ranking and household surveys : An integrative approach. Q-Squared Working paper No. 38, Centre for International Studies, University of Toronto, Canada.

Kennedy, G., Ballard, T., & Dop, M., 2011. Guidelines for measuring household and individual dietary diversity. Rome, Italy: FAO.

Kennedy, G., Nantel, G., Shetty, P., 2003. The scourge of 'hidden hunger': global dimensions of micronutrient deficiencies. Food, Nutrition and Agriculture 32, 8-16.

Köbrich, C., Rehman, T., Khan, M., 2003. Typification of farming systems for constructing representative farm models: two illustrations of the application of multi-variate analyses in Chile and Pakistan." Agricultural Systems 76(1): 141-157.

Kostrowicki, J., 1977. Agricultural typology concept and method. Agricultural Systems 2, 33-45.

Kuivanen, K., 2014. Patterns of persistence – characterizing smallholder farming system diversity in Ghana's Northern Region. MSc-thesis report, Farming Systems Ecology Group, Wageningen University, Wageningen.

Kumar, R., 2014. Research methodology: a step-by-step guide for beginners – fourth edition. Sage, Los Angelos. http://www.uk.sagepub.com/kumar4e/

Landais, E., 1998. Modelling farm diversity: new approaches to typology building in France. Agricultural Systems 58, 505-527.

Laurent, C., van Rooyen, C.J., Madikizela, P., Bonnal, P., Carstens, J., 1999. Household typology for relating social diversity and technical change, Agrekon: Agricultural Economics Research, Policy and Practice in Southern Africa, 38: S1, 190-208.

Madry, W., Mena, Y., Roskowska-Madra, B., Gozdowski, D., Hryniewski, R., Castel, J.M., 2013. An overview of farming system typology methodologies and its use in the study of pasture-based farming system: a review. Spanish Journal of Agricultural Research 11, 316-326.

Maton, L., Leenhardt, D., Goalard, M., Bergez, J.E., 2005. Assessing the irrigation strategies over a wide geographical area from structural data about farming systems. Agricultural Systems 86, 293-311.

Michielsens, C.G.J., Lorenzen, K., Phillips, M.J., Gauthier, R., 2002. Asian carp farming systems: towards a typology and increased resource use efficiency. Aquaculture Research 33, 403-413.

Moreno-Pérez, O., Arnalte-Alegre, E., Ortiz-Miranda, D., 2011. Breaking down the growth of family farms: A case study of an intensive Mediterranean agriculture. Agricultural Systems 104, 500-511.

Molua, E., 2011. Farm income, gender differentials and climate risk in Cameroon: typology of male and female adaptation options across agroecologies. Sustainability Science 6(1): 21-35.

Muthayya, S., Rah, J.H., Sugimoto, J.D., Roos, F.F., Kraemer, K., Black, R.E., 2013. The global hidden hunger indices and maps: An advocacy tool for action. PLoS ONE 8(6): e67860..

Pacini, G.C., Colucci, D., Baudron, F., Righi, E., Corbeels, M., Tittonell, P., Stefanini, F.M., 2013. Combining multi-dimensional scaling and cluster analysis to describe the diversity of the rural households. Experimental Agriculture FirstView, 1-22.

Remans, R., Flynn, D.F.B., DeClerck, F., Diru, W., Fanzo, J., 2011 Assessing Nutritional Diversity of Cropping Systems in African Villages. PLoS ONE 6(6):e21235.

Righi, E., Dogliotti, S., Stefanini, F.M., Pacini, G.C., 2011. Capturing farm diversity at regional level to up-scale farm level impact assessment of sustainable development options. Agriculture, Ecosystems & Environment 142, 63-74.

Rueff, C., Choisis, J.P., Balent, G., Gibon, A., 2012. A Preliminary Assessment of the Local Diversity of Family Farms Change Trajectories Since 1950 in a Pyrenees Mountains Area, Journal of Sustainable Agriculture

Sanogo, O.M., de Ridder, N., van Keulen, H., 2010. Diversité et dynamique des exploitations agricoles mixtes agriculture-élevage sud du Mali. Cahier Agriculture. 19, 185-193.

Thornton, P.K., Herrero, M., 2001. Integrated crop–livestock simulation models for scenario analysis and impact assessment. Agricultural Systems 70, 581-602.

Tittonell, P., Muriuki, A., Shepherd, K.D., Mugendi, D., Kaizzi, K.C., Okeyo, J., Verchot, L., Coe, R., Vanlauwe, B., 2010. The diversity of rural livelihoods and their influence on soil fertility in agricultural systems of East Africa – A typology of smallholder farms. Agricultural Systems 103, 83-97.

Tittonell, P., 2014, Categorising diversity of smallholder farming systems Household typologies. FSE staff seminar, 30 Jan 2014, Wageningen.

Tittonell, P., 2014. Livelihood strategies, resilience and transformability in African agroecosystems. Agricultural Systems 126(0): 3-14

Valbuena D., Groot, J., Mukalama, J., Gérard, B.,Tittonell, P., 2014, Improving rural livelihoods as a "moving target": trajectories of change in smallholder farming systems of Western Kenya. Regional Environmental Change, 1-13.

Van de Ven, G., Mellisse, B.T. (2014). Historic analysis of home garden agroforestry system changes in Ethiopia. Plant Production Systems Group, Wageningen University. [presentation]

Whatmore, S., Munton, R., Little, J.O., Marsden, T., 1987. Towards a typology of farm businesses in contemporary British Agriculture. Sociologia Ruralis 27, 21-37.

Appendix A: survey accuracy

From a report by Lotte Klapwijk (September 2013) on detailed data collection carried out in Tanzania for the AfricaRISING project:

"… sometimes there were big differences between the data reported in the surveys handed-in by enumerators, and the data found during the detailed data collection. For example, for several farmers, the total number of labor-days was surprisingly low, down to 40 days a year. After expressing surprise and re-asking the question, answers were very different. This could be the effect of several things, such as a different way of asking; questions were asked by a different person (factor = enumerator,) or a different answer, simply because it's a different day (factor = farmer). Even now, within the period of the detailed data collection, some differences were found when comparing the notes of the translator with the notes from the author of this report (for example 60k per bag of maize, against 80k), showing the difficulty to get clarity on data. Another problem is the fact that the definition of certain words or terms was not always clear, or agreed upon beforehand. For example 'last year' or 'last season' meant different things to different people.

According to survey-data, one farmer owned 39 acres of land, while during the second visit he reported a total of 22 acres. Only after minutes of re-asking questions in different ways, we managed to get the situation 'clear': the farmer had distributed 5acres of land to each of his 4 oldest sons, years ago. Throughout the rest of the interview it was not easy to separate the story of the farmer from that of those four oldest sons; as in many African families, they were deeply intertwined. For example, harvests of the sons were stored in separated piles in a sort of kraal, but when one pile runs out, the father makes sure there will still be food to eat.

Next to this, the basis of data collection is questionable; farmers, without knowing in advance, are asked to recall numbers and figures about almost everything going on in their lives, while most of them do not keep track of any numbers. Related to this, some questions might almost be impossible to answer; people seemed to have great difficulty to think in percentages. Also, it is not unthinkable that people sometimes give answers for social reasons, or to not look like a fool or because it's easier (for example, a farmer claimed his livestock needed 12hrs of labour/day, while the animals were still in the kraal during the interview, at mid-day, and another claimed to spend 12hrs/day on 7 pigs). As a result, data went from 100% home consumption of a certain crop product to only 25%. For another farmer we went from two fields of 1.5 acre, to 1 field of 2 acre, or for again another from 2 fields, of 7 and 8 acre, to 3 fields, of 5, 10 and 11 acres. It needs to be said that other changes are sad, but true, and mainly show how variable, and vulnerable, lives of smallholder farmers are; one lost 39 of his 45 chickens to New Castle Disease…"

Appendix B: contact persons

Table A1: Contact persons for typology making. These persons have developed and/or want to develop further experience and expertise on typology making.

Name	Email address	Specific interest
Mark van Wijk	m.vanwijk@cgiar.org	
Piet van Asten	p.vanasten@cgiar.org	
Diego Valbuena	d.valbuena@cgiar.org	Spatial explicit typologies
Anne Rietveld	a.rietveld@cgiar.org	Gender (and value chains)
Catherine Pfeifer	c.pfeifer@cgiar.org	Typologies and scaling-up
Stéphanie Alvarez	stephanie.alvarez@wur.nl	Multivariate analysis
Wim Paas	wim.paas@wur.nl	Typology literature
Gatien Falconnier	falconniergatien@yahoo.fr	Farm dynamics
Flemming Nielsen	fnielsen@bananahill.net	Drivers of farmer's decision making

Appendix C: example PCA and HC

This appendix provides an example of the steps to perform a Principal Component Analysis (PCA) and a Hierarchical Clustering (HC) for the typology construction using with R software, with a case study of a farm dataset from Tanzania.

The software R is open source and available online: http://cran.r-project.org/. In this example, the PCA and HC are run on R (version 3.0.3) using the R-package ade4.

Preparation of the dataset

In order to run the PCA on R, the dataset should be organised as a table with the observations (farms) in rows and the variables in columns (Table A2). The missing values could be expressed by "NA" or as empty cells and will be deleted in the first step of the data control for the PCA.

Table A1: Dataset from Tanzanian farms (n=174) used for the PCA and HC example; dataset called 'tanza'.

obs	region	hhsize	area	Totlabor	femratio	hiredratio	tlu	ncrop
1	1	13	6.27	3202	0.46	0.2	17.8	2
2	1	5	1.01	976	0.38	0.15	2.7	3
3	1	3	7.69	2455	0.29	0.83	19.3	3
4	1	7	1.42	1225	0.54	0.09	4.5	3
5	1	3	0.81	954	0.32	0.2	0.73	2
…								
174	2	5	4.96	2208	0.5	0.5	0.1	4

It is recommended to choose short variable names (Table A3), without space and accent to facilitate the work on R.

Table A2: Variables from tanza dataset

	Code	Variable
1	obs	number of the farm
2	region	region in Tanzania
3	hhsize	household size (number of member in the household)
4	area	land area (ha)
5	totlabor	total labor ($h.year^{-1}$)
6	femratio	female labor ratio (female labor/total labor)
7	hiredratio	hired labor ratio hired labor/total labor)
8	tlu	tropical livestock unit
9	ncrop	number of crop

Data control for the PCA and HC

One of the first steps for the PCA running is to check the data, i.e. find missing values, potential errors, outliers, "strong" correlations and control the variables distribution.

To delete of all the missing values in *tanza* dataset:

```
tanza <- na.omit(tanza)
```

Potential errors, outliers, "strong" correlations from *tanza* can be detected graphically using X-Y graphics (plots) or distribution graphics (Figure A1 and Figure A2).

To create a matrix of X-Y plots for the variables (Figure A1) of dataset *tanza* (except variable 1 (*obs*) and 2 (*region*):

```
pairs(tanza[,-c(1,2)], panel=panel.smooth)
```

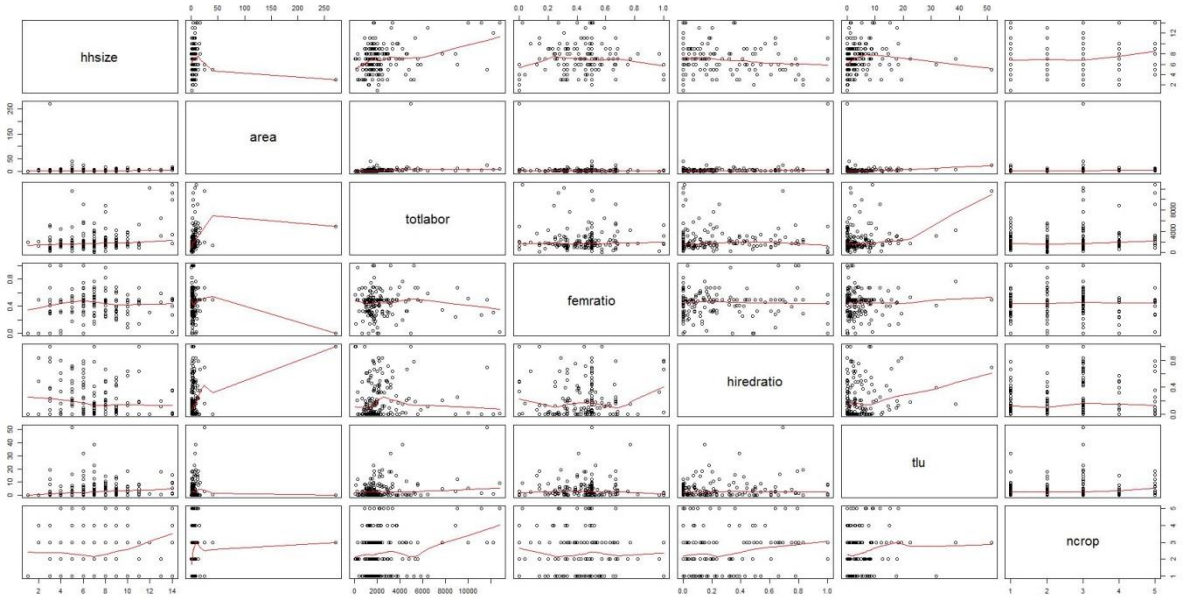

Figure A1: Matrix of X-Y plots for the tanza variables

To create a graphic with all distributions of the variables (Figure A2) from *tanza* (except variables 1 and 2, respectively obs and region):

```
hist(tanza[,-c(1,2)])
```

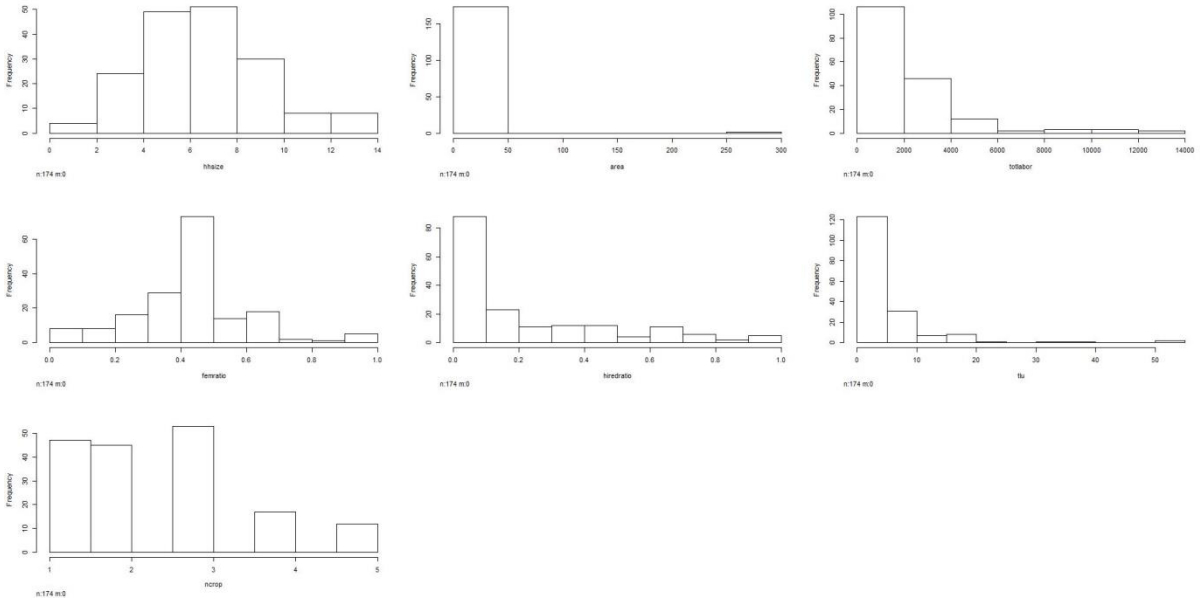

Figure A1: Variables distributions from tanza dataset

In Figures A1 and A2, we observe the existence of outliers in terms of field area (area > 50 ha) and in terms of livestock (tlu > 30). These outliers could be potential errors or could be existing cases in the area. In both cases it is advised to remove them for the PCA because of their strong impact on the results. However if these outliers are existing farms, they could be mentioned as outstanding/exceptional farms in the global result from the typology.

For studying the outliers in more detail, it is possible to create a boxplot per variable (for *area* and *tlu* for example, Figure A3):

```
boxplot(tanza$area)
boxplot(tanza$tlu)
```

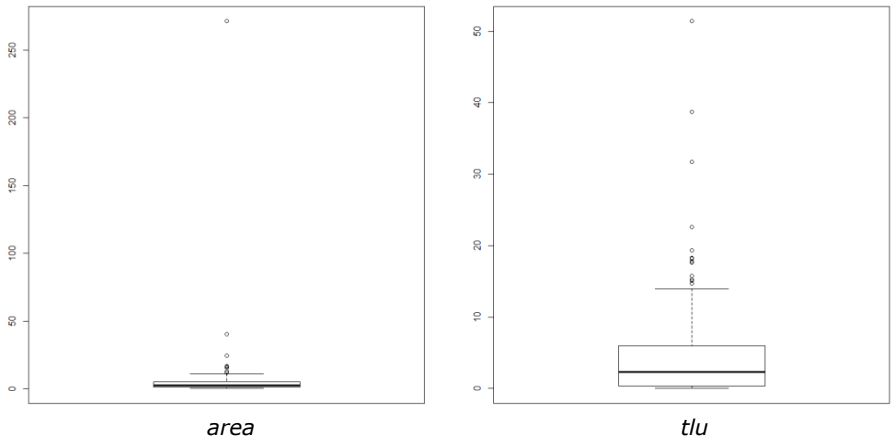

Figure A2: Boxplots for area and tlu variables from tanza dataset

To remove the outliers from the *tanza* dataset:

```
tanza<- tanza[tanza$area < 50,]
tanza<- tanza[tanza$tlu < 30,]
```

The new boxplots resulting from the reduction of *tanza* are presented in Figure A4; this type of figure can help to identify more precisely other outliers that have to be excluded for the PCA running.

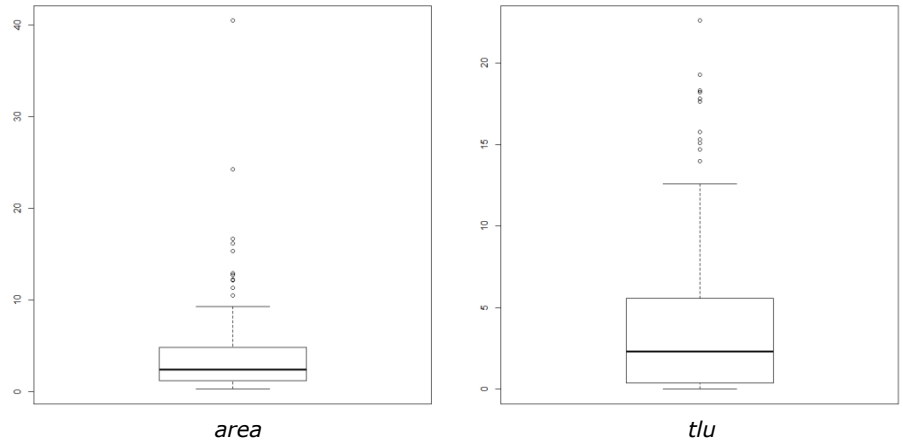

Figure A3: New boxplots for area and tlu variables from tanza dataset

Based on what we observe in the new boxplots, we choose to conduct the PCA with a *tanza* subset with *area* < 20 ha and *tlu* < 20:

```
tanza<- tanza[tanza$area < 20,]
tanza<- tanza[tanza$tlu < 20,]
```

Remarks: here there are still outliers, but we judge that they remain sufficiently grouped together to form a farm type.

Since it is recommended for the PCA to use quantitative variables with normal or at least symmetric distribution, some variables from *tanza* need to be transformed (Figure A5). Therefore we create new variables (e.g. *area2*, *totlabor2*, *femratio2*) applying logarithm (`log10`), square root (`sqrt`) or other functions to the original variables:

```
tanza$area2 <- log10(tanza$area)
tanza$totlabor2 <- log10(1+tanza$totlabor)
tanza$femratio2 <- log10(1+tanza$femratio)
tanza$hiredratio2 <- sqrt(tanza$hiredratio)
tanza$tlu2 <- sqrt(tanza$tlu)
tanza$ncrop2 <- sqrt(tanza$ncrop)
```

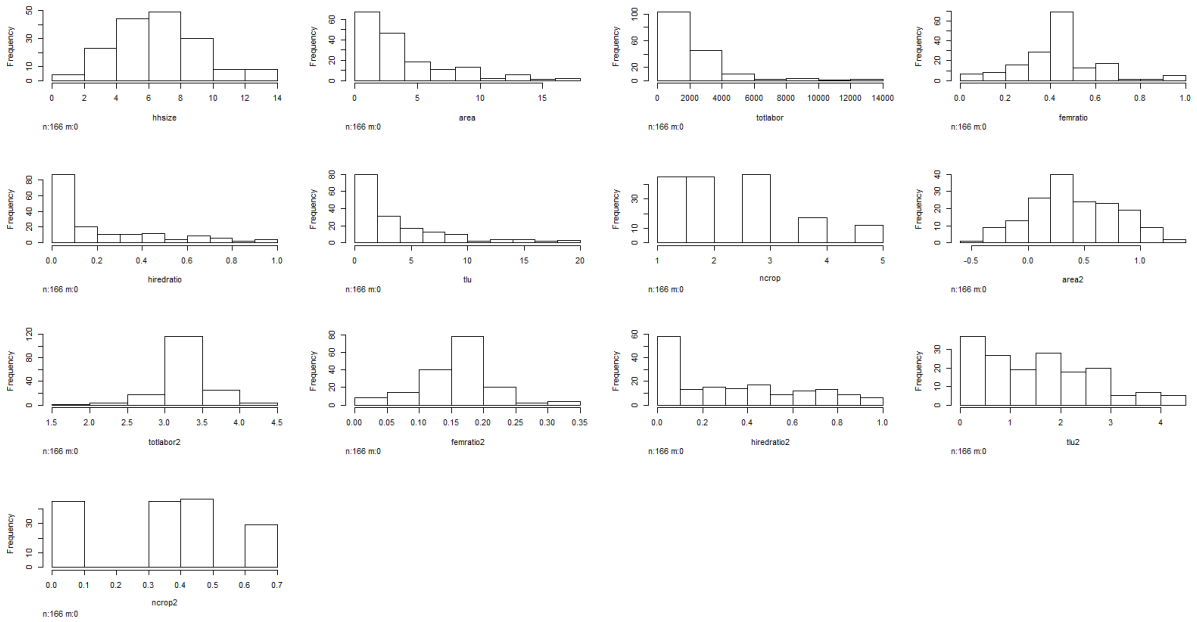

Figure A4: Variables distributions from 'tanza' dataset with the transformed variables.

Remarks: even if we improved the variables distribution, to reach a "good" symmetry could be difficult for some variables, such as *hiredratio2* and *tlu2* (Figure A5).

PCA on the transformed data

The R-package used for the PCA analysis is 'ade4'; so it should be installed and loaded:

```
install.packages("ade4")
library(ade4)
```

PCA No. 1

The function `dudi.pca()` is used to run the PCA on the *tanzaT* dataset which contains the transformed variables:

```
tanzaT <- tanza[,match(c("hhsize", "area2", "totlabor2", "femratio2",
"hiredratio2", "tlu2" ,"ncrop2" ), dimnames(tanza)[[2]])]
tanza.pca <- dudi.pca(tanzaT, center=T, scale=T, scannf=T, nf=5)
> Select the number of axes: 3
```

The `dudi.pca()` function displays automatically the barplot of the eigenvalues (Figure A6) in order to help select the number of principal components (PC) to keep for the further analysis. The use of the Kaiser criterion encourages to select all the axis having an eigenvalues greater than 1:

```
tanza.pca$eig
> 1.8996519 1.1906887 1.0549658 0.9962744 0.7951804 0.6305626 0.4326761
```

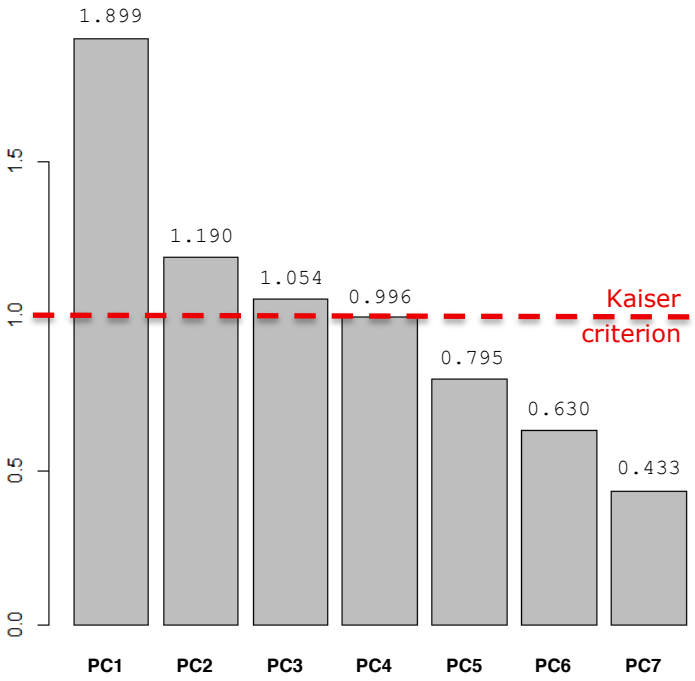

Figure A5: Barplot of the eigenvalues per principal component (PC)

The selection of the number of PC could be also determined by the percentage of variability explained we would like to represent. To calculate the cumulated percentage of variability explained by the PC:

```
cumsum(tanza.pca$eig) / sum(tanza.pca$eig)
> 0.2713788 0.4414772 0.5921866 0.7345116 0.8481088 0.9381891 1.0000000
```

Here, it means that with PC1 and PC2 we explain about 44.1% of the variability of the farms and with PC1, PC2 and PC3, we explain about 59.2% of the variability of the farms. At that step we choose to keep 3 axes (PC1, PC2 and PC3).
It should be noted that with more PCs the interpretation of the PCA and HC final results becomes more difficult.
The interpretation of the PCs is based on the correlation circles (Figure A7) and the correlation matrix (Table A4).

The function `s.corcircle()` is used to create the correlation circles:

```
s.corcircle(tanza.pca$co, xax=1, yax=2 )
s.corcircle(tanza.pca$co, xax=1, yax=3 )
```

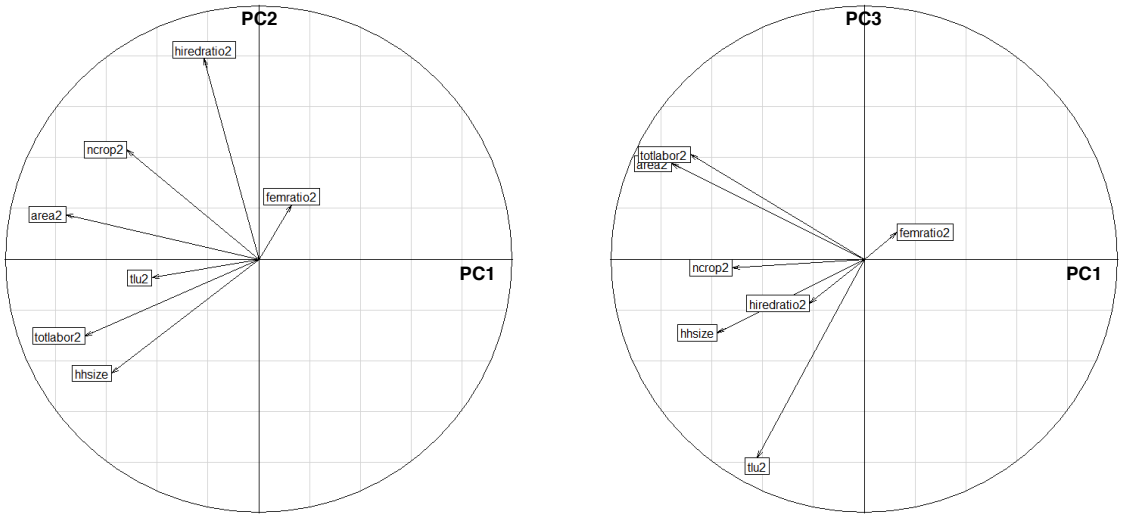

Figure A6: Correlation circles for the principal components PC1-PC2 and PC1-PC3.

The correlation coefficients between the PCs and the variables are contained in the object `tanza.pca$co` .

Table A3: Correlation matrix between the principal components (PC) and the variables from tanzaT dataset.

Variables	PC1	PC2	PC3
hhsize	-0.581	-0.450	-0.291
area2	***-0.761***	0.176	0.377
totlabor2	***-0.686***	-0.303	0.412
femratio2	0.130	0.214	0.106
hiredratio2	-0.216	***0.793***	-0.175
tlu2	-0.422	-0.074	***-0.785***
ncrop2	-0.520	0.431	-0.034

Here the variables *area2, totlabor2* and *hhsize* are (negatively) correlated to PC1, *hiredratio2* is correlated to PC2 and *tlu2* is correlated to PC3 (Figure A7 and Table A4). In brief, PC1 expresses global information about the farm size (in terms of land and labour), PC2 the relative importance of the hired labour and PC3 the size of the livestock herd.

Here the variable *femratio2* is not well represented on any of the planes PC1-PC2 and PC1-PC3 (short arrow on Figure A7). The variable *ncrop2* seems not to bring additional information for the PCA: it is correlated to *area2* and could provide redundancy on PC1, which is already well defined by *area2 and totlabor2*. Moreover, from the structure of the variable, *ncrop2* seems to be a categorical variable (Figure A1). Hence, we could consider deleting one of these two variables or both for a next PCA try. For this example, it is chosen to exclude the variable *ncrop2* in the next PCA running in order to test if it increases the percentage of the variability explained.

The function `s.label()` is used to represent the observations (farms) in the plane PC1-PC2 or PC1-PC3 (Figure A8):

```
s.label(tanza.pca$li, xax=1, yax=2)
s.label(tanza.pca$li, xax=1, yax=3)
```

Figure A7: Location of farms in the principal components planes PC1-PC2 and PC1-PC3.

Figure A8 helps to identify potential outliers; here for example farms 11 and 170 are very isolated from other farms. The outliers have a strong effect on the PCA results due to the PCA procedure, which linearly correlates the PCs and the dataset variables; these outliers therefore strongly affect the slope of the linear regression line (Figure A9). In the next PCA run, these two farms could be deleted.

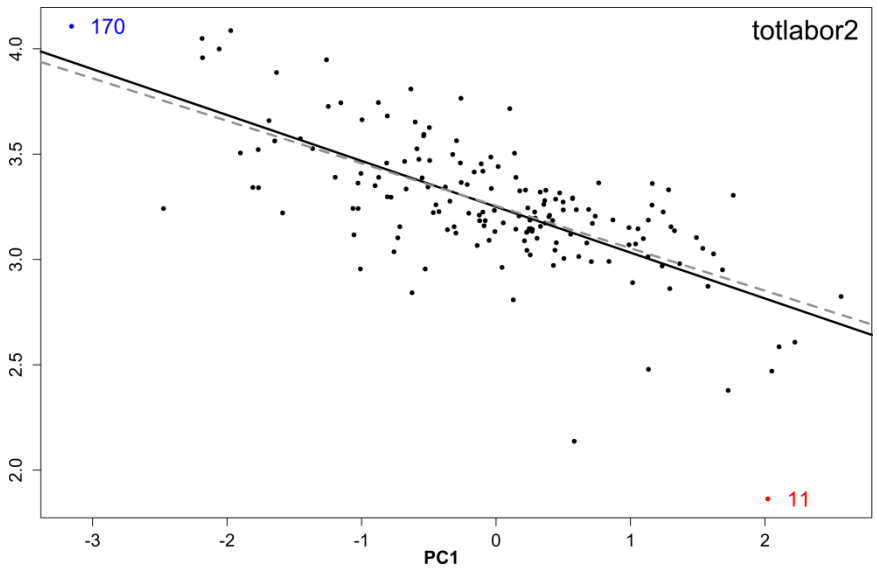

Figure A8: Linear regression between the variable totlabor2 of tanzaT and the Principal Component PC1; the black line is the regression line with the farms 11 and 170, and the dashed line represents the slope change due to the removal of the farms 11 and 170.

PCA No. 2

In this example, the second running of the PCA is performed without the farms 11 and 170 (outliers according Figure A8) and without the variable *ncrop2*.

To delete the farms 11 and 170 from *tanza* dataset (one by one):

```
tanza <- tanza[tanza$obs!=11,]
tanza <- tanza[tanza$obs!=170,]
```

To create a new tanzaT without the variable *ncrop* and run again the PCA:

```
tanzaT <- tanza[,match(c("hhsize", "area2", "totlabor2", "femratio2", "hiredratio2", "tlu2" ), dimnames(tanza)[[2]])]
tanza.pca <- dudi.pca(tanzaT, center=T, scale=T, scannf=T, nf=5)
> Select the number of axes: 4
```

To verify the eigenvalues:

```
tanza.pca$eig
> 1.7208478 1.1131360 1.0632992 1.0017315 0.6261152 0.4748704
```

To confirm of the percentage of the variability explained by the PCs:

```
cumsum(tanza.pca$eig) / sum(tanza.pca$eig)
> 0.2868080 0.4723306 0.6495472 0.8165024 0.9208549 1.0000000
```

Now PC1 and PC2 together explain 47.2% of the variability of the farms, PC1, PC2 and PC3 together explain 64.9% of the variability and PC1, PC2, PC3 and PC4 together explain 81.6% of the variability.

Here the Kaiser criterion advised to use four PCs for the analysis, but it also possible to decide to use only three PCs, with which about 65% of the variability can be explained.

We use the correlation circle (Figure A10) and the correlation matrix (Table A5) to interpret the meaning of PC1, PC2, PC3 and PC4:

```
s.corcircle(tanza.pca$co, xax=1, yax=2)
s.corcircle(tanza.pca$co, xax=1, yax=3)
s.corcircle(tanza.pca$co, xax=1, yax=4)
```

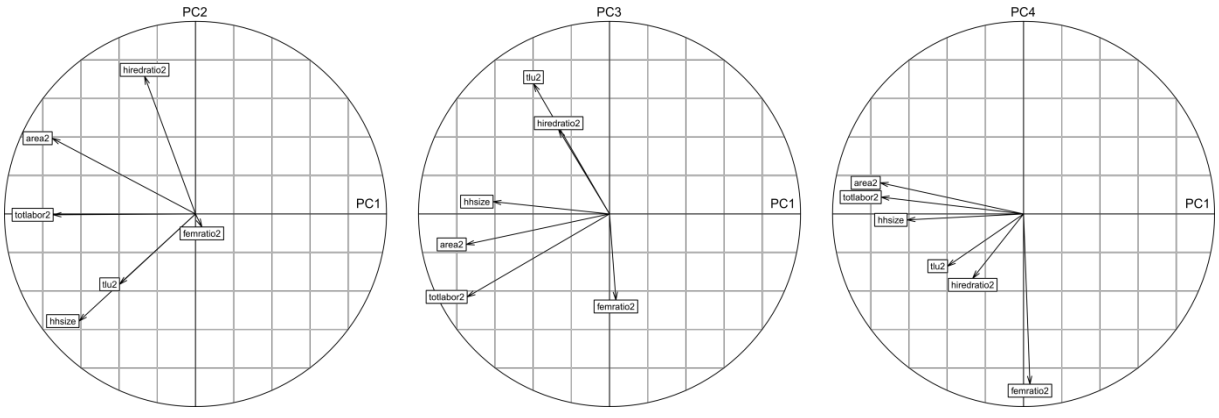

Figure A9: Correlation circles for the principal components PC1-PC2, PC1-PC3 and PC1-PC4.

Table A4: Correlation matrix between the principal components (PC) and the variables from tanzaT dataset.

Variables	PC1	PC2	PC3	PC4
hhsize	*-0.610*	-0.556	0.064	-0.031
area2	***-0.750***	0.393	-0.158	0.162
totlabor2	***-0.745***	-0.004	-0.431	0.087
femratio2	0.033	-0.067	-0.447	***-0.884***
hiredratio2	-0.267	***0.715***	0.438	-0.335
tlu2	-0.398	-0.365	***0.676***	-0.271

It is convenient to check again the likely outliers (Figure A11):

```
s.label(tanza.pca$li, xax=1, yax=2)
s.label(tanza.pca$li, xax=1, yax=3)
s.label(tanza.pca$li, xax=1, yax=4)
```

Figure A10: Farmers illustrated in the principal components planes PC1-PC2, PC1-PC3 and PC1-PC4.

Figure A11 helps to identify new plausible outliers; in the next running of the PCA it was chosen to delete the farm 47 isolated in the plane PC1-PC3 and farms 129 and 29 isolated in the plane PC1-PC4.

PCA No. 3

The third running of the PCA is done without the farms 47, 129 and 29:

```
tanza <- tanza[tanza$obs!=47,]
tanza <- tanza[tanza$obs!=129,]
tanza <- tanza[tanza$obs!=29,]

tanzaT <- tanza[,match(c("hhsize" ,"area2",
"totlabor2","femratio2","hiredratio2","tlu2"), dimnames(tanza)[[2]])]
tanza.pca <- dudi.pca(tanzaT, center=T, scale=T, scannf=T, nf=5)
> Select the number of axes: 3

tanza.pca$eig
> 1.7695090 1.1235104 1.0241344 0.9887622 0.6217878 0.4722962

cumsum(tanza.pca$eig) / sum(tanza.pca$eig)
> 0.2923875 0.4841733 0.6562492 0.8196921 0.9197690 1.0000000
```

To show the new circles of correlation (Figure A12) and the correlation matrix (Table A6):

```
s.corcircle(tanza.pca$co, xax=1, yax=2)
s.corcircle(tanza.pca$co, xax=1, yax=3)

tanza.pca$co
```

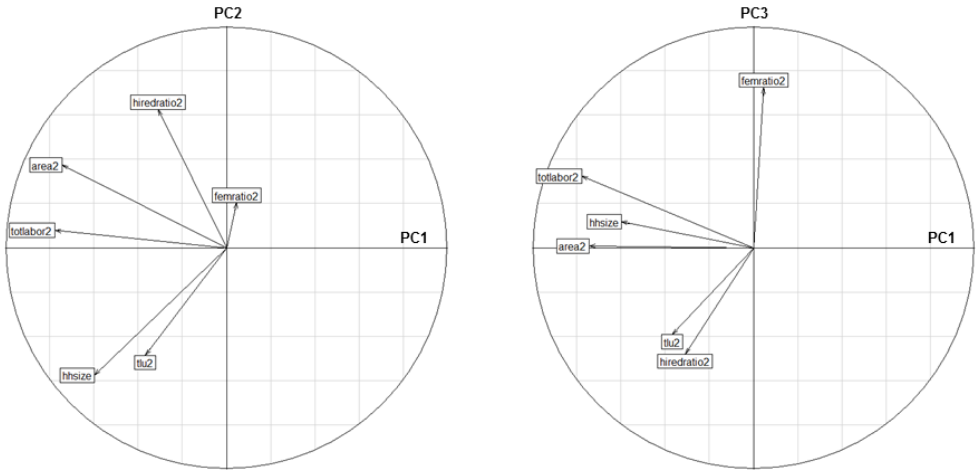

Figure A11: Correlation circles for the principal components PC1-PC2 and PC1-PC3.

Table A5: Correlation matrix between the principal components (PC) and the variables from tanzaT dataset.

	PC1	PC2	PC3
hhsize	*-0.600*	-0.577	0.117
area2	***-0.745***	0.374	0.008
totlabor2	***-0.778***	0.079	0.324
femratio2	0.047	0.205	***0.726***
hiredratio2	-0.310	***0.627***	-0.481
tlu2	-0.368	-0.487	-0.394

According to the PCA results (Figure A12 and Table A6), here the discriminating variables (segregating the farms best) are *area2*, *totlabor2*, *hiredratio2* and *femratio2*.
In brief, PC1 still expresses global information about the farm size, PC2 the relative importance of the hired labour and PC3 the relative importance of the female labour.
Remarks: deleting the three outlier farms on the previous PCA (Figure A11) had a strong effect on the new PCA results (*tlu2* is no more strongly correlated to the main PCs).

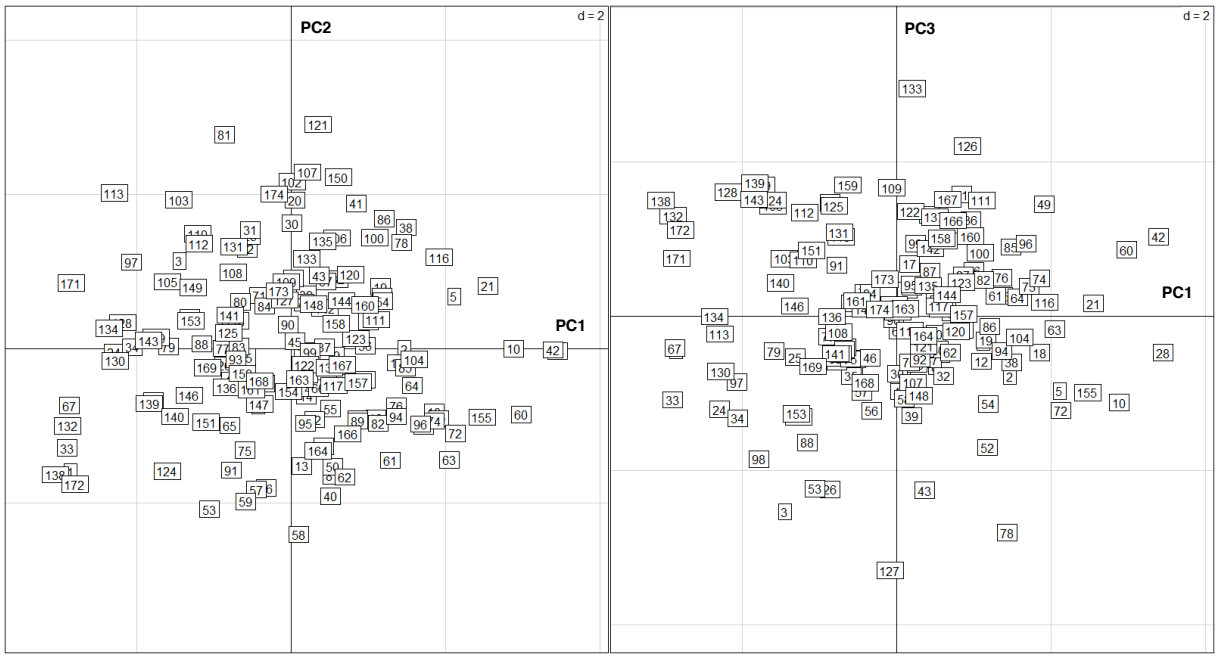

Figure A12: Farmers illustrated in the principal components planes PC1-PC2 and PC1-PC3.

To superpose the circle of correlation and the observation in the plane PC1-PC2 (Figure A13):

```
scatter.dudi(tanza.pca)
```

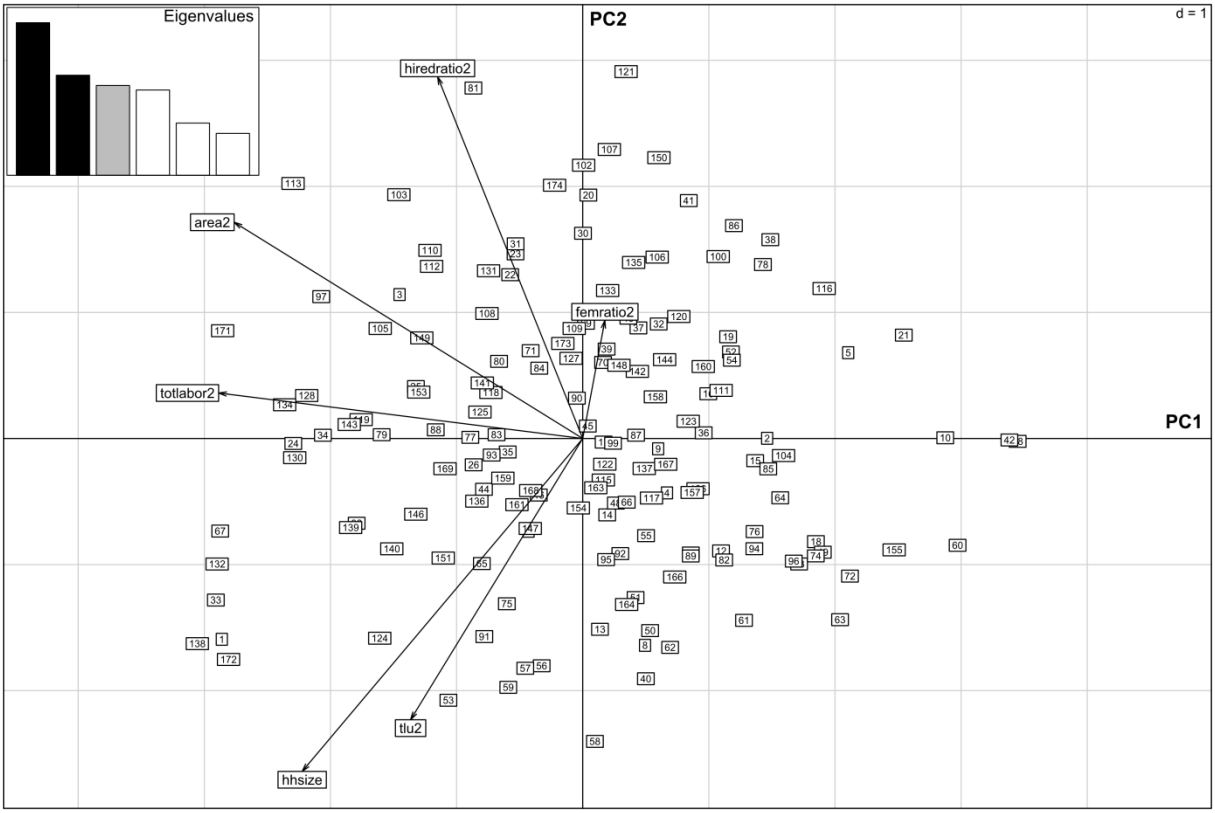

Figure A13: Farmers and variables illustrated in the principal components plane PC1-PC2.

Cluster Analysis on the PCA results

Here we applied the Hierarchical Clustering (HC) on the PCA results with the function `hclust()` using the Ward method:

```
tanza.cah <- hclust(dist(tanza.pca$li), method="ward")
```

The dendrogram and the barplot help to choose the number of clusters (or types) to use (Figure A15):

```
barplot(tanza.cah$height)

plot(tanza.cah)
```

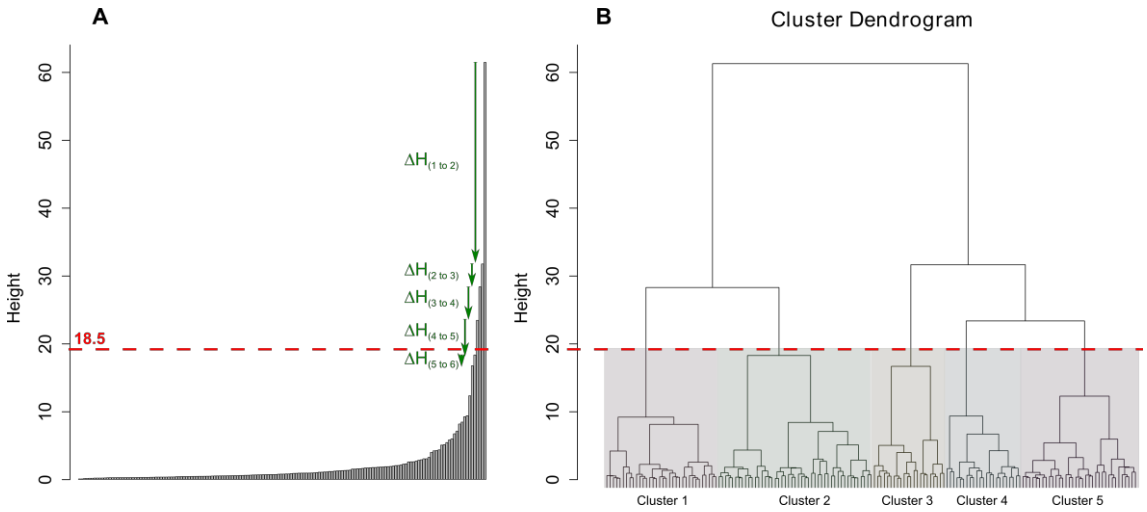

Figure A14: Barplot of the height (A) and dendrogram (B). $\Delta H_{(n \text{ to } n+1)}$ correspond to the delta (decrease) of Height passing *n* clusters to *n+1* clusters.

In Figure A15, Height is an indicator of the dissimilarity within clusters related to the number of clusters. The bar on the far right shows the maximum dissimilarity, i.e. the dissimilarity when all the farms are grouped in one cluster. The examination of the heights decreases from right to left in the barplot figure ($\Delta H_{(4 \text{ to } 5)} > \Delta H_{(5 \text{ to } 6)}$) and the overall structure of the dendrogram suggest to make the partition of the dendrogram ("cut the tree") at a Height of about 18.5 (the dotted line in Figure A15) leading to partition of the dendrogram into 5 clusters.

To separate the observations (farms) into in five clusters (k is the number of clusters we want):

```
tanza.type <- cutree(tanza.cah, k=5)
```

Finally to visualise and interpret the clusters in the PC1-PC2 and PC1-PC3 planes (Figure A16):

```
s.corcircle(tanza.pca$co, xax=1, yax=2)
s.class(tanza.pca$li, fac = as.factor(tanza.type))
s.corcircle(tanza.pca$co, xax=1, yax=3)
s.class(tanza.pca$li, xax=1, yax=3, fac = as.factor(tanza.type))
```

To add to the *tanza* dataset a column (*typo*) containing the type numbers for each farm:

```
tanza$typo <- tanza.type
```

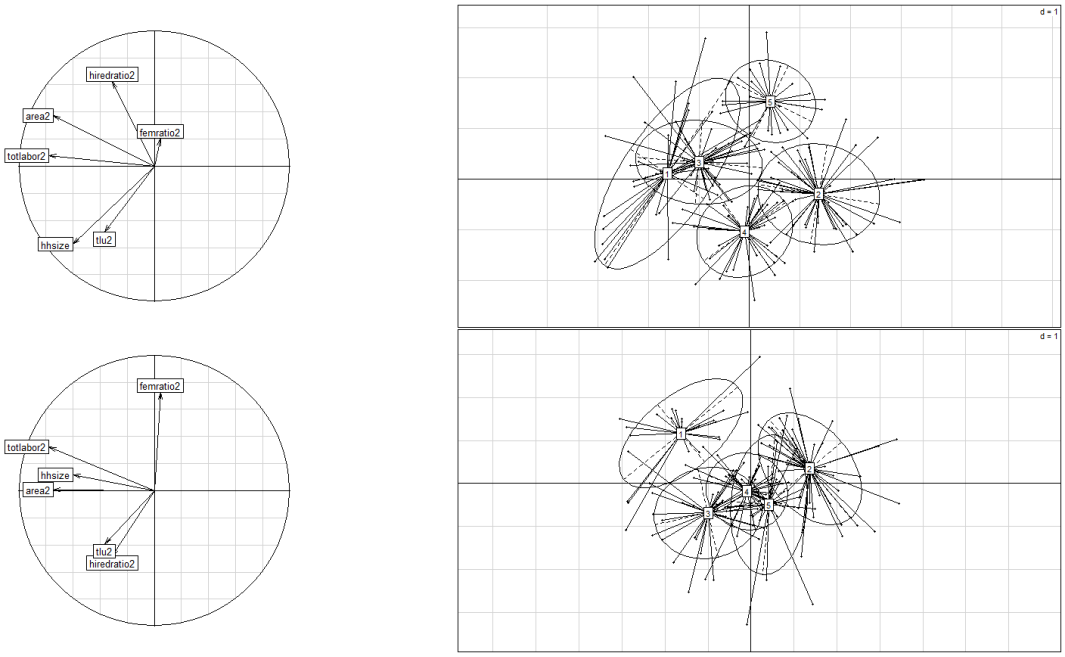

Figure A15: Result of the Principal Component Analysis and the Hierarchical Cluster.

Next step for the typology

The next step is to interpret each cluster (or type) using the graphics/results from the PCA (Figure A16) and statistic calculations for each cluster (e.g. means, ranges).

Then, it is required to compare the meaning of each type (i.e. results of the clusters interpretation) with the knowledge about the area (Hypothesis). The farms excluded could be mentioned as atypical farms existing in the area. If the clusters are meaningless according the local knowledge (e.g. role of the variable *femratio* too "strong") the PCA should be restarted.

Finally, a validation of the typology results by local experts is desired.