Food and Agriculture Organization
of the United Nations

Shiny RIMA

Preparing your dataset for a successful Shiny RIMA analysis

RESILIENCE INDEX MEASUREMENT AND ANALYSIS II ◀ RIMA II

Co-funded by
the European Union

Shiny RIMA

Preparing your dataset for a successful Shiny RIMA analysis

Food and Agriculture Organization of the United Nations
Rome, 2023

Required citation:
FAO. 2023. *Shiny RIMA – Preparing your dataset for a successful Shiny RIMA analysis*. Rome. https://doi.org/10.4060/cc6901en

This publication has been produced with the assistance of the European Union. The contents of this publication are the sole responsibility of FAO and can in no way be taken to reflect the views of the European Union.

Contents

Figures

Tables

1 Introduction

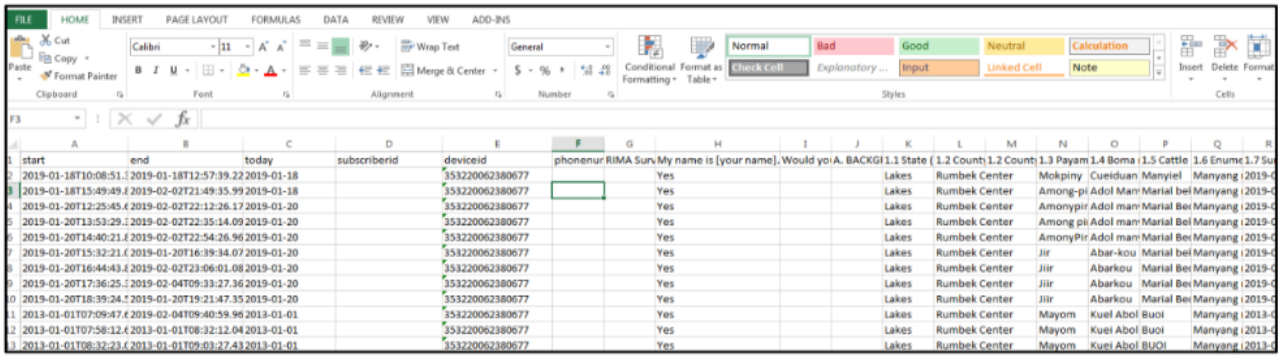
Shiny RIMA is a web-based tool designed to simplify the estimation of resilience capacity without compromising the scientific rigour of the Resilience Index Measurement and Analysis (RIMA). This technical manual will guide the user in cleaning and preparing their dataset for a successful analysis on the Shiny RIMA tool. In this manual, we provide examples based on Microsoft Excel but it is possible to use any other software allowing data cleaning (Stata, R, SPSS, etc.). This manual is designed and developed for monitoring and evaluation (M&E) officers, evaluators, government officials, researchers and other operators without an advanced background in econometrics and statistical tools (such as Stata). The simplicity of this document will be adequate to assist them in conducting resilience analyses. The information contained in this document refers solely to the variables employed in the Short RIMA questionnaire (FAO, 2020). However, statistical principles and standard procedures drafted in this note may apply to any other dataset containing the necessary variables for a resilience analysis with RIMA (FAO, 2022).

2 Preparing the dataset

2.1 Getting started – cleaning the data

Once the raw dataset is available, it can be imported in different statistical programs for the cleaning and analysis. In this manual, we will provide an example using Microsoft Excel for data cleaning and data management. To obtain a correct estimated resilience capacity index (RCI) with Shiny RIMA, the data needs to be cleaned from inaccurate records, inconsistency, misspelled words, improper cases, missing values and outliers.

Figure 1. Dataset export in Microsoft Excel format



Source: Screenshot of Microsoft Excel test sample.

Generally, data cleaning should follow the steps listed below.

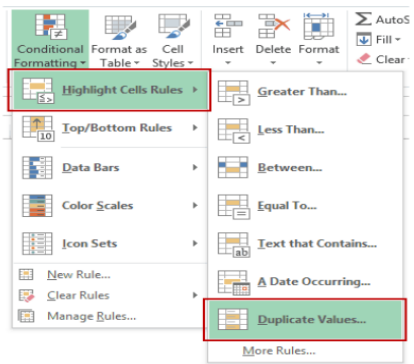
REMOVING DUPLICATES

Make sure there are no duplicates in the unique household identifier observations: one number for each household. It is also possible that one or more observation(s) have been misnumbered. It this case, the observation should not be deleted.

In Microsoft Excel, the following steps are required:

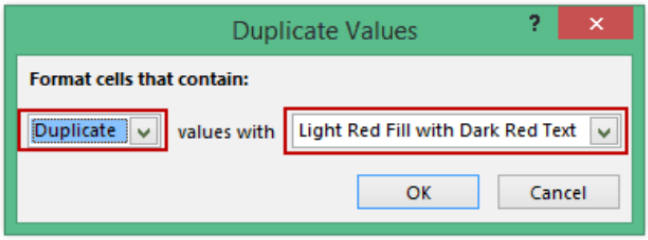
1. Select the data for which you want to find duplicates.
2. Go to Home in the Microsoft Excel ribbon → Conditional formatting → Highlight cell rules → Duplicate values (Figure 2).
3. In the "Duplicate values" dialog box, select "Duplicate" in the drop down on the left, and specify the format in which you want to highlight the duplicate values. You can choose from the readymade format options (in the drop down on the right) or specify your own format (Figure 3).

Figure 2. Checking duplicates in Microsoft Excel



Source: Screenshot of Microsoft Excel test sample.

Figure 3. Checking duplicates in Microsoft Excel



Source: Screenshot of Microsoft Excel test sample.

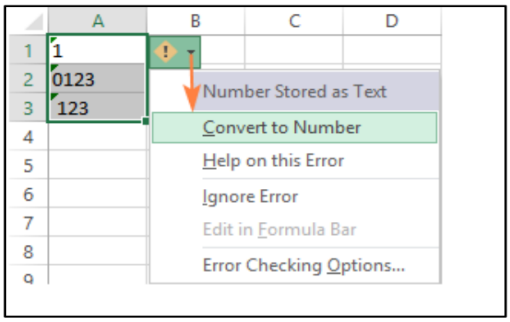
The values that have duplicates will be highlighted in red.

FIXING ERRORS

Numbers stored as text can cause problems. This is indicated by Microsoft Excel with a green triangle in the top left corner of the problematic cell. To convert text strings to numbers, two steps are required (see Figure 4):

1. Select all the cells containing numbers formatted as text.
2. Click the warning sign and select "Convert to number".

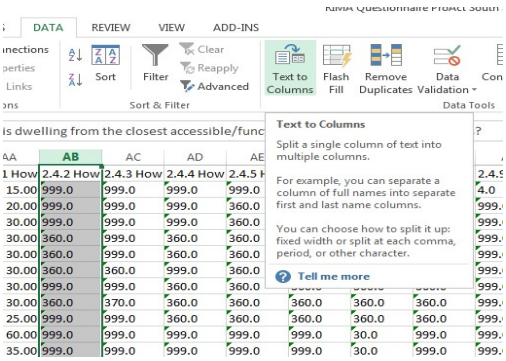
Figure 4. Microsoft Excel warning sign



Source: Screenshot of Microsoft Excel test sample.

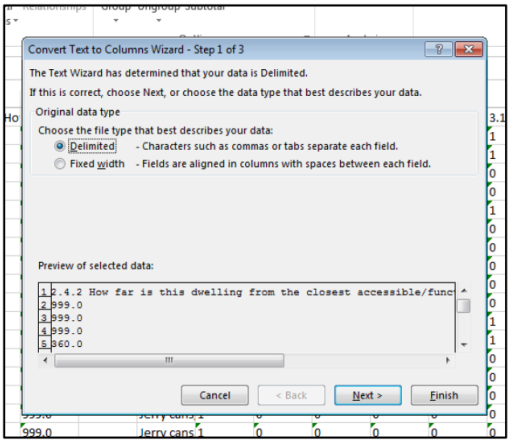
There is another way to convert text to number: using the "Text to columns" option (see Figure 5). The "Text to columns" button is used to split a column, but it can also be used to convert a single column of text to numbers. On the Data tab, click "Text to columns". Then, select "Delimited" and click "Finish" (see Figure 6).

Figure 5. Changing format "Text to columns"



Source: Screenshot of Microsoft Excel test sample.

Figure 6. Changing format "Convert text to columns"

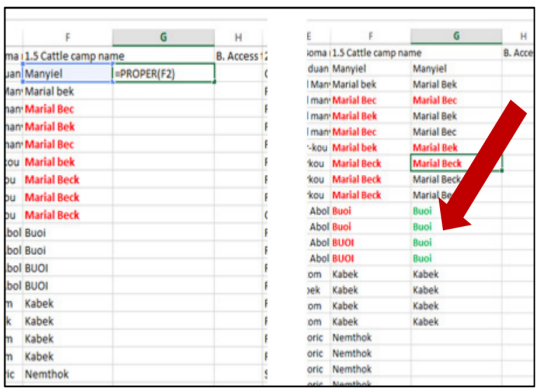


Source: Screenshot of Microsoft Excel test sample.

2.2 Checking for misspelled words and other text errors

Misspelled words, trailing spaces, unwanted prefixes, improper cases, and nonprinting characters should be detected and addressed. In Microsoft Excel, the spell checker finds misspelled words and values that are not used consistently, such as names of livestock, types of income generating activities (if reported as text), etc. To run the spellcheck, select column by column the variables to be checked, then click on the "Review tab" and select "Spelling" (see Figure 7). There are variables containing text that are specifically spelled. For instance, region names, province names, districts, etc. In this case, use the PROPER function, which reformats text capitalizing the first letter in each word, while, at the same time, it converts the rest of the word into lowercase.

Figure 7. Checking misspelled words

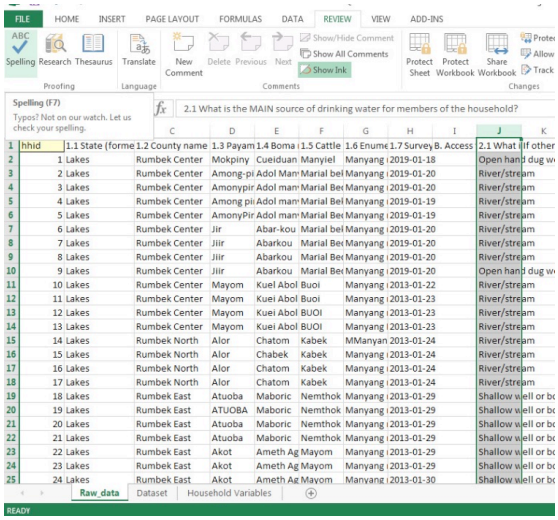


Source: Screenshot of Microsoft Excel test sample.

These are the steps to follow:

1. Add a new column close to the variable you need to work with.
2. In the new column write the function for each cell (=PROPER ("cells number")) (see Figure 8).
3. As for Figure 9, the reformatting did not correct the typos; indeed, these needs to be addressed one-by-one.
4. Once all the mistakes and typos are addressed, the original variable can be deleted, being replaced by the corrected one. Before deleting the original variable, select the new one, press Ctrl+C to copy it (Figure 10).
5. Then, click on the first cell of the new variable, open the "Paste special" dialog (Figure 11) and click on "Value", then ok.

Figure 8. Adjusting for upper case

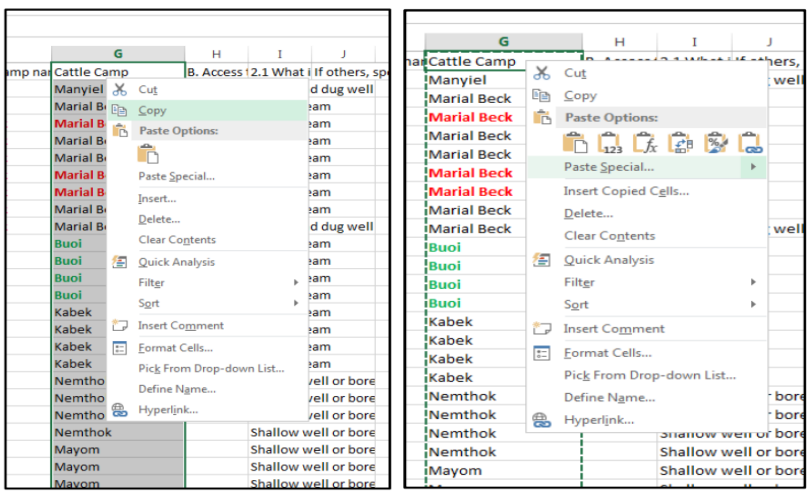


Source: Screenshot of Microsoft Excel test sample.

Figure 9. Addressing typos

Source: Screenshot of Microsoft Excel test sample.

Figure 10. Filling the household variable sheet



Source: Screenshot of Microsoft Excel test sample.

IMPUTING OUTLIERS

Checking outliers: those can cause problems with some models (for example, linear regression models are not robust to outliers). Before filtering them, confirm case-by-case that large numbers are really outliers (for example, a household owning 3 000 cows could be an outlier, while a household owning 100 chickens could not). Numbers that are stored as text can cause unexpected results. If cells display an error indicator (green triangle in the top left corner), number formatted, as text could be the problem. To convert text strings to numbers two steps are required:

1. Select all the cells containing numbers formatted as text.
2. Click the warning sign and select "Convert to Number". There is another way to convert text to number: using the "Text to Columns" option. The "Text to Columns" button is used to split a column, but it can also be used to convert a single column of text to numbers. On the Data tab, click "Text to Columns". Then, select "Delimited" and then click "Finish". For any variables showing the warning sign, the two procedures mentioned above need to be done variable by variable.

HANDLING MISSING VALUES

The identification and verification of missing values is an important step of data cleaning. Some values are supposed to reflect the skipping pattern in the questionnaire. These should have missing codes (or empty cells) and are very easy to find in the dataset. To indicate missing data/lack of response, 999 is the usual code; however, depending on the question, 999 could have different meanings. For example, the number of cattle 999 can be easily translated into 0; yet, in this particular case 999 could imply that the household is not engaged in livestock activity or that they are engaged but do not have cattle. For distance variables, the explanation is slightly different and we cannot just replace the variable with zero, as we will see in the next section. For all other variables, missing values are usually replaced with zero or with the mean of the variables, according to the type of variable (continuous vs categorical).

3 Behind the pillars: constructing the resilience variables

This section describes the steps to follow in order to construct the variables that make up the pillars of resilience.[1]

ACCESS TO BASIC SERVICES (ABS) PILLAR VARIABLES

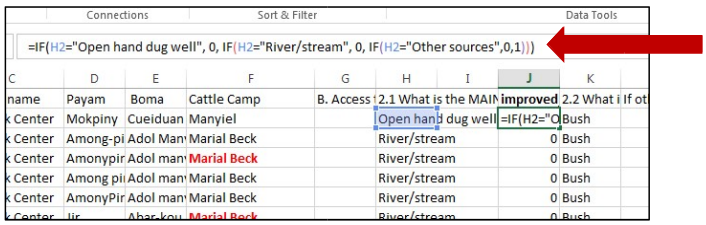
Before starting with the variable cleaning/creation, please keep in mind these recommendations:

- Always create a new column close to the variable you are cleaning and keep only the clean variable.
- Using colour highlights to distinguish between clean variable and variable that need to be cleaned is a good idea.

Questions 1 to 3 are straightforward. These are three different dummy variables, taking value 1 (respectively) if the household has access to an improved source of water, to improved toilet facilities, and to electricity, and 0 otherwise. You can rename those columns as abs_ water, abs_ sanitation, abs_electricity.
If questions regarding improved water and improved toilet facilities have been expanded to include different options, make sure to end up with dummy variables capturing all these options. Keeping in mind those categories, the dummy variable needs to be constructed as follows:

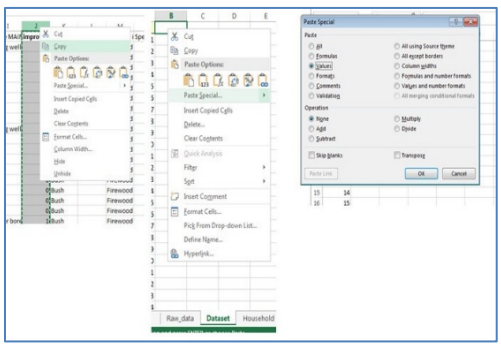
- Insert a new column close to the one containing the water source categories and call it abs_ water.
- Use the IF function to create the dummy as for Figure 11.
- Once the new variable is created, select it and then copy (Ctrl+c).
- Go to the "Dataset" sheet, with the "Paste Special" option, chose "Value" and then click "OK" (see Figure 12).

Figure 11. Paste special option



Source: Screenshot of Microsoft Excel test sample.

Figure 12. Creating improved water categories



Source: Screenshot of Microsoft Excel test sample.

The same steps are required for improved toilet facilities, always keeping in mind the WHO categories for improved sanitation facilities. Concerning the source of energy, one must highlight those households with electricity. In this case, when creating the dummy variable, indicate 1 if a household has electricity and 0 otherwise, then follow the steps in Figure 12.

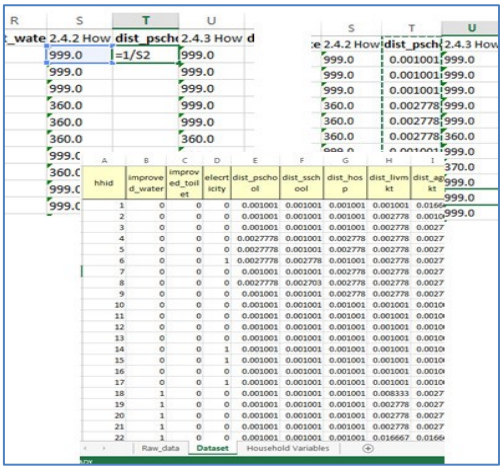
Question 4 refers to the distance to basic services. For distance here, the respondent was asked to quantify the walking distance in minutes, one way. The following rule applies when constructing pillars: all variables need to go in the same direction; for the numerical variables, this means that an increasing number is associated with higher household resilience. Since being far from a particular service contributes negatively to the household's resilience, we use the inverse of the distance variables (1/variable) to obtain variables indicating *closeness* to services. The procedure is the following:

- Next to each distance variable column, add a new column and call it *abs_closeness_servicename* (for example, for distance to school, *abs_closeness_school* is the corresponding variable)
- In each of the new columns, write "=1/" and select the cell or the variable to be inverted

[1] For more information on the Short RIMA questionnaire, see FAO (2020).

Apply the rule as for the copy and paste process (see Figure 13).

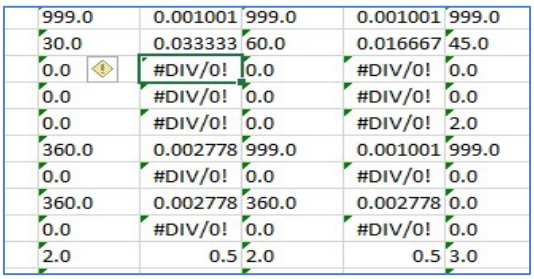
Figure 13. Inverting distance variables



Source: Screenshot of Microsoft Excel test sample.

With the inversion process, the values of 999 will be close to zero, indicating remoteness from the service. Note that Microsoft Excel shows the "#DIV/0!" error when the household is so close to the service that its value is 0 (a number cannot be divided by 0). In this case, replace by 1 for all cells containing "#DIV/0!" error (see Figure 14). This can be done either by using the *"iferror"* condition or by the *search and replace* function.

Figure 14. Dealing with error messages



999.0	0.001001	999.0	0.001001	999.0
30.0	0.033333	60.0	0.016667	45.0
0.0	#DIV/0!	0.0	#DIV/0!	0.0
0.0	#DIV/0!	0.0	#DIV/0!	0.0
0.0	#DIV/0!	0.0	#DIV/0!	2.0
360.0	0.002778	999.0	0.001001	999.0
0.0	#DIV/0!	0.0	#DIV/0!	0.0
360.0	0.002778	360.0	0.002778	0.0
0.0	#DIV/0!	0.0	#DIV/0!	0.0
2.0	0.5	2.0	0.5	3.0

Source: Screenshot of Microsoft Excel test sample.

After this process, the user may want to create an infrastructure index by summing the inverted distances, and then rescaling the aggregated inverted distance variables from 0–1.

ASSETS (AST) PILLAR VARIABLES

Question 5 contains all the information about the durable assets owned by the household. There is a row for each asset containing the number of those owned. Make sure to separate the agricultural tools from the other assets, and to separate the animals owned, because three different indices will be constructed: (i) the wealth index (durable assets), (ii) the agricultural wealth index (agricultural assets and tools), and (iii) the Tropical Livestock Unit (TLU),[2] or any other livestock unit index based on the context.

[2] Livestock Units are livestock numbers converted to a common unit. For more information about Household Dietary Diversity Score, see Swindale and Bilinsky (2006).

Figure 15. Assets data

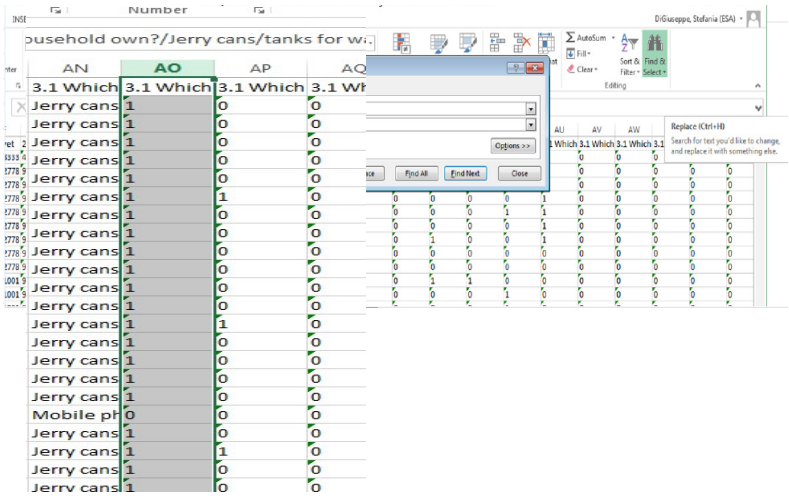
T	U	V	W	X
ast_ubt	ast_wealth_index	ast_agasset_index	ast_area	ast_seed
1.25	0.181818187	0	3	1
2.74	0.181818187	0	2	1
2.6	0.363636374	0.050000001	3	0

Source: Screenshot of Microsoft Excel test sample.

Replacing procedure for 999:

- Select the specific column, then go to the "Home" Microsoft Excel ribbon and select "Find&Select" options.
- In the "Find what" type 999 and in the "replace with" type 0, click on "Replace all".

Figure 16. Replacing procedure for 999



Source: Screenshot of Microsoft Excel test sample.

Please make sure to address the outliers/extreme values (for example reporting 34 rooms in a household is considered as an extreme value) with the steps below:

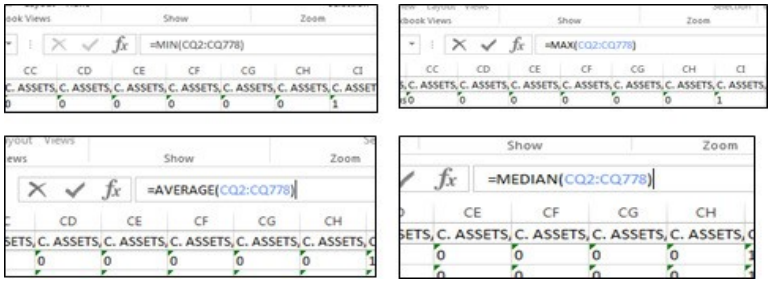
- Go at the end of the variable to be checked for outliers, on the first available cell write the function for the average "=AVERAGE" (select the column).
- After the average function, other function needs to be added and those are respectively: the median the, minimum value and the maximum value. Copy and paste special the value of the function in a separate sheet in order to analyse them (see Table 1).

Table 1. Table of statistics

Min	0.00
Max	67 200
Average	1 007
Median	3.00

Source: Author's elaboration.

Figure 17. Microsoft Excel functions for outliners checking

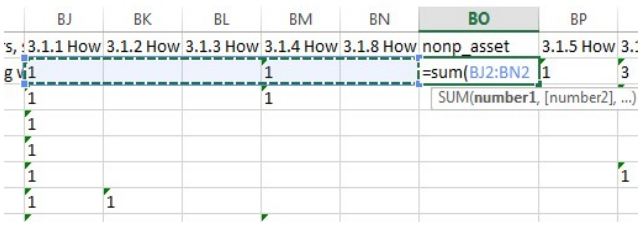


Source: Screenshot of Microsoft Excel test sample.

From this parameter, it is possible to identify outliers. Table 1 shows the statistics for a random variable. There is a huge discrepancy between the average and the median value with an extremely high maximum value, which suggests that the mean is heavily biased by extreme values. In this case, in order to address the extreme values:

- Select the variable, sort it in ascending order (on the "Data" Microsoft Excel ribbon click on "Sort A to Z")
- Scroll until the end of the variable column and check the extreme values; once detected, replace them with the median value already calculated.
- Go back to the unique identifier, select it and sort it in ascending order.
- Make sure to repeat the steps above for all the variables in the asset category (both productive and non-productive assets).
- Once data checking is done, next step consists in creating the indices. Two different options are available for the creation of an index:
- The summation of all the household assets (simple column summation of all the household assets). Create a new variable and call it ast_durable_assets, then write the summation function (see Figure 18).

Figure 18. Microsoft Excel column summation



Source: Screenshot of Microsoft Excel test sample.

Starting from the new column containing all the assets owned by the household, in order to have an index in a range of (0, 1), then the column (variable) needs to be rescaled according to the rules described below:

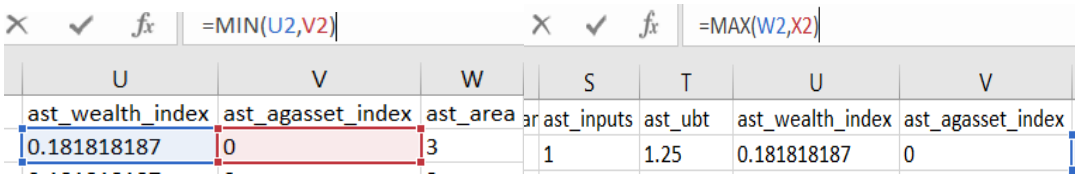
Rescaling a Variable. Dataset often contain variables that highly vary in magnitude, unit and range. To compare different variables or to make data readable, rescaling the variables is recommended. Different techniques can be used to rescale a variable; one of them is the Min-Max standardization. The formula is the following:

$$X_{rescaled} = \frac{X - min(X)}{max(X) - min(X)}$$

Where: *min (X)* is the minimum value of the variable X and *max (X)* is the maximum value of the variable X. Steps to follow in Microsoft Excel:

- Calculate the minimum value of the columns at the bottom of the variable of interest.
- Calculate the maximum value of the column at the bottom of the variable of interest right the minimum value (see Figure 19).

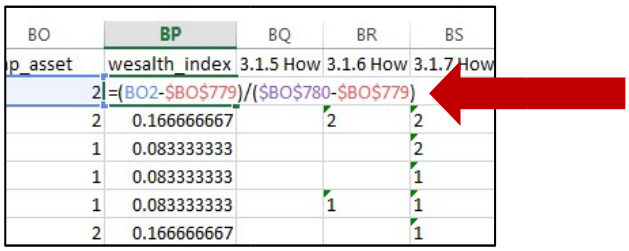
Figure 19. Microsoft Excel minimum and maximum function



Source: Screenshot of Microsoft Excel test sample.

Note that the analysis may be unbalanced if the user employs variables that differ greatly in their range. For instance, a dummy variable on whether the household has access to electricity will take values from 0 to 1, whereas another variable on formal transfers received might take values from 0 to 1,000,000 depending on the currency. In this case, it is recommended to convert the currency or use the Min-Max standardization procedure, according to the type of analysis to be made as well as the context of the study. In general, a variable is rescaled if the distribution is abnormal. With the kdensity function/graph, one may observe whether the distribution is skewed, suggesting that th variable should be rescaled. Add another column right after ast_durable_assets, name it ast_wealth_index and type the function as in Figure 20 (remember to fix the minimum value and the maximum value with the use of "$").

Figure 20. Creating the wealth index under asset pillar



Source: Screenshot of Microsoft Excel test sample.

The same procedure is required for **Questions 6 and 7** that contain information about agricultural assets and inputs. To construct agricultural inputs, the different options need to be summed together. For example, 3 different categories (seeds, pesticides, fertilizer) summed together will give a maximum value of 3. Where the variables are formatted as string variables, new numeric variables need to be constructed using the procedure Logical test and IF function. The Microsoft Excel IF function is useful to test a condition and to return one value if the condition is met, and another value if the condition is not met. The syntax for Microsoft Excel IF could found in Table 3.

Figure 21. The Microsoft Excel IF function

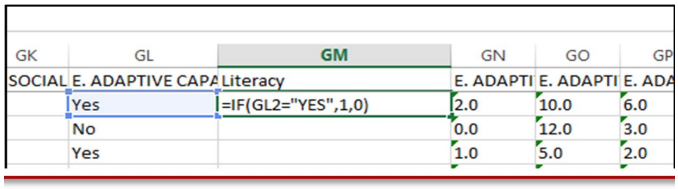
```
IF(logical_test, [value_if_true], [value_if_false])
```

Source: Screenshot of Microsoft Excel test sample.

In order to create a numeric variable from a string one, consider the following steps:

- Insert a new column close to the one containing string (YES/NO)
- Write the IF function as for Figure 22.

Figure 22. IF function in action



Source: Screenshot of Microsoft Excel test sample.

Question 8 contains information about the land owned by the household.

- Make sure to get rid of inaccurate recording or 999 – see Section 2.3, Section 2.4 and Section 2.5.
- Make sure to check for outliers and missing values – see previous sections

Question 9 contains information about the number of animals owned by the household. For each type of animal owned, make sure to impute the outliers, missing values and 999 values, if any.

In order to calculate the Tropical Livestock Unit (TLU) (or the other context-specific livestock unit index), conversion weights are required for each livestock category. Please refer to Table 2 for conversion weights.

Table 2. Tropical livestock unit – conversion weights

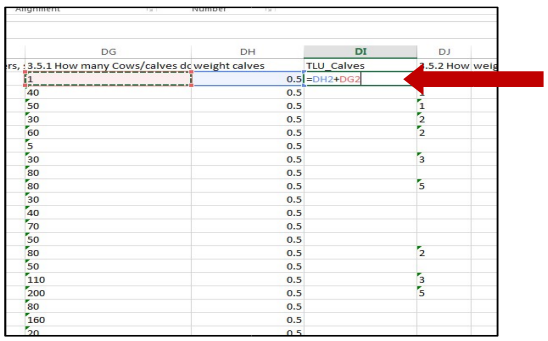
Livestock category	Weight
Cattle	0.5
Sheep	0.1
Goat	0.1
Pig	0.2
Chicken*	0.01
Mule**	0.4
Donkey	0.3

Notes: * The category chicken includes all kind of poultry. ** The category mule includes all draught animals (horses' bullock).
Source: Author's elaboration.

For the TLU calculation:

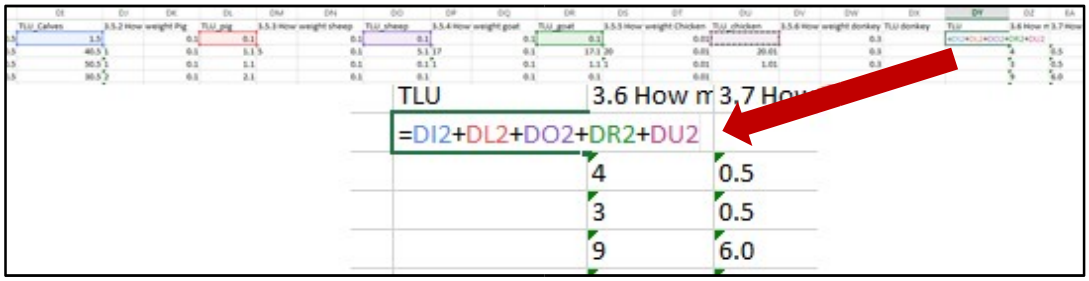
- Create a new column next to the variable to be cleaned and multiply by its respective weight (Figure 23).
- Once the weighted variable for each animal is generated, generate the TLU by summing up the different weighted categories (see Figure 24).

Figure 23. TLU construction (1)



Source: Screenshot of Microsoft Excel test sample.

Figure 24. TLU construction (2)

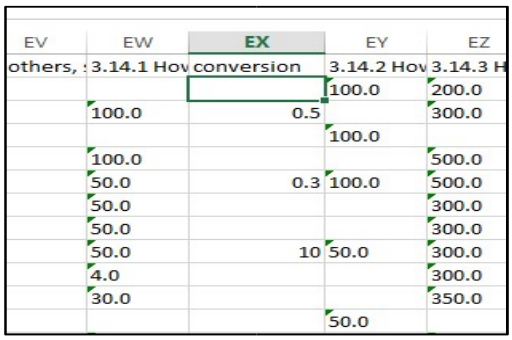


Source: Screenshot of Microsoft Excel test sample.

Question 9 reports for each crop the harvested or the expected harvest. The unit of measurement is kilograms as default; where a different unit of measure is used (for example local units) make sure to convert them in kilograms, following the steps below:

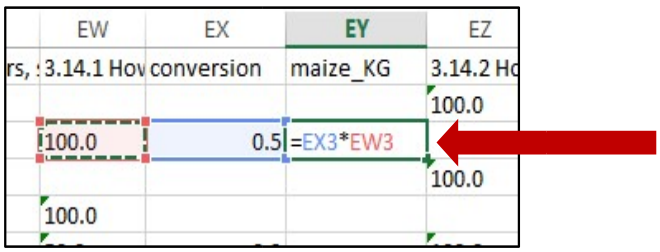
- Add a new column next to the one of unit of measure. Fill it with the conversion measure (Figure 25).
- Add another column and multiply the columns conversion for the original weight (Figure 26).
- The procedure needs to be replicated for the crops in the list. Once all the variables are constructed, use the Copy and Paste Process. The user may proceed to aggregate the various crop productions into one variable/index or keep them separate depending on the analysis they wish to make.

Figure 25. Conversion weight (1)



Source: Screenshot of Microsoft Excel test sample.

Figure 26. Conversion weight (2)



Source: Screenshot of Microsoft Excel test sample.

SOCIAL SAFETY NETS (SSN) PILLAR VARIABLES

Questions 10 to 13 report the monetary value of transfers received (formal and informal) and the value of loans a household received in the previous 12 months. For these variables the user needs to check for outliers, missing

values; also, it is recommended to translate them into dummy variables (=1 if >0, 0 otherwise) instead of using the numerical values.

Question 14 reports information about members of this household formally participating in a local group/association.

Question 15 reports how many relatives/friends/family members the household members can rely on in case of need.

ADAPTIVE CAPACITY (AC) PILLAR VARIABLES

Question 16 reports information about the literacy of the household head. The variable must be numeric: 1 if the household can read and write, or 0 if not.

Question 17 to 19 refer to the average years of education of the household members aged (15–64) (Table 3).

Table 3. Years of school of each household members

	Age	Years of school
Member 1	45	10
Member 2	42	13
Member 3	15	10
Member 4	10	5 (not included)
Member 5	60	0
Average household		(33/4) = 8.25

Source: Author's elaboration.

Question 18 refers to the years the household member has attended formal school with the highest level of education.

Question 19 refers to the years the household member of working age (>14 and <65 years old) has attended formal school with the highest level of education.

Question 20 reports the percentage of the household's overall income. And if it was generated by: Agriculture, animal breeding, fishing, Family business (other than agriculture) Government wage and salary, Private sector wage and salary, Transfers and social assistance, other.

Question 21 reports the total value of loan(s) received by household members.

Question 22 reports the number of crops grown by a household during the previous season.

Question 23 reports if the household members used improved quality seeds during the last season.

Question 24 reports if the household members received any training in the last 12 months, and if "yes" specify which type of training.

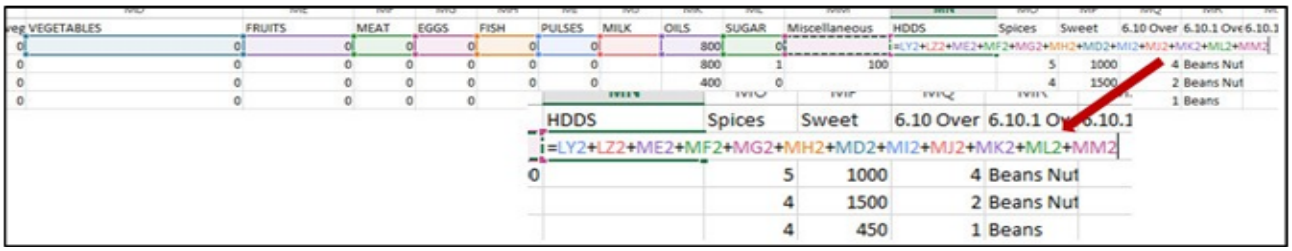
Question 25 reports if the livestock owned by the household received any vaccination in the last 12 months.

FOOD EXPENDITURE AND CONSUMPTION

With **Question 39** it is possible to construct the Household Dietary Diversity Score.[3] Generally, 12 food groups are the basis of the indicator.

[3] For more information about the Household Dietary Diversity Score, see Swindale and Bilinsky (2006).

Figure 27. Household dietary diversity score

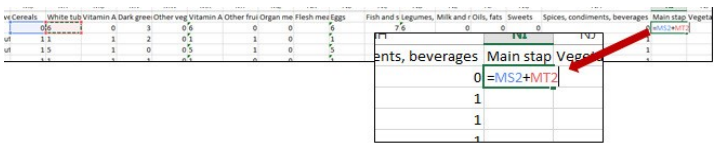


Source: Screenshot of Microsoft Excel test sample.

If in addition to a YES/NO answer, also frequency (number of days) over a seven-day period are collected, the Food Consumption Score (FCS) can be calculated.[4] Following the categories in the basic Short RIMA Questionnaire, those are the steps to construct the food groups:

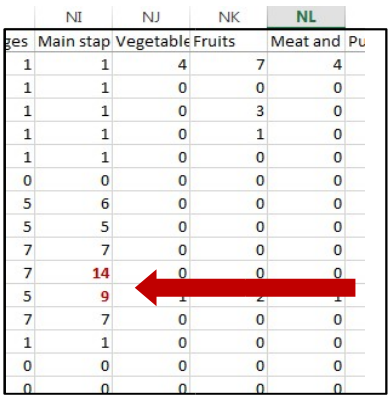
- Add nine columns one after the other following the food variables, rename the columns according to the nine FCS food groups.
- Aggregate the Short RIMA food groups into the FCS food groups.
- Once the groups have been created, recode the value of each group >7 to 7.
- Multiply each cell of the food groups by the corresponding weight (see Figure 28 for weights and Figure 29 for the instruction).

Figure 28. Aggregating food groups



Source: Screenshot of Microsoft Excel test sample.

Figure 29. Recoding higher values



Source: Screenshot of Microsoft Excel test sample.

Questions 34, 35, 36, and 38 relate to the household food expenditures. As these are monetary variables.

SHOCKS

Question 40 and 41 reports the most severe shocks faced by the household in the last 12 months and what did the household members do to cope with the shocks.

HOUSEHOLD DEMOGRAPHIC CHARACTERISTICS

Question 42 refers to the gender of household head. The answer can be numerical or string. In both cases you have to copy and paste it in the "Household variable" sheet.

[4] For more information about the Food Consumption Score (FCS), see Food security cluster Bangladesh (2014).

Question 43 refers to the household size; make sure to take into account outliers following the steps *Addressing extreme values*.

Question 44 refers to the number of household members in working age. This variable is particularly important for the construction of the work ratio, a variable that contributes to the Adaptive Capacity pillar. In order to construct the variable:

- Add a new column close to the variable.
- Rename the variable, in our case work_ratio.
- To calculate the work ratio, divide the total number of people in working age by the household size.

Questions 45, and 46 are useful for the RCI post-estimation ranking; check for typos and misspelled words following the rules in 2.4 Action 4.

4 Final dataset

In order to facilitate the dataset manipulation, it is preferable to keep only variables that will be used for the RCI calculation and profiling. This "workfile" will just be merged with the global one for more extended analysis. Once all the variables are cleaned, make sure that all the variables are on the same sheet (comprising the demographics and household identification variable). For the next steps on uploading the dataset into Shiny RIMA and using the tool for RCI calculation, please refer to the guidelines document available at: https://doi.org/10.4060/cc3353en (FAO, 2022).

References

FAO. 2020. *Resilience Index Measurement and Analysis (RIMA) - Short questionnaire*. Rome. www.fao.org/3/cb2348en/cb2348en.pdf

FAO. 2022. *Shiny RIMA – Guidelines*. Rome. https://doi.org/10.4060/cc3353en

Swindale, A. & Bilinsky, P. 2006. *Household Dietary Diversity Score (HDDS) for Measurement of Household Food Access: Indicator Guide (v.2).* Washington, DC. www.fantaproject.org/sites/default/files/resources/HDDS_v2_Sep06_0.pdf

Food security cluster Bangladesh. 2014. *Guidelines on food consumption score (FCS) calculation for Bangladesh.* https://fscluster.org/sites/default/files/documents/FSC%20Food%20Consumption%20Score%20Guideline%20%202014.docx

CONTACTS

Agrifood Economics Division – Economic and Social Development
ESA-Director@fao.org

Resilience index measurement and analysis (RIMA)
FAO-RIMA@fao.org

Food and Agriculture Organization of the United Nations
Rome, Italy

CC6901EN/1/08.23