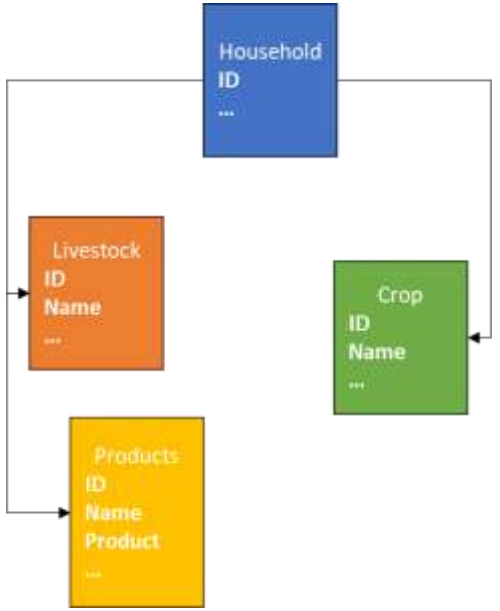
Data preparation for the harmonized database

Farmhousehold Workshop

2024

Romain Frelat



https://github.com/rfrelat/farmhousehold

Warning

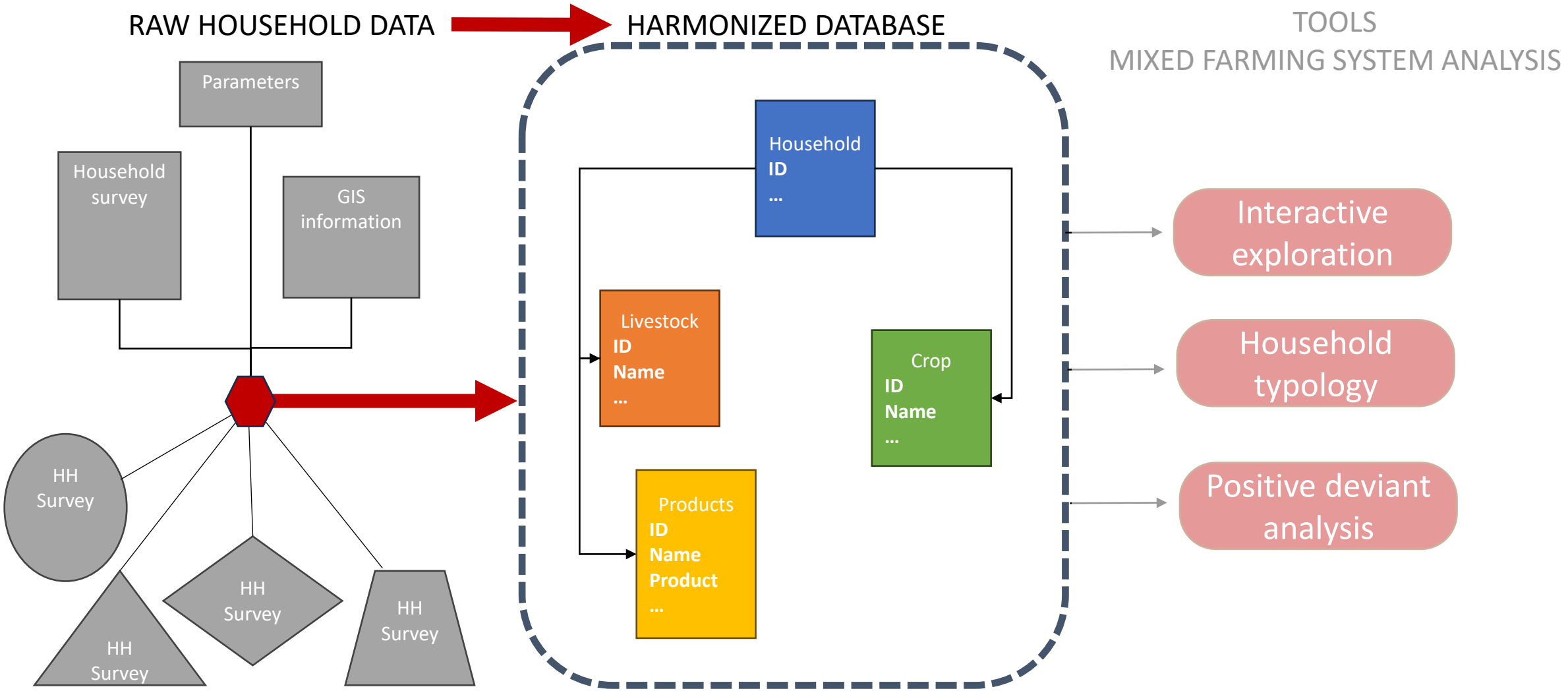
This is an intensive R scripting step, for advanced R users only.

Ideally it is followed by a practical afterward

To fully and properly transform a dataset into the harmonized format, it takes between 1 and 7 days (if data is previously cleaned).

No script/programing in the slides, but the essential steps are presented.

Objectif


RAW HOUSEHOLD DATA
HARMONIZED DATABASE
TOOLS
MIXED FARMING SYSTEM ANALYSIS
Parameters
Household survey
GIS information
HH Survey
HH Survey
HH Survey
HH Survey
Household
ID
...
Livestock
ID
Name
...
Crop
ID
Name
...
Products
ID
Name
Product
...
Interactive exploration
Household typology
Positive deviant analysis

Structure of the tutorial

1. Setting up
2. Crop information
3. Livestock information
4. Household information
5. Add spatial information (if any)
6. Create a proper farmhousehold object

Setting up

Preliminary checks

Before starting the data processing, make sure the household survey collected (reliable) information on:

- **Crop planted**: including the area per crop, the quantities harvested and the usage (whether sold, consumed or other) *(at least for the most important crops)*
- **Livestock herd**: the number of livestock kept by the household
- **Livestock production**: information on milk/eggs/meat annual production (quantities harvested and usage)
- **Household composition** (with age and gender)
- (Optional) Off farm occupation and income
- (Optional) Food security indicators (FIES/HDDS/HFIAS/number of month food insecure)
- (Optional) GPS coordinates

If any of these information is missing, you will not be able to fit your data into the farmhousehold format.

Install/Update R packages

Make sure you are using the latest package update(and R version).

If using RHoMIS survey tool, update the rhomis package:

```
> devtools::install_github("l-gorman/rhomis-R-package")
```

Install or update the farmhousehold package:

```
> devtools::install_github("rfrelat/farmhousehold")
```

Files organization

- Main folder
  - Data folder
    - Data files in csv (preferred format) or spss/sas/excel format
  - Script folder
    - Get_XX.R: the main script to transform the dataset XX into farmhousehold format
    - Param_XX.R: set the parameters for the dataset XX
    - Info_XX.Rmd: the markdown file to document the choices made when transforming the dataset into the farmhousehold format

Your turn: Get and drop the 3 R files in the script folder (NewDataset)

You can check out examples of these files for the data transformation in the farmhousehold package (Load_Data)

Data transformation steps

Seven independent steps:

1. *Set an household ID*
2. Crop information
3. Livestock herd information
4. Livestock production information
5. Household information
6. *Add GIS data (if any)*
7. *Merge all in a proper farmhousehold object*

Set an Household ID

- Use the one that exists if any
- Make sure they are unique to every household:
  ```
  > length(unique(ID)) == nrow(data)
  ```

- For ease of merging, it contains a prefix with the dataset name, country, and year; e.g. LSMS_MWI2019_245

Crop information

Crop table

Name	Definition	Unit	Example
hh_id	household id		*ke_2018_53*
name	name of the crop	in English	*maize*
land_area_ha	land area	ha	*0.5*
harvest_kg	quantity harvested	kg	*420*
consumed_kg	quantity consumed	kg	*250*
sold_kg	quantity sold	kg	*170*
income_lcu	income from sold crop	lcu	*2500*

lcu= local currency unit

Crop name and parameters

1. Make sure the names are in English without spelling issues
   `> table(data$name)`
   Especially be careful with crops that are listed only once, or listed with different names (e.g. maize and corn)
2. Get energy conversion factors for all the listed crops
   As a starter, you can use the crops parameter that are already in Github
   [inst/load_data/Parameters/crop_param.csv]

3. Load the crop energy conversion in the file Param_XX.R

Crop quantities

1. Get the land size per crop in ha

   Usually it needs a set of parameter to transform the local unit in ha (in Param_XX.R)
   Be sure to consider all cropping seasons
   Intercropping are divided into separate crops, with estimate of land use percentage

2. Get the quantity harvested per crop
3. Get the information on crop usage (sells vs consumption)

   Make sure the total quantity harvested is not lower than the quantity sold+consumed

4. *Add any other relevant information on crop (possibly crop residue use, irrigation, fertilizer use, intercropping)*

Livestock herd information

Livestock table

Name	Definition	Unit	Example
hh_id	household id		*ke_2018_53*
name	name of the livestock	in English	*cattle*
n	number of livestock kept	ha	*8*

Livestock herd information

1. Make sure the names are in English without spelling issues
   `> table(data$name)`
   you might want to group some of the categories (e.g. donkey and horses)
2. Get TLU parameters for all the listed livestock categories
   As a starter, you can use the livestock parameter in Github
   inst/load_data/Parameters/lstk_param.csv
3. Load the livestock TLU parameters in the file Param_XX.R
4. Get the number of livestock per species
5. *Add any other relevant information on livestock management (breed, where is the livestock kept, manure management, feeds)*

Livestock production information

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	lcu	3500

* Livestock products are everything that is consumed or sold from livestock herd: milk, eggs, honey, meat, manure, power force, wool, etc…

lcu= local currency unit

Livestock production information

1. Get the quantity harvested per product and per year

   Be careful that milk or egg production are often reported per day and per animal, so it needs to be converted as total per household and per year

   For whole animal sales, use 1TLU=250kg and convert the quantities in kg.

2. Get the information on production usage (sells vs consumption)

   Make sure the total quantity harvested is not lower than the quantity sold+consumed

3. Add energy conversion factors for all livestock products in Param_XX.R (0kcal/kg if not human food)
4. *Add any other relevant information on livestock production*

Suggested break

Household information

Household table

Name	Definition	Unit	Example
hhid	household id		ke_2018_53
country	country of the survey		*kenya*
year	year of the survey		2018
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
currency_conversion	conversion from local currenty to power parity purchase usd	lcu/usd	*41.9*
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*

Household information

1. Get the household size in male adult equivalent

   Male adult equivalent coefficient should be loaded in Param_XX.R

Age	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

2. Get the number of off-farm activities and estimate of total off-farm income per year (in local currency unit)
3. Get GPS coordinates *(if available)*
4. *Add any other relevant information on household (e.g. gender and age of household head, education level)*

Food security

1. For scores made of multiple categories (HDDS, FIES, HFIAS), keep all the columns and the agglomerated score.

   For instance, HDDS_Fruits, HDDS_Meat, … and hdds_score

2. If available, get the number of months when household felt food insecure in two columns

   the number of months in foodshortage_count

   the 3-letter abbreviate months, separated by space in foodshortage_months

Spatial information

Household table

Name	Definition	Unit	Example
hhid	household id		ke_2018_53
country	country of the household		kenya
large_region	large region		Eastern Africa
region	region of the household		nandi
gps_lat	latitude in decimal degrees	°N	-0.7
gps_lon	longitude in decimal degrees	°E	35.1
farming_system	size of the household in number of persons		5. Highland perennial
population_2020	size of the household in male adult equivalent	MAE	350
travel_time_cities	off farm income per year	lcu	37
koeppen	diversity of off farm activities		Af Tropical, rainforest

GIS information (optional)

Get the GIS information per household using terra package and gis data from geodata package

- farming_system: Dixon farming system map for Sub-Saharan Africa
- population_2020: population density map for 2020, at 30 arc-second resolutions
- travel_time_cities: travel time to the closest city in minutes
- koeppen: Koeppen-Geiger climate classification of present time at 1km resolution

```
library(terra)
p <- vect(cbind(hhinfo$gps_lon, hhinfo$gps_lat))
gis <-rast("gis.tif")
gisp <- terra::extract(gis, p)
```

* It is recommended to extract GIS information with the original GPS coordinates, not the rounded ones

farmhousehold object

Farmhousehold class

1. Create a new farmhousehold object

```
hhdb<- farmhousehold(crop, lstk, lstk_prod, hhinfo, conv_tlu, conv_energy)
```

2. Calculate the farm statistics

```
hhdb<- update_farmhousehold(hhdb)
```

3. Export the dataset as RDS file

```
saveRDS(hhdb, file="hhdb_xx.rds")
```

Automated crop summary

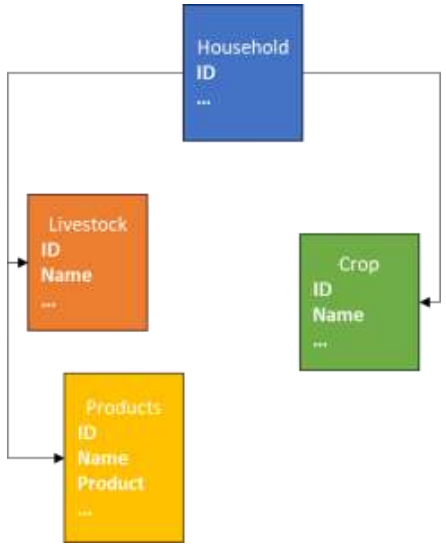
Name	Definition	Unit	Example
land_cultivated_ha	land area cultivated	ha	sum(crop$land_area_ha)
crop_div	number of different crops		length(unique(crop$name))
crop_harvest_kg	total harvest	kg/year	sum(crop$harvest_kg)
crop_yield_kg_per_ha	average crop yield	kg/ha/year	crop_harvest_kg/land_cultivated_ha
crop_sold_kg	total production sold	kg/year	sum(crop$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)

Automated livestock summary

Name	Definition	Unit	Example
livestock_tlu	livestock herd size	tlu	sum(lstk$n*conv_tlu)
lstk_div	number of different livestock		length(unique(lstk$name))
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)

Automated farm summary

Name	Definition	Unit	Example
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



End of the data processing
Data analysis can start ☺