Farm household data platform

Romain Frelat, James Hammond, and Mark T. van Wijk

Last update: 14/06/2024

Overview

This document provides a detailed description of the farm household data platform that was developed in 2023-2024. The objective was to build a **flexible data environment** that could host the large diversity of farm household data (collected by different surveys/scientists/institutions) and provide **tools to run mixed farming system analysis**, including data exploration, farm typologies, ex-ante scenario assessment, or positive deviance analysis.

In short, the **data platform aims to provide scientists with a set of tools to analyze their farm household survey data** with minimum effort in statistical analysis or programming skills. In addition, having a platform hosting different household surveys in one common data environment will allow **comparisons and compilations of multiple data sources**. The compilation of farm household data will allow large-scale analysis (across farming systems or across regions) and possibly temporal analysis too.

The data platform is in the format of an open-source R-package hosted on the Github platform (https://github.com/rfrelat/farmhousehold). It allows everyone to install it easily on their computer. Moreover, the Github platform has tools to report issues, document on-going development, and allow the contributions of multiple users. Following the R-package format helps structure the documentation.

In addition to the R-package, we developed three interactive dashboards, in the format of shinyapps. These dashboards allow users to directly see and interact with the dataset without having to open or write an R-command.

We hope that this documentation will help spread the use of this farm household data platform and further continue its development. This is an evolving tool that requires a diversity of contributors to ensure that it fits the needs of most farming system scientists.

Reminder: All information in the household data platform was based on responses given by smallholder farmers to an (often long and tiring) interview. It was entered and evaluated with the best care, yet this compilated information might contain errors and biases. All output should be evaluated with care and local expertise. Be extremely careful when extrapolating; be sure you understand the sampling strategy of the survey and its possible limitations, and avoid using outliers as models.

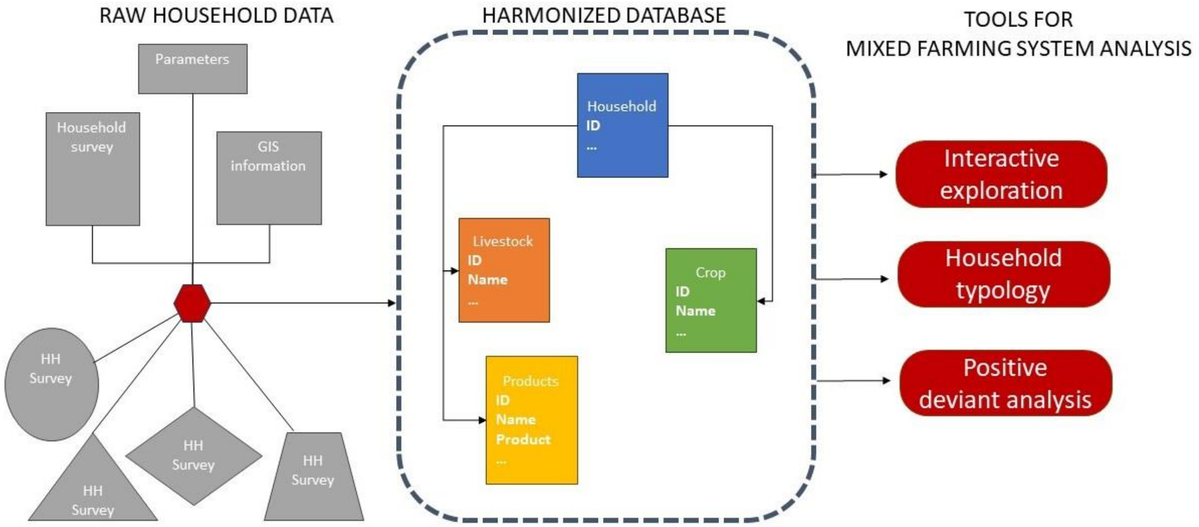


Figure 1: Overview of the household data platform. It can host different format of household data (the different shapes represents different surveys) into one common and harmonized database made of 4 related tables. From this common format, tools to run mixed farming system analysis are available as interactive dashboard or as functions in the R-package.

Table of content

Description of the relational database

Four related tables

The main innovation of the farm household data platform is the format of the database. We organized the information collected on farm households into four distinct and related tables: crop, livestock herd, livestock production, and household information (Figure 2). Together, this four-table relational database creates a new class of R objects called `farmhousehold`.

In the following description of the tables, we describe the key variables that need to be present in order to create new `farmhousehold` objects and run the farm system analysis. Yet, if some information is missing (e.g., it was not asked during the survey), fill in NA (and keep in mind that this information was not recorded when analyzing the results). On the contrary, if you want to keep track of other information that is relevant to your project, you can add more variables to each of these tables, with your own naming of the variables.

One key variable of the related database is the household ID called `hhid`. This ID must be unique and identify unequivocally a single household. With the variable `hhid`, we will be able to know the crop production or the livestock production of each household. It is recommended to use human-readable ID, using a combination of the country, the name of the project, the year, and add a unique number for each household (e.g., *ke_2018_53*)

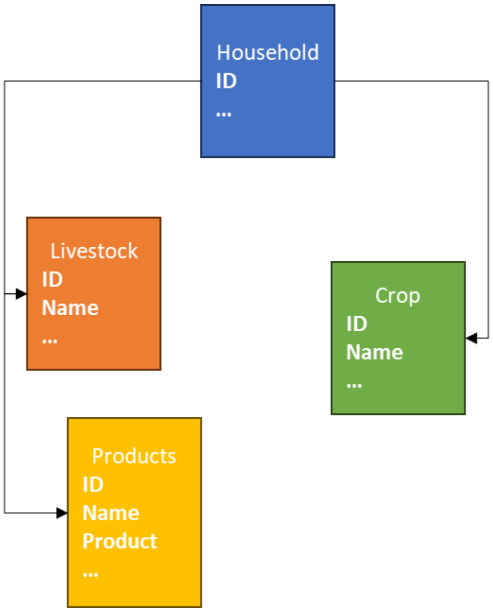


Figure 2: Format of the farm household database with four related tables: Crop, Livestock herd, Livestock production, and Household information.

Crop table

The crop table contains information about the crops cultivated by each household (Table 1). It contains information about the land dedicated to each crop (in ha), the quantity harvested (in kg), and its usage (either for consumption, sale, or other purposes). Please verify that the sum of the quantity sold and the quantity consumed is not higher than the quantity harvested.

Table 1: Definition of the variables in the crop table

Name	**Definition**	**Unit**	**Example**
hhid	household id		*ke_2018_53*
name	name of the crop		*maize*
land_area_ha	land cultivated	hectare	*0.5*
harvest_kg	amount harvested	kg	*420*
consumed_kg	amount consumed	kg	*250*
sold_kg	amount sold	kg	*170*
income_lcu	income from sells	local currency unit	*2500*

When intercropping, the information for each crop is reported in different lines. For instance, maize and beans intercropping equally on a field of 1 ha will make two lines in the crop table, one with maize on 0.5 ha and one with beans on 0.5 ha. When known, we keep the proportions of the intercropping when dividing the land area. If not, we assume that the land is equally divided among crops. It is important that the sum of the land area of the intercropped crops is equal to the area of the field intercropped.

Usually, the information is reported for one complete year (the previous 12 months). If there are more than one season, it is important to report the two (or three) seasons (in different lines). For instance, if a farmer cultivates onions in the wet season on 0.6 ha, and in the dry season on 0.2 ha; then it is recommended to keep two separate lines with the two cultivations of onions (the usage might depend on the season).

In general, we keep the information with the maximum precision and the least aggregation possible. If information on crop harvest and usage is available per field and per season, then this information will directly feed into the crop table. Unfortunately, in some cases, the crop usage is aggregated per crop for the whole year. In these cases, the crop harvested has to be aggregated per crop as well in order to have both the crop production and the crop usage in the same format.

Livestock table

The livestock table contains information about the livestock that are kept by each household (Table 2). It contains information about how many animals are kept by the household. Usually, this information is recorded at the time of the survey, disregarding the dynamics of the livestock herd (how much was sold or bought during the year).

Table 2: Definition of the variables in the livestock table

Name	Definition	Unit	Example
hhid	household id		*ke_2018_53*
name	name of the livestock		*cattle*
n	number of livestock kept		*8*

If needed, this table can host more information. For instance, one can add information about the breed, or how it is fed and how it is kept (inside and/or outside). Currently, this information is not needed for the farm system analysis that has been developed.

Livestock production table

The livestock production table contains information about the production of the livestock, e.g., milk, eggs, meat, etc. (Table 3). It contains information about the quantity produced (in kg) and its usage (either for consumption, sale, or other purposes). It is important that the information reported in the name match the name of the livestock table. Please verify that the sum of the quantity sold and the quantity consumed is not higher than the quantity harvested.

Table 3: Definition of the variables in the livestock production table

Name	Definition	Unit	Example
hhid	household id		*ke_2018_53*
name	name of the livestock		*cattle*
prod	livestock product	category	*milk*
harvest_kg	amount harvested	kg	*237*
consumed_kg	amount consumed	kg	*150*
sold_kg	amount sold	kg	*87*
income_lcu	income from sells	local currency unit	*3500*

The most common livestock products are milk, eggs, honey, and meat, but it can also contain information about manure, wool, power force, or any other benefit that households get from their livestock. It is important to transform the quantity reported into a total quantity per year (= the last 12 months). For instance, the milk production per cow and per day need to be multiplied by the number of milking cows and the number of milking days per cow to get an estimate of the milk production per year and per household. Similarly, the quantities consumed and sold, as well as the income from sales should be reported per year.

The sales of whole livestock are also reported in this table as product `whole`. In this case, the number of livestock sold has to be transformed into a quantity harvested in kg, using an estimate of the weight per animal (see tlu conversions, Table 7, with 1TLU=250kg). The quantity consumed will be 0 kg, and the quantity sold will be the quantity harvested.

Household table

The household table contains information about the location, the household size, the off-farm activities, and the food security (Table 4).

Table 4: Definition of the variables in the household table

Name	Definition	Unit	Example
hhid	household id		*ke_2018_53*
country	country of the survey		*kenya*
large_region	geopolitical region (continental scale)		*Eastern Africa*
region	region or province (national scale)		*nandi*
year	year of the survey		*2018*
gps_lat	latitude in decimal degrees	°N	*-0.7*
gps_lon	longitude in decimal degrees	°E	*35.1*
hh_size_members	size of the household in number of persons		*5*
hh_size_mae	size of the household in male adult equivalent	MAE	*3.8*
off_farm_lcu	off farm income per year	LCU	*1200*
off_farm_div	diversity of off farm activities		*1*
hdds_score	household diet diversity score based on 10 groups		*6*
fies_score	food insecurity experience scale based on 8 questions		*3*
foodshortage_count	number of months with food shortage		*2*
foodshortage_months	name of the months with food shortage		*oct nov*
currency_conversion	conversion from local currenty to power parity purchase usd	lcu/usd	*41.9*

Based on the household composition, we report the number of members (hh_size_members) and the size of the household expressed in Male Adult Equivalent (MAE). To know the size of households in MAE, we classify the members into gender categories and age classes, and each group has a coefficient based on its food requirement (Table 5).

Table 5: Male Adult equivalent coefficient per gender and age classes

Age classes	**0-4**	**5-10**	**11-24**	**25-50**	**51+**
Male	0.5	0.75	0.925	1	0.73
Female	0.5	0.75	0.75	0.86	0.6

The location of the household must contain the country, and possibly the large regions (continental scale) and the region or province (national scale). The gps coordinates should be rounded (at least at 0.01 decimal degrees) to preserve the anonymity of the household.

The off-farm activities should contain an estimate of the amount earned in the previous 12 months (here year is defined as with the crop and livestock table, it is not necessary a calendar year). The number of different off-farm activities is reported under the variable `off_farm_div`.

The food security status can be reported as the number of months with food shortages, estimated with the household diet diversity score (HDDS) and/or the food insecurity experience scale (FIES). It is recommended to keep recorded in which month households experience food shortages in the variable (foodshortage_months) with the three-letter abbreviation of the months, separated by a space (e.g., 'oct nov' if the household experienced food shortage in October and November). It is also recommended to keep the individual answers of the HDDS and the FIES scores, in the columns starting with 'HDDS_' and 'FIES_' respectively. For instance, one column can be `HDDS_meat` containing the binary variable stating whether the household eats meat regularly (=1) or not (=0). The sum of the HDDS_[X] column should be equal to the score in the column `hdds_score`. The same applies for the FIES score. If any of the three food security indicators are missing, fill the corresponding column with NA.

Parameters

There are two important parameters for the calculations of farm productions: the energy content (`conv_energy`) and the conversion to the tropical livestock unit (`conv_tlu`).

The energy content is a vector with the energy conversion in kcal/kg. A good source of information is the FoodData Central of the U.S.Departement of Agriculture (https://fdc.nal.usda.gov/).

Table 6: Example of energy content in kcal/kg

Item	maize	bean	sorghum	onion	groundnut	milk	eggs
Energy	3650	1480	3390	720	5670	597	1550

The conversion of tropical livestock unit is related to the weight of the animal (1TLU = 250kg).

Table 7: Example of the tropical livestock unit conversion factors

Livestock	camel	cattle	donkey_horse	goat	sheep	chicken	pig
TLU	0.7	0.7	0.7	0.1	0.1	0.01	0.3

It is important that all livestock listed in the livestock table (under the variable `name`) have a conversion for the tropical livestock unit (with the exact same name, including the same case and singular/plural form). Similarly, it is important that all crop and livestock products listed in the crop and livestock production tables have an energy conversion factor. Parameters used for the RHoMIS dataset are available in the Github folder inst/load_data/Parameters.

GIS information

To help with the comparison of the households, we added GIS information based on the localization of the household. We use four different sources of information, which creates four variables:

- farming_system: from the Dixon farming system map for Sub-Saharan Africa (Dixon et al. 2019).
- population_2020: population density map for 2020, at 30 arc-second resolutions (GPWv4, CIESIN, 2018)
- travel_time_cities: travel time to the closest city in minutes (Nelson, 2019)
- koeppen: Koeppen-Geiger climate classification of present time at 1km resolution (Beck et al. 2018)

Be careful that the GPS coordinates are often deteriorated (rounded at 0.1 or 0.01 decimal) to safeguard the anonymity of the households. It means that the small-scale variations will not be captured (especially for population density and travel time to the closest cities). It is recommended to use the exact GPS coordinates (before rounding) for the extraction of GIS information.

Description of the R-package

Structure of the package

The farmhousehold package follows the recommended structure of R packages (R Core Team, 1999, Wickham and Bryan 2023). There are four main folders with:

- data: the dataset provided with the package as an example dataset
- inst: supporting information for the package, including the shinyapps, the documentation, and script to transform data into farmhousehold objects
- man: the documentation of all the functions of the package
- R: the definition of the function as R scripts

Naming convention

In the file `tools.R`, there are two functions to help rename the crop and livestock following a common convention: `cleaname`(), `bestname`(). The function `cleaname`() makes sure that all strings are in lowercase and do not contain any brackets or numbers. The spaces or the minus ('-') are replaced by an underscore ('_'). The function `bestname`() compares unknown names with a list of known/accepted names. If the difference is only an underscore or a plural/singular form, it replaces the unknown name with the matching accepted name.

Automated update of the household farm production

In the file `calc_farm_prod.R`, we defined the function that summarizes the annual farm household productions. It requires the four tables as described above (crop, livestock, livestock production, and household information) and the two sets of parameters (energy and TLU conversion).

It automatically calculates 34 variables and adds them into the household table: 11 about crop production, 11 about livestock production, and 12 about farm production (Table 8).

Table 8: Farm characteristics calculated automatically from crop, livestock and livestock production information

Name	Definition	Unit	Formula
Crop production			
land_cultivated_ha	land area cultivated	ha	sum(crop$land_area_ha)
crop_div	number of different crops		length(unique(crop$name))
crop_harvest_kg	total harvest	kg/year	sum(crop$harvest_kg)
crop_yield_kg_per_ha	average crop yield	kg/ha/year	crop_harvest_kg/land_cultivated_ha
crop_sold_kg	total production sold	kg/year	sum(crop$sold_kg)
crop_sold_perc	percentage of crop production sold	%	crop_sold_kg/crop_harvest_kg *100
crop_income_div	number of crops sold		length(unique(crop$name)) with sold_kg>0
crop_income_lcu	total income from crop production	lcu/year	sum(crop$income_lcu)
crop_value_lcu	value of the crop consumed	lcu/year	sum(crop$consumed_kg*price)
crop_consumed_kcal	energy content of the crop consumed	kcal/year	sum(crop$consumed_kg*conv_energy)
crop_yield_lcu_per_ha	crop yield in monetary term	lcu/ha/year	(crop_income_lcu+ crop_value_lcu)/ land_cultivated_ha

Livestock production			
livestock_tlu	livestock herd size	tlu	sum(lstk$n*conv_tlu)
lstk_div	number of different livestock species		length(unique(lstk$name)) with lstk$n>0
lstk_harvest_kg	total livestock harvest	kg/year	sum(lstk$harvest_kg)
lstk_yield_kg_per_tlu	average livestock yield	kg/tlu/year	lstk_harvest_kg/livestock_tlu
lstk_sold_kg	total livestock production sold	kg/year	sum(lstk$sold_kg)
lstk_sold_perc	percentage of livestock production sold	%	lstk_sold_kg/lstk_harvest_kg*100
lstk_income_div	number of livestock products sold		length(unique(lstk$name)) with sold_kg>0
lstk_income_lcu	total income from livestock production	lcu/year	sum(lstk$income_lcu)
lstk_value_lcu	value of the livestock consumed	lcu/year	sum(lstk$consumed_kg*price)
lstk_consumed_kcal	energy content of the livestock consumed	kcal/year	sum(lstk$consumed_kg*conv_energy)
lstk_yield_lcu_per_tlu	livestock yield in monetary term	lcu/tlu/year	(lstk_income_lcu+ lstk_value_lcu)/ livestock_tlu
Farm characteristics			
farm_div	number of different crop and livestock species		crop_div+lstk_div
farm_harvest_kg	total farm harvest	kg/year	crop_harvest_kg+lstk_harvest_kg
farm_sold_perc	percentage of farm production sold	%	(crop_sold_kg+ lstk_sold_kg)/farm_harvest_kg *100
farm_income_div	number of crop and livestock products sold		crop_income_div+lstk_income_div
farm_income_lcu	total income from farm production	lcu/year	crop_income_lcu+lstk_income_lcu
farm_consumed_kcal	energy content of the farm production consumed	kcal/year	crop_consumed_kg+ lstk_consumed_kg
income_lcu	total annual income	lcu/year	farm_income_lcu+ off_farm_lcu
income_usd	annual income in USD equivalent	usd/year	income_lcu/currency_conversion
income_usd_pppd	income per person per day	usd/person/day	income_lcu/hh_size_members/365
off_farm_perc	percentage of income from off-farm sources	%	off_farm_lcu/income_lcu *100
pop_pressure_mae_per_ha	population pressure per crop area	MAE/ha	hh_size_mae/land_cultivated_ha
lstk_pressure_tlu_per_ha	livestock pressure per crop area	TLU/ha	livestock_tlu/land_cultivated_ha

To estimate the value of the production not sold, we needed to estimate the price of all products (crop or livestock production). When available, we used the median price of the given product; else we used the first quarter of all prices (crop and livestock product separated).

The function `update_farmhousehold()` calls the calculations in `calc_farm_prod()` and updates the farm household characteristics. This feature comes in handy when making scenarios: only the table of crop or livestock production can be modified, and the consequences for every household are calculated rapidly.

Graphics and interactive plots

In the file `plot.R` we defined multiple functions to plot the household data (Table 9). We used the base R plotting functions for the non-interactive plots, and the package `plotly` for the interactive plots (Sievert 2020). Every function can be plotted interactively, or not (parameter `interplot`). Most functions can also compare the distribution among groups with the parameter `seg`.

One mapping function was developed using the `tmap` package (Tennekes, 2018) and defined in the file `map.R`.

These graphical functions are directly used in the dashboards (as interactive plots), but can be used directly by users when needed. Please see the documentation for the functions for more information.

Table 9: Graphical functions developed for the interactive visualization of household data

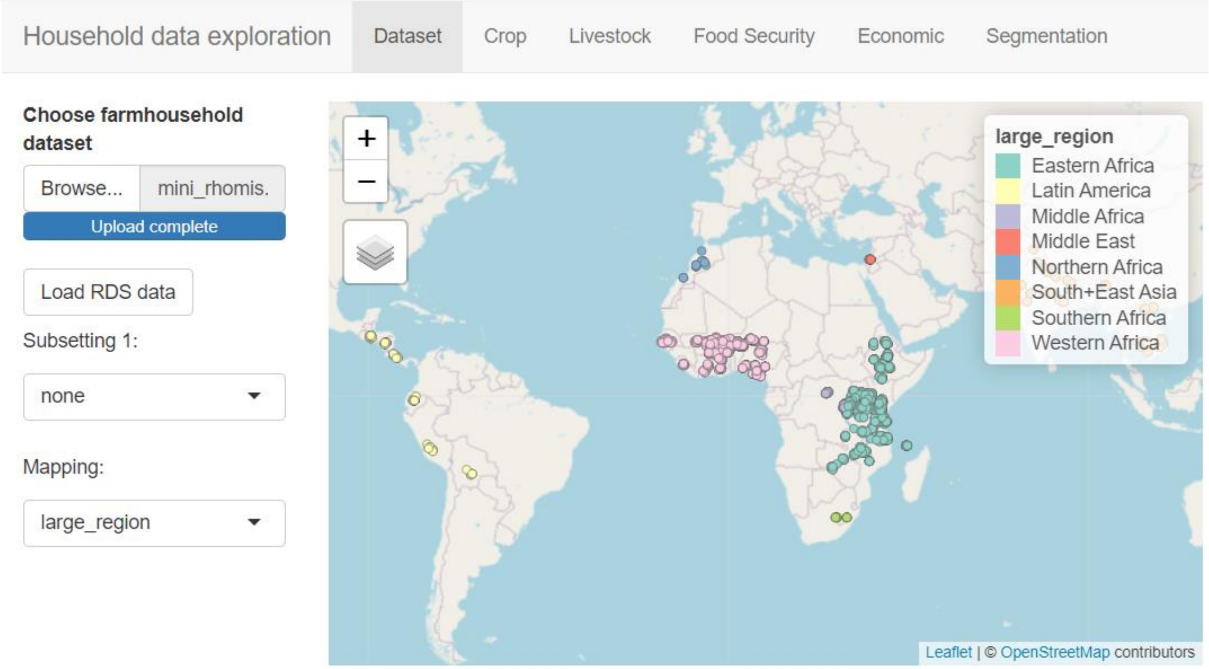
Name	**Short description**
`bar_div()`	Create a barplot showing the most popular crop or livestock in the dataset
`bar_hist()`	Distribution of quantities such as land cultivated or family size.
`bar_box()`	Distribution of count values such as the months with food insecurity
`bar_score()`	Distribution of the scores built on multiple categories (e.g. HDDS, FIES)
`bar_months()`	Barplot with the months with food insecurity
`bar_income()`	Show the value of production per sources, ordered per household income
`plot_density()`	Plot the distribution of a variable and the threshold separating the population in two
`tmap_ind()`	Make an interactive map showing the household location and the value of one selected variable

Household data exploration

The sources of the dashboard for data exploration are in the Github folder inst/app_explo. It can be run directly in R using the function `runExplo()`. A version of the dashboard is available online: https://startistic.shinyapps.io/farmhousehold_explo/

The household data exploration dashboard includes six panels that explore different aspects of farm households.

- **Dataset**: load and map your household data and select the relevant households (Figure 3)
- **Crop**: Show the most important crops and the land distribution for the whole population or per group (Figure 4)
- **Livestock**: Similar to the crop panel, show the most important livestock species and the livestock herd size distribution for the whole population or per group
- **Food security**: Show the HDDS, FIES or number of months with food insecurity for the whole population or per group (Figure 5)
- **Economic**: Show the value of production and its distribution per household and per source (Figure 6)
- **Segmentation**: Define the different groups within the population of interest. You can then navigate back to the previous 5 panels and see how crop, livestock, food security, or economic activities differ among groups.

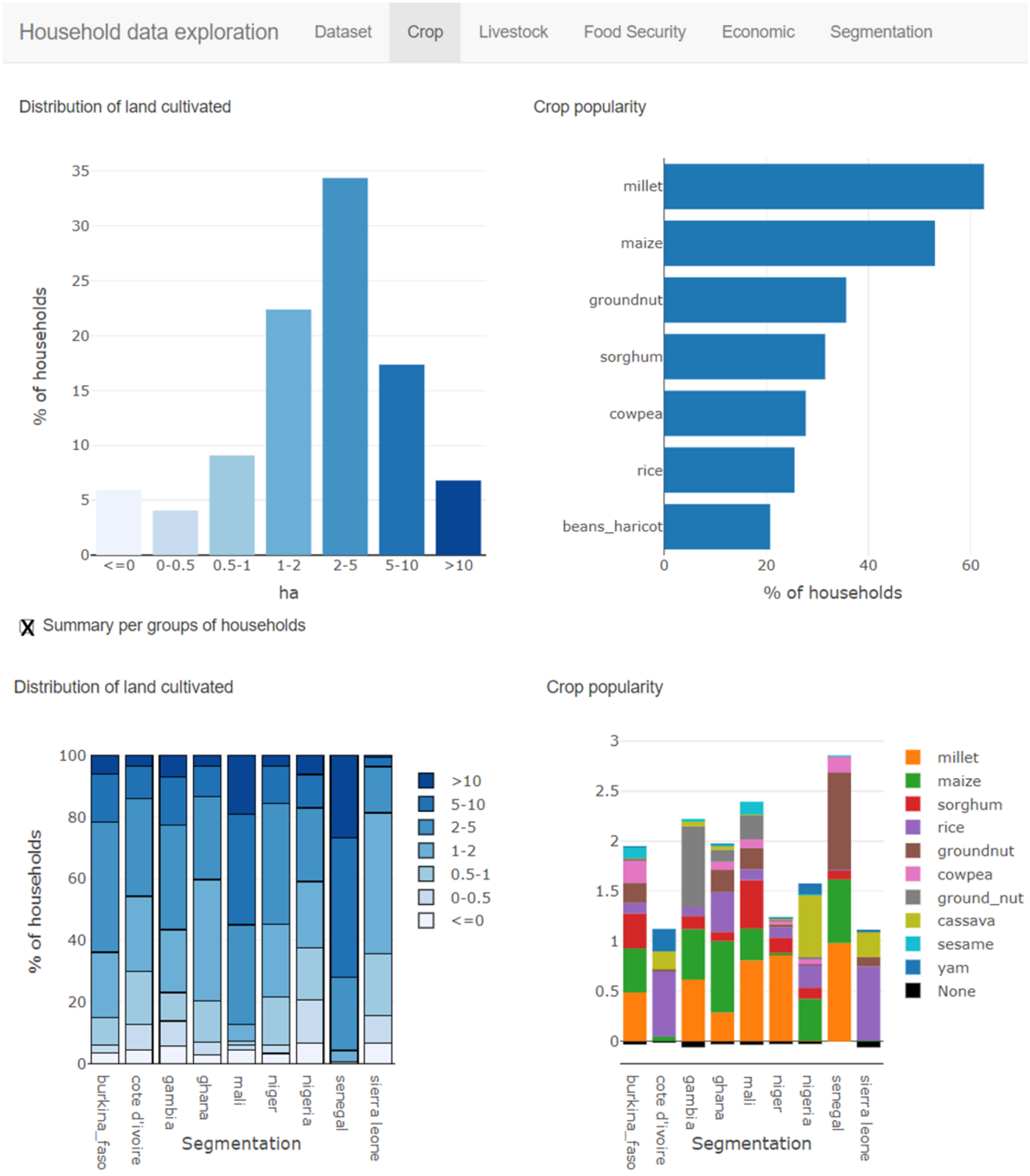


Choose farmhousehold dataset: mini_rhomis.

Subsetting 1: none

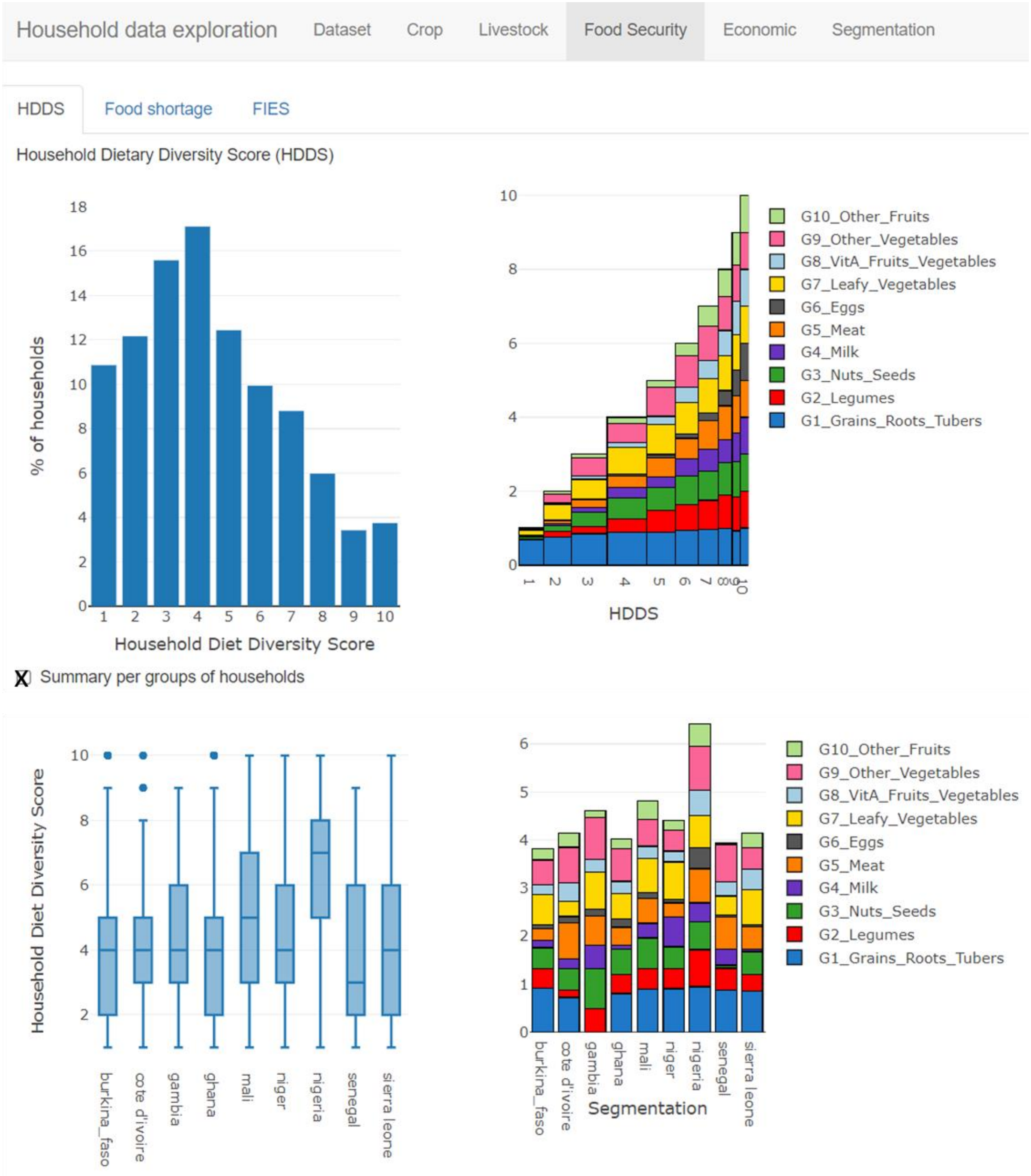
Mapping: large_region

Figure 3: The starting panel of the data exploratory dashboard for loading and subsetting the dataset



☒ Summary per groups of households

Figure 4: The crop panel of the data exploratory dashboard showing the land distribution and the crop preferences.



☒ Summary per groups of households

Figure 5: The food security panel of the data exploratory dashboard showing the diet diversity score

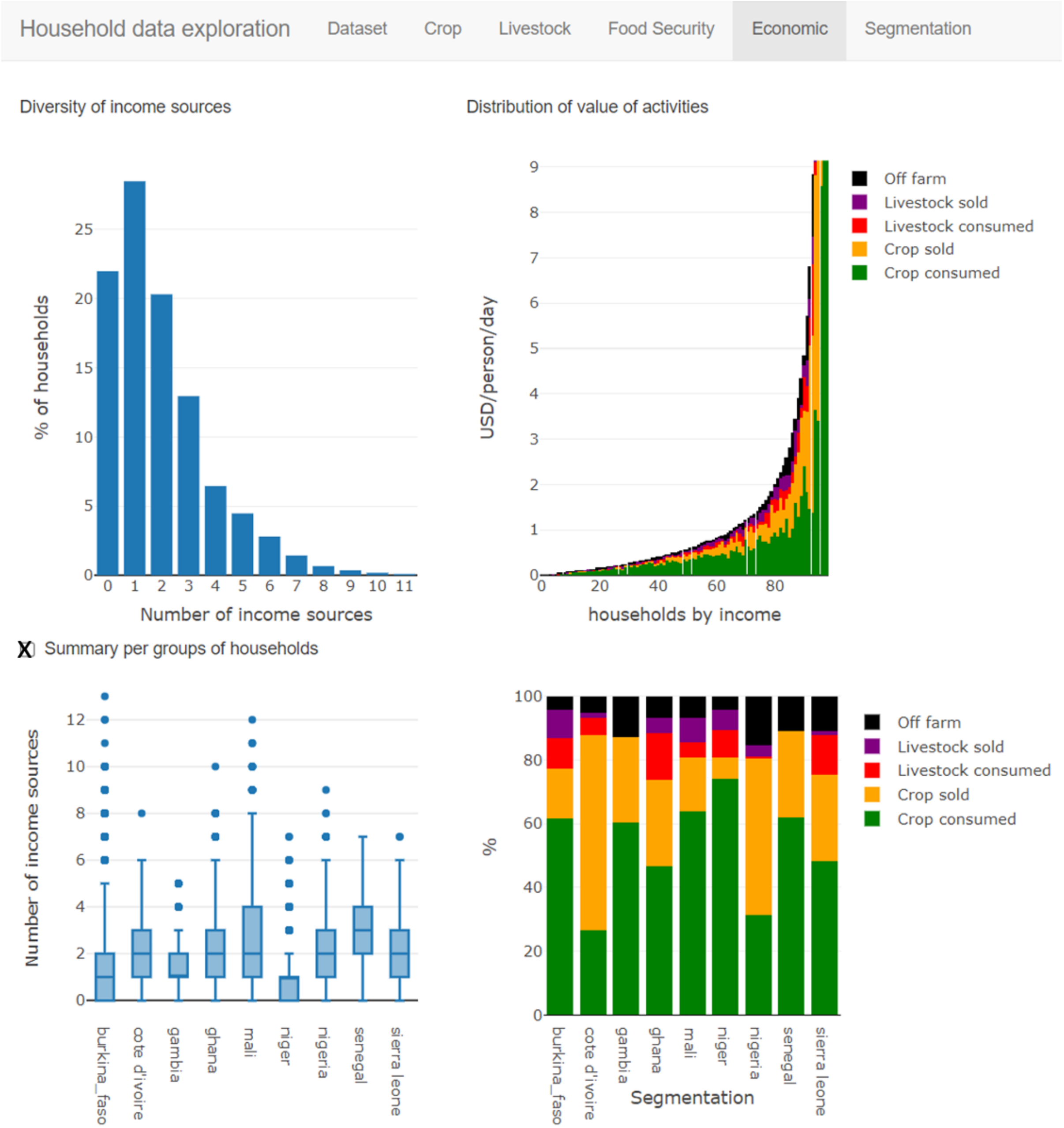


Figure 6: The economic panel of the data exploratory dashboard showing the sources of income and the distribution of the value of activities.

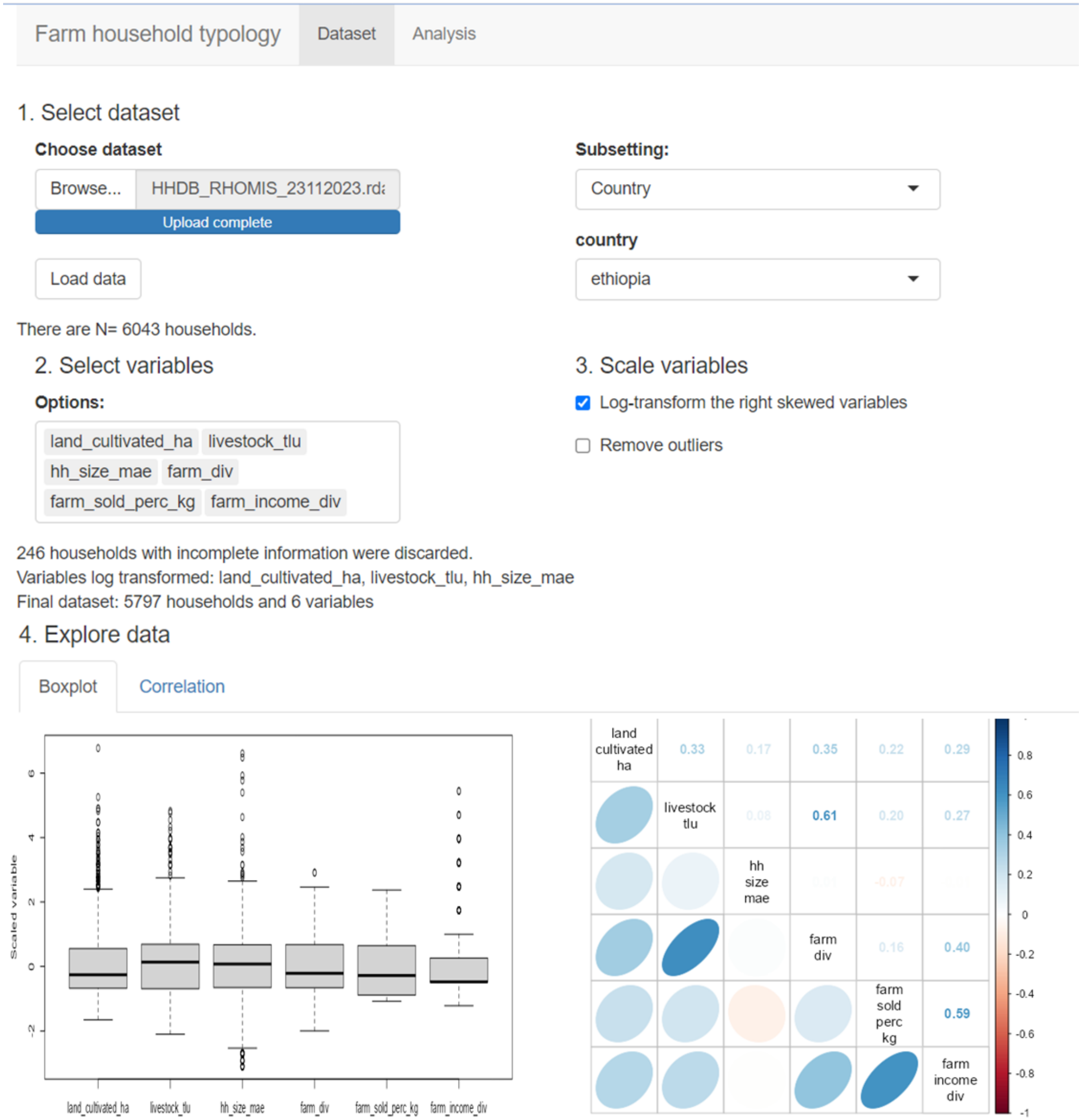
Household typologies

The sources of the dashboard for household typology are in the Github folder inst/app_cluster. It can be run directly in R using the function `runCluster()`. A version of the dashboard with the RHoMIS dataset is available online: https://startistic.shinyapps.io/farmhousehold_cluster/

The household typology dashboard has two panels that are sequential and need to be considered one after the other.

1. **Dataset**: load your own dataset (as a csv or rds file) and select, if needed, the relevant households. Then select the variables that will be used to classify the households. Third, decide how to treat the outliers, by log-transforming the right-skewed variables and/or removing the outliers. Finally, you can explore the dataset to see the distribution of the variables (make sure the distribution is *almost* Gaussian) and identify possible correlations among variables (to be avoided) (Figure 7)
2. **Analysis**: Run the multivariate analysis and define the number of principal components to be considered. Then define the clustering algorithm (hierarchical clustering or K-means) and the number of clusters. Finally, you can compare the distribution of the variables per group (Figure 8)

The main advantage of this dashboard is to explore the consequences of the choices made in variable selection, and in statistical considerations. Ideally, the cluster should make sense for the end-user and should be robust to the statistical method (e.g., K-means and hierarchical clustering should create almost similar groups).



1. Select dataset

Choose dataset

Browse... HHDB_RHOMIS_23112023.rd

Upload complete

Load data

Subsetting: Country

country ethiopia

There are N= 6043 households.

2. Select variables

Options:

land_cultivated_ha livestock_tlu hh_size_mae farm_div farm_sold_perc_kg farm_income_div

3. Scale variables

☑ Log-transform the right skewed variables

☐ Remove outliers

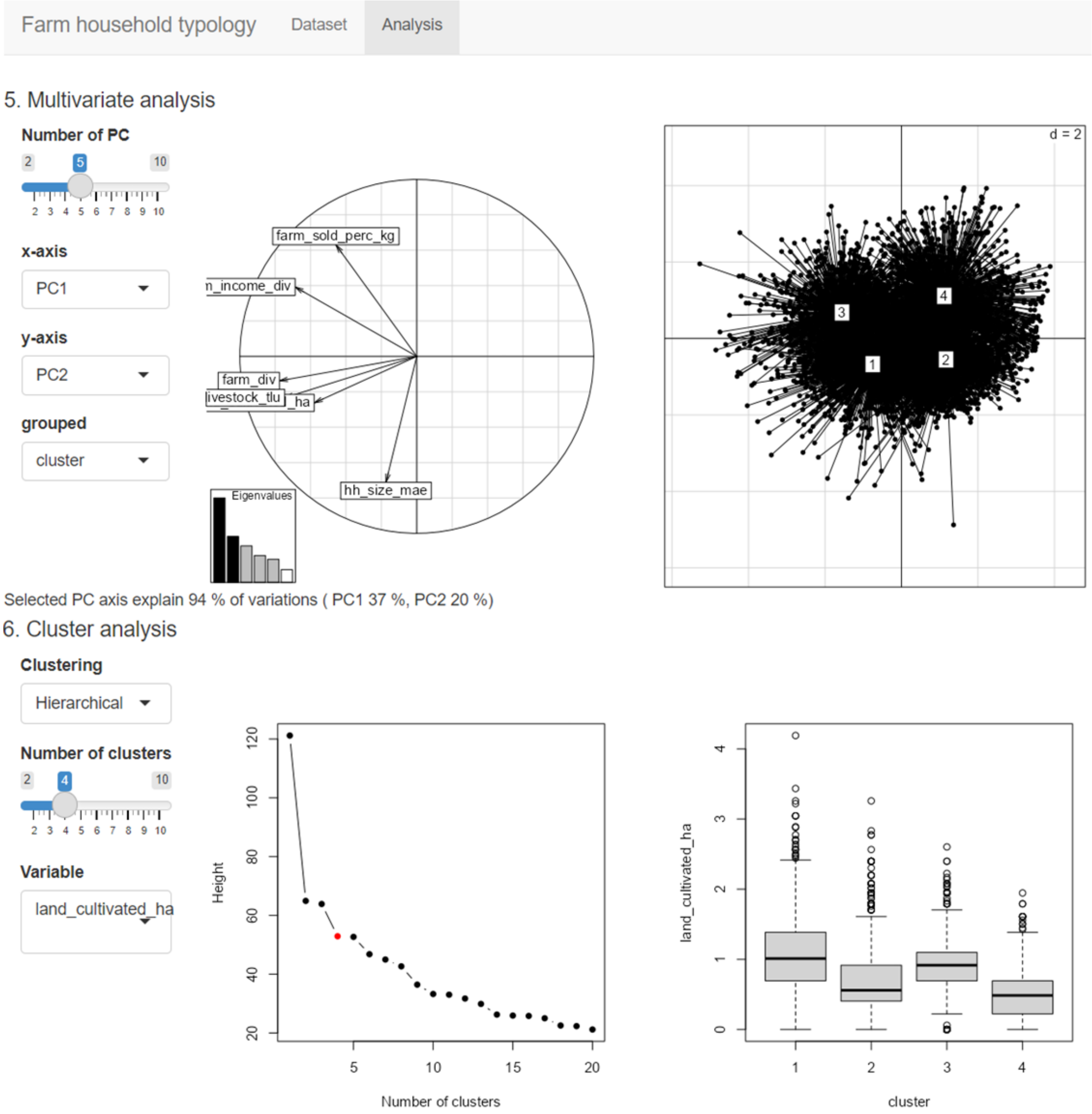
246 households with incomplete information were discarded.

Variables log transformed: land_cultivated_ha, livestock_tlu, hh_size_mae

Final dataset: 5797 households and 6 variables

4. Explore data

Figure 7: The dataset panel of the household typology dashboard.



5. Multivariate analysis

Number of PC: 5 (2–10)

x-axis: PC1

y-axis: PC2

grouped: cluster

Selected PC axis explain 94 % of variations (PC1 37 %, PC2 20 %)

6. Cluster analysis

Clustering: Hierarchical

Number of clusters: 4 (2–10)

Variable: land_cultivated_ha

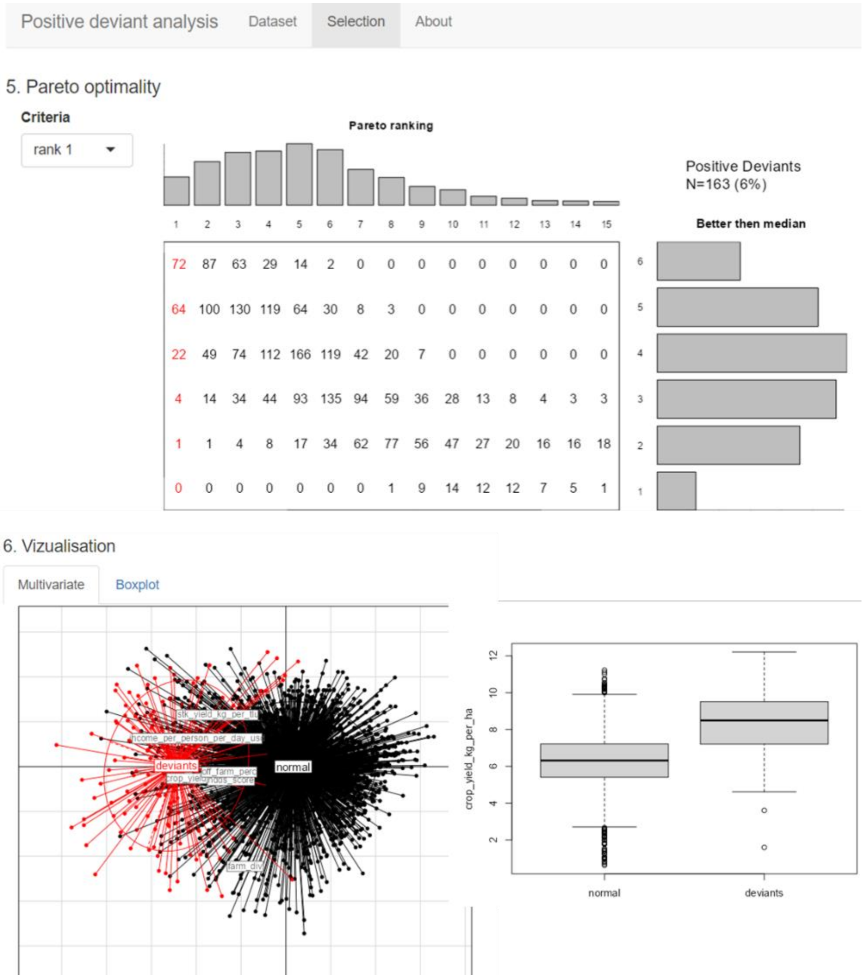
Figure 8: The analysis panel of the household typology dashboard.

Positive deviance analysis

The sources of the dashboard for positive deviance analysis are in the Github folder inst/app_posdev. It can be run directly in R using the function `runPosdev()`. A version of the dashboard with the RHoMIS dataset is available online: https://startistic.shinyapps.io/farmhousehold_posdev/

The positive deviance analysis dashboard has two panels that are sequential and need to be considered one after the other.

1. **Dataset**: load your own dataset (as a csv or rds file) and select, if needed, the relevant households. Then select the variables that will be maximized or minimized in the Pareto optimality algorithm. Third, decide how to treat the outliers, by log-transforming the right-skewed variables and/or removing the outliers. Finally, you can explore the dataset to see the distribution of the variables (make sure the distribution is *almost* Gaussian) and identify possible correlations among variables (to be avoided as much as possible)
2. **Analysis**: Run the Pareto optimality algorithm and define how to select the positive deviants. Finally, you can compare the distribution of the variables between the positive deviants and the rest of the population (Figure 9).



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
6	72	87	63	29	14	2	0	0	0	0	0	0	0	0	0
5	64	100	130	119	64	30	8	3	0	0	0	0	0	0	0
4	22	49	74	112	166	119	42	20	7	0	0	0	0	0	0
3	4	14	34	44	93	135	94	59	36	28	13	8	4	3	3
2	1	1	4	8	17	34	62	77	56	47	27	20	16	16	18
1	0	0	0	0	0	0	0	1	9	14	12	12	7	5	1

Figure 9: The analysis panel of the positive deviance analysis dashboard.

Way forward

Below is a list of on-going or potential adds-on that could be built on the farm household data environment.

- Add variables about the livestock breed, feed and management in the livestock table and create graphical functions to show them in the data exploration dashboard.
- Add variables that capture crop residue management, the fertilizer, manure and irrigation use in the crop table and create graphical functions to show them in the data exploration dashboard.
- Create a dashboard for ex-ante farm scenarios that is flexible to create various scenarios
- Calculation of greenhouse gas emissions per household (using by default parameters per region and/or more detailed information on livestock management and crop residue management)

References

Beck, H.E., N.E. Zimmermann, T.R. McVicar, N. Vergopolan, A. Berg, E.F. Wood (2018). Present and future Köppen-Geiger climate classification maps at 1-km resolution, *Scientific Data* 5:180214, https://doi.org/10.1038/sdata.2018.214.

Center for International Earth Science Information Network - CIESIN - Columbia University. 2018. Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11. Palisades, New York: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H49C6VHW.

Dixon, J., Garrity, D.P., Boffa, J.-M., Williams, T.O., Amede, T., Auricht, C., Lott, R., & Mburathi, G. (Eds.). (2019). Farming Systems and Food Security in Africa: Priorities for Science and Policy Under Global Change (1st ed.). Routledge. https://doi.org/10.4324/9781315658841

Hijmans R (2024). terra: Spatial Data Analysis. R package version 1.7-78, https://CRAN.R-project.org/package=terra

Nelson, A. A suite of global accessibility indicators for sustainable rural development. (2019) A report prepared for the CGIAR Consortium for Spatial Information, https://geodata.ucdavis.edu/geodata/travel/

R Core Team (1999) Writing R extensions. R Foundation for Statistical Computing, p. 1-208. https://cran.r-project.org/doc/manuals/R-exts.html

Sievert C (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC. ISBN 9781138331457, https://plotly-r.com.

Tennekes M (2018). “tmap: Thematic Maps in R.” *Journal of Statistical Software*, 84(6), 1-39. https://doi.org/10.18637/jss.v084.i06

Wickham, H., & Bryan, J. (2023). R packages. " O'Reilly Media, Inc. https://r-pkgs.org/

Appendix A: How to load my data into the farmhousehold package?

This is the most time consuming and difficult step, it should be done carefully. As with any data analysis, data preparation can represent up to 80% of the work. With the proposed database structure, we simplify the analysis steps, as well as optimize the storage, sharing and re-use of the collected household data.

There are four required preliminary steps:

1. Check for data entry errors and outliers
2. Transform all quantities into kg
3. Transform the land areas into ha
4. Check the spellings of the crop and livestock names

If you have used the RHoMIS survey tool, the rhomis package (https://github.com/RHoMIS/rhomis-R-package) can help you making these steps as easily as possible. If not, you will have to make these steps by yourself. In all cases, make sure to keep a copy of the original raw dataset (before cleaning) and document (preferably with a script and/or in a written document) all the changes you make (e.g. how you detect outliers, how many were they, and how you treat them).

Once the data is cleaned, and the quantities are in kg and ha, and names are verified, you can start renaming the variables with the standard names described above and fill in the corresponding tables. In the Github folder, we provide three examples on how to transform household survey data into the `farmhousehold` format. These examples can be found in the folder inst/load_data. The first example uses the rhomis dataset, an example of raw data as a single file (different crops are in different columns). The second and third examples are the LSMS-ISA dataset for Malawi and the SIMLESA dataset for Malawi. The two are examples of datasets split in multiple data files that need to be merged. With these three examples. and the parameters used (in the folder inst/load_data/Parameters), you will find answers to all your questions.

For extracting the GIS information, we use the terra package (Hijman 2024). First, a vector of GPS coordinates is created with the function `vect()`, and then the values for each household is extracted with the function `extract()`.

Please contact us if you need further help.

Appendix B: Abbreviations

FIES: Food insecurity experience scale
ha: hectare
HDDS: household diet diversity score
kg: kilogram
LCU: Local currency unit
LSMS-ISA: Living Standards Measurement Study - Integrated Surveys on Agriculture (https://www.worldbank.org/en/programs/lsms/initiatives/lsms-ISA)
lstk: livestock
MAE: Male Adult Equivalent
RHoMIS: Rapid Household Multi-Indicator Survey (https://www.rhomis.org/)
SIMLESA: Sustainable Intensification of Maize and Legumes in East and Southern Africa (https://simlesa.cimmyt.org/)
TLU: Tropical Livestock Unit