Computers and Electronics in Agriculture 212 (2023) 108074

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Construction of a generalised farm typology to aid selection, targeting and scaling of onfarm research

Kirsty L. Hassall [a,*], Frédéric Baudron [b], Chloe MacLaren [a,d], Jill E. Cairns [b], Thokozile Ndhlela [b], Steve P. McGrath [c], Isaiah Nyagumbo [b], Stephan M. Haefele [c]

[a] Intelligent Data Ecosystems, Rothamsted Research, Harpenden AL5 2JQ, United Kingdom
[b] International Maize and Wheat Improvement Center (CIMMYT), Harare, Zimbabwe
[c] Sustainable Soils and Crops, Rothamsted Research, Harpenden AL5 2JQ, United Kingdom
[d] Department of Crop Production Ecology, Swedish University of Agricultural Sciences, Almas Alle 8, Uppsala 750 07, Sweden

ARTICLE INFO

Keywords:
Typology
Clustering
Anosim
RShiny app
Farms
Agriculture

ABSTRACT

Farm typologies are often used to reduce the complexity in categorising diverse farming systems, particularly in sub-Saharan Africa. The resulting typologies can then be used in multiple ways including designing efficient sampling schemes that capture the diversity in smallholder farms, prescribing the selection of certain farm types to which interventions can be targeted or upscaled, or to give context into derived relationships. However, the construction of farm typologies consists of many subjective decisions that are not always obvious or evident to the end-user. By developing a generalized framework for constructing farm typologies, we clarify where these subjective decisions are and quantify the impact they have on the resulting typologies. Further, this framework has been encapsulated in the open source RShiny App: *TypologyGenerator* to enable users to focus on the decisions and not the underlying implementation.

1. Introduction

On-farm studies have and always will form a critical part of agricultural research. Not only do they enable a broader spectrum of environmental conditions to be exploited than otherwise available in on-station field trials, but they are also more representative of the practicalities of real-life implementation. This is particularly pertinent when studying farming systems in sub-Saharan Africa as they largely consist of smallholder farms crossing diverse cultural and landscape contexts (Giller et al. 2011). It is well-recognised that a "one-size fits all" approach is misplaced in these highly heterogeneous landscapes. Nonetheless, such heterogeneity can be efficiently captured through the construction of simplified typologies. Specifically, by allocating farms to a certain typology, farming systems can be grouped into similar classes according to a set of diverse characteristics as measured by multiple indicators consisting of environmental, socio-economic, demographic and agronomic factors. The resulting typologies can then be used in multiple ways including designing efficient sampling schemes that capture the diversity in smallholder farms, prescribing the selection of certain farm types to which interventions can be targeted or upscaled, or to give context into derived relationships and aid interpretation of observed responses to interventions (Alvarez et al. 2014).

Applications of farm typologies have been broad with many different foci. For example, resource endowment typologies have been used to group households into those with more or less resources available, to help understand differences between farms in terms of productivity, poverty and constraints (Rusere et al. 2019, Hammond et al. 2020, MacLaren et al. 2022). Functional typologies of livelihood strategies have been used to understand differences in soil fertility management and nutrient resource flows (Tittonell et al. 2005, Tittonell et al. 2010), to assess resource use efficiency to target agricultural interventions (Kansiime et al., 2018), to understand differences in food security and farm incomes (MacLaren et al. 2022), and to understand diversity across regions of Zambia (Alvarez et al. 2018, Silva et al. 2023). Combined structural and functional typologies have been used to identify farm types based on household opportunities and constraints for the targeting of agricultural interventions and innovations (Kuivanen et al., 2016a; Berre et al., 2019), to what extent ecological intensification practices need to be focussed on specific farm types (Kansiime et al., 2021), to understand diversity in small wetland farming systems (Sakané et al.,

* Corresponding author.
E-mail address: kirsty.hassall@rothamsted.ac.uk (K.L. Hassall).

https://doi.org/10.1016/j.compag.2023.108074
Received 16 March 2023; Received in revised form 28 June 2023; Accepted 14 July 2023
Available online 20 July 2023

2013), to understand diversity in African farming systems in the context of soil fertility management (Giller et al. 2011), and in scaling-up of field and farm-level model results to a regional level (Righi et al. 2011). Hammond et al. (2020) used typologies to identify farm types for prioritising engagement and geographical locations for investment and targeting decision support systems whilst Rusere et al. (2019) used them to identify and link specific ecosystem services with suitable ecological intensification options. In each case, typologies have given context to the wider research question by simplifying the highly heterogeneous landscape of farming systems and farms.

However, along with the broad application of typology construction, is the wide array of methods available to construct such typologies. In many cases, on-farm studies need to be representative of a wide range of different characteristics. Established survey protocols have addressed this issue by using multivariate statistical techniques to define distinct farm typologies (Alvarez et al. 2014). The most common approach is to apply a principal components analysis to derive a reduced dimensional representation of the data followed by a hierarchical cluster analysis to identify the distinct typologies (Kansiime et al., 2021; Sakané et al., 2013; Alvarez et al. 2018, Blazy et al. 2009, Rueff et al. 2012). Variations on this approach have included the use of non-hierarchical cluster methods such as k-means (Kansiime et al., 2018; Kuivanen et al. 2016b) or PAM (partitioning around mediods) (Kuivanen et al., 2016a). Additional post-hoc refinement of the identified clusters has been done manually (Tittonell et al. 2010) through e.g. additional categorical variables such as gender of the household head (Kansiime et al., 2021). Alternatively, variations on the dimension reduction step include using multiple correspondence analysis (for categorical variables) or multi-dimensional scaling (MDS) (Pacini et al., 2014; Righi et al. 2011), or to bypass the dimension reduction step entirely (MacLaren et al. 2022, Hammond et al. 2020). The latter two (MDS or no dimension reduction) have the advantage of allowing both quantitative and qualitative data to be included in the typology definitions. Despite this, for those studies that could include qualitative variables, they have often been relegated to ancillary roles in refining clusters rather than in the primary construction step (Kansiime et al., 2021; Pacini et al., 2014).

In contrast to the statistical definition of typologies, farm typologies have also been defined through expert knowledge with the intention to picking out nuanced relationships and capturing the dynamic nature of farming characteristics (e.g., farm and herd size may vary year-to-year, Rusere et al. 2019, Kuivanen et al. 2016b). In both the expert derived typologies and those defined through statistical techniques, the literature emphasises the importance of variable selection and ensuring this is related to the driving hypotheses of interest. We do not focus on variable selection in this paper but refer the reader to e.g., Hammond et al. (2020), Alvarez et al. (2014) and Kuivanen et al. (2016b).

Despite the emphasis on the importance of variable selection in the construction of typologies, established methods are primarily focussed on quantitative variables. There are many qualitative variables that might be of interest and, importantly, quantitative proxies for such variables may not exist, such as the gender of the farm manager (Molua 2011, Cairns et al. 2021). Furthermore, not all variables will have equal importance in distinguishing farm types. If an on-farm study explicitly aims to investigate gender differences (perhaps as a secondary outcome measure) then such a variable will be highly important compared to others. In contrast, for other studies it might only be of peripheral interest and in such cases one would not wish for this to be a main driving factor in the typology definition. Consequently, in this paper, we extend established protocols to a) include both qualitative and quantitative variables and b) incorporate a weighting structure to give *a priori* ranks of each variable in the typology definition. Furthermore, we highlight the need for typologies to be constructed through an iterative approach and therefore present an R-shiny application: *TypologyGenerator* (Hassall, 2023) to implement the methods. The proposed methods for constructing typologies are demonstrated on a case study of maize production in Murehwa District of Zimbabwe.

2. Materials and methods

2.1. Case study: Maize production in Murehwa District, Zimbabwe

This section describes a case study related to the project "Addressing malnutrition with biofortified maize in Zimbabwe: from crop management to policy and consumers" (IATI Identifier: GB-GOV-13-FUND–GCRF-BB_T009047_1). The District of Murehwa was selected for the implementation of this project, as maize is the predominant crop and malnutrition has remained high in this district, despite significant reduction in other districts (ZimVac, 2020). Further, two contrasted wards – in particular in terms of soil texture and elevation – were selected within the District of Murehwa: Ward 4 and Ward 27. In September 2020, a total of 306 farmers representing around 7.5% of the population were selected at random, using an adaptation of the Y sampling, in Ward 4 and Ward 27 of Murehwa, and heads of households interviewed by a team of 10 trained enumerators using a structured questionnaire programmed with the software KoboToolbox (https://support.kobotoolbox.org/welcome.html) and uploaded on remotely controlled mobile devices (model Famoco FX100, https://www.famoco.com/android-devices/handheld-devices/fx100/). The questionnaire addressed the following: characteristics of the head of the household, size and composition of the household, production capital (e.g., land, equipment), land allocation, livestock numbers, livestock production and management, crop production and management in homefields and outfields, food security and dietary diversity, and income generating and food producing activities. The full questionnaire is provided as a supplementary file.

The goal of the typology was to delineate between relatively homogeneous groups of farms in terms of their structures, their functioning, and their diet to target interventions around biofortified maize. For the analysis, we used:

- five continuous structural variables (age of the head of the household, family size, total cropped area, cattle ownership, and sheep and goats ownership)
- two discrete structural variables each with a binary yes/no outcome (female-headed household, and education of the head of the household higher than primary level)
- three continuous functional variables (total maize produced during the 2019–20 season, total area cultivated to maize during the 2019–20 season, and total quantity of fertilizer applied to maize during the 2019–20 season)
- four discrete functional variables each with a binary yes/no outcome (own production as main source of food, crop sales as main source of income, use of intercropping, and use of manure)
- two continuous variables related to nutrition (total number of months during which food security was assessed as medium or high over the 12 months preceding the interview, and the household dietary diversity score in the 24 h preceding the interview)
- three discrete variables related to nutrition each with a binary yes/no outcome (consumption of plant-based vitamin A rich food in the 24 h preceding the interview, consumption of animal-based vitamin A rich food in the 24 h preceding the interview, and consumption of iron rich food in the 24 h preceding the interview)

All continuous variables except age of the head of the household and family size had a skewed distribution and were log-transformed to approximately follow a symmetric distribution.

3. Typology formation

Typology formation consists of four key steps; variable selection, dimension reduction, cluster formation and validation. The main process as implemented in *TypologyGenerator* is outlined in Fig. 1 and described below.

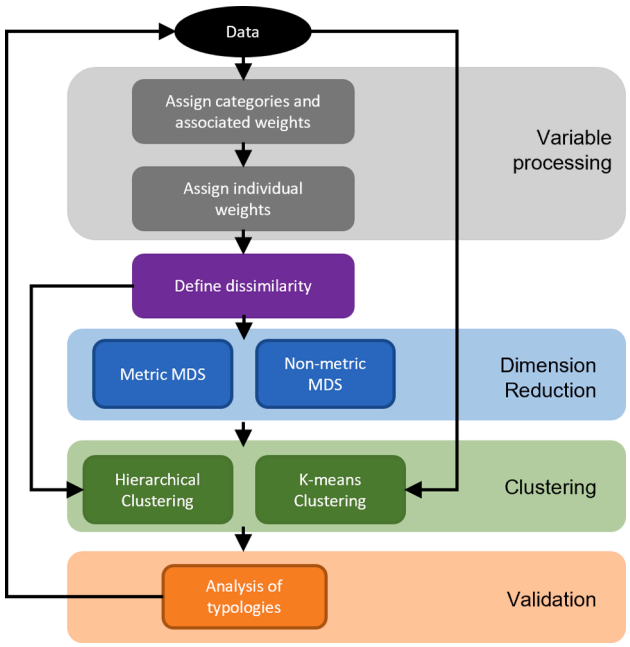


Fig. 1. Schematic of typology formation.

R-shiny web app: Typology Generator. We have developed an open-source R-shiny application (Hassall, 2023) implementing the workflow in Fig. 1 as described below. In brief, data can be loaded into the app from a standard CSV file. It is assumed that the data have been appropriately processed, i.e., any transformation has been performed and for categorical variables, appropriate amalgamation of categories has been done, etc.

Variable Selection. With the ever-increasing availability of data, the number of potential variables to include in typology construction seems limitless. By weighting the input variables, a user can emphasise different features in the typology. Variables can generally be grouped into similar "types" or categories. These might include, for example, structural versus functional variables. Weights can take two forms; i) weight variables based on the type of variable and ii) weight individual variables within each type. For instance, in the first case, if a typology definition was to focus on structural aspects, one might weight these components higher than functional variables in a e.g., 2:1 ratio. Additionally, weights may be defined for individual variables, for instance of the structural variables gender may be of most interest and therefore have higher weight than the other components. It is typical in the literature to discard highly correlated variables to prevent over-emphasising certain characteristics. Instead, one can simply down-weight the correlated variables and retain all components and associated nuances.

Both the high-level weighting of categories of variable, the "main weights", and the low-level "individual weighting" are implemented in *TypologyGenerator*. A screenshot of the weights used for the Murehwa case study is given in Fig. 2 along with the graphical visualization.

Dimension Reduction enables a simplification of the, potentially many, variables used to define the resulting typologies. To proceed with dimension reduction, a dissimilarity or distance matrix is constructed. For generality, let s_{ij}^k be the similarity measure between observations i and j for variable k and d_{ij}^k the corresponding distance defined by,

$$\left(d_{ij}^k\right)^2 = 2\left(1 - s_{ij}^k\right)$$

The overall distance between two observations is then given by the weighted sum,

$$D_{ij} = \sqrt{\sum_k w_k \left(d_{ij}^k\right)^2}$$

where w_k is the associated weight to variable k and $\sum_k w_k = 1$. Given matrix D, multidimensional scaling (MDS) can be used to reduce the dimensionality. Where all individual distance measures are metric, classical multi-dimensional scaling (principal coordinates analysis) can be used. If non-metric measures, such as Bray-Curtis dissimilarities, are included, non-metric multidimensional scaling should be used instead. If D is given by a matrix of squared (not scaled) Euclidean distance, the resulting classical MDS is equivalent to a principal components analysis. Choosing the number of dimensions to retain is subjective, but generally it is preferable to keep a relatively small number that captures the majority of the variation of the original data. In addition, if multicollinearity has not been addressed in the weighting options of the previous step, a user may wish to scale the resulting orthogonalized variables. Let $D^{'}$ be the Euclidean distance of the reduced dimensional representation of the data.

Multiple dissimilarity measures are available in *TypologyGenerator* and can be selected for each variable in turn (Fig. 1). The dimension reduction options as implemented in the Murehwa case study are illustrated in Fig. 3.

Cluster formation. A plethora of clustering methods exist within the literature. Here, we have implemented two distinct methods; hierarchical clustering and k-means. For hierarchical clustering either the matrix D or $D^{'}$ are used as inputs. The resulting typologies are obtained by "cutting" the associated dendrogram at a user defined point to define distinct clusters. For k-means, either the original variables or the reduced dimensional representation of them are used as inputs. For k-means clustering, the number of groups is defined in advance of the clustering. This is again a subjective choice and different values can be inputted to investigate how the allocations differ. The aim of k-means clustering is to minimise the within group sums of squares (equivalently to maximise the between group sums of squares) which is displayed in a diagnostic plot within *TypologyGenerator (*Fig. 4*)*.

Validation is the final step to define the typologies. ANOSIM (analysis of similarity) can be used to measure the degree of separation between groups (Righi et al. 2011; Pacini et al., 2014. ANOSIM provides an analogous approach to ANOVA where the response variable is itself the dissimilarity matrix. Here, we implement the approach of permutation MANOVA (multivariate analysis of variance) as implemented by the adonis2 function in the vegan R package (Oksanen et al. 2022 and Anderson, 2001). The result is an ANOVA-like table with statistical significance assessed by permutation of whether the groups are statistically different or not.

In addition, a number of visual assessments can be made to a) investigate how well each variable is represented in the reduced dimensional space (if a dimension reduction step is chosen) b) investigate how well each variable is separated between clusters and c) what the clusters represent in terms of the original variables. These are illustrated in Figs. 7 and 8 for the Murehwa case study with a full set of graphical assessments given in Appendix A2.

For the purposes of the typology comparison study, two quantitative assessments regarding the cluster definition are made. These are i) the evenness of the resulting groups and ii) the "clarity" of the groups. To calculate the evenness, we used Pielou's evenness index often used as a measure of species evenness in biodiversity studies. This is given by,

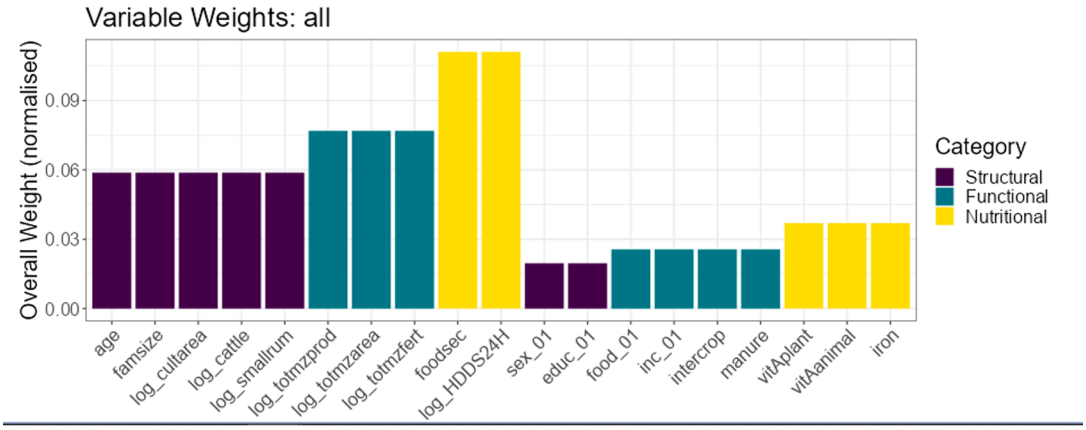
$$J = \frac{H}{H_{max}},$$

where H is the Shannon diversity index given by $H = -\sum_i p_i \ln p_i$ where p_i is the proportion of observations belonging to group i and H_{max} is given by the maximum possible value of H (i.e., if every group was of equal size). J is constrained between 0 and 1. The less even the clustering, i.e.,

Variable Selection

Variable	Include	Binary	Category	Main Weights	Individual Weights	Overall Weight	Overall Weight (normalised)	Dissimilarity
age	Yes	No	Structural	1	3	0.1765	0.0588	Scaled Euclidean
famsize	Yes	No	Structural	1	3	0.1765	0.0588	Scaled Euclidean
log_cultarea	Yes	No	Structural	1	3	0.1765	0.0588	Scaled Euclidean
log_cattle	Yes	No	Structural	1	3	0.1765	0.0588	Scaled Euclidean
log_smallrum	Yes	No	Structural	1	3	0.1765	0.0588	Scaled Euclidean
log_totmzprod	Yes	No	Functional	1	3	0.2308	0.0769	Scaled Euclidean
log_totmzarea	Yes	No	Functional	1	3	0.2308	0.0769	Scaled Euclidean
log_totmzfert	Yes	No	Functional	1	3	0.2308	0.0769	Scaled Euclidean
foodsec	Yes	No	Nutritional	1	3	0.3333	0.1111	Scaled Euclidean
log_HDDS24H	Yes	No	Nutritional	1	3	0.3333	0.1111	Scaled Euclidean
sex_01	Yes	No	Structural	1	1	0.0588	0.0196	SMC
educ_01	Yes	No	Structural	1	1	0.0588	0.0196	SMC
food_01	Yes	No	Functional	1	1	0.0769	0.0256	SMC
inc_01	Yes	No	Functional	1	1	0.0769	0.0256	SMC
intercrop	Yes	No	Functional	1	1	0.0769	0.0256	SMC
manure	Yes	No	Functional	1	1	0.0769	0.0256	SMC
vitAplant	Yes	No	Nutritional	1	1	0.1111	0.037	SMC
vitAanimal	Yes	No	Nutritional	1	1	0.1111	0.037	SMC
iron	Yes	No	Nutritional	1	1	0.1111	0.037	SMC

Show Plot



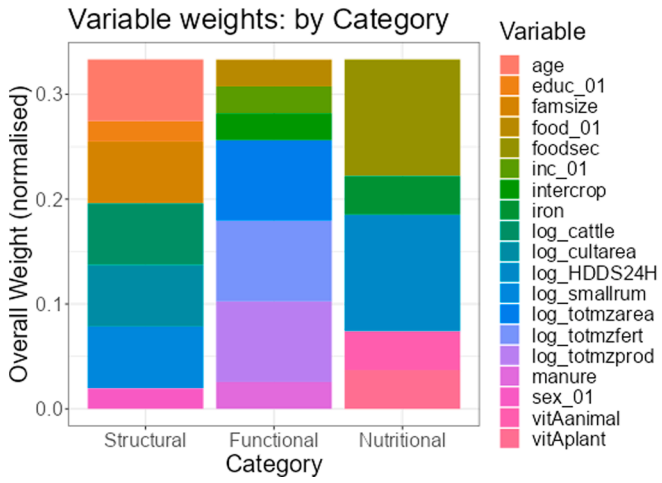


Fig. 2. Screenshots from *TypologyGenerator* illustrating how variables can be categorised and individually weighted. The variables can be assigned to each of the 3 defined categories. In addition, the individual variables can be weighted differently within each category. The overall weight of each variable is updated automatically to show the relative weights between all variables in the data.

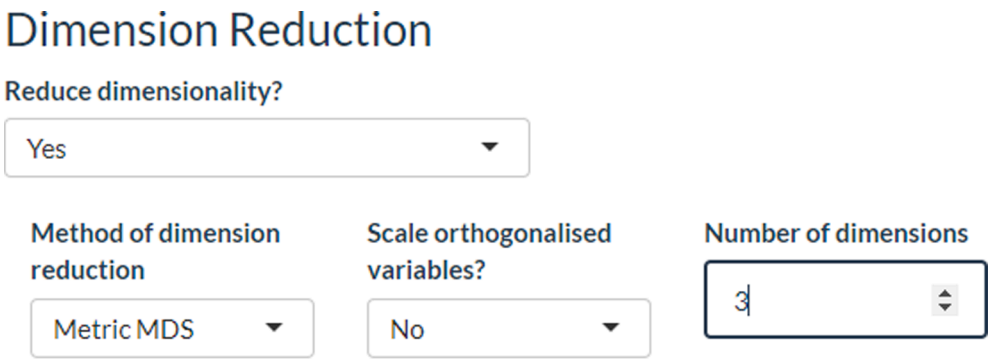
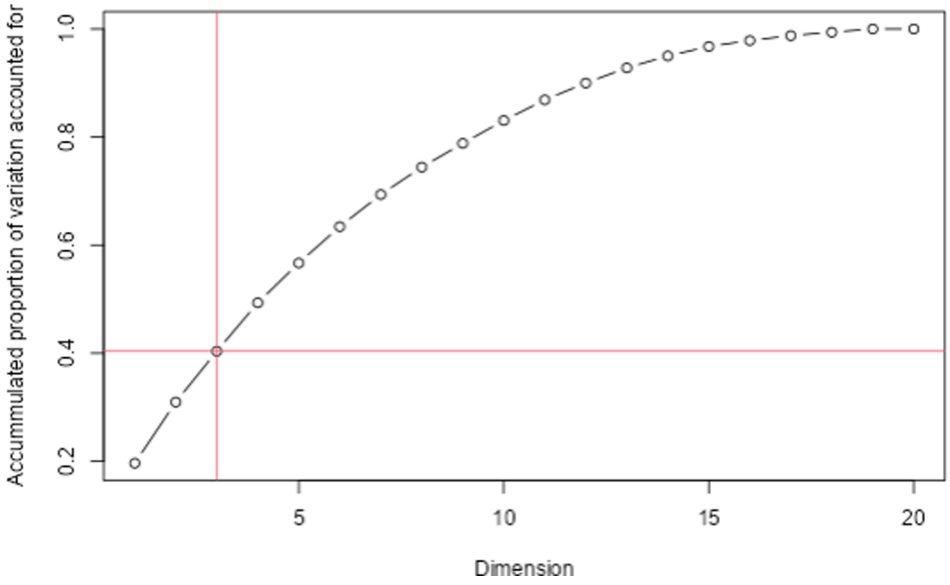


Fig. 3. Screenshot from *TypologyGenerator* illustrating the options available for dimension reduction. Once a user has decided to implement the dimension reduction step along with whether metric or non-metric multi-dimensional scaling is to be used, a diagnostic plot showing the accumulated variance accounted for by each dimension is shown allowing the user to select an appropriate number of dimensions for downstream analysis. The user can choose to scale the resulting orthogonalized variables.



more dominated by 1 group, the lower the value of *J*. To calculate the "clarity", we calculate the range of dendrogram heights for which the chosen number of clusters would be obtained.

4. Results

Here we investigate the effect of different options on the construction of typologies for the Murehwa sampling survey. In total, 36 different parameter sets were explored investigating the effect of variable categorisation (3 categories vs 1 category), high level weighting of the categories (equal weight vs doubling the impact of structural variables), binary variables (included or not), low level individual weighting of the binary variables (equal to continuous variables or downweighted by a factor of 2 or 3) and dimension reduction (either no reduction, or a PCO reduction satisfying a 40% or 50% variance explained threshold). In all cases, a hierarchical clustering approach with complete linkage was used with a visual assessment of the dendrogram to define the final number of clusters. The results are shown in Table 1 and summarised in Fig. 5.

It is clear that the impact of different typology options differs depending on the input data and in combination with other options. For instance, where binary variables are included, in general greater separation between groups (larger F-statistics) is obtained when binary variables are given equal weighting to continuous variables. The exception, however, occurs when a PCO is used and dimensions are retained that explain 40% of the total variation. Here, it is preferable to downweight binary variables but this in turn seems to depend on the overall weighting chosen for the variable categories. It is also interesting to note that the impact of the high-level weighting differs in each scenario (e.g., to include or exclude a dimension reduction step). This is perhaps unsurprising due to the different amounts of data being fed into the algorithms in each case and highlights the need to investigate the typology options each time. In general, including a categorisation of the different input variables improves the distinctness of the typology definition, certainly when the binary variables are included. It is tempting to conclude that including a dimension reduction step improves the distinctness of the typology groups, however, direct quantitative comparisons using the F-statistics from an analysis of similarity cannot be made as the input data are different. Having said this, the fact the F-statistics increase as the dimension reduction is included implies that more background noise is being removed and the algorithms are finding it easier to define the clusters. This does, however come at the cost of interpretation as it can be difficult to identify the variables that are principally responsible for the resulting classification due to the additional step of dimension reduction.

Although the F-statistics give an assessment of whether the identified typologies are statistically distinct, one may also wish to have other features in the typologies. For instance, the cluster evenness gives an indication of how equal the group sizes are in the resulting typologies and the cluster clarity indicates how "easy" it was to define the

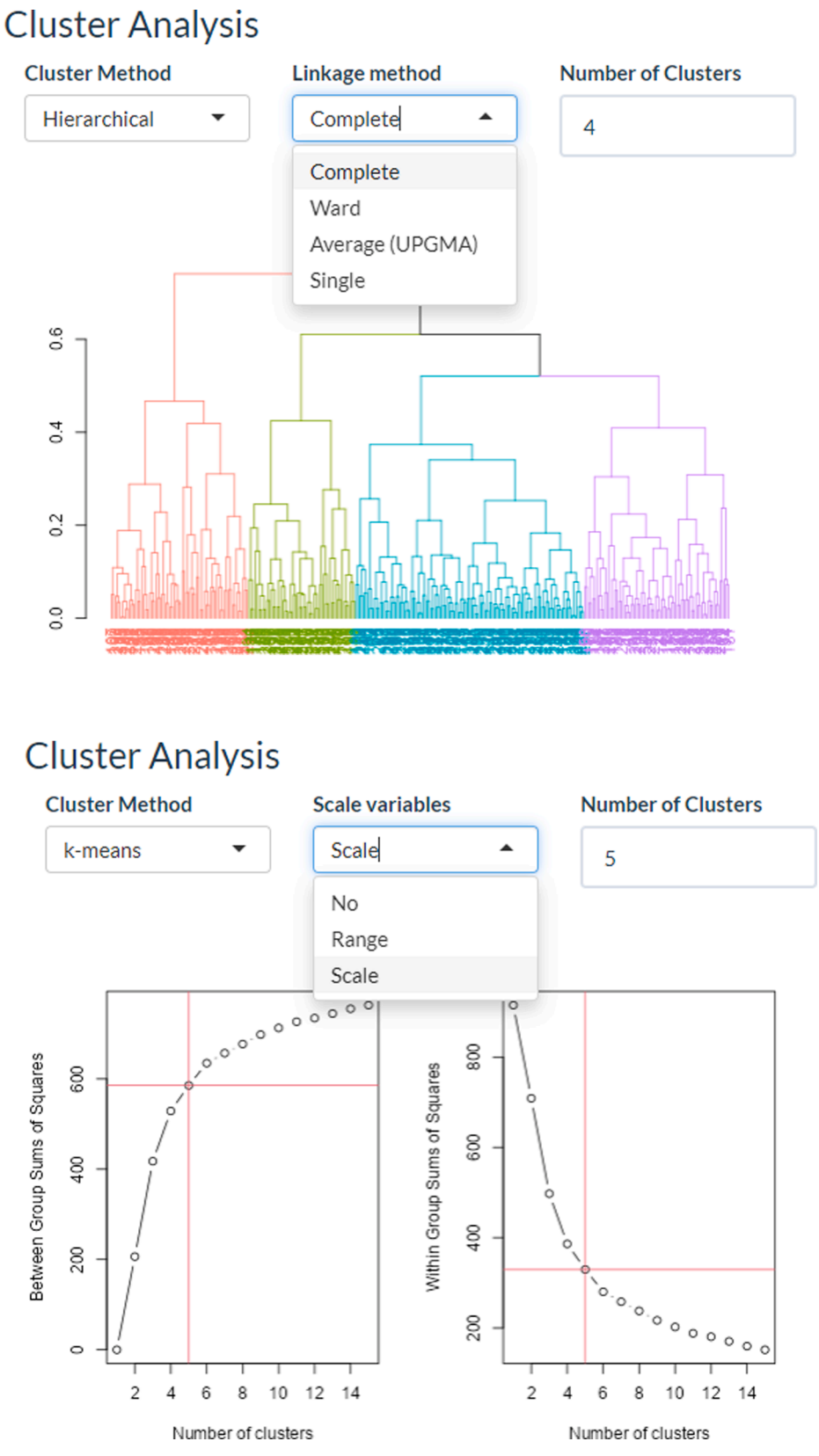


Fig. 4. Screenshots from *TypologyGenerator* illustrating the options for clustering. Two methods of clustering are currently available in the software i) hierarchical clustering (with four possible linkage methods) and ii) k-means clustering. In each case, figures are produced to aid the user in selecting an appropriate number of clusters. For the hierarchical clustering, a dendrogram is shown indicating how the observations are grouped. For k-means clustering, the between group sums of squares and within group sums of squares criteria are shown. The user then selects an appropriate number of clusters, ideally identifying an 'elbow' in the criteria as the number of clusters change.

appropriate cut point on the dendrogram. These are shown in Fig. 6. Cluster evenness is generally best when more data (binary variables included and no dimension reduction) is used. Cluster clarity appears to be greater when binary variables are excluded but when they are included, clarity is generally improved under a dimension reduction step. Examples of different dendrograms exhibiting different levels of evenness and clarity are shown in Appendix A3.

Table 1
Results of typologies constructed under varying parameterisations. The number of dimensions retained in the PCO was defined by the number taken to explain a specific threshold of variance (either 40 or 50%). F-values and associated p-values were obtained from an analysis of dissimilarity of the resulting clusters.

Binary Variables	Binary weights (downweighted factor)	Variable categorisation	Category weights	Dimension Reduction	Threshold for dimension reduction	Number of dimensions retained	Number of clusters	F value (Analysis of dissimilarity)	P value (Analysis of dissimilarity)
excluded	NA	No	NA	No	NA	NA	3	41.30	0.001
excluded	NA	Yes	1	No	NA	NA	4	44.26	0.001
excluded	NA	Yes	2	No	NA	NA	4	37.61	0.001
excluded	NA	No	NA	Yes	exceed 40%	2	4	196.62	0.001
excluded	NA	Yes	1	Yes	exceed 40%	2	3	168.35	0.001
excluded	NA	Yes	2	Yes	exceed 40%	2	3	208.56	0.001
excluded	NA	No	NA	Yes	exceed 50%	3	4	88.05	0.001
excluded	NA	Yes	1	Yes	exceed 50%	2	3	168.35	0.001
excluded	NA	Yes	2	Yes	exceed 50%	3	3	91.76	0.001
included	1	No	NA	No	NA	NA	5	26.52	0.001
included	1	Yes	1	No	NA	NA	3	32.19	0.001
included	1	Yes	2	No	NA	NA	4	33.50	0.001
included	2	No	NA	No	NA	NA	3	22.60	0.001
included	2	Yes	1	No	NA	NA	4	21.59	0.001
included	2	Yes	2	No	NA	NA	3	25.05	0.001
included	3	No	NA	No	NA	NA	4	18.11	0.001
included	3	Yes	1	No	NA	NA	3	26.78	0.001
included	3	Yes	2	No	NA	NA	4	23.35	0.001
included	1	No	NA	Yes	exceed 40%	3	4	99.36	0.001
included	1	Yes	1	Yes	exceed 40%	3	4	92.42	0.001
included	1	Yes	2	Yes	exceed 40%	3	3	125.14	0.001
included	2	No	NA	Yes	exceed 40%	3	5	104.47	0.001
included	2	Yes	1	Yes	exceed 40%	3	4	98.32	0.001
included	2	Yes	2	Yes	exceed 40%	3	4	80.37	0.001
included	3	No	NA	Yes	exceed 40%	4	3	61.46	0.001
included	3	Yes	1	Yes	exceed 40%	3	4	113.43	0.001
included	3	Yes	2	Yes	exceed 40%	4	3	62.66	0.001
included	1	No	NA	Yes	exceed 50%	4	4	60.78	0.001
included	1	Yes	1	Yes	exceed 50%	4	4	69.59	0.001
included	1	Yes	2	Yes	exceed 50%	4	4	68.39	0.001
included	2	No	NA	Yes	exceed 50%	5	3	44.31	0.001
included	2	Yes	1	Yes	exceed 50%	5	3	50.40	0.001
included	2	Yes	2	Yes	exceed 50%	5	3	41.59	0.001
included	3	No	NA	Yes	exceed 50%	5	3	40.81	0.001
included	3	Yes	1	Yes	exceed 50%	5	4	54.58	0.001
included	3	Yes	2	Yes	exceed 50%	5	5	45.52	0.001

5. Case study

To understand farming household heterogeneity in Murehwa, we developed a delineation through the *TypologyGenerator*. The final typology selected for the Murehwa case study consisted of the following options. Within each category of variables (functional, nutritional and structural), continuous variables were given three times as much weight as binary variables. A PCO was done to retain three dimensions and four cluster were delineated from the resulting dendrogram (Fig. 7).

As seen in Fig. 8, the four identified typologies represent different characteristics of the farming landscape. Specifically,

- Type 1 are relatively large farms, with relatively large herds, depending on crop production for both food and income
- Type 2 are medium-sized farms with diversified livelihoods (crop production is not the main source of food and income for a large proportion of farms in that category)
- Type 3 are relatively small farms, with a high proportion of female-headed household, depending on crop production for both food and income
- Type 4 are vulnerable households, with small farms and small herds, dependent on off-farm activities.

Thus, although wealth is a clear demarcation between typologies, we see a much more nuanced description of resource endowment through these typologies. In particular, Type 1 and 2 are relatively food secure but it is only Type 1 that demonstrates a high dietary diversity score with food rich in iron or animal sourced vitamin A. Education scores are relatively consistent across the four Types although a greater propensity for male-headed households is seen in Types 1 and 2. Type 1 tends to have quite distinct agronomical characteristics compared to Types 2–4 with higher use of manure, compost and fertilizer and larger herds of both cattle and small ruminants.

These typologies formed the basis of a stratified random sampling scheme where participants were selected for a program of targeted interventions around biofortified maize. The results of this study will be the focus of separate papers and it is anticipated that the derived typologies will give context to the wider research question by simplifying the highly heterogeneous landscape of smallholder farms.

6. Discussion

The selection of variables to be used in the construction of farm typologies will primarily be context driven and it is the belief of the authors that the mathematical constraints of the typology construction should have as little impact on the choice of variables as possible. Specifically, no constraint on whether data should be quantitative or qualitative should be imposed. It is the experience of the authors that choosing to exclude qualitative variables can have large impact on the resulting typologies and as such, the decision to exclude or include them should be done in an informed way as implemented in the *TypologyGenerator*.

Of course, not all variables may be suitable candidates for selection. Data quality, as ever, remains one of the most critical assessments for inclusion. More subtly, variables may need editing in various ways; examples include combining categories of a qualitative variable if

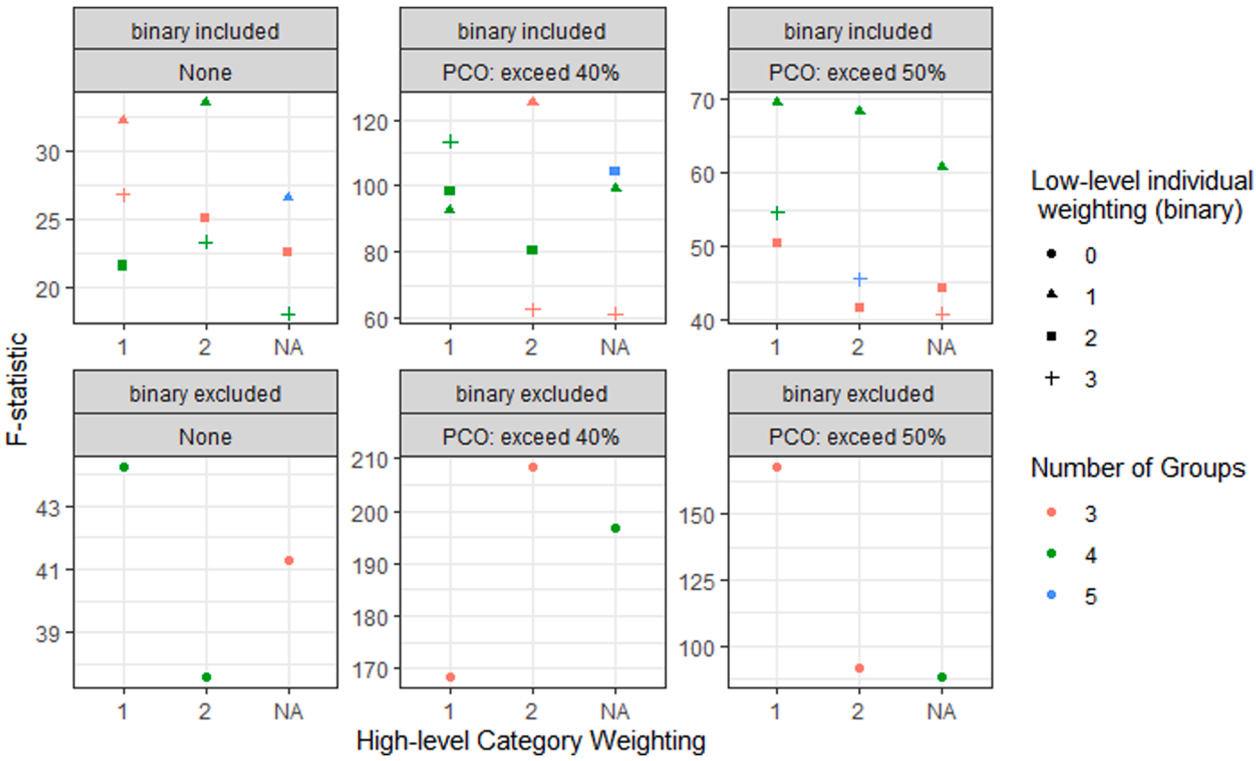


Fig. 5. F-statistics as obtained from the analysis of similarity for each combination of input options to the *TypologyGenerator*. Note, F-statistics are only directly comparable for typologies based on the same input data and should therefore only be compared within each panel.

individual group sizes are too small or transforming variables to ensure efficient representation of the spread. These edits will result from a thought-driven process and several iterations, influenced by the impact on the downstream analysis, are likely to be needed. One further consideration is the presence of (multi-)collinearity between variables. Where multiple variables are highly co-linear, a greater emphasis is placed on this feature in the resulting cluster analysis. For example, consider the extreme case where two variables are perfectly correlated, then the same feature will have twice as much contribution to the distance measure used for clustering compared to any variable to which it is not correlated. To address this, studies often run a pre-processing step to identify correlated variables (e.g., via PCA) and to select a subset that are independent. By providing a framework of weighting variables, we allow users to input all variables (collinear or not) with equal weight, i.e., each of two collinear variables should be given an individual weight of 0.5. Note, a direct replacement of these variables with an orthogonalized version (e.g., the PC scores) will not address the underlying issue unless the orthogonalized variables are also subsequently standardised.

More generally, selection of weights will be driven by *a priori* knowledge of what the typologies should represent but as it is a subjective decision, an iterative procedure should be implemented to understand the sensitivity of the resulting typologies. There are many subjective choices in the process to reach a typology definition, some of which will be more clearcut than others. A thorough validation process will give an assessment of the robustness of the resulting typologies to these decisions. Validation will consist of i) investigating how well each variable is represented in the dimension reduction step, ii) investigating how robust the cluster groups are defined, and iii) what the clusters represent in terms of the original variables. *TypologyGenerator* allows a user to investigate these features through a visual assessment and it remains an open challenge to provide a more automated quantitative assessment.

Creating farm typologies involves a series of both subjective and objective decisions as well as the use of multiple statistical methods (dimension reduction, clustering, and multivariate significance testing). Consequently, typologies may be a somewhat intimidating undertaking for many researchers in agriculture, who are often expert agronomists, ecologists, and/or social scientists (among others). Similarly, to readers of published typology studies who are unfamiliar with the approach, it may be difficult to disentangle the subjective from objective steps when considering the strength and relevance of a study's findings. To help overcome these challenges, this paper and the associated *TypologyGenerator* lay out a clear framework of the steps involved in typology generation and the possible options at each step, alongside visualisations and diagnostic tools to help researchers validate their decisions. We therefore hope to help researchers both utilise the full potential of the typology method and to report their steps taken and decisions made in a systematic way that ensures robust conclusions.

Deriving farm typologies remains a highly relevant technique for describing the highly heterogeneous landscape of small-holder farmers in rural sub-Saharan Africa. Previous studies have demonstrated that different farm types in the study area (and in other areas in rural sub-Saharan Africa) have different soil fertility management practices, leading to different level of crop productivity (Chikowo et al. 2014). In particular, through the delineation of a farm typology we are able to target interventions (different farm types being characterized by different resources, constraints and opportunities, potentially affecting their adoption of technologies) and to scale technologies (i.e., to understand how representative a particular farm is of the larger farm population).

7. Conclusion

This paper has described a unified framework for constructing typologies and provided an open-source software application that implements this framework. By doing so, we provide the research community with an easily accessible route to develop typologies, emphasising the need for an iterative approach investigating the impact of individual methodologies, e.g. to include a dimension reduction step or not. We have demonstrated how the myriad of options available in the steps of

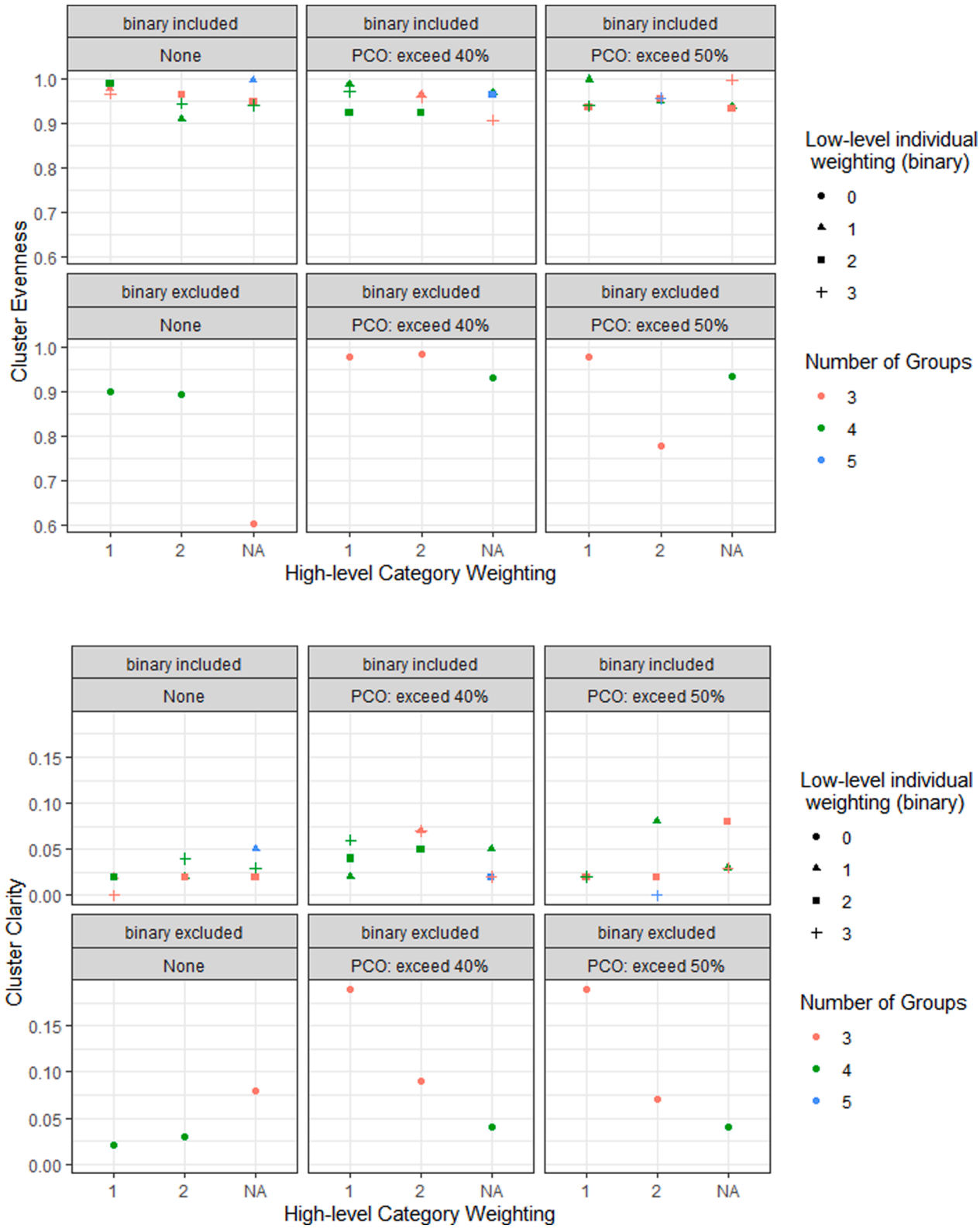


Fig. 6. Cluster evenness (top panel) and clarity (bottom panel) for each of the scenarios listed in Table 1.

constructing farm typologies can have great influence on the resulting groups through the use of statistical summaries describing the separation of clusters.

Defining a small number of distinct farm typologies can be very efficient for capturing the main sources of diversity between different farming systems and we have applied our methods to defining a four-group typology over 306 households in Murehwa district of Zimbabwe. The four groups provide a nuanced separation based primarily on resource endowment but with additional emphasis on gender, diet and agronomy. This typology has formed the basis of a stratified random sampling scheme that was utilised in a subsequent on-farm study of maize variety performance, the results of which will be

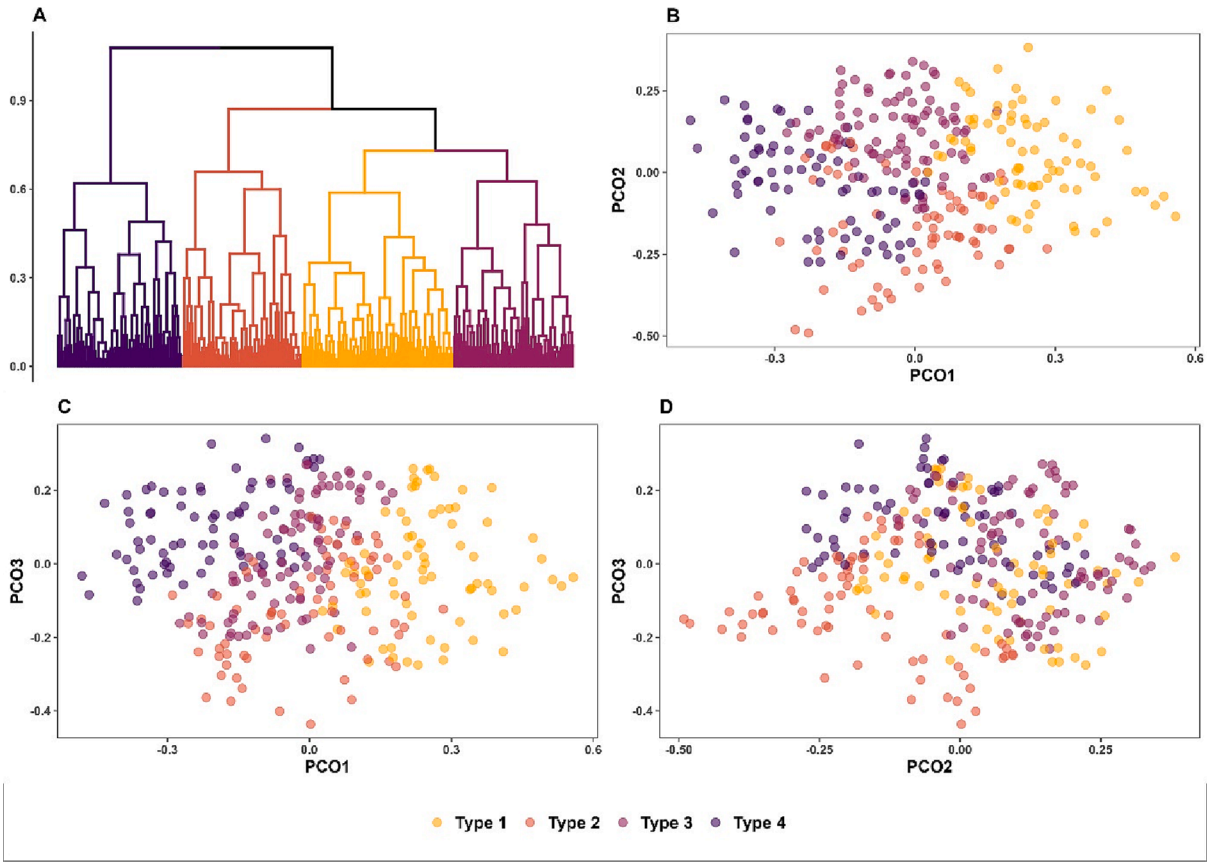


Fig. 7. (A) Dendrogram representing the hierarchical agglomerative clustering using Ward's method (four clusters were identified), and representation of the four farm types identified (B) on the plane defined by the first two principal components, (C) on the plane identified by the first and the third principal component, and (D) on the plane identified by the second and the third component.

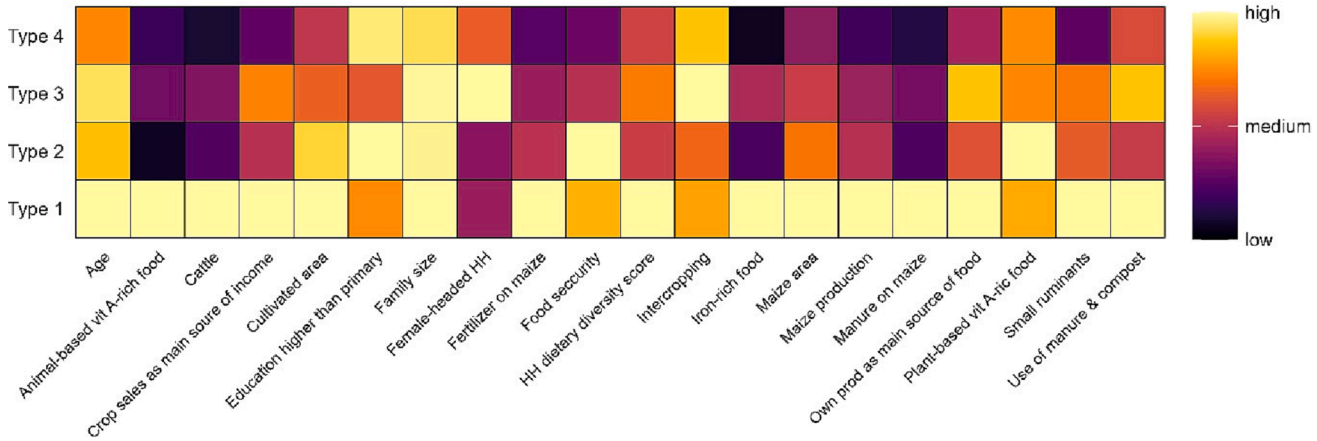


Fig. 8. Heatmap representing the mean value (age of the head of the household, number of cattle, cultivated area, family size, quantity of fertilizer used on maize, household dietary diversity score, maize area, maize production, quantity of manure used on maize, and number of small ruminants) and the proportion of households concerned (consumption of animal-based vitamin A-rich food, crop sales as the main source of income, education of the head of the household higher than primary level, female-headed household, proportion of the year the household as adequate food security, consumption of iron-rich food, own production as the main source of food, consumption of plant-based vitamin A-rich food, and use of manure and compost) for key variables.

presented in a subsequent manuscript.

8. Software and/or data availability section

All methods described in this paper have been incorporated into an open source software *TypologyGenerator*, freely available at https://github.com/KirstyLHassall/TypologyGenerator with DOI https://doi.org/10.5281/zenodo.7727862(Hassall, 2023). Alternatively, the users can access the app directly at https://kirstylhassall.shinyapps.io/TypologyGenerator/.

The Murehwa case study data is available to download from the GitHub repository under the file TP.csv.

Ethical compliance

We compiled with all relevant ethical regulations regarding human research participants. The survey was approved by the ethics committee of the International Maize and Wheat Improvement Center (CIMMYT). Informed consent was obtained from all farmer participants.

Funding

This study was supported through the UK Global Challenges Research Fund administered by the Biotechnology and Biological Sciences Research Council for the project "Addressing malnutrition with biofortified maize in Zimbabwe: from crop management to policy and consumers (IATI Identifier: GB-GOV-13-FUND–GCRF-BB_T009047_1)". Rothamsted Research receives strategic funding from the Biotechnology and Biological Sciences Research Council of the United Kingdom. We acknowledge support from the Growing Health (BB/X010953/1) Institute Strategic Programme

CRediT authorship contribution statement

Kirsty L. Hassall: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Frédéric Baudron:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Chloe MacLaren:** Writing – review & editing. **Jill E. Cairns:** Writing – review & editing. **Thokozile Ndhlela:** Writing – review & editing. **Steve P. McGrath:** Writing – review & editing. **Isaiah Nyagumbo:** Writing – review & editing. **Stephan M. Haefele:** Writing – review & editing, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All code and data are available at https://github.com/KirstyLHassall/TypologyGenerator.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2023.108074.

References

Alvarez, S., Paas, W., Descheemaeker, K., Tittonell, P., Groot, J., 2014. Typology construction, a way of dealing with farm diversity. In: Report for the CGIAR Research Program on Integrated Systems for the Humid Tropics., Plant Science (December), pp. 1–37.

Alvarez, S., Timler, C.J., Michalscheck, M., Paas, W., Descheemaeker, K., Tittonell, P., Groot, J.C.J., 2018. Capturing farm diversity with hypothesis-based typologies: an innovative methodological framework for farming system typology development. PLoS One 13 (5), 1–24. https://doi.org/10.1371/journal.pone.0194757.

Anderson, M.J., 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 26, 32–46. https://doi.org/10.1111/j.1442-9993.2001.01070. pp. x.

Berre, David, Frédéric Baudron, Menale Kassie, Peter Craufurd, And Santiago Lopez-Ridaura (2019). Different ways to cut a cake: comparing expert-based and statistical typologies to target sustainable intensification technologies, a case-study in southern Ethiopia. Experimental Agriculture 55 (S1). Cambridge University Press: 191–207. doi:10.1017/S0014479716000727.

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. Agr. Syst. 101 (1–2), 30–41. https://doi.org/10.1016/j.agsy.2009.02.004.

Cairns, J., Baudron, F., Hassall, K., Ndhlela, T., Nyagumbo, I., McGrath, S., Haefele, S., 2021. Revisiting strategies to incorporate gender intentionality into maize breeding in southern Africa. Outlook on Agriculture. https://doi.org/10.1177/00307270211045410.

Chikowo, R., Zingore, S., Snapp, S., et al., 2014. Farm typologies, soil fertility variability and nutrient management in smallholder farming in Sub-Saharan Africa. Nutr Cycl Agroecosyst 100, 1–18. https://doi.org/10.1007/s10705-014-9632-y.

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., 2011. Communicating complexity: Integrated assessment of trade-offs concerning soil fertility management within African farming systems to support innovation and development. Agric. Syst. 104 (2), 191–203. https://doi.org/10.1016/j.agsy.2010.07.002.

Hammond, J., Rosenblum, N., Breseman, D., Gorman, L., Manners, R., van Wijk, M.T., Sibomana, M., Remans, R., Vanlauwe, B., Schut, M., 2020. Towards actionable farm typologies: Scaling adoption of agricultural inputs in Rwanda. Agric. Syst. 183, 102857 https://doi.org/10.1016/j.agsy.2020.102857.

Hassall, K.L., 2023. TypologyGenerator. https://doi.org/10.5281/zenodo.7727862.

Kansiime, M.K., van Asten, P., Sneyers, K., 2018. Farm diversity and resource use efficiency: Targeting agricultural policy interventions in East Africa farming systems. NJAS - Wageningen Journal of Life Sciences 85, 32–41. https://doi.org/10.1016/j.njas.2017.12.001.

Kansiime, M.K., Girling, R.D., Mugambi, I., Mulema, J., Oduor, G., Chacha, D., Ouvrard, D., Kinuthia, W., Garratt, M.P.D., 2021. Rural livelihood diversity and its influence on the ecological intensification potential of smallholder farms in Kenya. Food Energy Secur. 10 (1), 1–13. https://doi.org/10.1002/fes3.254.

Kuivanen, K.S., Alvarez, S., Michalscheck, M., Adjei-Nsiah, S., Descheemaeker, K., Mellon-Bedi, S., Groot, J.C.J., 2016a. Characterising the diversity of smallholder farming systems and their constraints and opportunities for innovation: a case study from the Northern Region, Ghana. NJAS - Wageningen Journal of Life Sciences 78, 153–166. https://doi.org/10.1016/j.njas.2016.04.003.

Kuivanen, K.S., Michalscheck, M., Descheemaeker, K., Adjei-Nsiah, S., Mellon-Bedi, S., Groot, J.C.J., Alvarez, S., 2016b. A comparison of statistical and participatory clustering of smallholder farming systems - a case study in Northern Ghana. J. Rural. Stud. 45, 184–198. https://doi.org/10.1016/j.jrurstud.2016.03.015.

MacLaren, C., Aliyu, K.T., Waswa, W., Storkey, J., Claessens, L., Vanlauwe, B., Mead, A., 2022. Can the right composition and diversity of farmed species improve food security among smallholder farmers? Front. Sustain. Food Syst. 6, 1–19. https://doi.org/10.3389/fsufs.2022.744700.

Molua, E., 2011. Farm income, gender differentials and climate risk in Cameroon: typology of male and female adaptation options across agroecologies. Sustain. Sci. 6 (1), 21–35. https://doi.org/10.1007/s11625-010-0123-z.

Oksanen, Jari, Simpson, Gavin L., Guillaume Blanchet, F., Kindt, Roeland, Legendre, Pierre, Minchin, Peter R., O'Hara, R.B., Solymos, Peter, Henry, M., Stevens, H., Szoecs, Eduard, Wagner, Helene, Barbour, Matt, Bedward, Michael, Bolker, Ben, Borcard, Daniel, Carvalho, Gustavo, Chirico, Michael, Caceres, Miquel De, Durand, Sebastien, Antoniazi Evangelista, Heloisa Beatriz, FitzJohn, Rich, Friendly, Michael, Furneaux, Brendan, Hannigan, Geoffrey, Hill, Mark O., Lahti, Leo, McGlinn, Dan, Ouellette, Marie-Helene, Ribeiro Cunha, Eduardo, Smith, Tyler, Stier, Adrian, Ter Braak, Cajo J.F., Weedon, James, 2022. vegan: Community Ecology Package. R package version 2.6-4. https://CRAN.R-project.org/package=vegan.

Pacini, G.C., Colucci, D., Baudron, F., Righi, E., Corbeels, M., Tittonell, P., Stefanini, F. M., 2014. Combining multi-dimensional scaling and cluster analysis to describe the diversity of rural households. Exp. Agric. 50 (3), 376–397. https://doi.org/10.1017/S0014479713000495.

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. Agric. Ecosyst. Environ. 142 (1–2), 63–74. https://doi.org/10.1016/j.agee.2010.07.011.

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. J. Sustain. Agric. 36 (5), 564–590. https://doi.org/10.1080/10440046.2012.672547.

Rusere, F., Mkuhlani, S., Crespo, O., Dicks, L.V., 2019. Developing pathways to improve smallholder agricultural productivity through ecological intensification technologies in semi-arid Limpopo, South Africa. Afr. J. Sci. Technol. Innov. Dev. 11 (5), 543–553. https://doi.org/10.1080/20421338.2018.1550936.

Sakané, N., Becker, M., Langensiepen, M., Van Wijk, M.T., 2013. Typology of smallholder production systems in small east-African wetlands. Wetlands 33 (1), 101–116. https://doi.org/10.1007/s13157-012-0355-z.

Silva, J.V., Baudron, F., Ngoma, H., et al., 2023. Narrowing maize yield gaps across smallholder farming systems in Zambia: what interventions, where, and for whom? Agron. Sustain. Dev. 43, 26. https://doi.org/10.1007/s13593-023-00872-1.

Tittonell, P., Vanlauwe, B., Leffelaar, P.A., Rowe, E.C., Giller, K.E., 2005. Exploring diversity in soil fertility management of smallholder farms in western Kenya: I. Heterogeneity at region and farm scale. Agr. Ecosyst. Environ. 110 (3–4), 149–165. https://doi.org/10.1016/j.agee.2005.04.001.

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. Agr. Syst. 103 (2), 83–97. https://doi.org/10.1016/j.agsy.2009.10.001.

ZimVac, 2020. Food and nutrition security update report, February 2020. Vulnerability Assessment Committee, Harare, Zimbabwe.